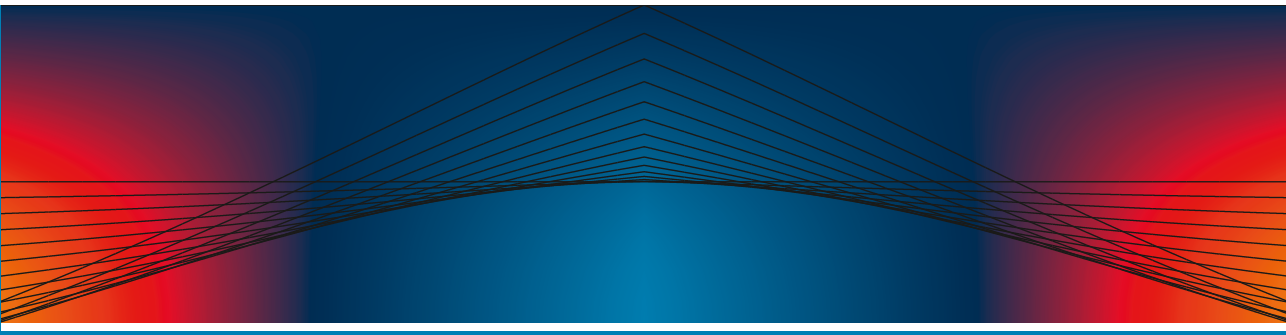
Universität Potsdam



Christoph Sawade

Active Evaluation of Predictive Models

Universitätsverlag Potsdam

Christoph Sawade
Active Evaluation of Predictive Models

Christoph Sawade

Active Evaluation of Predictive Models

Universitätsverlag Potsdam

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de/ abrufbar.

Universitätsverlag Potsdam 2013
http://verlag.ub.uni-potsdam.de/

Am Neuen Palais 10, 14469 Potsdam
Tel.: +49 (0)331 977 2533 / Fax: 2292
E-Mail: verlag@uni-potsdam.de

Zugl.: Potsdam, Univ., Diss., 2013
Gutachter:
Prof. Dr. Francis Bach, Département d'Informatique, Ecole Normale Supérieure
Prof. Dr. Tobias Scheffer, Institut für Informatik, Universität Potsdam
Prof. Dr. Max Welling, Department of Computer Science, University of California
Datum der Disputation: 13.05.2013

Dieses Werk ist unter einem Creative Commons Lizenzvertrag lizenziert:
Namensnennung – Weitergabe unter gleichen Bedingungen 3.0 Deutschland
Um die Bedingungen der Lizenz einzusehen, folgen Sie bitte dem Hyperlink:
http://creativecommons.org/licenses/by-sa/3.0/de/

Online veröffentlicht auf dem Publikationsserver der Universität Potsdam:
URL http://pub.ub.uni-potsdam.de/volltexte/2013/6558/
URN urn:nbn:de:kobv:517-opus-65583
http://nbn-resolving.de/urn:nbn:de:kobv:517-opus-65583

Zugleich gedruckt erschienen im Universitätsverlag Potsdam:
ISBN 978-3-86956-255-1

Abstract

The field of machine learning studies algorithms that infer predictive models from data. Predictive models are applicable for many practical tasks such as spam filtering, face and handwritten digit recognition, and personalized product recommendation. In general, they are used to predict a target label for a given data instance. In order to make an informed decision about the deployment of a predictive model, it is crucial to know the model's approximate performance. To evaluate performance, a set of labeled test instances is required that is drawn from the distribution the model will be exposed to at application time. In many practical scenarios, unlabeled test instances are readily available, but the process of labeling them can be a time- and cost-intensive task and may involve a human expert.

This thesis addresses the problem of evaluating a given predictive model accurately with minimal labeling effort. We study an *active model evaluation process* that selects certain instances of the data according to an instrumental sampling distribution and queries their labels. We derive sampling distributions that minimize estimation error with respect to different performance measures such as error rate, mean squared error, and F-measures. An analysis of the distribution that governs the estimator leads to confidence intervals, which indicate how precise the error estimation is. Labeling costs may vary across different instances depending on certain characteristics of the data. For instance, documents differ in their length, comprehensibility, and technical requirements; these attributes affect the time a human labeler needs to judge relevance or to assign topics. To address this, the sampling distribution is extended to incorporate instance-specific costs. We empirically study conditions under which the active evaluation processes are more accurate than a standard estimate that draws equally many instances from the test distribution.

We also address the problem of comparing the risks of two predictive models. The standard approach would be to draw instances according to the test distribution, label the selected instances, and apply statistical tests to identify significant differences. Drawing instances according to an instrumental distribution affects the power of a statistical test. We derive a sampling procedure that maximizes test power when used to select instances, and thereby minimizes the likelihood of choosing the inferior model. Furthermore, we investigate the task of comparing several alternative models; the objective of an evaluation could be to rank the models according to the risk that they incur or to identify the model with lowest

risk. An experimental study shows that the active procedure leads to higher test power than the standard test in many application domains.

Finally, we study the problem of evaluating the performance of ranking functions, which are used for example for web search. In practice, ranking performance is estimated by applying a given ranking model to a representative set of test queries and manually assessing the relevance of all retrieved items for each query. We apply the concepts of active evaluation and active comparison to ranking functions and derive optimal sampling distributions for the commonly used performance measures Discounted Cumulative Gain (DCG) and Expected Reciprocal Rank (ERR). Experiments on web search engine data illustrate significant reductions in labeling costs.

Kurzfassung

Maschinelles Lernen befasst sich mit Algorithmen zur Inferenz von Vorhersagemodelle aus komplexen Daten. Vorhersagemodelle sind Funktionen, die einer Eingabe – wie zum Beispiel dem Text einer E-Mail – ein anwendungsspezifisches Zielattribut – wie „Spam“ oder „Nicht-Spam“ – zuweisen. Sie finden Anwendung beim Filtern von Spam-Nachrichten, bei der Text- und Gesichtserkennung oder auch bei der personalisierten Empfehlung von Produkten. Um ein Modell in der Praxis einzusetzen, ist es notwendig, die Vorhersagequalität bezüglich der zukünftigen Anwendung zu schätzen. Für diese Evaluierung werden Instanzen des Eingaberaums benötigt, für die das zugehörige Zielattribut bekannt ist. Instanzen, wie E-Mails, Bilder oder das protokollierte Nutzerverhalten von Kunden, stehen häufig in großem Umfang zur Verfügung. Die Bestimmung der zugehörigen Zielattribute ist jedoch ein manueller Prozess, der kosten- und zeitaufwendig sein kann und mitunter spezielles Fachwissen erfordert.

Ziel dieser Arbeit ist die genaue Schätzung der Vorhersagequalität eines gegebenen Modells mit einer minimalen Anzahl von Testinstanzen. Wir untersuchen aktive Evaluierungsprozesse, die mit Hilfe einer Wahrscheinlichkeitsverteilung Instanzen auswählen, für die das Zielattribut bestimmt wird. Die Vorhersagequalität kann anhand verschiedener Kriterien, wie der Fehlerrate, des mittleren quadratischen Verlusts oder des *F-measures*, bemessen werden. Wir leiten die Wahrscheinlichkeitsverteilungen her, die den Schätzfehler bezüglich eines gegebenen Maßes minimieren. Der verbleibende Schätzfehler lässt sich anhand von Konfidenzintervallen quantifizieren, die sich aus der Verteilung des Schätzers ergeben. In vielen Anwendungen bestimmen individuelle Eigenschaften der Instanzen die Kosten, die für die Bestimmung des Zielattributs anfallen. So unterscheiden sich Dokumente beispielsweise in der Textlänge und dem technischen Anspruch. Diese Eigenschaften beeinflussen die Zeit, die benötigt wird, mögliche Zielattribute wie das Thema oder die Relevanz zuzuweisen. Wir leiten unter Beachtung dieser instanzspezifischen Unterschiede die optimale Verteilung her. Die entwickelten Evaluierungsmethoden werden auf verschiedenen Datensätzen untersucht. Wir analysieren in diesem Zusammenhang Bedingungen, unter denen die aktive Evaluierung genauere Schätzungen liefert als der Standardansatz, bei dem Instanzen zufällig aus der Testverteilung gezogen werden.

Eine verwandte Problemstellung ist der Vergleich von zwei Modellen. Um festzustellen, welches Modell in der Praxis eine höhere Vorhersagequalität aufweist, wird eine Menge von Testinstanzen ausgewählt und das zugehörige Zielattribut

bestimmt. Ein anschließender statistischer Test erlaubt Aussagen über die Signifikanz der beobachteten Unterschiede. Die Teststärke hängt von der Verteilung ab, nach der die Instanzen ausgewählt wurden. Wir bestimmen die Verteilung, die die Teststärke maximiert und damit die Wahrscheinlichkeit minimiert, sich für das schlechtere Modell zu entscheiden. Des Weiteren geben wir eine Möglichkeit an, den entwickelten Ansatz für den Vergleich von mehreren Modellen zu verwenden. Wir zeigen empirisch, dass die aktive Evaluierungsmethode im Vergleich zur zufälligen Auswahl von Testinstanzen in vielen Anwendungen eine höhere Teststärke aufweist.

Im letzten Teil der Arbeit werden das Konzept der aktiven Evaluierung und das des aktiven Modellvergleichs auf Rankingprobleme angewendet. Wir leiten die optimalen Verteilungen für das Schätzen der Qualitätsmaße *Discounted Cumulative Gain* (DCG) und *Expected Reciprocal Rank* (ERR) her. Eine empirische Studie zur Evaluierung von Suchmaschinen zeigt, dass die neu entwickelten Verfahren signifikant genauere Schätzungen der Rankingqualität liefern als die untersuchten Referenzverfahren.

Acknowledgments

I would like to take the opportunity of expressing my thanks to everyone who supported me in my research the last years and contributed to this thesis. First and foremost, I gratefully acknowledge my advisor Tobias Scheffer for his invaluable support and guidance. Thank you so much for your motivation, critical examination, and your urge to focus on the essentials.

Special thanks go to Niels Landwehr for sharing the office and his ideas with me. I enjoyed the numerous and never-ending but, however, very fruitful discussions. I am obliged for pleasant working atmosphere and your constant willingness for supporting and advising me whenever possible.

I especially want to thank Michael Brückner, Laura Dietz, Uwe Dick, Peter Haider, and Ulf Brefeld for introducing me to the field of machine learning during as well as after working hours. In particular, I greatly appreciate Niels Landwehr, Michael Brückner, and Laura Dietz for constantly reviewing and proofreading my thesis. I am also grateful to work with Paul Prasse and Michael Großhans.

Moreover, I like to thank Arvid Terzibaschian for having fun on hardcore debugging. I am thankful for your companionship for many years. At this point, I wish to thank the team of nugg.ad AG for a successful cooperation. The cooperation motivated new problem settings and I gave me the chance to learn a lot outside the academic environment. I deeply appreciate it.

Last but not least, I want to thank my family and friends for all the support they have given me all the time.

Christoph Sawade
Potsdam, Germany, May 2013

Contents

Chapter 1

Introduction

Predictive models play a central role in many practical domains, such as spam filtering, face or handwritten digit recognition, and personalized product recommendation. In spam filtering, for example, they are used to classify incoming and outgoing emails as spam or non-spam in order to reduce the amount of unwanted emails reaching a user's mailbox. In general, a predictive model is a function that maps an instance to a target label. The true relationship between instances and the corresponding labels is typically unknown or hard to describe by an explicit rule. Research in the area of *machine learning* is concerned with algorithms that use a finite set of examples that represent the underlying relationship to infer a predictive model. The goal is to identify the model with the highest predictive performance, that is, the model that predicts the label of a new and so far unseen instance as accurately as possible.

A set of labeled instances is essential to build and to evaluate predictive models. Unlabeled instances are typically inexpensive and readily available, but acquiring the corresponding label is often a costly process, which may involve a human expert. For example, email service providers receive a huge amount of emails every day, which can be used to build a spam filter. However, these emails have to be examined manually, since they do not come with the required target label "spam" or "non-spam". The predictive performance generally depends on the number of instance-label pairs that are available to the learning algorithm; an increased number of training instances yields a more accurate model. However, an exhausted labeling of all seen instances can become costly. *Active learning* algorithms are designed to produce accurate models with minimal labeling effort. The idea is to build a sequence of intermediate models with increasing predictive performance. In each step, the algorithm identifies most valuable instances, which would give the highest improvement to the model learned so far. These instances are labeled at a cost. Afterwards, the current model will be updated using all labeled data up to this time, and so on. Depending on the learning task, these strategies can save considerable labeling effort.

Before a predictive model can be deployed in practice, its predictive performance has to be assessed. For this purpose a set of instance-label pairs is required that is governed by the distribution the model will be exposed to at application time. If the training data are distributed according to the test distribution, an estimate of the performance is typically obtained by cross-validation. In practice, however, training data are often unavailable or do not reflect the desired test distribution. In the following, we present examples of such application scenarios motivating the problem setting of the thesis.

Confidential Training Data. When a readily trained model is shipped and deployed, the training data—which are usually used to estimate the model's risk—may be held confidential by the supplier of the model. For instance, a medical diagnosis system would not typically come with the medical records that have been used to train it. Another example are credit scoring models, which predict creditworthiness. Since they are based on confidential data like credit history, loan application, customer data, etc. the training data are also held back in this case for privacy reasons. The supplier may provide a risk estimate, but such estimate might be biased because it is obtained without access to the test distribution. In order to estimate the predictive performance of these models accurately, a set of labeled instances is needed that reflects the test distribution.

Training and Test Distribution Differ. Using cross-validation in order to obtain consistent performance estimates requires that the training data reflect the test distribution. This condition is often not met. Off-the-shelf models such as commercial spam filters or face recognition systems are trained without the knowledge of the distribution the model will be exposed to after deployment. In domains in which the distribution of instances changes over a period of time, one may wish to monitor the risk of the model in order to determine at which point an update becomes necessary. For a reliable estimate one needs access to the current distribution. As an example, commercial email spam filters have to be updated with an additional labeled sample in intervals that depend on the extent to which spammers impose shift on the distribution by employing new strategies to generate messages. As another example, ranking models often cannot be evaluated accurately on held-out training data, because query distributions and item relevance change over time. Instead, considerable effort is spent on manually labeling the relevance of documents for test queries in order to track ranking performance.

Actively Trained Model. Active learning algorithms are used in situation in which no labeled instances are available in advance. In order to minimize the labeling effort, active learners query the labels that they predict least confidently. These instances are not governed by the test distribution. Hence, the resulting labeled data are a biased sample which would incur a pessimistic bias on any cross-validation estimate. In order to obtain an unbiased estimate of the risk, additional test instances have to be labeled.

In these scenarios, estimates are either communicated from the model provider or result from hold-out evaluations on outdated or biased samples; they can be arbitrarily inaccurate. In order to evaluate the model accurately, new instances have to be drawn and labeled at a cost. This thesis addresses the problem of estimating the performance of a given predictive model accurately at minimal labeling costs. The standard approach is to draw instances directly from the test distribution, label these data, and calculate an empirical estimate of the model's performance. Instead, we study an *active evaluation process* that, in analogy to active learning, queries the labels of the most informative instances. Instances are selected according to an instrumental sampling distribution. We derive sampling distributions that minimize the estimation error with respect to a certain performance measure such as error rate, mean squared error, and F-measures. A related problem is to compare two models as confidently as possible on a fixed labeling budget. We devise an *active comparison method* that selects instances according to the instrumental distribution that maximizes the power of a statistical test that compares the performance of two predictive models. Finally, we investigate active evaluation methods for ranking functions. Empirically, we observe that all derived procedures outperform the traditional approach on several classification and regression data sets. Section 1.1 lists own previously published work and summarizes the main contribution of this thesis. An overview is given in Section 1.2.

1.1 Contributions

In this thesis, we develop new evaluation methods to estimate and to compare the performance of predictive models. We now summarize the main results and discuss the relation to own publications.

Optimal Sampling Distribution for Risk Estimation. We introduce the concept of active risk estimation: Instances are selected from a pool of unlabeled instances in order to evaluate the risk of a given model accurately at minimal labeling costs. We analyze sources of estimation error of the empirical risk, and derive the sampling distribution that asymptotically minimizes the estimation error. The optimal sampling distribution depends on unknown quantities. We derive an empirical sampling distribution that uses the model to decide on instances whose labels are queried. The resulting active evaluation process can be applied immediately with a probabilistic prediction model and yields a consistent estimate of the true risk. An analysis of the distribution that governs the estimator leads to confidence intervals. We empirically study conditions under which the active risk estimate is more accurate than a standard risk estimate that draws equally many instances from the test distribution. These results have been published in

[*Sawade et al., 2010a*] Christoph Sawade, Niels Landwehr, Steffen Bickel, and Tobias Scheffer. Active Risk Estimation. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.

Generalization to F-measures. We generalize the regular risk functional to incorporate F-measures, which are common performance measures in information retrieval tasks. We show that the commonly used statistics constitute consistent estimators of that generalized risk. On this basis, we derive an evaluation process that actively estimates a generalized risk by sampling test instances from an instrumental distribution. An analysis of the sources of estimation error leads to the instrumental distribution that minimizes estimator variance. Our empirical study supports the conclusion that the advantage of active over passive evaluation is particularly strong for skewed classes. These results have been published in

[*Sawade et al., 2010b*] Christoph Sawade, Niels Landwehr, and Tobias Scheffer. Active Estimation of F-Measures. In *Proceedings of the 24th Annual Conference on Neural Information Processing Systems*, 2010.

Optimal Sampling Distribution for Hypotheses Testing. We address the problem of comparing the predictive performance of two given models as confidently as possible given a fixed labeling budget. We lift the active evaluation principle to hypothesis testing and derive a sampling distribution that maximizes test power when used to select instances, and thereby minimizes the likelihood of choosing the inferior model. Empirically, we observed that the resulting active comparison method consistently outperforms a traditional comparison based on a uniform sample of test instances. Active comparison identifies the model with lower true risk more often, and is able to detect significant differences between the risks of two given models more quickly. We perform experiments under the null hypothesis that both models incur identical risks, and verified that active comparison does not lead to increased false-positive significance results. These results have been published in

[*Sawade et al., 2012b*] Christoph Sawade, Niels Landwehr, and Tobias Scheffer. Active Comparison of Prediction Models. In *Proceedings of the 26th Annual Conference on Neural Information Processing Systems*, 2012.

Cost-Optimal Sampling for Ranking Functions. We study active estimation of ranking performance. A novel aspect of active estimation in a ranking setting is that labeling costs vary according to the number of items that are relevant for a query. We derive a cost-optimal sampling distributions for the estimation of DCG and ERR. Naïve computation of the sampling distributions is exponential in the number of items, we derive polynomial-time solutions by dynamic programming. Experiments on web search engine data illustrate significant reductions in labeling costs when estimating the performance of a single ranking model or comparing different types of ranking models. These results have been published in

[*Sawade et al., 2012a*] Christoph Sawade, Steffen Bickel, Timo von Oertzen, Tobias Scheffer, and Niels Landwehr. Active Evaluation of Ranking Functions based on Graded Relevance. In *Proceedings of the 22nd European Conference on Machine Learning*, 2012. Best Paper Award.

1.2 Outline

In this thesis we study evaluation processes for predictive models. The first two chapters recapitulate the foundations of learning theory and evaluation methods. The principle ideas of predictive models and how they can be inferred from data are presented in Chapter 2. In Chapter 3, we introduce the concept of risk functionals, their essential estimators, and statistical tests which are used to estimate the absolute and relative performance of predictive models. Additionally, we state a new generalization of the traditional risk and derive statements on the estimators.

In the following two chapters, we focus on scenarios in which test instances have to be drawn and labeled to obtain an estimate. We study performance measures which can be expressed as a generalized risk, such as error rate, mean squared error, and F-measures. The sampling distribution that minimizes the estimation error on a fixed labeling budget with respect to a generalized risk is derived in Chapter 4. Furthermore, we extend our results to the case in which labeling costs vary over different instances; the derived sampling distribution involves instance-specific labeling costs and is optimal for constrained overall costs. In Chapter 5, we address the problem of comparing the risks of predictive models as confidently as possible. To this end, we analyze the statistical testing process, which is resulting in a sampling procedure that maximizes test power. Chapter 6 studies active evaluation in the context of ranking functions. Many ranking measures can be formulated as risks, however, the optimal sampling distributions involve exponential sums. We show how they can be computed in polynomial time for two important ranking measures using dynamic programming.

In all three chapters, we experimentally study conditions under which the active evaluation is more accurate than the standard passive procedure that draws equally many instances from the test distribution. Finally, Chapter 7 concludes.

Chapter 2

Learning Predictive Models from Data

The concept of predictive models plays a central role in this thesis. In this chapter, we present principle ideas of predictive models and summarize the state of the art of probabilistic learning algorithms. In many practical tasks one aims at identifying a target label $y \in \mathcal{Y}$ of a given instance $x \in \mathcal{X}$, where $\mathcal{Y}$ is referred to as label space and $\mathcal{X}$ as instance space. An unknown test distribution $p(x, y) = p(y|x)p(x)$ is defined over $\mathcal{X} \times \mathcal{Y}$. The conditional distribution $p(y|x)$ describes the true relationship between an instance x and a label y. An estimate of $p(y|x)$ enables us to infer the most probable label y for an instance x as well as to derive a confidence value for any prediction. The process of estimating the true conditional distribution $p(y|x)$ is also referred to as *learning* the relationship between x and y. For this purpose, a finite set of instance-label pairs

$$T_n = \{(x_i, y_i)|i = 1, \ldots, n\} \tag{2.1}$$

is given, which is called the *training data*. The training instances (x_i, y_i) are assumed to be *independent and identically distributed* (i.i.d.) according to $p(x, y)$, that is, the probability of observing the training set T_n can be decomposed into a product over the distribution of instance-label pairs:

$$p(\{(x_i, y_i)|i = 1, \ldots, n\}) = \prod_{i=1}^{n} p(y_i|x_i)p(x_i). \tag{2.2}$$

A non-parametric estimate of the distribution $p(y|x)$ is non-trivial, since the instance space $\mathcal{X}$ is often a high dimensional vector space (see, *e.g.,* Bishop, 2006, Chapter 1.4). One possible way to tackle this problem is to model $p(y|x)$ by a fixed family of distributions $p(y|x; \boldsymbol{\theta})$, which is parameterized by a vector $\boldsymbol{\theta} \in \Theta$, and estimate the corresponding parameters from T_n. The set of all possible pa-

rameterizations Θ is called the model space. Let us assume that the model space contains a model $\boldsymbol{\theta}^*$, that completely characterizes the true relationship between a label y and a given instance x. Therefore, we replace the conditional distribution $p(y|x)$ by $p(y|x;\boldsymbol{\theta}^*)$. In order to estimate $\boldsymbol{\theta}^*$, we now study the *posterior predictive distribution* $p(y|x,T_n)$ that quantifies the likelihood of a label y given an instance x and the training set T_n under the model assumption. Applying the law of total probability, the i.i.d.-assumption (see Equation 2.2), and the assumption, that the probability of a model $\boldsymbol{\theta}$ depends only on the observed training set T_n, the posterior predictive distribution can be expressed as weighted average over the *model-based predictive distributions* $p(y|x;\boldsymbol{\theta})$:

$$p(y|x,T_n) = \int p(y|x;\boldsymbol{\theta})p(\boldsymbol{\theta}|T_n)\mathrm{d}\boldsymbol{\theta}. \tag{2.3}$$

Each of the predictive distributions is weighted by the *posterior* distribution $p(\boldsymbol{\theta}|T_n)$, that is, the probability of the model $\boldsymbol{\theta}$ after having seen the training set T_n. The posterior distribution can be decomposed further into a *likelihood*, *prior*, and *marginal likelihood* term using Bayes' rule:

$$p(\boldsymbol{\theta}|T_n) = \frac{p(T_n|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(T_n)}. \tag{2.4}$$

Under the assumption given by Equation 2.2, the likelihood can be expressed as

$$p(T_n|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|x_i;\boldsymbol{\theta})p(x_i). \tag{2.5}$$

It captures how well the model $\boldsymbol{\theta}$ fits the training data T_n. The data independent prior $p(\boldsymbol{\theta})$ quantifies the likelihood of a model $\boldsymbol{\theta}$ independently of the data. The remaining normalization term $p(T_n) = \int p(T_n|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$ is known as marginal likelihood (see Section 2.2).

In order to evaluate Equation 2.3, we need to define the family $p(y|x;\boldsymbol{\theta})$ and the prior distribution $p(\boldsymbol{\theta})$. This choice is specific to the learning task; it depends on the label space $\mathcal{Y}$ and assumptions about the data. In general, a prediction based on Equation 2.3 is known as *Bayesian model averaging*. It can be seen as the optimal decision for an unseen instance x, since the posterior predictive distribution accounts for the model uncertainty $p(\boldsymbol{\theta}|T_n)$ caused by the finiteness of T_n (see, *e.g.,* Domingos, 2000; Davidson & Fan, 2006). However, the Bayes optimal solution is intractable for many choices of the model class. In such a case, the posterior distribution (see Equation 2.4) can be approximated by some

point estimate $\hat{\boldsymbol{\theta}}$. The *maximum a posteriori* (MAP) estimate is obtained by replacing the expectation $\mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|T_n)}[p(y|x;\boldsymbol{\theta})]$ over all models by the prediction of the most probable model after having seen the training set T_n:

$$p(y|x,T_n) \approx p\big(y\big|x;\hat{\boldsymbol{\theta}}^{map}\big), \text{ where } \hat{\boldsymbol{\theta}}^{map} = \arg\max_{\boldsymbol{\theta}\in\Theta} p(\boldsymbol{\theta}|T_n). \tag{2.6}$$

If a uniform prior over the model parameters is defined, the MAP estimate reduces to the *maximum likelihood* (ML) estimate. It is given by

$$p(y|x,T_n) \approx p\big(y\big|x;\hat{\boldsymbol{\theta}}^{ml}\big), \text{ where } \hat{\boldsymbol{\theta}}^{ml} = \arg\max_{\boldsymbol{\theta}\in\Theta} p(T_n|\boldsymbol{\theta}). \tag{2.7}$$

In Equation 2.7, the posterior predictive distribution is approximated by the predictive distribution of the model that gives the observed data the highest probability.

The quality of an approximation $p(y|x;\boldsymbol{\theta})$ can be assessed by the *theoretical label likelihood*, which is defined as

$$\mathcal{L}(\boldsymbol{\theta}) = \exp\left(\mathbb{E}_{(x,y)\sim p(y|x;\boldsymbol{\theta}^*)p(x)}\left[\log p(y|x;\boldsymbol{\theta})\right]\right). \tag{2.8}$$

The theoretical label likelihood is the exponentiated expected value of the per-instance label likelihood in the logarithmic space. It can be estimated by the geometric mean of $p(y|x;\boldsymbol{\theta})$ taken over a set T_n sampled i.i.d. from $p(x,y)$:

$$\hat{\mathcal{L}}_n(\boldsymbol{\theta}) = \sqrt[n]{\prod_{i=1}^{n} p(y_i|x_i;\boldsymbol{\theta})}. \tag{2.9}$$

Equation 2.9 is referred to as the *per-instance label likelihood*. Analogically to Wasserman (2004, Chapter 9.5), it can be shown that $\hat{\boldsymbol{\theta}}^{map}$ maximizes asymptotically the theoretical label likelihood $\mathcal{L}$ for any prior distributions $p(\boldsymbol{\theta})$. Thus, Bayesian model averaging (see Equation 2.3) and the maximum a posteriori estimate (see Equation 2.7) are optimal as well for $n \to \infty$. In order to analyze the quality of an estimate $\hat{\boldsymbol{\theta}}$, it can be useful to study also the distance between the distributions $p(y|\mathbf{x};\hat{\boldsymbol{\theta}})$ and $p(y|\mathbf{x};\boldsymbol{\theta}^*)$. The distance between two arbitrary distributions p and p' can be measured by the *Kullback-Leibler divergence*. It is defined as

$$\mathrm{KL}[p||p'] = \int \log\frac{p(x)}{p'(x)} p(x)\mathrm{d}x.$$

The Kullback-Leibler divergence is non-negative and vanishes if and only if $p = p'$. We now show that maximizing $\mathcal{L}(\boldsymbol{\theta})$ is equivalent to minimizing the Kullback-Leibler divergence of $p(y|\mathbf{x};\boldsymbol{\theta})$ from $p(y|\mathbf{x};\boldsymbol{\theta}^*)$ in expectation over $\mathbf{x}$. In Equation 2.10, we make use of the monotonicity of the logarithm and add constants $\log p(y|\mathbf{x};\boldsymbol{\theta}^*)$:

$$\begin{aligned}
\arg\max_{\boldsymbol{\theta}\in\Theta} \mathcal{L}(\boldsymbol{\theta}) &= \arg\max_{\boldsymbol{\theta}\in\Theta} \exp\left(\iint \log p(y|x;\boldsymbol{\theta}) p(y|x;\boldsymbol{\theta}^*) p(x) \mathrm{d}y\, \mathrm{d}x\right) \\
&= \arg\min_{\boldsymbol{\theta}\in\Theta} \iint \log \frac{p(y|x;\boldsymbol{\theta}^*)}{p(y|x;\boldsymbol{\theta})} p(y|x;\boldsymbol{\theta}^*) p(x) \mathrm{d}y\, \mathrm{d}x \qquad (2.10) \\
&= \arg\min_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{x\sim p(x)} \left[\mathrm{KL}\left[p(y|x;\boldsymbol{\theta}^*)||p(y|x;\boldsymbol{\theta})\right]\right]. \qquad (2.11)
\end{aligned}$$

Since $\boldsymbol{\theta}^* \in \Theta$, it follows that the theoretical likelihood is maximized by $\boldsymbol{\theta}$ if and only if $p(y|x;\boldsymbol{\theta}) = p(y|x;\boldsymbol{\theta}^*)$. Consequently, the predictive distributions of the presented estimators (see Equation 2.3, 2.6, and 2.7) converge indeed to $p(y|x;\boldsymbol{\theta}^*)$.

A *predictive model* is a function

$$f_{\boldsymbol{\theta}}(x) = \arg\max_{y\in\mathcal{Y}} p(y|x;\boldsymbol{\theta}),$$

which assigns a label $y \in \mathcal{Y}$ to a given instance $x \in \mathcal{X}$ based on a model-based predictive distribution $p(y|x;\boldsymbol{\theta})$. The task of determining $f_{\boldsymbol{\theta}}$ in the case of a finite label space $\mathcal{Y}$ is referred to as a *classification* problem, whereas if $\mathcal{Y} = \mathbb{R}$ the learning task is called *regression*. In Section 2.1, we present logistic regression for classification tasks and Bayesian linear regression for continuous label spaces in Section 2.2. In order to learn a predictive model a set of labeled instances is required, which represents the true probability $p(y|x;\boldsymbol{\theta}^*)$ of the label y for an instance x. In the absence of labeled training data, new training instances have to be labeled at a cost. If the labeling budget is limited, the choice of instances which will be labeled is crucial to obtain a model with high predictive performance. In Section 2.3, we discuss active learning strategies, which are used to determine only a small subset of instances that have to be labeled.

2.1 Learning a Classification Model

This section recapitulates the logistic regression model following Jordan (1995) and Bishop (2006, Chapter 4.3). Therefore, we specify the assumptions that are made about the data and state the prior distribution over the model parameters. As a result, we present the optimization criterion of the maximum a posteriori estimate of the model parameters; it can be solved efficiently by standard solvers. Finally, we introduce a kernelized version of the logistic regression model.

Assumptions of the Logistic Regression Model

We now derive the optimization problem of logistic regression. In general, it can be interpreted as finding the maximum a posteriori estimate in a classification setting. Let $\mathbf{x} \in \mathbb{R}^d$ be a numerical Euclidean vector representation of an instance x and let $\mathcal{Y}$ be a finite space. Inserting Equation 2.4 and 2.5 into Equation 2.6 leads to

$$\begin{aligned}\hat{\boldsymbol{\theta}}^{map} &= \arg\max_{\boldsymbol{\theta}\in\Theta} \frac{\prod_{i=1}^{n} p(y_i|\mathbf{x}_i;\boldsymbol{\theta})p(\mathbf{x}_i)}{\prod_{i=1}^{n} p(y_i|\mathbf{x}_i)p(\mathbf{x}_i)} p(\boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}\in\Theta} \prod_{i=1}^{n} p(y_i|\mathbf{x}_i;\boldsymbol{\theta})p(\boldsymbol{\theta}). \end{aligned} \tag{2.12}$$

In the following, we need to specify the distribution $p(y|\mathbf{x};\boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$. Using Bayes' theorem, the model-based predictive distribution $p(y|\mathbf{x};\boldsymbol{\theta})$ can be reformulated in terms of class-conditional distributions of the instances and a marginal distribution of the labels:

$$p\left(y|\mathbf{x};\boldsymbol{\theta}\right) = \frac{p\left(\mathbf{x}|y;\boldsymbol{\theta}'\right) p\left(y|\boldsymbol{\theta}''\right)}{\sum_{\bar{y}\in\mathcal{Y}} p\left(\mathbf{x}|\bar{y};\boldsymbol{\theta}'\right) p\left(\bar{y}|\boldsymbol{\theta}''\right)}. \tag{2.13}$$

In Equation 2.13, we have subdivided the parameter vector $\boldsymbol{\theta}$ into parameters $\boldsymbol{\theta}'$ which specify the distribution $p\left(\mathbf{x}|y;\boldsymbol{\theta}'\right) = p\left(\mathbf{x}|\boldsymbol{\theta}'_y\right)$ for instances belonging to class y and parameters $\boldsymbol{\theta}''$ which correspond to the label distribution $p\left(y|\boldsymbol{\theta}''\right)$.

Since $\mathcal{Y}$ is finite, the labels follow a categorical distribution, which is given by

$$p\left(y|\boldsymbol{\theta}''\right) = \left(\sum_{\bar{y}\in\mathcal{Y}} \boldsymbol{\theta}''_{\bar{y}}\right)^{-1} \prod_{\bar{y}\in\mathcal{Y}} \left(\boldsymbol{\theta}''_{\bar{y}}\right)^{[\![y=\bar{y}]\!]}, \tag{2.14}$$

where $[\![\cdot]\!]$ denotes the indicator function and $\boldsymbol{\theta}'' = \left(\boldsymbol{\theta}''_y\right)_{y\in\mathcal{Y}}$ with $\boldsymbol{\theta}''_y \geq 0$ is a vector of parameters that represents the probability of observing the class $y \in \mathcal{Y}$. The former term of Equation 2.14 ensures that the distribution is normalized.

Instances belonging to a class $y \in \mathcal{Y}$ are assumed to be drawn from an *exponential family*. An exponential family (see, *e.g.,* Bishop, 2006, Chapter 2.4) is a set of parameterized distributions which can be expressed in the form

$$p\left(\mathbf{x}|\boldsymbol{\theta}'\right) = h(\mathbf{x}) \exp\left(\phi(\mathbf{x})^\mathsf{T}\boldsymbol{\theta}' - \ln g(\boldsymbol{\theta}')\right). \tag{2.15}$$

A certain class of distributions is obtained by instantiating the *feature mapping* $\phi : \mathbb{R}^d \to \mathbb{R}^e$ and the non-negative base measure $h : \mathbb{R}^d \to \mathbb{R}$, where e is the number of parameters. The feature mapping $\phi(\mathbf{x})$ projects the input vector $\mathbf{x}$ into the parameter space and provides all information needed to derive the probability of $\mathbf{x}$; it is also known as *sufficient statistic* in the statistics literature. Finally, the partition function $g(\boldsymbol{\theta}')^{-1}$ must be chosen in such a way as to ensure that the probability distribution is normalized.

Distributions that belong to an exponential family are, for example, the multinomial, Poisson, and, in particular, the Gaussian distribution. A random vector $\mathbf{x} \in \mathbb{R}^d$ is said to be Gaussian if it has the density function

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right), \tag{2.16}$$

with mean vector $\boldsymbol{\mu} \in \mathbb{R}^d$ and positive semidefinite covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$. Equation 2.16 can be expressed in the form of an exponential family (see Equation 2.15) using the quantities

$$\begin{aligned} \boldsymbol{\theta}' &= \begin{pmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\operatorname{vec}\left(\boldsymbol{\Sigma}^{-1}\right) \end{pmatrix}, & \phi(\mathbf{x}) &= \begin{pmatrix} \mathbf{x} \\ \mathbf{x}\otimes\mathbf{x} \end{pmatrix}, \\ h(\mathbf{x}) &= (2\pi)^{-d/2}, & g(\boldsymbol{\eta}) &= \sqrt{|\boldsymbol{\Sigma}| \exp\left(\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)}, \end{aligned} \tag{2.17}$$

where $\mathbf{u}\otimes\mathbf{v}$ denotes the Kronecker product multiplying each component of $\mathbf{u}$ by each component of $\mathbf{v}$, $\operatorname{vec}(\mathbf{U})$ stacks the column vectors of a matrix $\mathbf{U}$ below one another, and $|\mathbf{U}|$ is the determinant of a square matrix $\mathbf{U}$.

Having established and motivated the exponential family, we can now derive the model-based predictive distribution for logistic regression. Let all class-conditional distributions $p\left(\mathbf{x}|y;\boldsymbol{\theta}'\right)$ be a member of some exponential family.

Then, by using Equation 2.14 and 2.15, the model-based predictive distribution $p(y|\mathbf{x};\boldsymbol{\theta})$ given by Equation 2.13 can be expressed as

$$\begin{aligned}&\frac{h(\mathbf{x})\exp\left(\phi(\mathbf{x})^\mathsf{T}\boldsymbol{\theta}'_y - \ln g(\boldsymbol{\theta}'_y)\right)\left(\sum_{\tilde{y}\in\mathcal{Y}}\boldsymbol{\theta}''_{\tilde{y}}\right)\prod_{\tilde{y}\in\mathcal{Y}}\left(\boldsymbol{\theta}''_{\tilde{y}}\right)^{[\![y=\tilde{y}]\!]}}{h(\mathbf{x})\sum_{\bar{y}\in\mathcal{Y}}\exp\left(\phi(\mathbf{x})^\mathsf{T}\boldsymbol{\theta}'_{\bar{y}} - \ln g(\boldsymbol{\theta}'_{\bar{y}})\right)\left(\sum_{\tilde{y}\in\mathcal{Y}}\boldsymbol{\theta}''_{\tilde{y}}\right)\prod_{\tilde{y}\in\mathcal{Y}}\left(\boldsymbol{\theta}''_{\tilde{y}}\right)^{[\![\bar{y}=\tilde{y}]\!]}}\\ =&\frac{\exp\left(\phi(\mathbf{x})^\mathsf{T}\boldsymbol{\theta}'_y + b_y\right)}{\sum_{\bar{y}\in\mathcal{Y}}\exp\left(\phi(\mathbf{x})^\mathsf{T}\boldsymbol{\theta}'_{\bar{y}} + b_{\bar{y}}\right)},\end{aligned} \tag{2.18}$$

where we have defined $b_y = \ln\boldsymbol{\theta}''_y - \ln g(\boldsymbol{\theta}'_y)$. Notice that, the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_y)_{y\in\mathcal{Y}}$ comprises all class-wise parameters $\boldsymbol{\theta}_y = ((\boldsymbol{\theta}'_y)^\mathsf{T}, b_y)^\mathsf{T}$. Equation 2.18 is known as a *generalized linear model* (McCullagh & Nelder, 1989).

The set of points $\mathbf{x}$, for which it holds that $p(y|\mathbf{x};\boldsymbol{\theta}) = p(\bar{y}|\mathbf{x};\boldsymbol{\theta})$, or equivalently,

$$\begin{aligned}0 &= \log\frac{p(y|\mathbf{x};\boldsymbol{\theta})}{p(\bar{y}|\mathbf{x};\boldsymbol{\theta})}\\ &= \phi(\mathbf{x})^\mathsf{T}(\boldsymbol{\theta}_y - \boldsymbol{\theta}_{\bar{y}}) + (b_y - b_{\bar{y}})\end{aligned} \tag{2.19}$$

is called *decision boundary*. The decision boundary between two arbitrary classes $y, \bar{y} \in \mathcal{Y}$ under a generalized linear model (see Equation 2.19) is given by a linear combination of $\phi(\mathbf{x})$ and the model parameter $\boldsymbol{\theta}$. In particular, it can be shown that the decision boundary is affine in $\phi(\mathbf{x})$ if and only if the class-conditional distribution $p(\mathbf{x}|y;\boldsymbol{\theta})$ and $p(\mathbf{x}|\bar{y};\boldsymbol{\theta})$ belong to the same exponential family (Banerjee, 2007). An interesting special case occurs when the class-conditional distributions $p(\mathbf{x}|y;\boldsymbol{\theta})$ are assumed to be Gaussian with identical covariance matrices for all classes $y \in \mathcal{Y}$. Then, the identity mapping $\phi(\mathbf{x}) = \mathbf{x}$ is sufficient to characterize the predictive distribution $p(y|\mathbf{x};\boldsymbol{\theta})$. The corresponding predictive model $f_{\boldsymbol{\theta}}(\mathbf{x})$ is referred to as a *linear model* (see, *e.g.,* Bishop, 2006, Chapter 4). In the following, we omit b_y since it can be encoded by augmenting the statistic $\phi(\mathbf{x})$ by one.

If the prior distribution $p(\boldsymbol{\theta})$ is assumed to be Gaussian $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}|\mathbf{0}, \bar{\boldsymbol{\Sigma}})$ with zero mean and covariance matrix $\bar{\boldsymbol{\Sigma}}$, the MAP estimate is given by Proposition 2.1. Equation 2.20 follows by inserting Equation 2.18 and the Gaussian prior into Equation 2.13 and the monotonicity of the logarithm. A detailed proof is given by, *e.g.,* Karsmakers et al. (2007).

Proposition 2.1 (Maximum a Posteriori for Logistic Regression)**.** *If $p(\mathbf{x}|y;\boldsymbol{\theta})$ is an exponential family, the maximum a posteriori estimate with a Gaussian prior is given by*

$$\begin{aligned}\hat{\boldsymbol{\theta}}^{map} &= \arg\max_{\boldsymbol{\theta}\in\Theta} p(\boldsymbol{\theta}|T_n) \\ &= \arg\min_{\boldsymbol{\theta}\in\Theta} \sum_{i=1}^{n} \ell_{log}\left(\boldsymbol{\theta}, \phi(\mathbf{x}_i), y_i\right) + \frac{1}{2}\boldsymbol{\theta}^{\mathsf{T}}\bar{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\theta}, \end{aligned} \tag{2.20}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_y)_{y\in\mathcal{Y}}$ denotes the vector of all parameters and the logistic loss

$$\ell_{log}(\boldsymbol{\theta}, \phi(\mathbf{x}), y) = \log \sum_{y'\in\mathcal{Y}} \exp\left(\phi(\mathbf{x})^{\mathsf{T}}\boldsymbol{\theta}_{y'}\right) - \phi(\mathbf{x})^{\mathsf{T}}\boldsymbol{\theta}_y$$

measures the disagreement between the prediction and the true label.

Equation 2.20 is also known as penalized log-likelihood estimation. It can been seen as a minimization of a regularized empirical risk. From this perspective, the former term constitutes a sum over an instance-specific loss function ℓ_{log} whereas the latter penalizes the model's complexity. An isotropic covariance matrix $\bar{\boldsymbol{\Sigma}} = \sigma^2\mathbf{I}$ with $\sigma > 0$ corresponds to a standard L^2-norm regularization of the decision function $f_{\boldsymbol{\theta}}$ (Tikhonov & Arsenin, 1977).

The optimization problem given by Equation 2.20 is convex and continuously differentiable for all fixed $y \in \mathcal{Y}$ and can thus be minimized, for example, by stochastic gradient descent (also known as Robbins-Monro algorithm; see, *e.g.,* Spall, 2003, Chapter 4). The partial gradient with respect to $\boldsymbol{\theta}_y \in \mathbb{R}^e$ is given by

$$-\frac{\partial}{\partial\boldsymbol{\theta}_y}\log p(\boldsymbol{\theta}|T_n) = \sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}_y}\ell_{log}\left(\boldsymbol{\theta}, \phi(\mathbf{x}_i), y_i\right) + \sum_{\bar{y}\in\mathcal{Y}}\bar{\boldsymbol{\Sigma}}^{-1}_{y,\bar{y}}\boldsymbol{\theta}_{\bar{y}}, \tag{2.21}$$

where the gradient of the logistic loss is given by

$$\frac{\partial}{\partial\boldsymbol{\theta}_y}\ell_{log}\left(\boldsymbol{\theta}, \phi(\mathbf{x}), y'\right) = \left(\frac{\exp(\phi(\mathbf{x})^{\mathsf{T}}\boldsymbol{\theta}_y)}{\sum_{\bar{y}\in\mathcal{Y}}\exp(\phi(\mathbf{x})^{\mathsf{T}}\boldsymbol{\theta}_{\bar{y}})} - [\![y = y']\!]\right)\phi(\mathbf{x})$$

and the inverse covariance matrix $\bar{\boldsymbol{\Sigma}}^{-1} = (\bar{\boldsymbol{\Sigma}}^{-1}_{y,\bar{y}})_{y,\bar{y}\in\mathcal{Y}}$ is given by a block matrix of pairwise inverse covariances $\bar{\boldsymbol{\Sigma}}^{-1}_{y,\bar{y}}$.

The MAP estimate depends on the covariance matrix $\bar{\boldsymbol{\Sigma}}$ of the Gaussian prior, which was assumed to be fixed during the derivation in this section. In practice, the covariance matrix $\bar{\boldsymbol{\Sigma}} = \sigma^2\mathbf{I}$ is often assumed to be isotropic; the parameter σ^2 can be estimated by cross validation to perform well on so far unseen test instances (Weiss & Kulikowski, 1990).

Kernel Functions and Implicit Feature Mappings

The feature mapping ϕ maps instances into a potentially high-dimensional space, which in turn affects the number of parameters e which have to be estimated when solving the optimization problem given by Proposition 2.1. Depending on the assumption about $p(\mathbf{x}|y;\boldsymbol{\theta})$, the parameter space may be large and solving the optimization problem can become inefficient. However, the *representer theorem* (Kimeldorf & Wahba, 1971; Schölkopf et al., 2001) states, that the maximizer of Equation 2.20 can be equivalently written as a linear combination over the mapped training instances, that is, there exists $\alpha_{i,y} \in \mathbb{R}$ such that

$$\phi(\mathbf{x})^\mathsf{T}\hat{\boldsymbol{\theta}}_y^{map} = \sum_{i=1}^{n} \alpha_{i,y}\phi(\mathbf{x}_i)^\mathsf{T}\phi(\mathbf{x}). \tag{2.22}$$

Substituting Equation 2.22 into Equation 2.20 and 2.21, respectively, leads to the dual formulation of the multi-class logistic regression. This optimization problem depends on the parameters $\alpha_{i,y}$ rather than $\boldsymbol{\theta}_y$; the number of optimization parameters per class is equal to the number of observed instances n, which can be much smaller than the number of dimensions e of the mapped instances $\phi(\mathbf{x})$.

The dual formulation depends on the mapped data only through inner products. The inner product can often be computed quite efficiently using *kernel functions*. In general, a kernel is referred to a function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ that constitutes the inner product $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\mathsf{T}\phi(\mathbf{x}')$ in some Hilbert space induced by a feature mapping ϕ. It can be seen as a similarity measure between two instances $\mathbf{x}$ and $\mathbf{x}'$. Evaluating a kernel function does not necessarily require an explicit mapping of the instances. For example, a Gaussian distribution assumption of $p(\mathbf{x}|y;\boldsymbol{\theta})$ yields the feature mapping $\phi(\mathbf{x}) = (\mathbf{x}, \mathbf{x}\otimes\mathbf{x}, 1)^\mathsf{T}$ (see Equation 2.17). An explicit computation of the inner product $\phi(\mathbf{x})^\mathsf{T}\phi(\mathbf{x}')$ requires $\mathcal{O}(d^2)$ multiplication and addition operations. However, the inner product is equivalent to the *polynomial*

kernel function

$$k_{poly}(\mathbf{x}, \mathbf{x}') = \left(\mathbf{x}^\mathsf{T}\mathbf{x}' + 1\right)^a \tag{2.23}$$

with degree $a = 2$ (see, *e.g.,* Schölkopf & Smola, 2002, Chapter 2.1). It can be computed in time $\mathcal{O}(d)$.

Although it is useful to know which transformation ϕ has to be applied to an instance $\mathbf{x}$ to implement a specific distribution assumption $p(\mathbf{x}|y;\boldsymbol{\theta})$, the true distribution class of $\mathbf{x}$ is often unknown in practice; often implicit mappings are used, which are only represented by a kernel function. This raises the question how kernel functions can be identified. Using the concept of *reproducing kernel Hilbert spaces* or *Mercer's theorem*, it can be shown, that for any positive semi-definite function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ a mapping ϕ can be constructed such that the inner product between two mapped instances is equal to k (see, *e.g.,* Schölkopf & Smola, 2002, Chapter 2.2.2 and 2.2.4). This justifies to use flexible classes of distributions, that are only implicitly represented by an inner product k; the corresponding statistic ϕ can be high- or even infinite-dimensional. An example of a kernel function for which an explicit form of ϕ is unknown, is the *radial basis function* (RBF) kernel. It is given by

$$k_{rbf}(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2\varsigma^2}\left\|\mathbf{x} - \mathbf{x}'\right\|^2\right), \tag{2.24}$$

with bandwidth parameter $\varsigma > 0$. The concept of kernel function can be generalized to abstract instances $x \in \mathcal{X}$ such as graphs, sequences, or texts. A more detailed discussion is given by Schölkopf & Smola (2002).

2.2 Learning a Regression Model

In the previous section, we state the MAP estimate of $\boldsymbol{\theta}^*$ under the logistic model assumption for finite label spaces $\mathcal{Y}$. This section presents Bayesian linear regression following Rasmussen & Williams (2006) for the case that $\mathcal{Y}$ is continuous.

Assumptions of the Bayesian Linear Regression Model

Let assume that the label y of a given instance $\mathbf{x}$ is generated by a linear model $f_{\boldsymbol{\theta}^*}(\mathbf{x}) = \mathbf{x}^\mathsf{T}\boldsymbol{\theta}^*$ and perturbed by additive Gaussian noise with zero mean and fixed but unknown variance σ^2:

$$f_{\boldsymbol{\theta}^*}(\mathbf{x}) - y \sim \mathcal{N}(0, \sigma^2). \tag{2.25}$$

The true model parameters $\boldsymbol{\theta}^*$ are unknown but assumed to be drawn from a normal distribution $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{0}, \boldsymbol{\Sigma})$. We now derive the predictive distribution $p(y|\mathbf{x}, T_n)$ (see Equation 2.3) under these assumptions. Following Equation 2.25, the model-based predictive distribution under the perturbed Gaussian model is given by

$$p(y|\mathbf{x}; \boldsymbol{\theta}) = \mathcal{N}(y|f_{\boldsymbol{\theta}}(\mathbf{x}), \sigma^2).$$

Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{d \times n}$ define the matrix of instances and $(y, \ldots, y_n)^\mathsf{T} \in \mathcal{Y}^n$ the vector of the corresponding labels. Then, the label likelihood can be expressed by a multivariate Gaussian distribution (see Equation 2.16) by suitable algebraic manipulation:

$$\begin{aligned}
\prod_{i=1}^{n} p(y_i|\mathbf{x}_i; \boldsymbol{\theta}) &= \prod_{i=1}^{n} \mathcal{N}(y_i|f_{\boldsymbol{\theta}}(\mathbf{x}_i), \sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - \mathbf{x}_i^\mathsf{T}\boldsymbol{\theta}\right)^2\right) \\
&= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i^\mathsf{T}\boldsymbol{\theta}\right)^2\right) \\
&= (2\pi)^{-n/2}\left|\sigma^2\mathbf{I}\right|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{X}^\mathsf{T}\boldsymbol{\theta})^\mathsf{T}(\sigma^2\mathbf{I})^{-1}(\mathbf{y} - \mathbf{X}^\mathsf{T}\boldsymbol{\theta})\right) \\
&= \mathcal{N}(\mathbf{y}|\mathbf{X}^\mathsf{T}\boldsymbol{\theta}, \sigma^2\mathbf{I}),
\end{aligned} \tag{2.26}$$

where $\mathbf{I} \in \mathbb{R}^{d \times d}$ denotes the identity matrix of size d. Since $\boldsymbol{\theta}$ is assumed to be drawn from a Gaussian distribution, the posterior is also normally distributed. This yields a closed-form solution of the MAP estimate. The following propositions state the MAP estimate of $\boldsymbol{\theta}^*$ and the posterior predictive distribution.

Proposition 2.2 (Maximum a Posteriori for Bayesian Linear Regression)**.** *If the likelihood* $p(T_n|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}|\mathbf{X}^\mathsf{T}\boldsymbol{\theta}, \sigma^2\mathbf{I})$ *and the prior* $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{0}, \boldsymbol{\Sigma})$ *are Gaussian, the posterior distribution* $p(\boldsymbol{\theta}|T_n) = \mathcal{N}(\boldsymbol{\theta}|\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\Sigma}})$ *is also Gaussian. The maximum a posteriori estimate of* $\boldsymbol{\theta}^*$ *is given by*

$$\begin{aligned}\hat{\boldsymbol{\theta}}^{map} &= \arg\max_{\boldsymbol{\theta} \in \Theta} p(\boldsymbol{\theta}|T_n) \\ &= \bar{\boldsymbol{\theta}},\end{aligned}$$

where $\bar{\boldsymbol{\theta}} = \sigma^{-2}\bar{\boldsymbol{\Sigma}}\mathbf{X}\mathbf{y}$ *and* $\bar{\boldsymbol{\Sigma}} = \left(\sigma^{-2}\mathbf{X}\mathbf{X}^\mathsf{T} + \boldsymbol{\Sigma}^{-1}\right)^{-1}$.

The predictive distribution reflects the remaining uncertainty about y caused by the label noise σ^2 and the uncertainty as a result of estimating the model parameters from a finite sample $\mathbf{X}$. It is given by Proposition 2.3.

Proposition 2.3 (Posterior Predictive Distribution for Bayesian Linear Regression)**.** *If the posterior* $p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}) = \mathcal{N}(\boldsymbol{\theta}|\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\Sigma}})$ *is Gaussian, the predictive distribution for a new* $\mathbf{x} \sim p(\mathbf{x})$ *is given by*

$$p(y|\mathbf{x}, T_n) = \mathcal{N}\left(y|\mathbf{x}^\mathsf{T}\bar{\boldsymbol{\theta}}, \tau_\mathbf{x}^2\right), \tag{2.27}$$

where $\tau_\mathbf{x}^2 = \sigma^2 + \mathbf{x}^\mathsf{T}\bar{\boldsymbol{\Sigma}}\mathbf{x}$.

Proposition 2.2 and 2.3 can be proven by making use of the Gaussian identities (see, *e.g.,* O'Hagan, 1978).

Since the posterior distribution (see Proposition 2.3) is symmetric and unimodal, the mode and the mean coincide. Therefore, the Bayes optimal solution and the maximum a posteriori estimate lead to the same prediction $f_{\boldsymbol{\theta}}(\mathbf{x})$. However, the predictive distribution provides us with an estimate $\tau_\mathbf{x}^2$ of the variance at instance $\mathbf{x}$.

In contrast to Section 2.1, the maximum a posteriori estimate is given by a closed-form solution and can be calculated efficiently. However, hyperparameters, such as the degree of label noise σ^2, have to be determined when the model is applied in practice. Therefore, many selection criteria have been proposed in the statistics

literature, such as AIC (Akaike, 1974) and BIC (Schwarz, 1978); they assess both the likelihood and the complexity of the model. Alternatively, hyperparameters can be tuned by cross-validation (Weiss & Kulikowski, 1990) or by maximizing the marginal likelihood

$$p(\mathbf{y}|\mathbf{X}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \sigma^2\mathbf{I} + \mathbf{X}^\mathsf{T}\boldsymbol{\Sigma}\mathbf{X})$$

as described, for example, by Mardia & Marshall (1984). In the context of Gaussian processes the gradient of the marginal likelihood can be analytically derived (Rasmussen & Williams, 2006, Chapter 5.4).

Kernel Functions for Regression

Non-linear relationships are modeled in analogy to Section 2.1: The predictive distribution given by Equation 2.27 is reformulated, such that it depends on the data only through inner products. Replacing the inner product by any arbitrary kernel maps the data points implicitly into another Hilbert space. In the context of regression, the kernel function is referred to as covariance function.

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a covariance function. A kernelized version of the predictive distribution is given by

$$p(y|\mathbf{x}, T_n) = \mathcal{N}\left(y \middle| \mathbf{k}_\mathbf{x}^\mathsf{T}\left(\mathbf{K} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{y}, k(\mathbf{x},\mathbf{x}) - \mathbf{k}_\mathbf{x}^\mathsf{T}\left(\mathbf{K} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{k}_x\right),$$

where $\mathbf{K} = (k(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1,\ldots,n}$ denotes the kernel matrix of the training data and $\mathbf{k}_\mathbf{x} = (k(\mathbf{x}, \mathbf{x}_i))_{i=1,\ldots,n}$ the vector of covariances between an instance $\mathbf{x}$ and the training instances (see, *e.g.,* Rasmussen & Williams, 2006, Chapter 2.2). The linear covariance function

$$k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\mathsf{T}\Sigma\mathbf{x}_j$$

corresponds to the result of Proposition 2.3.

For regression a squared exponential covariance, which is closely related to the RBF kernel in a classification setting, underestimates the variance of the predictive distribution (Stein, 1999, Chapter 2.7). Instead, a popular choice is the more general class of Matérn kernel functions

$$k_{mat}(\mathbf{x}, \mathbf{x}') = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}}{l}\|\mathbf{x} - \mathbf{x}'\|\right)^\nu K_\nu\left(\frac{\sqrt{2\nu}}{l}\|\mathbf{x} - \mathbf{x}'\|\right), \tag{2.28}$$

where K_ν is the modified Bessel function of the second kind (Abramowitz & Stegun, 1964, Chapter 9.6) of degree ν and Γ is the gamma function. The parameter $\nu > 0$ controls the degree of smoothness and $l > 0$ is the characteristic length-scale. The Matérn kernel coincides with the squared exponential kernel as the degrees of freedom approach infinity.

2.3 Active Learning

In the previous sections, we have discussed how to estimate the conditional distribution $p(y|\mathbf{x})$ for classification and regression. The conditional distribution $p(y|\mathbf{x})$ was approximated by a model-based distribution $p(y|\mathbf{x};\boldsymbol{\theta})$; the optimal model parameters $\boldsymbol{\theta}^*$ are estimated from a set of labeled instances. In many application scenarios that require learning a predictive model, unlabeled instances $\mathbf{x}$ are readily available whereas acquiring labels y that are distributed according to the true conditional distribution $p(y|\mathbf{x};\boldsymbol{\theta}^*)$ is a costly process. Throughout this thesis, we focus on *pool-based* settings in which a large pool D_m of m unlabeled instances is available. The pool is assumed to be drawn i.i.d. according to the distribution $p(\mathbf{x})$. Instances from this pool can be sampled and then labeled according to $p(y|\mathbf{x};\boldsymbol{\theta}^*)$ by an oracle at a cost. If the pool D_m is too large to label the complete set or an exhausted labeling does not justify the costs, a limited labeling budget $n \ll m$ is typically defined. An obvious approach is to label n instances drawn uniformly from the pool D_m and use these instances as training set T_n to learn the predictive model; this strategy is referred to as *passive learning*. However, instances need not necessarily be drawn uniformly from the pool.

The research field of *active learning* in the machine learning literature (MacKay, 1992; Cohn, 1996) and *optimal experimental design* in the statistics literature (Fedorov, 1972) address the problem of selecting a subset of instances from the pool D_m that yields a more accurate model than passive learning. The optimal strategy would be to choose the subset which yields the model with highest predictive performance. However, to calculate this strategy the unknown test distribution $p(\mathbf{x}, y; \boldsymbol{\theta}^*) = p(y|\mathbf{x};\boldsymbol{\theta}^*)p(\mathbf{x})$ needs to be known. In a pool-based setting, the *empirical distribution*

$$\hat{p}(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m}[\![\mathbf{x} = \mathbf{x}_i]\!] \tag{2.29}$$

defined over the pool D_m provides an estimate of the distribution $p(\mathbf{x})$. Es-

Algorithm 1: Active Learning

input Pool D_m, labeling budget n.

1: Initialize estimate of model parameters $\hat{\boldsymbol{\theta}}_0$
2: **for** $i = 0, \ldots, n-1$ **do**
3: Draw $\mathbf{x}_i \sim q_{\hat{\boldsymbol{\theta}}_i}(\mathbf{x})$ from D_m based on the estimate $\hat{\boldsymbol{\theta}}_i$.
4: Query label $y_i \sim p(y|\mathbf{x}_i)$ from oracle.
5: Update estimate $\hat{\boldsymbol{\theta}}_{i+1}$.
6: **end for**

output Estimate of the model parameter $\hat{\boldsymbol{\theta}}_n$.

timating the conditional distribution $p(y|\mathbf{x}; \boldsymbol{\theta}^*)$ is difficult, since it is precisely the quantity we want to estimate by learning a predictive model. To solve this "chicken and egg" problem, active learning algorithms typically alternate between selecting instances to label and estimating the model parameters. The selection strategy can be defined by an instrumental distribution $q_{\boldsymbol{\theta}}(\mathbf{x})$ that describes the probability for choosing the next instance $\mathbf{x}$ to label based on a model $\boldsymbol{\theta}$. Algorithm 1 summarizes the typical protocol: Given an initial estimate $\hat{\boldsymbol{\theta}}_0$ of $\boldsymbol{\theta}^*$, an instance can be drawn from $q_{\hat{\boldsymbol{\theta}}_0}(\mathbf{x})$. After the label y is queried from an oracle, a learning algorithm is applied to the enlarged set of labeled instances to obtain a refined estimate $\hat{\boldsymbol{\theta}}_{i+1}$. This procedure is repeated until the labeling budget n is exhausted.

In the following, we give a brief overview of popular active learning algorithms. They can be differentiated as to whether the sampling strategy $q_{\boldsymbol{\theta}}(\mathbf{x})$ is deterministic or probabilistic. In Section 2.3.1, we analyze active learning algorithms which choose the next instance to label using a deterministic criterion. In contrast, Section 2.3.2 presents sampling distributions minimizing some trade-off between the variance and the bias of the parameters estimate.

2.3.1 Active Learning with a Deterministic Sampling Strategy

The choice of the selection strategy $q_{\boldsymbol{\theta}}(\mathbf{x})$ in Algorithm 1 (Line 3), which is applied to query the next instance to label, is crucial for the success of active over passive learning. Let $\hat{\boldsymbol{\theta}}_i$ be the estimate of the model parameters $\boldsymbol{\theta}^*$ after having seen i instance-label pairs and let $\hat{\boldsymbol{\theta}}_i^{\mathbf{x},y}$ be the estimate which is based on an additional instance $\mathbf{x}$ with label y. Schohn & Cohn (2000), Roy & McCallum

(2001), and Chapelle (2005) propose to query the label $\bar{y}$ of that instance $\bar{\mathbf{x}}$ such that the model $f_{\hat{\boldsymbol{\theta}}_i^{\bar{\mathbf{x}},\bar{y}}}$ is most accurate in expectation over the unknown label $\bar{y}$, that is,

$$\bar{\mathbf{x}} = \arg\min_{\dot{\mathbf{x}} \in D} \sum_{\dot{y} \in \mathcal{Y}} \mathbb{E}_{(\mathbf{x},y)\sim p(\mathbf{x},y;\boldsymbol{\theta}^*)} \left[y \neq f_{\hat{\boldsymbol{\theta}}_i^{\dot{\mathbf{x}},\dot{y}}}(\mathbf{x}) \right] p(\dot{y}|\dot{\mathbf{x}};\boldsymbol{\theta}^*). \tag{2.30}$$

In order to implement the query strategy, the authors approximate the distribution $p(y|\mathbf{x};\boldsymbol{\theta}^*)$ by the estimate $p(y|\mathbf{x};\hat{\boldsymbol{\theta}}_i)$ provided by the current model; the marginal distribution $p(\mathbf{x})$ is estimated by $\hat{p}(\mathbf{x})$ given by Equation 2.29 in a pool-based setting. Then, Equation 2.30 can be evaluated by learning a model with parameters $\boldsymbol{\theta}_i^{\dot{\mathbf{x}},\dot{y}}$ for each instance $\dot{\mathbf{x}} \in D$ and each feasible label $\dot{y} \in \mathcal{Y}$. Having labeled $\bar{\mathbf{x}}$ with $\bar{y}$, the estimate $\hat{\boldsymbol{\theta}}_i$ is replaced by the new estimate $\hat{\boldsymbol{\theta}}_i^{\bar{\mathbf{x}},\bar{y}}$, and so on. If the pool or the label space is large, this approach becomes computationally intractable.

An alternative strategy is to query the label of the instance $\bar{\mathbf{x}}$, for which the prediction of the current model $f_{\hat{\boldsymbol{\theta}}_i}(\bar{\mathbf{x}})$ is least likely to be correct, that is

$$\bar{\mathbf{x}} = \arg\min_{\mathbf{x} \in D} \mathbb{E}_{y\sim p(y|\mathbf{x};\boldsymbol{\theta}^*)} \left[y = f_{\hat{\boldsymbol{\theta}}_i}(\mathbf{x}) \right]. \tag{2.31}$$

Approximating $p(y|\mathbf{x};\boldsymbol{\theta}^*)$ by the current model leads to a simple sampling heuristic that selects instances for which the prediction of the model is least confident. This strategy is known as *uncertainty sampling*. In a classification scenario, the instance with lowest confidence is given by

$$\bar{\mathbf{x}} = \arg\min_{\mathbf{x} \in D} \max_{y \in \mathcal{Y}} p\big(y\big|\mathbf{x};\hat{\boldsymbol{\theta}}_i\big).$$

For regression problems under the assumption of a Gaussian predictive distribution $p(y|\mathbf{x};\boldsymbol{\theta}) = \mathcal{N}(y|f_{\boldsymbol{\theta}}(\mathbf{x}),\tau_{\mathbf{x}}^2)$ with predictive variance $\tau_{\mathbf{x}}^2$ for instance $\mathbf{x}$ (see Section 2.2), the most informative instance in the sense of Equation 2.31 is given by the maximal label variance of the predictive distribution

$$\bar{\mathbf{x}} = \arg\max_{\mathbf{x} \in D} \tau_{\mathbf{x}}^2. \tag{2.32}$$

Uncertainty sampling is studied for several learning algorithms as SVMs (Tong & Koller, 2002), logistic regression (Lewis & Gale, 1994), and Gaussian processes (Kapoor et al., 2007). Further uncertainty measures are also examined in the literature: Dagan & Engelson (1995) use the entropy and Scheffer et al. (2001) consider the difference between most and second most likely label; they coincide with Equation 2.31 in the case of a binary classification task. *Query by committee* (Seung et al., 1992) is a related approach, where the label uncertainty of an instance is assessed by measuring the disagreement among a committee of models rather than using the predictive distribution of a single model. The committee of models can either be sampled from the posterior distribution (McCallum & Nigam, 1998) or obtained by boosting and bagging techniques (Abe & Mamitsuka, 1998).

In general, there is a trade-off between selecting instances to refine the current model and exploring the whole support of $p(x)$. Uncertainty sampling algorithms focus only on instances whose labels are most uncertain and thus tend to discover only a small region of the instance space. Consequently, the resulting model may approximate the conditional distribution $p(y|\mathbf{x};\boldsymbol{\theta}^*)$ poorly for some regions. If such a region has high density $p(\mathbf{x})$, this is known as *missed-cluster effect* (Schütze et al., 2006). To tackle this problem, several dual strategies were proposed in the literature (Osugi et al., 2005; Pandey et al., 2005; Donmez et al., 2007). Basically, they decide in each iteration between passive and uncertainty sampling following some heuristic. As another approach, Nguyen & Smeulders (2004) and Dasgupta & Hsu (2008) propose to first cluster the instance space $p(\mathbf{x})$ and then use the cluster assignments to ensure that informative instances from all regions are be labeled. Finally, active learning strategies have been proposed for a broad range of other learning tasks such as information extraction (Scheffer et al., 2001), ranking (Long et al., 2010), and time-series analysis (Singh et al., 2005). Convergence bounds for active learning with deterministic sampling strategy are derived by, *e.g.,* Dasgupta (2006), Castro & Nowak (2007), and Hanneke (2011). A detailed overview is given by Settles (2009).

In this section, we have addressed strategies to decide which instance has to be labeled. The presented approaches focus on instances whose label is least likely to be predicted correctly. If examples are selected deterministically in order to minimize the labeling effort, the resulting sample is irreversible biased according to the test distribution. Hence, any estimate of the model's performance on such an actively drawn sample is pessimistically biased (see, *e.g.,* Schütze et al., 2006).

2.3.2 Active Learning with an Instrumental Distribution

In the previous section, we considered deterministic sampling strategies in order to determine most informative instance-label pairs. An alternative procedure is to draw instances to label randomly from an instrumental distribution $q(\mathbf{x})$ rather than from the input distribution $p(\mathbf{x})$. Situations in which the training instances are governed by a distribution that differs from the test distribution $p(\mathbf{x})$ are known as learning under *covariate shift*. Intuitively, the instrumental distribution $q(\mathbf{x})$ should be chosen such that the resulting estimate $p(y|\mathbf{x};\hat{\boldsymbol{\theta}})$ is as close as possible to the conditional distribution $p(y|\mathbf{x};\boldsymbol{\theta}^*)$ for a fixed labeling budget n. The best-performing model maximizes the theoretical label likelihood (see Equation 2.8). We have seen (see Equation 2.11), that maximizing the theoretical likelihood is equivalent to minimizing the expected Kullback-Leibler divergence:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\Theta} \exp\left(\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\mathrm{KL}\left[p(y|\mathbf{x};\boldsymbol{\theta}^*)||p(y|\mathbf{x};\boldsymbol{\theta})\right]\right]\right). \tag{2.33}$$

Since $\mathrm{KL}[p||p']$ is zero if and only if $p = p'$, the minimum of optimization problem given by Equation 2.33 is attained independently of the marginal distribution of x if the model space is correctly specified, that is, $\boldsymbol{\theta}^* \in \Theta$. Thus, $\hat{\boldsymbol{\theta}}^{ml}$ maximizes asymptotically the theoretical label likelihood even if the instances are drawn from $q(\mathbf{x}) \neq p(\mathbf{x})$. In practice, however, it cannot be ensured that the model space contains the true model $\boldsymbol{\theta}^*$, because the model might be misspecified. In this case, Equation 2.33 is no longer independent of the marginal distribution $p(\mathbf{x})$ and thus the maximum likelihood estimate does not necessarily converge to the optimal parameters with respect to Equation 2.8.

In order to estimate $p(y|\mathbf{x};\boldsymbol{\theta}^*)$ accurately by a misspecified model under covariate shift, we can make use of the *weighted maximum likelihood* (WML) estimate

$$\hat{\boldsymbol{\theta}}^{wml} = \arg\max_{\boldsymbol{\theta}\in\Theta} \sum_{i=1}^{n} w(\mathbf{x}_i) \log p(y_i|\mathbf{x}_i;\boldsymbol{\theta}). \tag{2.34}$$

The non-negative weighting function $w : \mathbb{R}^d \to \mathbb{R}^+$ is fixed and quantifies the relative importance of an instance $\mathbf{x}$. Shimodaira (2000) and Wiens (2000) show that the weighted maximum likelihood $\hat{\boldsymbol{\theta}}^{wml}$ converges to the maximal theoretical label likelihood if the weighting function $w_u(\mathbf{x}) = \frac{p(\mathbf{x})}{q(\mathbf{x})}$ is chosen to be the ratio of the input and the instrumental distribution. Under this choice, Wiens (2000) and Kanamori & Shimodaira (2003) derive an instrumental distribution $q^*(\mathbf{x})$ that minimizes the variance of $\hat{\boldsymbol{\theta}}^{wml}$ for regression.

Although the weights $w_u(\mathbf{x})$ are asymptotically optimal, non-homogeneous importance weights increase the variance of $\hat{\boldsymbol{\theta}}^{wml}$ for finite training set sizes. Therefore, Bach (2006) studies a smoothed weighting function

$$w(\mathbf{x}) = \left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right)^{\eta},$$

with trade-off parameter $\eta \in [0, 1]$. He derives a two-step method to estimate the parameters of a generalized linear model. In each iteration, the instrumental distribution $q^*(\mathbf{x})$ that maximizes the expected performance gain is computed. After an instance is drawn from $q^*(\mathbf{x})$ and labeled, η is determined by a grid search such that the predictive performance is maximal.

A closely related task to active learning with an instrumental distribution is learning from streaming data, where the learning algorithm observes in each step an unlabeled instance and has to decide whether to query the label or not. Beygelzimer et al. (2009) use importance weighting to correct the sampling bias and propose a rejection sampling distribution, which controls the parameters' variance. The presented theoretical optimal distributions again involve unknown quantities depending on the true conditional $p(y|\mathbf{x}; \boldsymbol{\theta}^*)$. To determine if the label of the $(i + 1)$-th incoming example have to be queried, the authors propose to approximate $p(y|\mathbf{x}; \boldsymbol{\theta}^*) \approx p(y|\mathbf{x}; \hat{\boldsymbol{\theta}}_i)$ by the model $\hat{\boldsymbol{\theta}}_i$ learned so far.

In this chapter, we have seen how a predictive model can be learned from data. Before a predictive model can be shipped and deployed, an estimate of the predictive performance is required. In the next chapter, we formalize the concept of predictive performance and show what conclusion about this quantity can be drawn from an estimate, which is based on a finite set of instances.

Chapter 3

Evaluation of Predictive Models

This thesis addresses the problem of evaluating a given predictive model as accurately as possible in situations in which labeled instances, which reflect the test distribution, are unavailable. Before we investigate the case in which labels have to be queried, this chapter introduces the fundamental concepts of model evaluation and comparison based on a given sample.

Learning and evaluating can both be seen as instances of statistical inference. In principle there are two perspectives on statistical inference. From a Bayesian point of view, one considers one fixed data set; the underlying parameters are unknown and the subjective beliefs about them are described probabilistically. In the last chapter, we have followed this perspective in the context of learning predictive models, since it is considered as natural, when combining prior knowledge of domain experts and observations; the goal was to infer the best model using the available data. In this chapter, we turn to the frequentistic view. Frequentists assume that the parameters are fixed and consider the observed data set, which is drawn from some underlying distribution, as random variable. This view might be more appropriate when evaluating a model, since the analysis is unconditioned on the current data set and thus corresponds to multiple settings in which a model is used; the uncertainty about the performance statements is derived from the fact that we have observed only one data set. Note that the philosophical differences between the Bayesian and frequentist paradigm have no impact on the proposed evaluation methods; they can be applied to any statistical model.

In contrast to Chapter 2, we make no model assumptions about the data generating process in this context. Therefore, we denote the true distribution of the observed data by $p(x, y) = p(y|x)p(x)$ instead of $p(y|x; \boldsymbol{\theta}^*)p(x)$; the notation $p(y|x; \boldsymbol{\theta})$ refers to the model-based predictive distribution induced by a

certain predictive model $\boldsymbol{\theta}$. The disagreement between a prediction $f_{\boldsymbol{\theta}}(x)$ and a true label y for an instance x is measured by a problem-specific loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. For classification, the zero-one loss $\ell_{0/1}(y, \bar{y}) = [\![y \neq \bar{y}]\!]$ is a widely-used choice; it equals one if prediction and true label differ, and is zero otherwise. For regression, the quadratic loss $\ell_2(\bar{y}, y) = (\bar{y} - y)^2$ is a standard choice. The *risk* functional constitutes a common theoretical quantity to measure predictive performance of a predictive model $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}$ with respect to the test distribution $p(x, y)$. It is defined as the expectation of the loss function taken over $p(x, y)$:

$$\begin{aligned} R[f_{\boldsymbol{\theta}}] &= \mathbb{E}_{(x,y)\sim p(x,y)} \left[\ell(f_{\boldsymbol{\theta}}(x), y)\right] \\ &= \iint \ell(f_{\boldsymbol{\theta}}(x), y) p(x, y) \mathrm{d}y \, \mathrm{d}x. \end{aligned} \tag{3.1}$$

In a classification setting, the integral over $\mathcal{Y}$ reduces to a finite sum. If the context is clear, we refer to $R[f_{\boldsymbol{\theta}}]$ simply by R.

Since the true risk depends on the unknown test distribution $p(x, y)$, the performance of a predictive model $f_{\boldsymbol{\theta}}$ is typically estimated from a sample of labeled instances. Common estimators are presented in Section 3.1. Furthermore, we state confidence intervals, which quantify the estimation uncertainty caused by the finiteness of the sample. In order to compare models reliably, we give a brief introduction into testing theory in Section 3.2. Finally, in Section 3.3, we introduce a new generalization of the traditional risk functional and show that the theoretical findings can be extended to its estimator. The generalized risk functional has recently been studied (Sawade et al., 2010b); it additionally captures the F_η-measure which is a commonly used performance measure for prediction problems with skewed class distributions.

3.1 Estimating the Model's Risk

In general, an estimate $\hat{R}_n$ is an approximation of a quantity R based on a set of instances $x_1, \ldots, x_n$ drawn from a distribution $p(x)$. The procedure of calculating an estimate is called estimator. Since, sampling instances x_i from a distribution $p(x)$ is a random process, an estimate is a random variable, whose distribution depends on $p(x)$. The quality of the approximation $\hat{R}_n$ can be quantified by the squared deviation of the estimator $\hat{R}_n$ from the true value R in

expectation over the drawn sample:

$$\mathrm{MSE}_{x_i \sim p(x)} \left[\hat{R}_n \right] = \mathbb{E}_{x_i \sim p(x)} \left[\left(\hat{R}_n - R \right)^2 \right]. \tag{3.2}$$

Equation 3.2 is referred to as *estimation error*. A minimum requirement for an estimator is *consistency*. Intuitively, an estimator is consistent, if it indeed calculates the quantity to be estimated. In order to define the concept of consistency formally, we need to introduce the notions of convergence of random variables:

Definition 3.1 (Convergence of Random Variables)**.** *Let $X_1, \ldots, X_n$ be a sequence of random variables, X a single random variable, and $F_i(x)$ and $F(x)$ their cumulative distribution functions.*

- *The sequence is said to converge* almost surely *to X, if $\lim_{n\to\infty} X_n = X$ holds with probability one. Almost sure convergence is denoted by $X_n \xrightarrow{as} X$.*
- *The sequence is said to converge* in distribution *to X, if $\lim_{n\to\infty} F_n(x) = F(x)$. Convergence in distribution is denoted by $X_n \xrightarrow{d} X$, or $X_n \xrightarrow{d} p_X$, where p_X is the distribution of X.*

Almost sure convergence implies convergence in distribution (see, *e.g.,* Van der Vaart, 2000, Theorem 2.7). An estimate $\hat{R}_n$ is (strongly) consistent, if the random sequence of estimates $\hat{R}_1, \ldots, \hat{R}_n$ converges almost surely to R. Thus, the estimation error MSE vanishes for $n \to \infty$ and $\hat{R}_n$ can indeed be seen as an estimate of R.

Another two quantities to investigate the sources of the estimation error for finite sample sizes are the *bias* and the *variance* of an estimator. The variance

$$\mathrm{Var}_{x_i \sim p(x)} \left[\hat{R}_n \right] = \mathbb{E}_{x_i \sim p(x)} \left[\left(\hat{R}_n - \mathbb{E}_{x_i \sim p(x)} \left[\hat{R}_n \right] \right)^2 \right] \tag{3.3}$$

measures the amount of variation of the estimator and the bias

$$\mathrm{Bias}_{x_i \sim p(x)} \left[\hat{R}_n \right] = \mathbb{E}_{x_i \sim p(x)} \left[\hat{R}_n \right] - R \tag{3.4}$$

quantifies the systematic deviation from the value being estimated. If the bias is zero, the estimate is said to be unbiased. The estimation error can be expressed in terms of the bias and the variance: In Equation 3.6, we make use of the definition of the estimation error (see Equation 3.2), expand the square, and

add and subtract the expected value of $\hat{R}_n$. Reordering terms and inserting the definition of the variance (see Equation 3.3) and the bias (see Equation 3.4) yield Equation 3.7.

$$\begin{aligned} &\mathrm{MSE}_{x_i \sim p(x)} \left[\hat{R}_n \right] && (3.5) \\ &= \left(\mathbb{E}_{x_i \sim p(x)} \left[\hat{R}_n^2 - 2R\hat{R}_n + R^2 \right] \right) + \mathbb{E}_{x_i \sim p(x)} \left[\hat{R}_n \right]^2 - \mathbb{E}_{x_i \sim p(x)} \left[\hat{R}_n \right]^2 && (3.6) \\ &= \left(\mathrm{Bias}_{x_i \sim p(x)} \left[\hat{R}_n \right] \right)^2 + \mathrm{Var}_{x_i \sim p(x)} \left[\hat{R}_n \right]. && (3.7) \end{aligned}$$

Equation 3.7 is known as *bias-variance decomposition* (Geman et al., 1992).

An estimate of the risk R (see Equation 3.1) can be obtained by replacing the unknown distribution $p(x, y)$ by an empirical distribution. Given n instance-label pairs (x_i, y_i) drawn from $p(x, y)$, the joint empirical distribution over the pairs can be defined in analogy to Equation 2.29:

$$\hat{p}(x, y) = \frac{1}{n} \sum_{i=1}^{n} [\![(x, y) = (x_i, y_i)]\!]. \tag{3.8}$$

The empirical distribution converges uniformly to the test distribution, in the sense that $\hat{p}(x, y)$ converges for any pair (x, y) almost surely to $p(x, y)$, whereby the speed of convergence is independent of the considered pair (see, *e.g.,* Van der Vaart, 2000, Theorem 19.1). Inserting Equation 3.8 into Equation 3.1 yields the *empirical risk*, given by an average over the instance-specific losses $\ell(f_{\boldsymbol{\theta}}(x_i), y_i)$:

$$\hat{R}_n[f_{\boldsymbol{\theta}}] = \frac{1}{n} \sum_{i=1}^{n} \ell(f_{\boldsymbol{\theta}}(x_i), y_i). \tag{3.9}$$

The empirical risk $\hat{R}_n$ is an unbiased estimate of R. This can be seen by using the linearity of the expected value and the definition of R:

$$\begin{aligned} \mathbb{E}_{(x_i, y_i) \sim p(x,y)} \left[\hat{R}_n \right] &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{(x,y) \sim p(x,y)} \left[\ell(f_{\boldsymbol{\theta}}(x), y) \right] \\ &= R. \end{aligned}$$

Thus, it follows from Equation 3.7 that the estimation error results solely from the estimator's variance. In the following, we present the self-normalized importance sampling estimator, which yields a consistent estimate of R based on

a sample drawn from an almost arbitrarily but known instrumental distribution $q(x)$. Consistency means asymptotic unbiasedness; that is, the expected value of the estimate $\hat{R}$ converges almost surely to the true value R. Although introducing an asymptotically vanishing bias, this estimator gives us the opportunity to carefully choose instances to obtain an estimator with lower variance and thus a more accurate estimate.

3.1.1 Self-Normalized Importance Sampling Estimator

Estimating the expected value of a rarely occurring outcome of a random variable based on a set of instances drawn directly from the underlying distribution can be inadequate. Assume that a model $f_{\boldsymbol{\theta}}$ is assessed in terms of the zero-one loss $\ell_{0/1}$. If the risk $R[f_{\boldsymbol{\theta}}]$ is very low, it is unlikely that an instance x with $\ell(f_\theta(x), y) = 1$ occurs in the finite test set; it requires a large number of instances to estimate R with high confidence. Test instances (x_i, y_i) need not necessarily be drawn according to the distribution $p(x, y)$. An instrumental $q(x, y)$ may be available that highlights crucial instances. In this section, we introduce the concept of *importance sampling*. Importance sampling is a general technique to estimate an unknown quantity using test instances drawn from an instrumental distribution instead of $p(x, y)$. A precondition for the instrumental distribution is that any instance (x, y) that can be drawn from $p(x, y)$ can also be drawn from the instrumental distribution $q(x, y)$. This condition is formalized in Definition 3.2.

Definition 3.2 (Absolutely Continuous)**.** *Let $p(x)$ and $q(x)$ be distributions defined over a set $\mathcal{X}$. The distribution $q(x)$ is said to be* absolutely continuous *with respect to $p(x)$ if $p(x) > 0$ implies $q(x) > 0$ for all $x \in \mathcal{X}$.*

In the following, we derive a consistent estimator of the risk, when instances are selected according to an instrumental distribution. For the purpose of this thesis, we focus on instrumental distributions $q(x, y) = p(y|x)q(x)$ to select unlabeled instances x to label; the corresponding label will be drawn according to $p(y|x)$. Let $q(x)$ be absolutely continuous with respect to the distribution $p(x)$. Then, the risk defined over the test distribution can be expressed as expectation of $\ell(f_{\boldsymbol{\theta}}(x), y)$ taken over the instrumental distribution $q(x, y) = p(y|x)q(x)$ by weighting the instance-specific losses by the *Radon-Nikodym derivatives* $\frac{p(x)}{q(x)}$

of $p(x)$ with respect to $q(x)$:

$$\begin{aligned} R[f_{\boldsymbol{\theta}}] &= \iint \ell(f_{\boldsymbol{\theta}}(x), y) p(x, y) \mathrm{d}y \, \mathrm{d}x \\ &= \iint \ell(f_{\boldsymbol{\theta}}(x), y) p(y|x) p(x) \frac{q(x)}{q(x)} \mathrm{d}y \, \mathrm{d}x \\ &= \mathbb{E}_{(x,y)\sim q(x,y)} \left[\frac{p(x)}{q(x)} \ell(f_{\boldsymbol{\theta}}(x), y) \right]. \end{aligned} \tag{3.10}$$

Replacing the distribution $q(x, y)$ in Equation 3.10 by its empirical counterpart $\hat{q}(x, y)$ (see Equation 3.8) induced by n instance-label pairs (x_i, y_i) drawn from $q(x)p(y|x)$ yields an estimator of the true risk:

$$\hat{R}_{n,q}[f_{\boldsymbol{\theta}}] = \left(\sum_{i=1}^{n} \frac{p(x_i)}{q(x_i)} \right)^{-1} \sum_{i=1}^{n} \frac{p(x_i)}{q(x_i)} \ell(f_{\boldsymbol{\theta}}(x_i), y_i). \tag{3.11}$$

Equation 3.11 is referred to as a *self-normalized importance sampling estimator* in the statistics literature (see, *e.g.,* Geweke, 1989; Liu, 2001). The estimator $\hat{R}_n$ (see Equation 3.9) is a special case of $\hat{R}_{n,q}$, using the instrumental distribution $q(x) = p(x)$.

The choice of the instrumental distribution $q(x)$ affects the bias and the variance of the estimator $\hat{R}_{n,q}$ (see Equation 3.7). Hence, for certain sampling distributions $q(x)$, the estimator $\hat{R}_{n,q}$ of the risk R may be a more label-efficient than $\hat{R}_n$. In contrast to the empirical risk $\hat{R}_n$, the estimator $\hat{R}_{n,q}$ is biased, because both the numerator and the denominator depend on the drawn sample. To see this, consider the expected value of Equation 3.11:

$$\mathbb{E}_{(x,y)\sim q(x,y)} \left[\hat{R}_{n,q} \right] = \mathbb{E}_{(x,y)\sim q(x,y)} \left[\frac{\sum_{i=1}^{n} \frac{p(x_i)}{q(x_i)} \ell(f_{\boldsymbol{\theta}}(x_i), y_i)}{\sum_{i=1}^{n} \frac{p(x_i)}{q(x_i)}} \right].$$

The expected value of the numerator is nR, whereas the expectation of the denominator is n. Since the expectation of a ratio $\frac{X}{Y}$ is not necessarily equal to the ratio of the expectations of X and Y, the expected value of $\hat{R}_{n,q}$ differs in general from the risk. Although being biased, Equation 3.11 defines a consistent estimator of the true risk R because the weighting factors $\frac{p(x)}{q(x)}$ compensate for the discrepancy between test and instrumental distributions. To see this, note

that due to the *strong law of large numbers* the quantities

$$\frac{1}{n}\sum_{i=1}^{n}\frac{p(x_i)}{q(x_i)}\ell(f_{\boldsymbol{\theta}}(x_i),y_i) \xrightarrow{as} R \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n}\frac{p(x_i)}{q(x_i)} \xrightarrow{as} 1$$

converge almost surely (see Definition 3.2) to their expected values. Then, Slutsky's Theorem (see, *e.g.,* Cramér, 1946) applied to the numerator and denominator of Equation 3.11 implies that $\hat{R}_{n,q} \xrightarrow{as} R$.

The estimation error also depends on the variance of an estimator, and will play a central role when deriving a cost-efficient sampling distribution $q(x)$. Lemma 3.1 states that $\hat{R}_{n,q}$ is asymptotically normally distributed, and characterizes the variance of the self-normalized importance sampling estimator in the limit.

Lemma 3.1 (Asymptotic Distribution of Estimator)**.** *Let $\hat{R}_{n,q}$ be defined as in Equation 3.11 and let us assume that*

1. *the expected value $R = \mathbb{E}_{(x,y)\sim p(x,y)}\left[\ell(f_{\boldsymbol{\theta}}(x),y)\right]$ exists,*
2. *the variance $\mathrm{Var}_{(x,y)\sim p(x,y)}\left[\ell(f_{\boldsymbol{\theta}}(x),y)\right]$ is finite,*
3. *the distribution $q(x)$ is absolutely continuous with respect to $p(x)$, and*
4. *the weights $\frac{p(x)}{q(x)} \leq E$ are bounded from above by a constant $E < \infty$.*

Then, $\hat{R}_{n,q}$ is asymptotically normally distributed,

$$\sqrt{n}\left(\hat{R}_{n,q} - R\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_q^2\right),$$

with asymptotic variance

$$\begin{aligned}\sigma_q^2 &= \mathbb{E}_{x\sim q(x)}\left[\frac{p(x)}{q(x)}\right]^{-2} \mathbb{E}_{(x,y)\sim q(x)p(y|x)}\left[\left(\frac{p(x)}{q(x)}\right)^2 \left(\ell(f_{\boldsymbol{\theta}}(x),y) - R\right)^2\right] \\ &= \int \left(\frac{p(x)}{q(x)}\right)^2 \left(\int \left(\ell(f_{\boldsymbol{\theta}}(x),y) - R\right)^2 p(y|x)\mathrm{d}y\right) q(x)\mathrm{d}x, \end{aligned} \tag{3.12}$$

where $\xrightarrow{d}$ denotes convergence in distribution.

We omit this proof here and show a more general result in Section 3.3 instead.

Finally, it is worth mentioning that an unbiased estimator can be obtained if the normalizer—the sum of weights $\sum_{i=1}^{n} \frac{p(x_i)}{q(x_i)}$—is replaced by the number of instances n. This is known as the standard (not self-normalized) importance sampling estimator. Although being unbiased, in practice this estimator has often a higher estimation error caused by higher variance induced by the resampling weights (see, *e.g.,* Liu, 2001, Chapter 2.5).

3.1.2 Confidence Intervals

In practice a point estimate $\hat{R}_{n,q}$ of the true error is often not sufficient, since it does not quantify the estimation error. This section states confidence intervals for risk estimators presented in the previous section. Confidence intervals are indicating a region where the true error lies in with certain probability. Specifically, a two-sided confidence interval $[\hat{R}_{n,q} - \varepsilon_\alpha, \hat{R}_{n,q} + \varepsilon_\alpha]$ with coverage $1-\alpha$ indicates that $|R - \hat{R}_{n,q}| < \varepsilon_\alpha$ holds with a predefined probability $1-\alpha$, or equivalently that the probability of observing a deviation of ε_α, or a more extreme value, of the true risk is less than α.

Confidence Intervals for Normally Distributed Estimators

In order to estimate a confidence interval, we analyze the estimator's underlying distribution; the corresponding cumulative distribution quantifies the range capturing the true test error for a certain probability. The assumption of a normally distributed estimator yields the *Wald interval* which is closely related to the commonly used *t-test interval*. We now turn towards the problem of determining the Wald interval for the self-normalized importance sampling estimator $\hat{R}_{n,q}$. Following Lemma 3.1, the statistic

$$\sqrt{n}\frac{\hat{R}_{n,q} - R}{\sigma_q} \sim \mathcal{N}(0,1) \tag{3.13}$$

follows asymptotically a standard normal distribution. In practice, the asymptotic variance σ_q^2 of the estimator is unknown. Substituting the empirical for the true variance yields an observable statistic. The empirical variance is given by

$$S_{n,q}^2 = n \left(\sum_{i=1}^{n} \frac{p(x_i)}{q(x_i)} \right)^{-2} \sum_{i=1}^{n} \left(\frac{p(x_i)}{q(x_i)} \right)^2 \left(\ell(f_{\boldsymbol{\theta}}(x_i), y_i) - \hat{R}_{n,q} \right)^2 . \tag{3.14}$$

It can be derived by replacing the distribution $p(x, y)$ in Equation 3.12 by the empirical distribution function (see Equation 3.8) induced by a labeled sample $(x_1, y_1), \ldots, (x_n, y_n)$. In the case of $p(x) = q(x)$, the sample variance

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(\ell(f_{\boldsymbol{\theta}}(x_i), y_i) - \hat{R}_n \right)^2$$

is an unbiased estimator for σ_p^2 (see *Bessel's correction*). In contrast, the empirical variance $S_{n,q}^2$ is generally biased because both the numerator and the denominator depend on the drawn sample; see discussion in Section 3.1.1 about the estimator $\hat{R}_{n,q}$. However, Lemma 3.2 states that the empirical variance $S_{n,q}$ is a (strongly) consistent estimate of the asymptotic variance σ_q^2.

Lemma 3.2 (Consistency of Empirical Variance)**.** *Under the assumptions of Lemma 3.1 the empirical variance*

$$S_{n,q}^2 \xrightarrow{as} \sigma_q^2$$

converges almost surely to the asymptotic variance.

The claim follows by the strong law of large numbers and Slutsky's theorem. A detailed proof is given for example by Geweke (1989).

Since $S_{n,q}^2$ consistently estimates σ_q^2, the observable statistics

$$\sqrt{n} \frac{R - \hat{R}_{n,q}}{S_{n,q}} \sim \mathcal{N}(0, 1) \tag{3.15}$$

is also asymptotically normally distributed. Hence, the probability α of observing the Statistic 3.15 or a more extreme value is given by the cumulative distribution function of the standard normal distribution (see Figure 3.1, top); the boundaries of the confidence interval ε_α are given by the $100(1-\alpha)$-th percentile of the normal distribution. Lemma 3.3 states the size of the confidence interval for a given confidence level α.

Lemma 3.3 (Wald Interval for Normally Distributed Estimators)**.** *Let $\hat{R}_{n,q}$ be normally distributed with expected value R and estimated variance $S^2_{n,q}$. Then, a two-sided confidence interval $[\hat{R}_{n,q} - \varepsilon_\alpha, \hat{R}_{n,q} + \varepsilon_\alpha]$ with coverage $1 - \alpha$ is given by*

$$\varepsilon_\alpha = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\frac{S_{n,q}}{\sqrt{n}},$$

where Φ^{-1} is the inverse cumulative distribution function of the standard normal distribution.

Proof. Rewriting the coverage probability in terms of the Statistic 3.15 and resolving the absolute value by case differentiation according to the sign of $\hat{R}_{n,q} - R$ yields

$$\begin{aligned}\alpha &= p\left(|\hat{R}_{n,q} - R| > \varepsilon_\alpha\right)\\ &= p\left(\sqrt{n}\frac{\hat{R}_{n,q} - R}{S_{n,q}} > \sqrt{n}\frac{\varepsilon_\alpha}{S_{n,q}}\right) + p\left(\sqrt{n}\frac{R - \hat{R}_{n,q}}{S_{n,q}} > \sqrt{n}\frac{\varepsilon_\alpha}{S_{n,q}}\right).\end{aligned}$$

Let $Z \sim \mathcal{N}(0,1)$ be a random variable that is standard normally distributed. Then, the probability that Z is less than or equal to a certain value z is given by the cumulative distribution function of the the standard normal distribution $\Phi(z)$ and thus

$$\begin{aligned}\alpha &= p\left(Z > \sqrt{n}\frac{\varepsilon_\alpha}{S_{n,q}}\right) + p\left(Z > \sqrt{n}\frac{\varepsilon_\alpha}{S_{n,q}}\right)\\ &= 2\left(1 - p\left(Z \leq \sqrt{n}\frac{\varepsilon_\alpha}{S_{n,q}}\right)\right)\\ &= 2\left(1 - \Phi\left(\sqrt{n}\frac{\varepsilon_\alpha}{S_{n,q}}\right)\right).\end{aligned} \tag{3.16}$$

Finally, the claim follows by solving Equation 3.16 for ε_α. □

Confidence Intervals for Binomial Proportions

The standard confidence interval for $\hat{R}_{n,q}$ presented in Section 3.1.2 is only asymptotically correct (Wasserman, 2004, Section 6.3.2). Although the distribution of $\hat{R}_{n,q}$ converges independently of the choice of ℓ to a normal distribution

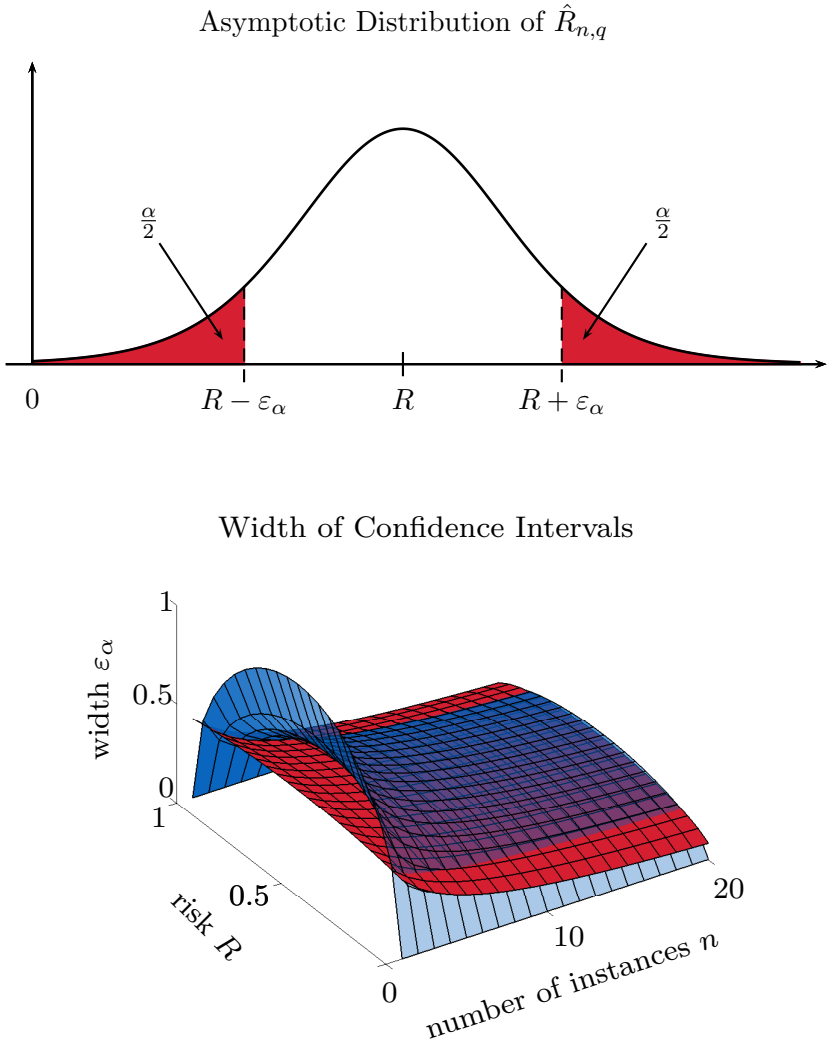


Figure 3.1: Asymptotic distribution of the estimator $\hat{R}_{n,q}$ and designated confidence interval of coverage $1 - \alpha$ (top). The width of the standard (blue) and Wilson (red) confidence interval as a function of the true risk R and the sample size n for $\alpha = 0.05$ (bottom).

(see Lemma 3.1), the convergence rate can be slow if the actual distribution is skewed (see *Berry-Esseen theorem*). This can be the case when a classification model is evaluated with respect to the zero-one loss. In this section, we study the empirical coverage of confidence intervals for $\hat{R}_n$ with $\ell = \ell_{0/1}$ and expectation $R \in [0, 1]$. We will see that the interval suffers strongly from the skewness and the discreteness of the actual binomial distribution of $\hat{R}_n$. Finally, we discuss alternative intervals.

The standard confidence interval (see Lemma 3.3) for a binomially distributed random variable $\hat{R}$ can be expressed as

$$\varepsilon_\alpha^{bin} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\frac{\hat{R}(1-\hat{R})}{n}},$$

where the sampling variance is calculated by $S_n^2 = \hat{R}(1-\hat{R})$. In order to assess the quality of a confidence interval, we define the *empirical coverage* as

$$\psi_R\left(\hat{R}\right) = [\![|\hat{R} - R| < \varepsilon_\alpha^{bin}]\!].$$

The empirical coverage indicates whether the confidence interval centered around the estimate $\hat{R}$ includes the quantity R or not. Since $\hat{R}$ depends on a set of randomly drawn instances, the empirical coverage is a random variable. For a reliable confidence interval, the expected value of ψ_R is $1 - \alpha$. The empirical coverage $\psi_R(\hat{R}_n)$ of the empirical risk $\hat{R}_n$ is negatively biased, that is, the probability that the interval covers the true risk $\psi_R(\hat{R}_n) = 1$ is less than $1 - \alpha$ (see, *e.g.,* Brown et al., 2002). Intuitively, if the value R being estimated is close to the boundaries and the sample size n is small, the estimator's distribution is very skewed and thus empirical estimates $\hat{R}_n$ of zero and one occur regularly. An empirical risk of zero and one, respectively, leads to an empirical variance S_n^2 of zero which in turn collapses the confidence interval into a single point. Another reason for the empirical coverage to be biased is that the standard interval is symmetric and centered around the estimate $\hat{R}_n$. For reasonable choices of α and n a negative lower bound $\hat{R}_n - \varepsilon_\alpha^{bin} < 0$ of the confidence interval can be obtained. Since $R \geq 0$ holds, the coverage of the feasible part of the interval is thus lower than $1 - \alpha$.

Surprisingly, even if R is not close to the boundaries, the empirical coverage of the confidence interval can be erratically poor. To see this, we now study the expected empirical coverage probability taken over the outcomes of the estimate $\hat{R}_n$. It is

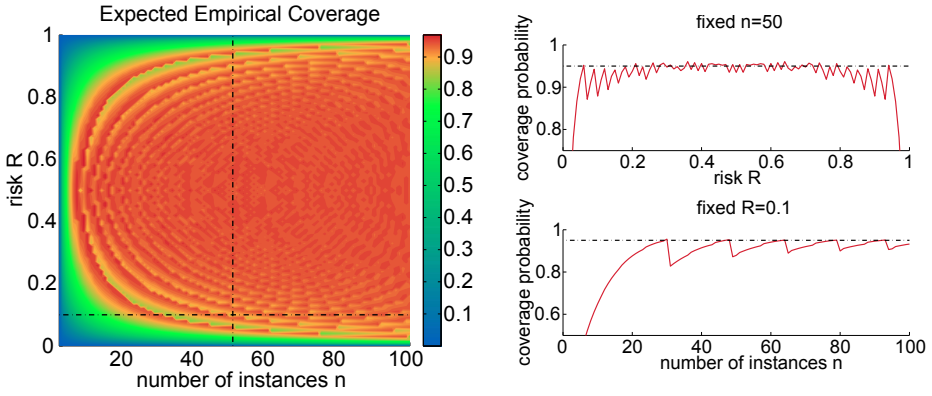


Figure 3.2: Heatmap of the empirical coverage of standard interval with coverage level $1-\alpha = 0.95$, plotted into a two-dimensional space with axes R and n (left). Detailed representation of the empirical coverage for fixed number of instances and fixed risk, respectively (right). Horizontal lines indicate the theoretical coverage.

given by

$$\mathbb{E}_{\hat{R}_n \sim p(\hat{R}_n|R,n)}\left[\psi_R\left(\hat{R}_n\right)\right] = \int \psi_R\left(\hat{R}_n\right) p(\hat{R}_n|R,n)\mathrm{d}\hat{R}_n.$$
$$= \sum_{i=0}^{n} \psi_R\left(\frac{i}{n}\right)\binom{n}{i} R^i\left(1-R\right)^{n-i}. \tag{3.17}$$

Equation 3.17 exploits that sampling from $p(x)$ leads to a finite number of possible outcomes in a classification setting; the estimate $\hat{R}_n$ is binomially distributed. Figure 3.2 (left) shows the expected empirical coverage as a function of the risk R and the number of instances n. The discrete lattice structure of $\hat{R}_n$ causes an oscillation in the coverage probability even for larger n. Hence, standard interval estimates governed by the corresponding estimate $\hat{R}_n$ may be appropriate or drastically poor depending on the choice of n for a given R. Figure 3.2 (right) illustrates the behavior of the coverage probability for fixed n and fixed R, respectively. For small n, the negative bias caused by collapsed intervals is dominating. For larger n, the oscillation effect due to the lattice structure becomes visible. A more formal investigation is given by Brown et al. (2001).

In the case in which ℓ is binomially distributed, it would seem more natural to derive confidence intervals by inverting the exact distribution rather than using a normal approximation. This idea was proposed by Clopper & Pearson (1934). By definition, the *Clopper-Pearson interval* is correct; it guarantees that the coverage is at least $1-\alpha$ for all sample sizes n and values of R. However, the discrete structure of the distribution results mostly in very conservative and too wide ranges. Therefore, many authors argue in favor of approximate confidence intervals (see, *e.g.,* Agresti & Coull, 1998). An alternative to the standard interval is the *Wilson interval* (Wilson, 1927). In contrast to the standard interval, we do not substitute the variance σ_q^2 by an empirical estimate. Instead, we study the squared deviation $(\hat{R}_n - R)^2$ for a given confidence level α under the assumption that $\hat{R}_n$ is normally distributed. From Lemma 3.3, it follows that

$$\sqrt{n}\frac{|\hat{R}_n - R|}{\sqrt{R(1-R)}} \leq \Phi^{-1}\left(1-\frac{\alpha}{2}\right)$$

holds with probability $1-\alpha$. Thus, the squared deviation is bounded with probability $1-\alpha$ by

$$\left(\hat{R}_n - R\right)^2 \leq \Phi^{-1}\left(1-\frac{\alpha}{2}\right)^2 \frac{R(1-R)}{n}. \tag{3.18}$$

The roots of this quadratic equation with respect to the unknown value R give rise to the Wilson confidence interval $[\bar{R}_n - \varepsilon_\alpha^{wil}, \bar{R}_n + \varepsilon_\alpha^{wil}]$, where we have defined

$$\bar{R}_n = \frac{\hat{R}_n + \frac{1}{2n}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)^2}{1+\frac{1}{n}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)^2} \tag{3.19}$$

$$\varepsilon_\alpha^{wil} = \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\frac{\sqrt{\hat{R}(1-\hat{R}) + \frac{1}{4n}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)^2}}{\sqrt{n}+\frac{1}{\sqrt{n}}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)^2}. \tag{3.20}$$

In contrast to the standard confidence interval $\hat{R}_n \pm \varepsilon_\alpha^{bin}$, the Wilson interval is not symmetric to $\hat{R}_n$; the center is shifted to $R = 0.5$. Furthermore, Figure 3.1, (bottom) shows that the size of ε_α^{wil} is generally larger for extreme values of R. Finally, an overview of alternative intervals is given, for example, by Henderson & Meyer (2001).

Although the coverage of the Wilson interval is closer to the expected coverage than the standard interval for a binomial loss function, the oscillating behav-

ior caused by the discreteness of the binomial distribution can not be avoided without an additional randomization of the estimation process (see, *e.g.,* Brown et al., 2001). However, note that, $\hat{R}_{n,q}$ is in general not binomial distributed for $p(x) \neq q(x)$ even if ℓ follows a Bernoulli distribution. The resampling weights $\frac{p(x)}{q(x)}$, that live in a potentially continuous space, soften the lattice structure of possible estimates. We will see in Section 4.4.3 that the empirical coverage of the corresponding confidence interval increases more smoothly with the number of observed instances.

3.2 Comparison of Prediction Models

In this section, we summarize the statistical foundations of testing theory which allow us to compare prediction models accurately. The standard approach to comparing models is to calculate their empirical risks based on instances that are governed by the test distribution $p(x)$ which the models are exposed to in practice. The underlying distribution of the estimator $\hat{R}_n$ provides information on whether the observed difference is significant or due to chance. If instances are drawn according to an instrumental distribution $q(x)$, this procedure also applies to a self-normalized importance sampling estimator $\hat{R}_{n,q}$. In Section 3.2.1, we detail a statistical test for estimates based on instances which are drawn according to an instrumental sampling distribution. Confidence intervals (see Section 3.1.2) and hypothesis testing are closely related. We discuss their relationship in Section 3.2.2. Finally, we present statistical tests that can be used to comparing multiple models (see Section 3.2.3).

3.2.1 A Statistical Test for Actively Drawn Instances

Given two models $f_{\boldsymbol{\theta}_1}$ and $f_{\boldsymbol{\theta}_2}$, our goal is to identify the one with lower risk R. Since the true risks are unknown, they are typically estimated from a sample of labeled test instances. Given estimates $\hat{R}_{n,q}[f_{\boldsymbol{\theta}_1}]$ and $\hat{R}_{n,q}[f_{\boldsymbol{\theta}_2}]$, the difference

$$\hat{\Delta}_{n,q} = \hat{R}_{n,q}[f_{\boldsymbol{\theta}_1}] - \hat{R}_{n,q}[f_{\boldsymbol{\theta}_2}] \tag{3.21}$$

provides evidence on which model is preferable; a negative sign of $\hat{\Delta}_{n,q}$ argues in favor of $f_{\boldsymbol{\theta}_1}$ whereas a positive sign makes $f_{\boldsymbol{\theta}_2}$ preferable.

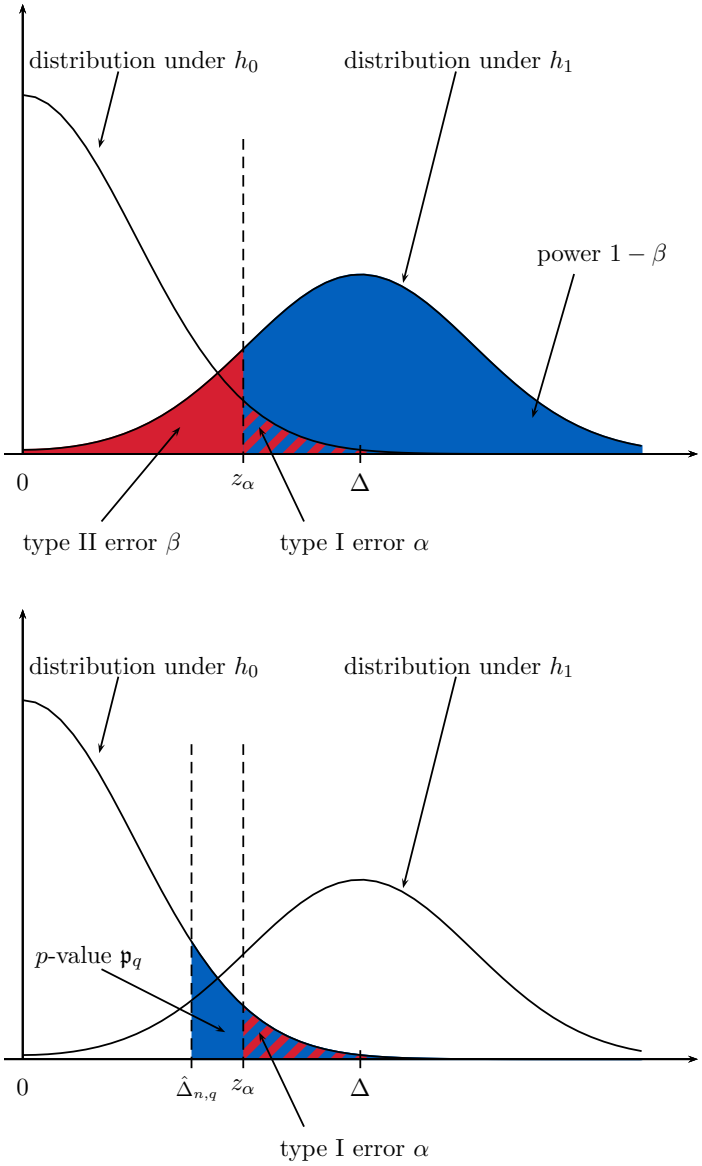


Figure 3.3: Type I and II of a statistical test induced by the distribution of the estimator $\hat{\Delta}_{n,q}$ under null and alternative hypothesis for a fixed critical value z_α (top). The p-value quantifies the likelihood of the observed statistic or a more extrem value under the null hypothesis (bottom).

In preferring one model over the other, one rejects the *null hypothesis* h_0 that the observed difference $\hat{\Delta}_{n,q}$ is only a random effect, and actually $\Delta = R[f_{\boldsymbol{\theta}_1}] - R[f_{\boldsymbol{\theta}_2}] = 0$ holds. Rejecting h_0 confidently allows us to conclude that the opposite (*alternative hypothesis* h_1) is true, that is $R[f_{\boldsymbol{\theta}_1}] \neq R[f_{\boldsymbol{\theta}_2}]$ justifying to choose the model with lower empirical risk. To quantify the evidence that can be gathered from the data, we now analyze the distribution of the test statistic under the null hypothesis, which leads to the *Wald-test* (see, *e.g.,* Wasserman, 2004, Chapter 10).

Lemma 3.1 implies that the risk estimates $\hat{R}_{n,q}[f_{\boldsymbol{\theta}_i}]$ and thus the difference $\hat{\Delta}_{n,q}$ are asymptotically normally distributed. Furthermore, under the null hypothesis the mean of $\hat{\Delta}_{n,q}$ is asymptotically zero and hence the statistic

$$\sqrt{n}\frac{\hat{\Delta}_{n,q}}{\sigma_{n,q}} \xrightarrow{d} \mathcal{N}(0,1),$$

where $\frac{1}{n}\sigma_{n,q}^2 = \mathrm{Var}_{(x,y)\sim q(x,y)}[\hat{\Delta}_{n,q}]$ denotes the variance of $\hat{\Delta}_{n,q}$, follows asymptotically a standard normal distribution. In practice, $\sigma_{n,q}$ is unknown. Let

$$\delta(x,y) = \ell(f_{\boldsymbol{\theta}_1}(x), y) - \ell(f_{\boldsymbol{\theta}_2}(x), y)$$

denote the difference in loss between the predictions of the two models for a test instance (x,y). Then, following Lemma 3.2 with loss function $\delta(x_i, y_i)$ a consistent estimate of $\sigma_{n,q}^2$ is obtained from the labeled sample $(x_1,y_1),\ldots,(x_n,y_n)$ drawn from $q(x)p(y|x)$ by computing empirical variance

$$S_{n,q}^2 = n\left(\frac{p(x_i)}{q(x_i)}\right)^{-2}\sum_{i=1}^{n}\left(\frac{p(x_i)}{q(x_i)}\right)^2\left(\delta(x_i,y_i) - \hat{\Delta}_{n,q}\right)^2. \tag{3.22}$$

Substituting the empirical for the true standard deviation yields an observable statistic $\sqrt{n}\frac{\hat{\Delta}_{n,q}}{S_{n,q}}$. Because $S_{n,q}^2$ consistently estimates $\sigma_{n,q}^2$ the observable statistic would be asymptotically standard normally distributed,

$$\sqrt{n}\frac{\hat{\Delta}_{n,q}}{S_{n,q}} \sim \mathcal{N}(0,1), \tag{3.23}$$

if the null hypothesis were true.

The *p-value* $\mathfrak{p}_q$ quantifies the likelihood of observing a test statistic or a more extreme value, by chance under the null hypothesis. The p-value of the two-sided

Wald test can be derived in analogy to the proof of Lemma 3.3. It is given by

$$\begin{aligned}\mathfrak{p}_q &= p\left(Z > \sqrt{n}\frac{|\hat{\Delta}_{n,q}|}{S_{n,q}}\right) \\ &= 2\left(1 - \Phi\left(\sqrt{n}\frac{|\hat{\Delta}_{n,q}|}{S_{n,q}}\right)\right),\end{aligned} \tag{3.24}$$

where $Z \sim \mathcal{N}(0,1)$ is standard normally distributed and Φ denotes the cumulative distribution function of the standard normal distribution. If it falls below a pre-defined confidence threshold α (admissible type I error), one can reject the null hypothesis and conclude that the models' risks are significantly different. Equivalently, the null hypothesis can be rejected if the test statistic exceeds the corresponding *critical value*

$$z_\alpha = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right).$$

The type II error rate of a statistical test is the probability of not accepting the alternative hypothesis h_1 although it in fact holds. It is given by

$$\begin{aligned}\beta_{\alpha,q} &= p\left(\mathfrak{p}_q > \alpha\right) \\ &= p\left(\sqrt{n}\frac{|\hat{\Delta}_{n,q}|}{S_{n,q}} < z_\alpha\right).\end{aligned} \tag{3.25}$$

Furthermore, $1-\beta_{\alpha,q}$ is known as the *power* of a statistical test; it is the likelihood that the p-value falls below α, if the alternative hypothesis truly does hold and the two models indeed incur different risks. The central concepts are summarized in Figure 3.3.

3.2.2 Relationship between Tests and Confidence Intervals

Confidence intervals and hypothesis testing are closely related. Specifically, the Wald test with confidence level $1-\alpha$ rejects the null hypothesis $\Delta = \Delta_0$ if and only if Δ_0 is not covered by the standard confidence interval, that is $|\Delta_0 - \hat{\Delta}_{n,q}| \geq \varepsilon_\alpha$ (Wasserman, 2004, Theorem 10.10). Therefore, the coverage of the confidence interval corresponds to confidence level $1-\alpha$ of a statistical test. As a consequence, employing a Wald-test for binary proportions may yield a poorly calibrated type I error (see Section 3.1.2).

As an alternative, Student's t-distribution can serve as an approximation of the distribution of a test statistic under the null hypothesis as well as for the derivation of confidence intervals in Section 3.1.2. This results in the widely used *Student's t-test* and the corresponding t-test interval, respectively. Note, however, that the statistic $(n-1)\frac{S^2_{n,q}}{\sigma^2_{n,q}}$ would have to be governed by a χ^2-distribution with $n-1$ degrees of freedom for the test statistic to be asymptotically governed by the t-distribution. This assumption is only satisfied if $S^2_{n,q}$ would be a sum of squared, normally distributed random variables which is reasonable for regression, but not for classification, and only for the case of $p(x) = q(x)$. Nevertheless, the normal distribution is often replaced by the Student's t-distribution even if the distribution assumption is not justified. Since the t-distribution has heavier tails for small n, the resulting confidences are more conservative and thus more robust to unlikely events. The t-distribution converges to the normal distribution. For the sample sizes n that are studied in this thesis, the difference is already negligible for the considered sample sizes: The corresponding confidence regions differ by a factor of $F^{-1}_{n-1}(1-\frac{\alpha}{2})/\Phi^{-1}(1-\frac{\alpha}{2}) = 1.012$ for $n = 101$, where Φ^{-1} and F^{-1}_{ν} are the inverse cumulative distribution functions of the Gaussian and the t-distribution with ν degrees of freedom.

3.2.3 Comparing Multiple Prediction Models

So far we have focused on the problem of comparing the risks of two prediction models, such as a baseline and a challenger. We might also compare several alternative models and rank the models according to their risks or to identify the model with lowest risk.

Comparing multiple prediction models is even more challenging than to evaluate whether there is any evidence that the performance difference Δ between two models is significantly different from zero (Demšar, 2006). A naïve strategy to evaluate the relative performance of k different models is to apply multiple pairwise Wald or t-test, respectively. However, when testing multiple hypotheses the probability α that at least one of the pairwise difference becomes falsely significant under the null hypothesis increases with an increasing number of performed tests. Hence, α exceeds considerably the type I error $\alpha_i = \alpha'$ of each single test i. This multiplicity effect can be countered by the *Bonferroni* correction. Using the principle of inclusion and exclusion, the probability that one individual null

hypothesis is rejected although no difference exists can be upper bounded by

$$\alpha \leq \sum_{i=1}^{k} \alpha_i = k\alpha'.$$

Then, a reliable overall type I error α can be ensured by testing all pairwise differences at a significance level of $\frac{\alpha'}{k}$. The Bonferroni correction is conservative, that is, the likelihood that any single null hypothesis is rejected by chance is much less than the pre-specified confidence threshold α. In particular, when many models k are considered, the power of this method might be impractically low (Salzberg, 1997). A popular choice of reliable, yet powerful, tests that compare multiple alternatives are, for example, within-subject *analysis of variance* (ANOVA; see, *e.g.,* Sheskin, 2004, Test 24) or the *Tukey range test* (see, *e.g.,* Sheskin, 2004, Test 21c); both try to reject the null hypothesis

$$h_0 : R[f_{\boldsymbol{\theta}_1}] = \ldots = R[f_{\boldsymbol{\theta}_k}]$$

that the risk of all considered models are equal. Rejection of h_0 does not imply that all empirically observed differences are significant; for example, the test could become significant because one of the alternatives performs clearly worst. When applying a Tukey test, the homogeneity of the risks is examined by the likelihood of the observed maximum range, that is

$$\varrho = \max_{i=1,\ldots,k} \hat{R}_{n,q}[f_{\boldsymbol{\theta}_i}] - \min_{i=1,\ldots,k} \hat{R}_{n,q}[f_{\boldsymbol{\theta}_i}].$$

Adjusting the range by a χ^2-distributed estimator of the variance

$$S_k^2 = \frac{1}{k(k-1)} \sum_{i \neq j} \left(S_{n,q}^{i,j}\right)^2$$

yields a statistic that is governed by a *studentized range distribution* with $\nu = (n-1)k$ degrees of freedom under the null hypothesis. Let T be a random variable which follows a studentized range distribution. Then, the cumulative distribution of the studentized range distribution is given by

$$\begin{aligned} \Omega_\nu(s) &= p(T \leq s) \\ &= \int_0^\infty \left(k \int_{-\infty}^{\infty} \varphi(t) \left[\Phi(t) - \Phi(t - su)\right]^{k-1} \mathrm{d}t \right) \chi(u|\nu) \mathrm{d}u, \end{aligned}$$

where χ and φ are the probability density function of the chi-squared and the

standard normal distribution, respectively. The likelihood of observing $\frac{\varrho}{S_k}$ or a more extreme value is given by $1 - \Omega_\nu(\varrho S_k)$.

If the likelihood of the observed range ϱ falls under a given confidence level α the null hypothesis h_0 can be rejected and one can conclude that at least the highest and the lowest risks are significantly different. However, any subrange $|\hat{R}_{n,q}[f_{\boldsymbol{\theta}_i}] - \hat{R}_{n,q}[f_{\boldsymbol{\theta}_j}]|$ that is already sufficiently unlikely implies that ϱ is also unlikely under the null hypothesis. Thus, testing all pairwise differences allows us to examine the risks that differ significantly. Evaluating the power of a multiple comparison procedure such as the Tukey test is challenging, because it generally requires repeated numerical calculations of a joint multivariate distribution.

The power of a statistical test describes how likely an effect such as $\Delta > 0$ can be identified. In Section 5, we study its dependence on the drawn sample $(x_1, y_1), \ldots, (x_n, y_n)$ in order to derive data selection strategies that increase the power for a fixed type I error.

3.3 A Generalized Risk Functional

Several performance measures cannot be expressed as a risk. Perhaps the most prominent such measure is the F_η-measure (van Rijsbergen, 1979). The F_η-measure is a weighted harmonic mean of *recall* and *precision*. For a given binary classifier $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}$ where $\mathcal{Y} = \{0, 1\}$ and a sample of size n, let

$$n_{tp} = \sum_{i=1}^{n} y_i f_{\boldsymbol{\theta}}(x_i) \text{ and}$$
$$n_{fp} = \sum_{i=1}^{n} (1 - y_i) f_{\boldsymbol{\theta}}(x_i)$$

denote the number of true and false positives, respectively, and

$$n_{fn} = \sum_{i=1}^{n} y_i (1 - f_{\boldsymbol{\theta}}(x_i))$$

the number of false negatives. Then, the classifier's F_η-measure on the sample is defined as

$$F_\eta = \frac{n_{tp}}{\eta(n_{tp} + n_{fp}) + (1 - \eta)(n_{tp} + n_{fn})}. \tag{3.26}$$

Precision and recall are special cases for $\eta = 1$ and $\eta = 0$, respectively. This class of measures takes the marginal distribution of the positive class $y = +1$ into account and is thus more appropriate than measuring the error rate of a model in domains with highly skewed class distributions. For example, in spam filtering problems we are often interested in measuring spam and non-spam recall. Furthermore, in information retrieval tasks F-measures for a designated class are used to assess the quality of text classifiers or the result of a web search.

The F_η-measure is defined as an estimator in terms of empirical quantities. This is unintuitive from a statistical point of view and raises the question which quantity of the underlying distribution the F_η-measure actually estimates. We will now introduce the class of *generalized risk* functionals. Like the risk functional (see Equation 3.1), the generalized risk is parameterized with a loss function ℓ. In addition, the generalized risk is parameterized with a function w that assigns a weight $w(x, y, f_{\boldsymbol{\theta}})$ to each instance. For example, precision sums over instances with $f_{\boldsymbol{\theta}}(x) = 1$ with constant weight and gives no consideration to other in-

stances. Equation 3.27 defines the generalized risk:

$$G[f_{\boldsymbol{\theta}}] = \frac{\mathbb{E}_{(x,y)\sim p(x,y)}\left[w(x,y,f_{\boldsymbol{\theta}})\ell(f_{\boldsymbol{\theta}}(x),y)\right]}{\mathbb{E}_{(x,y)\sim p(x,y)}\left[w(x,y,f_{\boldsymbol{\theta}})\right]} = \frac{\iint \ell(f_{\boldsymbol{\theta}}(x),y)w(x,y,f_{\boldsymbol{\theta}})p(x,y)\mathrm{d}y\,\mathrm{d}x}{\iint w(x,y,f_{\boldsymbol{\theta}})p(x,y)\mathrm{d}y\,\mathrm{d}x}. \tag{3.27}$$

Note that the generalized risk (see Equation 3.27) reduces to the regular risk for $w(x, y, f_{\boldsymbol{\theta}}) = 1$.

In analogy to the regular risk, a consistent estimator for the generalized risk functional can be obtained by substituting the true distribution $p(x, y)$ with the empirical distribution function $\hat{p}(x, y)$ (see Equation 3.8):

Proposition 3.1 (Consistency of Empirical Generalized Risks)**.** *Let the test sample* $(x_1, y_1), \ldots, (x_n, y_n)$ *be drawn* i.i.d. *according to* $p(x, y)$*. The quantity*

$$\hat{G}_n[f_{\boldsymbol{\theta}}] = \frac{\sum_{i=1}^{n} \ell(f_{\boldsymbol{\theta}}(x_i), y_i)w(x_i, y_i, f_{\boldsymbol{\theta}})}{\sum_{i=1}^{n} w(x_i, y_i, f_{\boldsymbol{\theta}})} \tag{3.28}$$

is a consistent estimate of the generalized risk G *defined by Equation 3.27.*

Proof. Due to the weak law of strong numbers the numerator and denominator converges almost surely to their expected values. Thus, the proposition follows from Slutsky's theorem (see, *e.g.,* Cramér, 1946) applied to the numerator and denominator of Equation 3.28. □

We now observe that F_η-measures—including precision and recall—are consistent empirical estimates of generalized risks for appropriately chosen functions w.

Corollary 3.1 (Consistency of F-measures)**.** *Let* $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}$ *be a predictive model.* F_η *of* $f_{\boldsymbol{\theta}}$ *is a consistent estimate of the generalized risk with* $\mathcal{Y} = \{0, 1\}$*,* $w(x, y, f_{\boldsymbol{\theta}}) = \eta f_{\boldsymbol{\theta}}(x) + (1 - \eta)y$ *and* $\ell = 1 - \ell_{0/1}$*, where* $\ell_{0/1}$ *denotes the zero-one loss.*

Proof. The claim follows from Proposition 3.1 since

$$\begin{aligned}\hat{G}_n &= \frac{\sum_{i=1}^{n}(1-\ell_{0/1}(f_{\boldsymbol{\theta}}(x_i),y_i))\,(\eta f_{\boldsymbol{\theta}}(x_i)+(1-\eta)y_i)}{\sum_{i=1}^{n}(\eta f_{\boldsymbol{\theta}}(x_i)+(1-\eta)y_i)}\\ &= \frac{\sum_{i=1}^{n} f_{\boldsymbol{\theta}}(x_i)y_i}{\eta\sum_{i=1}^{n} f_{\boldsymbol{\theta}}(x_i)+(1-\eta)\sum_{i=1}^{n} y_i}\\ &= \frac{n_{tp}}{\eta\,(n_{tp}+n_{fp})+(1-\eta)\,(n_{tp}+n_{fn})}.\end{aligned}$$

□

Intuitively, precision and recall can be interpreted as accuracy conditioned to instances which are predicted as positive $\psi = [\![f_{\boldsymbol{\theta}}(x) = 1]\!]$ or are truly positive $\psi = [\![y = 1]\!]$. The regular risk conditioned to ψ can be rewritten using Bayes' theorem (Equation 3.29) and the law of total probability (Equation 3.30):

$$\begin{aligned}&\iint \ell(f_{\boldsymbol{\theta}}(x),y)p(x,y|\psi)\mathrm{d}y\,\mathrm{d}x\\ &= \iint \ell(f_{\boldsymbol{\theta}}(x),y)\frac{p(\psi|x,y)p(x,y)}{p(\psi)}\mathrm{d}y\,\mathrm{d}x &(3.29)\\ &= \frac{\iint \ell(f_{\boldsymbol{\theta}}(x),y)p(\psi|x,y)p(x,y)\mathrm{d}y\,\mathrm{d}x}{\iint p(\psi|x,y)p(x,y)\mathrm{d}y\,\mathrm{d}x}. &(3.30)\end{aligned}$$

Note that $p(\psi|x,y)$ is deterministic in the sense that ψ has probability zero or one. This is captured by the weight function $w(x,y,f_{\boldsymbol{\theta}})$ in Equation 3.27.

Having established and motivated the generalized risk functional, we turn towards the case in which test instances are not be drawn according to the test distribution. In analogy to the regular risk R in Section 3.1.1, the generalized risk G with respect to the test distribution $p(x,y) = p(x)p(y|x)$ can be defined on instances drawn from an instrumental distribution $q(x,y) = p(x)q(y|x)$ by weighting the instance-specific losses:

$$\begin{aligned}&\frac{\mathbb{E}_{(x,y)\sim p(x,y)}\left[w(x,y,f_{\boldsymbol{\theta}})\ell(f_{\boldsymbol{\theta}}(x),y)\right]}{\mathbb{E}_{(x,y)\sim p(x,y)}\left[w(x,y,f_{\boldsymbol{\theta}})\right]}\\ &= \frac{\mathbb{E}_{(x,y)\sim q(x)p(y|x)}\left[\frac{p(x)}{q(x)}w(x,y,f_{\boldsymbol{\theta}})\ell(f_{\boldsymbol{\theta}}(x),y)\right]}{\mathbb{E}_{(x,y)\sim q(x)p(y|x)}\left[\frac{p(x)}{q(x)}w(x,y,f_{\boldsymbol{\theta}})\right]}.\end{aligned}$$

Then, the self-normalized importance sampling estimator of a generalized risk

can be defined as

$$\hat{G}_{n,q}[f_{\boldsymbol{\theta}}] = \frac{\sum_{i=1}^{n} \frac{p(x_i)}{q(x_i)} w(x_i, y_i, f_{\boldsymbol{\theta}}) \ell(f_{\boldsymbol{\theta}}(x_i), y_i)}{\sum_{i=1}^{n} \frac{p(x_i)}{q(x_i)} w(x_i, y_i, f_{\boldsymbol{\theta}})}, \tag{3.31}$$

where (x_i, y_i) are drawn from $q(x)p(y|x)$. Because of the weighting factors, Slutsky's Theorem again implies that Equation 3.31 defines a consistent estimator for G. Moreover, Lemma 3.4 states that the generalized risk estimator $\hat{G}_{n,q}$ is asymptotically normally distributed, and characterizes its variance in the limit.

Lemma 3.4 (Asymptotic Distribution of Estimator)**.** *Let $\hat{G}_{n,q}$ be defined as in Equation 3.31 and let us assume that*

1. *the expected values $\mathbb{E}_{(x,y)\sim p(x,y)}\left[w(x, y, f_{\boldsymbol{\theta}})\ell(f_{\boldsymbol{\theta}}(x), y)\right]$ and $\mathbb{E}_{(x,y)\sim p(x,y)}\left[w(x, y, f_{\boldsymbol{\theta}})\right]$ exist,*
2. *the expected value $\mathbb{E}_{(x,y)\sim p(x,y)}\left[w(x, y, f_{\boldsymbol{\theta}})\right]$ is non-zero,*
3. *the variances $\mathrm{Var}_{(x,y)\sim p(x,y)}\left[w(x, y, f_{\boldsymbol{\theta}})\ell(f_{\boldsymbol{\theta}}(x), y)\right]$ and $\mathrm{Var}_{(x,y)\sim p(x,y)}\left[w(x, y, f_{\boldsymbol{\theta}})\right]$ are finite,*
4. *the distribution $q(x)$ is absolutely continuous with respect to $p(x)$, and*
5. *the weights $\frac{p(x)}{q(x)} \leq E$ are bounded from above by a constant $E < \infty$.*

Then, $\hat{G}_{n,q}$ is asymptotically normally distributed,

$$\sqrt{n}\left(\hat{G}_{n,q} - G\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_q^2\right),$$

with asymptotic variance

$$\sigma_q^2 = \frac{\mathbb{E}_{(x,y)\sim q(x)p(y|x)}\left[\left(\frac{p(x)}{q(x)}\right)^2 w(x, y, f_{\boldsymbol{\theta}})^2 \left(\ell(f_{\boldsymbol{\theta}}(x), y) - G\right)^2\right]}{\mathbb{E}_{(x,y)\sim q(x)p(y|x)}\left[\frac{p(x)}{q(x)} w(x, y, f_{\boldsymbol{\theta}})\right]^2},$$

where $\xrightarrow{d}$ denotes convergence in distribution.

Proof. Let $(x_1, y_1), ..., (x_n, y_n)$ be drawn according to $q(x)p(y|x)$. In this proof, all expectations and variances are over the distribution $q(x)p(y|x)$. We omit the underlying distribution to keep the notation uncluttered. Let $\hat{G}^0_{n,q} = \sum_{i=1}^{n} v_i w_i \ell_i$

and $W_n = \sum_{i=1}^{n} v_i w_i$ denote the numerator and the dominator of $\hat{G}_{n,q}$, respectively, where we have defined

$$\begin{aligned} v_i &= \frac{p(x_i)}{q(x_i)}, \\ w_i &= w(x_i, y_i, f_{\boldsymbol{\theta}}), \text{ and} \\ \ell_i &= \ell(f_{\boldsymbol{\theta}}(x_i), y_i). \end{aligned}$$

Making use of the linearity of the expected value and the definition of G (see Equation 3.27), it follows that

$$\mathbb{E}\left[W_n\right] = n\mathbb{E}\left[v_i w_i\right] \text{ and } \mathbb{E}\left[\hat{G}^0_{n,q}\right] = nG\mathbb{E}\left[v_i w_i\right].$$

The random variables $v_1 w_1, \ldots, v_n w_n$ and $v_1 w_1 \ell_1, \ldots, v_n w_n \ell_n$ are i.i.d., therefore, under Condition 1 and 3, the central limit theorem implies that $\frac{1}{n}\hat{G}^0_{n,q}$ and $\frac{1}{n}W_n$ are asymptotically normally distributed with

$$\begin{aligned} \sqrt{n}\left(\frac{1}{n}\hat{G}^0_{n,q} - G\mathbb{E}\left[v_i w_i\right]\right) &\xrightarrow{d} \mathcal{N}(0, \mathrm{Var}[v_i w_i \ell_i]) \\ \sqrt{n}\left(\frac{1}{n}W_n - \mathbb{E}\left[v_i w_i\right]\right) &\xrightarrow{d} \mathcal{N}(0, \mathrm{Var}[v_i w_i]) \end{aligned} \tag{3.32}$$

where $\xrightarrow{d}$ denotes convergence in distribution. From Condition 2 and the convergence statement in Equation 3.32 it follows that

$$\lim_{n\to\infty} p\left(|W_n| > 0\right) = 1$$

and hence the denominator of $\hat{G}_{n,q}$ is non-zero for sufficiently large n. We now employ the multivariate *delta method* (see, *e.g.,* Wasserman, 2004, Chapter 5.5) to extend the convergence results for $\hat{G}^0_{n,q}$ and W_n to a convergence result for the normalized estimator $\hat{G}_{n,q}$. The delta method allows us to derive the asymptotic distribution of a differentiable function g whose input variables are asymptotically normally distributed. Applying it to the function $g(x, y) = \frac{x}{y}$ with $x = \frac{1}{n}\hat{G}^0_{n,q}$ and $y = \frac{1}{n}W_n$ yields

$$\begin{aligned} \sqrt{n}\left(\frac{\frac{1}{n}\hat{G}^0_{n,q}}{\frac{1}{n}W_n} - G\right) &\xrightarrow{d} \mathcal{N}(0, \sigma_q^2), \\ \sigma_q^2 &= \nabla g\left(G\mathbb{E}\left[v_i w_i\right], \mathbb{E}\left[v_i w_i\right]\right)^{\mathsf{T}} \Sigma \nabla g\left(G\mathbb{E}\left[v_i w_i\right], \mathbb{E}\left[v_i w_i\right]\right) \end{aligned}$$

where $\nabla g(x,y) = \left(\frac{1}{y}, -\frac{x}{y^2}\right)^\mathsf{T}$ denotes the gradient of g and Σ is the asymptotic covariance matrix of the input arguments

$$\Sigma = \begin{pmatrix} \operatorname{Var}[v_i w_i \ell_i] & \operatorname{Cov}[v_i w_i \ell_i, v_i w_i] \\ \operatorname{Cov}[v_i w_i \ell_i, v_i w_i] & \operatorname{Var}[v_i w_i] \end{pmatrix}.$$

Furthermore,

$$\begin{aligned} &\nabla g\left(G\mathbb{E}\left[v_i w_i\right], \mathbb{E}\left[v_i w_i\right]\right)^\mathsf{T} \Sigma \nabla g\left(G\mathbb{E}\left[v_i w_i\right], \mathbb{E}\left[v_i w_i\right]\right) \\ =&\frac{\operatorname{Var}\left[w_i \ell_i v_i\right] - 2G \operatorname{Cov}\left[w_i v_i, w_i \ell_i v_i\right] + G^2 \operatorname{Var}\left[w_i v_i\right]}{\mathbb{E}\left[v_i w_i\right]^2} \\ =&\frac{\mathbb{E}\left[w_i^2 \ell_i^2 v_i^2\right] - 2G\mathbb{E}\left[w_i^2 \ell_i v_i^2\right] + G^2 \mathbb{E}\left[w_i^2 v_i^2\right]}{\mathbb{E}\left[v_i w_i\right]^2} \\ =&\frac{\mathbb{E}\left[v_i^2 w_i^2 (\ell_i - G)^2\right]}{\mathbb{E}\left[v_i w_i\right]^2}. \end{aligned}$$

Condition 3 and 5 imply that the variance σ_q^2 is finite. From this, the claim follows by back substituting v_i, w_i, and ℓ_i. □

Using uniform weights $w(x_i, y_i, f_{\boldsymbol{\theta}}) = 1$, Lemma 3.4 particularly shows that $\hat{R}_{n,q}$ is normally distributed. Thus, the presented generalized risk functional G consistently extends the risk functional R and its estimators given in Section 3.1.1. Specifically, in analogy to Section 3.1.2 a two-sided confidence interval $[\hat{G}_{n,q} - \varepsilon_\alpha, \hat{G}_{n,q} + \varepsilon_\alpha]$ with coverage $1-\alpha$ is given by $\varepsilon_\alpha = \Phi_n^{-1}\left(1 - \frac{\alpha}{2}\right) \frac{S_{n,q}}{\sqrt{n}}$, where

$$\begin{aligned} S_{n,q}^2 = n \left(\sum_{i=1}^n \frac{p(x_i)}{q(x_i)} w(x_i, y_i, f_{\boldsymbol{\theta}}) \right)^{-2} \cdot \\ \sum_{i=1}^n \left(\frac{p(x_i)}{q(x_i)} \right)^2 w(x_i, y_i, f_{\boldsymbol{\theta}})^2 \left(\ell(f_{\boldsymbol{\theta}}(x_i), y_i) - \hat{G}_{n,q} \right)^2 \end{aligned}$$

is a consistent estimate of σ_q^2 based on the labeled sample $(x_1, y_1), \ldots, (x_n, y_n)$ drawn from the instrumental distribution $q(x)p(y|x)$ (see Lemma 3.2).

The estimators presented in this chapter can be used to evaluate the (generalized) risk of a given model consistently. The labeled instances are assumed to be drawn either directly from test distribution or from a known instrumental distribution. In the next chapter, we study the case, in which a labeled sample is *not* available in advance and instances have to be labeled at a cost.

Chapter 4

Active Model Evaluation

A predictive model is typically evaluated by exposing it to instances with known labels and determining the deviance between predicted and actual label. In order to achieve consistent estimates of the prediction performance, the set of instances have to reflect the input distribution at test time. However, in many application scenarios the test distribution diverges from the training distribution. Such a setting requires to estimate the risk of a given model on separately drawn test instances.

Recall the example scenario presented in Section 1, in which the model is estimated on confidential data and then provided to a customer. Neither party can accurately estimate the prediction performance; the customer has no access to the original training data and the model provider is lacking access to the test distribution. If the model provider's estimates are based on out-dated or biased samples, the risk estimation can be arbitrarily inaccurate. Consequently, in order to estimate the risk accurately, new test instances have to be drawn and labeled. The goal of this chapter is to alleviate this problem with a minimal labeling effort. We present an *active evaluation method*, in which, in analogy to active learning, unlabeled instances are drawn from an instrumental distribution and their labels are queried until a pre-defined budget is exhausted. Using a self-normalized importance sampling estimate the empirical risk on the actively selected sample is weighted appropriately to compensate for the discrepancy between instrumental and test distributions which leads to a consistent estimate.

The sampling distribution minimizing the estimation error for (generalized) risks is derived in Section 4.2. It is optimal when each instance is equally expensive to label. Section 4.3 extends the optimal sampling distribution to the case of individual instance-specific costs. In Section 4.4, we explore the relative benefits of active and regular risk estimates under varying problem characteristics empirically. We study the case of a shift between training and test distribution as well as the case in which an actively learned model has to be evaluated.

Finally, Section 4.5 concludes with a discussion concerning similarities and differences between active evaluation and active learning. Results of this chapter has previously been published (Sawade et al., 2010a,b, 2012a).

4.1 Problem Setting

Let $p(y|x;\boldsymbol{\theta})$ be a given $\boldsymbol{\theta}$-parameterized model of $p(y|x)$ and let $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}$ with

$$f_{\boldsymbol{\theta}}(x) = \arg\max_y p(y|x;\boldsymbol{\theta})$$

be the corresponding predictive function. We study the problem of estimating the predictive performance of $f_{\boldsymbol{\theta}}$ in terms of a generalized risk G in situations in which labeled data governed by the test distribution $p(x,y) = p(y|x)p(x)$ are not available. We assume that unlabeled data are readily available and acquiring labels for selected instances according to the true conditional distribution $p(y|x)$ is costly. Instances to label can be drawn from an instrumental distribution $q(x)$.

When instances are drawn according to an instrumental distribution $q(x)$ rather than the original distribution $p(x)$, the self-normalized importance sampling estimator $\hat{G}_{n,q}$ (see Equation 3.31) provides a consistent estimate of the generalized risk G of the given model (see Section 3.3). The estimation error (MSE) of $\hat{G}_{n,q}$ depends on the selected instances (x,y), which are drawn according to the distribution $q(x)p(y|x)$. Our goal is to find the instrumental distribution $q(x)$ such that the estimation error is minimal for fixed labeling costs n:

$$q^* = \arg\min_q \mathrm{MSE}_{(x,y)\sim q(x)p(y|x)} \left[\hat{G}_{n,q}\right]. \tag{4.1}$$

4.2 Minimizing the Estimation Error

We now turn towards the problem of deriving an optimal sampling distribution q^* according to Equation 4.1 that minimizes the estimation error when used to select instances. Section 4.2.1 analytically derives a sampling distribution that is asymptotically optimal. Section 4.2.2 discusses the empirical sampling distribution in a pool-based setting and presents the active estimation algorithm.

4.2.1 Asymptotically Optimal Sampling Distribution

We start our investigation with an analysis of the sources of estimation error. Recall the bias-variance decomposition of the estimation error (see Equation 3.7). It states that the estimation error can be expressed as a sum of the squared bias and the variance of the estimator

$$\begin{aligned}&\mathrm{MSE}_{(x,y)\sim q(x)p(y|x)}\left[\hat{G}_{n,q}\right]\\&= \mathrm{Bias}_{(x,y)\sim q(x)p(y|x)}\left[\hat{G}_{n,q}\right]^2 + \mathrm{Var}_{(x,y)\sim q(x)p(y|x)}\left[\hat{G}_{n,q}\right].\end{aligned}$$

Because $\hat{G}_{n,q}$ is consistent, both $\mathrm{Bias}[\hat{G}_{n,q}]$ and $\mathrm{Var}[\hat{G}_{n,q}]$ vanish for $n \to \infty$. More specifically, Lemma 4.1 shows that $\mathrm{Bias}[\hat{G}_{n,q}]^2$ is of order $\frac{1}{n^2}$.

Lemma 4.1 (Bias of Estimator)**.** *Let $\hat{G}_{n,q}$ be as defined in Equation 3.31. Then, under the assumptions of Lemma 3.4 there are constants $C \geq 0$ and n_0 such that*

$$\left|\mathbb{E}_{(x,y)\sim q(x)p(y|x)}\left[\hat{G}_{n,q}\right] - G\right| \leq \frac{C}{n}. \tag{4.2}$$

is satisfied for all $n \geq n_0$.

A proof is given, for example, by Liu (2001), Flueck & Holland (1976), and David & Sukhatme (1974). In order to quantify the bias, the standard approach is to approximate the ratio estimator using a Taylor series expansion around the expectation of the denominator up to the second power and evaluate the expectation term-by-term.

According to Lemma 3.4, the estimator $\hat{G}_{n,q}$ is asymptotically normally distributed:

$$\sqrt{n}\left(\hat{G}_{n,q} - G\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_q^2\right), \tag{4.3}$$

with asymptotic variance

$$\sigma_q^2 = \frac{\mathbb{E}_{(x,y)\sim q(x)p(y|x)}\left[\left(\frac{p(x)}{q(x)}\right)^2 w(x,y,f_{\boldsymbol{\theta}})^2 \left(\ell(f_{\boldsymbol{\theta}}(x),y) - G\right)^2\right]}{\mathbb{E}_{(x,y)\sim q(x)p(y|x)}\left[\frac{p(x)}{q(x)} w(x,y,f_{\boldsymbol{\theta}})\right]^2}. \tag{4.4}$$

Taking the variance of both sides of Equation 4.3, we obtain

$$n \operatorname{Var}_{(x,y)\sim q(x)p(y|x)} \left[\hat{G}_{n,q}\right] \xrightarrow{d} \sigma_q^2. \tag{4.5}$$

Because $n \operatorname{Var}[\hat{G}_{n,q}]$ converges to a constant, $\operatorname{Var}[\hat{G}_{n,q}]$ is of order $\frac{1}{n}$. As the bias term vanishes with $\frac{1}{n^2}$ and the variance term with $\frac{1}{n}$, the expected estimation error MSE will be dominated by the variance term $\operatorname{Var}[\hat{G}_{n,q}]$. Moreover, the asymptotic variance $\sigma_q^2 = \lim_{n\to\infty} n \operatorname{Var}[\hat{G}_{n,q}]$ exists. For large n, we can thus approximate

$$\operatorname{MSE}_{(x,y)\sim q(x)p(y|x)} \left[\hat{G}_{n,q}\right] \approx \frac{1}{n}\sigma_q^2.$$

In the following, we will consequently derive a sampling distribution q^* that minimizes the asymptotic variance σ_q^2 of the estimator $\hat{G}_{n,q}$, thereby approximately solving Problem 4.1. By minimizing estimator variance, selecting test instances according to $q^*(x)$ will yield (approximately) most accurate estimates for a given labeling budget. The following theorem derives the sampling distribution that minimizes the functional σ_q^2.

Theorem 4.1 (Optimal Sampling Distribution)**.** *The instrumental distribution that minimizes the asymptotic variance σ_q^2 of the generalized risk estimator $\hat{G}_{n,q}$ is given by*

$$q^*(x) \propto p(x)\sqrt{\int w(x, y, f_{\boldsymbol{\theta}})^2 \left(\ell(f_{\boldsymbol{\theta}}(x), y) - G\right)^2 p(y|x)\mathrm{d}y}. \tag{4.6}$$

Proof. To minimize the variance with respect to the function $q(x)$ under the normalization constraint $\int q(x)\mathrm{d}x = 1$ we state the Lagrangian

$$L\left[q, \gamma\right] = \sigma_q^2 + \gamma \left(\int q(x)\mathrm{d}x - 1\right),$$

with Lagrange multiplier $\gamma \in \mathbb{R}$. The Lagrangian can be reformulated as follows. In Equation 4.7, we insert the definition of the asymptotic variance (see Equation 4.4), apply the *law of total expectation* to the numerator of σ_q^2, and rewrite the expectation over $p(x)$ as an expectation over the instrumental distribution $q(x)$. Finally, all terms which are independent of $q(x)$ are subsumed into

a function c of x (see Equation 4.8):

$$L\left[q,\gamma\right] = \frac{\mathbb{E}_{x\sim p(x)}\left[\frac{p(x)}{q(x)}\mathbb{E}_{y\sim p(y|x)}\left[w(x,y,f_{\boldsymbol{\theta}})^2\left(\ell(f_{\boldsymbol{\theta}}(x),y)-G\right)^2\middle| x\right]\right]}{\mathbb{E}_{(x,y)\sim p(x,y)}\left[w(x,y,f_{\boldsymbol{\theta}})\right]^2} + \gamma\left(\int q(x)\mathrm{d}x - 1\right) \tag{4.7}$$

$$= \int \underbrace{\frac{c(x)}{q(x)} + \gamma\left(q(x)-p(x)\right)}_{=K[q,x]}\mathrm{d}x, \tag{4.8}$$

where

$$c(x) = p(x)^2\frac{\mathbb{E}_{y\sim p(y|x)}\left[w(x,y,f_{\boldsymbol{\theta}})^2\left(\ell(f_{\boldsymbol{\theta}}(x),y)-G\right)^2\middle| x\right]}{\mathbb{E}_{(x,y)\sim p(x,y)}\left[w(x,y,f_{\boldsymbol{\theta}})\right]^2}.$$

It will turn out that we need not constrain the distribution to be nonnegative $q(x) \geq 0$. The optimal function q^* for the constrained problem is given by the saddle point of the Lagrangian. Since the objective is a convex function of q and the constraint is affine in q, it follows from the *strong duality theorem* that

$$\min_q \max_\gamma L\left[q,\gamma\right] = \max_\gamma \min_q L\left[q,\gamma\right].$$

The solution of the inner optimization problem is given by the solution of the Euler-Lagrange equation

$$\frac{\partial}{\partial q}K[q,x] = -\frac{c(x)}{q(x)^2} + \gamma = 0,$$

that is

$$q(x) = \pm\sqrt{\frac{c(x)}{\gamma}}. \tag{4.9}$$

Note that we dismiss the negative solution, since $q(x)$ is a probability density function. Substituting Equation 4.9 into the Lagrangian L we obtain

$$\gamma^* = \arg\max_{\gamma} 2\sqrt{\gamma} \int \sqrt{c(x)}\mathrm{d}x - \gamma. \tag{4.10}$$

Setting the corresponding derivative to zero leads to

$$\gamma^* = \left(\int \sqrt{c(x)}\mathrm{d}x \right)^2. \tag{4.11}$$

Then, the optimal solution can be found by substituting Equation 4.11 into 4.9, leading to

$$q^*(x) = \frac{\sqrt{c(x)}}{\int \sqrt{c(x)}\mathrm{d}x}. \tag{4.12}$$

Finally, back substitution of c in Equation 4.12 implies the theorem. □

We will now detail the optimal sampling distribution for important instances of the generalized risk functional. They are characterized by their loss function ℓ and the instance-specific weights w. Firstly, we study the subfamily of regular risks. In general, the optimal sampling distribution for regular risks is given by the following corollary.

Corollary 4.1 (Optimal Sampling Distribution for Risks)**.** *The instrumental distribution that minimizes the asymptotic variance σ_q^2 of the regular risk estimator $\hat{R}_{n,q}$ is given by*

$$q^*(x) \propto p(x) \sqrt{\int \left(\ell(f_{\boldsymbol{\theta}}(x), y) - R\right)^2 p(y|x)\mathrm{d}y}. \tag{4.13}$$

Proof. The proof follows directly from Theorem 4.1 for $w(x, y, f_{\boldsymbol{\theta}}) = 1$. □

The zero-one error $\ell_{0/1}$ and the squared loss ℓ_2 are widely-used choices for classification and regression problems, respectively. Corollary 4.2 states the optimal sampling distribution for $\ell = \ell_{0/1}$.

Corollary 4.2 (Optimal Sampling for Zero-One Loss)**.** *The sampling distribution that minimizes σ_q^2 for the zero-one loss $\ell_{0/1}$ resolves to*

$$q^*(x) \propto p(x)\sqrt{(1-2R_{0/1})(1-p(f_{\boldsymbol{\theta}}(x)|x)) + R_{0/1}^2}, \tag{4.14}$$

where $R_{0/1}$ is the true error rate.

Proof. Rewriting the result of Theorem 4.1 for $\ell = \ell_{0/1}$ in a classification setting, we obtain

$$\begin{aligned} q^*(x) &\propto p(x)\sqrt{\sum_{y\in\mathcal{Y}} \left(\ell_{0/1}(f_{\boldsymbol{\theta}}(x), y) - R_{0/1}\right)^2 p(y|x;\boldsymbol{\theta})} \\ &= p(x)\sqrt{\sum_{y\neq f_{\boldsymbol{\theta}}(x)} (1-2R_{0/1}) p(y|x;\boldsymbol{\theta}) + R_{0/1}^2} \qquad (4.15) \\ &= p(x)\sqrt{(1-2R_{0/1})(1-p(f_{\boldsymbol{\theta}}(x)|x;\boldsymbol{\theta})) + R_{0/1}^2}. \qquad (4.16) \end{aligned}$$

In Equation 4.15 we make use of the definition of $\ell_{0/1}$ and factorize in Equation 4.16. □

Equation 4.14 constructs $q^*(x)$ such that it gives preference to instances whose loss has a high variance according to $p(y|x)$. If $R_{0/1} = \frac{1}{2}$, the optimal sampling distribution degenerates to sampling from $p(x)$.

In the following, we derive the optimal sampling distribution for regression problems and a squared loss function. In contrast to classification settings a model assumption about the label noise has to be made to end up with a closed-form solution. Corollary 4.3 assumes that the observed value y is Gaussian distributed (see Chapter 2.2):

Corollary 4.3 (Optimal Sampling for Squared Loss)**.** *Let the observed label y be normally distributed $p(y|x) = \mathcal{N}(y|\mu_x, \sigma_x^2)$. Then, the sampling distribution that minimizes σ_q^2 for the squared loss ℓ_2 resolves to*

$$q^*(x) \propto p(x)\sqrt{(f_{\boldsymbol{\theta}}(x) - \mu_x)^2 \left(6\sigma_x^2 - 2R_2 + (f_{\boldsymbol{\theta}}(x) - \mu_x)^2\right) + (3\sigma_x^2 - 2R_2)\sigma_x^2 + R_2^2},$$

where R_2 is the true mean squared error.

Proof. Rewriting the result of Theorem 4.1 for $\ell = \ell_2$ yields

$$\begin{aligned} q^*(x) &\propto p(x)\sqrt{\int \left((f_{\boldsymbol{\theta}}(x) - y)^2 - R_2\right)^2 p(y|x)\mathrm{d}y} \\ &= p(x)\sqrt{\int (f_{\boldsymbol{\theta}}(x) - y)^4 p(y|x)\mathrm{d}y - 2R_2 \int (f_{\boldsymbol{\theta}}(x) - y)^2 p(y|x)\mathrm{d}y + R_2^2} \\ &= p(x)\left((f_{\boldsymbol{\theta}}(x) - \mu_x)^4 + 6\sigma_x^2 (f_{\boldsymbol{\theta}}(x) - \mu_x)^2 + 3\sigma_x^4 \right. \\ &\qquad \left. -2R_2\left((f_{\boldsymbol{\theta}}(x) - \mu_x)^2 + \sigma_x^2\right) + R_2^2\right)^{1/2}. \end{aligned} \tag{4.17}$$

Equation 4.17 exploits that the two integrals over $\mathcal{Y}$ are raw moments of the Gaussian distribution $p(y|x)$. □

Since F-measures are estimators of generalized risks according to Corollary 3.1, we can now derive their variance-minimizing sampling distributions. Corollary 4.4 states the optimal sampling distribution for F_η and, in particular, for recall and precision.

Corollary 4.4 (Optimal Sampling for F_η)**.** *The sampling distribution that minimizes the asymptotic variance* σ_q^2

- *of the* F_η*-estimator resolves to*

$$q^*(x) \propto p(x) \cdot \begin{cases} \sqrt{p(f_{\boldsymbol{\theta}}(x)|x)(1 - G_{F_\eta})^2 + \eta^2 (1 - p(f_{\boldsymbol{\theta}}(x)|x)) G_{F_\eta}^2} & : f_{\boldsymbol{\theta}}(x) = 1 \\ (1-\eta)\sqrt{(1 - p(f_{\boldsymbol{\theta}}(x)|x)) G_{F_\eta}^2} & : f_{\boldsymbol{\theta}}(x) = 0, \end{cases} \tag{4.18}$$

- *for recall resolves to*

$$q^*(x) \propto p(x) \begin{cases} \sqrt{p(f_{\boldsymbol{\theta}}(x)|x)(1 - G_{rec})^2} & : f_{\boldsymbol{\theta}}(x) = 1 \\ \sqrt{(1 - p(f_{\boldsymbol{\theta}}(x)|x)) G_{rec}^2} & : f_{\boldsymbol{\theta}}(x) = 0, \end{cases} \tag{4.19}$$

- *for precision resolves to*

$$q^*(x) \propto p(x) f_{\boldsymbol{\theta}}(x) \sqrt{(1 - 2G_{prec}) p(f_{\boldsymbol{\theta}}(x)|x) + {G_{prec}}^2}, \tag{4.20}$$

where G_{F_η}*,* G_{rec}*, and* G_{prec} *are the* F_η*, recall, and precision, in the limit.*

Proof. According to Corollary 3.1, F_η estimates a generalized risk with $\mathcal{Y} = \{0,1\}$, $w(x,y,f_{\boldsymbol{\theta}}) = \eta f_{\boldsymbol{\theta}}(x) + (1-\eta)y$ and $\ell = 1 - \ell_{0/1}$. Starting from Theorem 4.1, we derive

$$\begin{aligned} q^*(x) &\propto p(x) \sqrt{\sum_{y\in\{0,1\}} \left(\eta f_{\boldsymbol{\theta}}(x) + (1-\eta)y\right)^2 \left(1 - \ell_{0/1}(f_{\boldsymbol{\theta}}(x), y) - G_{F_\eta}\right)^2 p(y|x)} \\ &= p(x) \Big(\eta^2 f_{\boldsymbol{\theta}}(x) \left((1 - f_{\boldsymbol{\theta}}(x)) - G_{F_\eta}\right)^2 p(y=0|x) \\ &\quad + (1 - \eta(1 - f_{\boldsymbol{\theta}}(x)))^2 \left(f_{\boldsymbol{\theta}}(x) - G_{F_\eta}\right)^2 p(y=1|x) \Big)^{\frac{1}{2}} . \end{aligned}$$

Equation 4.18 follows by case differentiation according to the value of $f_{\boldsymbol{\theta}}(x)$. Finally, using $\eta = 1$ and $\eta = 0$ implies Equation 4.19 and 4.20 immediately. □

4.2.2 Empirical Sampling Distribution

In the previous section, we derived the sampling distribution minimizing asymptotically the estimation error of $\hat{R}_{n,q}$ and $\hat{G}_{n,q}$, respectively. Unfortunately, Theorem 4.1 and Corollaries 4.1-4.4 depend on the test distribution $p(x)$ and the true conditionals $p(y|x)$, which are typically unknown. The typical approach of active learning algorithms is to choose an instance to label depending on the model learned far (see Section 2.3). In this section, we transfer this idea to approximate the unknown quantities in a pool-based setting and discuss its consequences to the resulting sampling distribution.

Instances governed by the test distribution can be obtained in various ways: They can either be synthetically generated, selected from a stream of data, or can be sampled from a given pool. The focus of this thesis is the setting in which a large pool D_m of unlabeled test instances is available. Instances from this pool can be sampled and then labeled at a cost. Drawing instances from the pool replaces generating them under the test distribution; that is, we approximate $p(x)$ by the empirical distribution $\hat{p}(x)$ over the pool D_m (see Equation 2.29).

The optimal sampling distribution depends on the unknown (generalized) risk G we want to estimate, and the unknown true conditional $p(y|x)$. In order to implement the method, we have to approximate these quantities. Note that as long as $p(x) > 0$ implies $q(x) > 0$ for all instances with $w(x,y,f_{\boldsymbol{\theta}}) \neq 0$, any choice of $q(x)$ will yield consistent risk estimates because weighting factors $\frac{p(x_i)}{q(x_i)}$ account for the discrepancy between sampling and test distribution (see Lemma 3.4). That is, $\hat{G}_{n,q}$ is guaranteed to converge to G as n grows large; any approximation

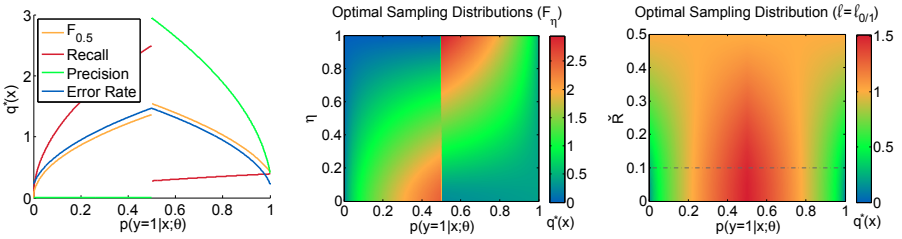


Figure 4.1: Instrumental distribution for F_η-measures and zero/one-loss over the predictive distribution in a pool-based setting for $\breve{G}_{rec} = \breve{G}_{prec} = 0.9 = 1 - \breve{R}_{0/1}$ (left). Heatmap of the sampling distribution when evaluating a model in terms of an F-measure, plotted into a two-dimensional space with axes $p(y = 1|x;\boldsymbol{\theta})$ and trade-off parameter η (center). Heatmap of the sampling distribution when evaluating the error rate of a model with axes $p(y = 1|x;\boldsymbol{\theta})$ and intrinsic error $\breve{R}$ (right).

employed to compute q^* will only affect the number of test examples required to reach a certain level of estimation accuracy.

To approximate the true conditional, we use the given predictive model $p(y|x) \approx p(y|x;\boldsymbol{\theta})$. Since G is equally dependent on $p(y|x)$ it is natural to replace it by an intrinsic risk calculated from Equation 3.1, in which the integral over $\mathcal{X}$ is replaced by a sum over the pool, $p(x) \approx \hat{p}(x)$, and $p(y|x) \approx p(y|x;\boldsymbol{\theta})$. The intrinsic generalized risk is given by

$$\breve{G}[f_{\boldsymbol{\theta}}] = \frac{\sum_{x \in D} \int w(x, y, f_{\boldsymbol{\theta}}) \ell(f_{\boldsymbol{\theta}}(x), y) p(y|x;\boldsymbol{\theta}) \mathrm{d}y}{\sum_{x \in D} \int w(x, y, f_{\boldsymbol{\theta}}) p(y|x;\boldsymbol{\theta}) \mathrm{d}y} \tag{4.21}$$

and the intrinsic regular risk by

$$\breve{R}[f_{\boldsymbol{\theta}}] = \frac{1}{|D|} \sum_{x \in D} \int \ell(f_{\boldsymbol{\theta}}(x), y) p(y|x;\boldsymbol{\theta}) \mathrm{d}y. \tag{4.22}$$

Figure 4.1 (left) shows the sampling distribution $q^*(x)$ for specific instances of the generalized risk functional as a function of the predictive distribution $p(y|x;\boldsymbol{\theta})$. If a binary classifier is evaluated with respect to the expected zero-one loss $\ell_{0/1}$, the empirical sampling distribution $q^*(x)$ is symmetric and prefers instances close

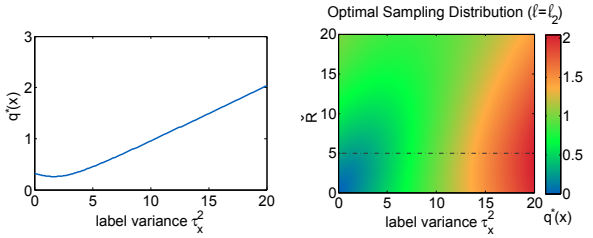


Figure 4.2: Instrumental distribution for squared-loss over the instance-specific variance τ_x^2 in a pool-based setting for $\breve{R}_2 = 5$ (left). Heatmap of the sampling distributions when evaluating the mean squared error of a model, plotted into a two-dimensional space with axes τ_x^2 and $\breve{R}$ (right).

to the decision boundary $p(f_{\boldsymbol{\theta}}(x) \neq y|x;\boldsymbol{\theta}) = 0.5$ (blue curve). We observe that the precision estimator (green curve) dismisses all examples which are classified as belonging to the negative class $f_{\boldsymbol{\theta}}(x) = 0$. This is intuitive because precision is a function of true-positive and false-positive examples only. Specifically, precision is equivalent to the accuracy of a predictive model $f_{\boldsymbol{\theta}}(x)$ conditioned to instances with $f_{\boldsymbol{\theta}}(x) = 1$ (see Section 3.3). Thus, the optimal sampling distribution for precision (see Equation 4.20) coincides with the sampling distribution for zero/one-loss (see Corollary 4.2) truncated to the range $p(y = 1|x, \boldsymbol{\theta}) > 0.5$. By contrast, the recall estimator (red curve) selects examples on both sides of the decision boundary, as it has to estimate both the true positive and false negative rate. Specifically, recall is entirely influenced by instances that belong effectively to the positive class $y = 1$. Hence, $q^*(x)$ prefers instances which are likely to be true positive ($y = 1 = f_{\boldsymbol{\theta}}(x)$) and false negative ($y = 1 \neq f_{\boldsymbol{\theta}}(x)$) in this case. In general, F-measures are a weighted harmonic mean of precision and recall, which is also reflected by the corresponding sampling distributions; the yellow curve corresponds to $\eta = 0.5$. Figure 4.1 (center) shows the influence of η; recall and precision belong to $\eta = 0$ and $\eta = 1$, respectively. Figure 4.1 (right) visualizes the asymptotical optimal sampling distribution for all possible values of the intrinsic risk $\breve{R}$ in the case of $\ell = \ell_{0/1}$. Note that the model approximation implies $\breve{R} \leq 0.5$. We observe that the intrinsic risk $\breve{R}$ acts as a smoothing parameter and ensures that all instances have some probability to be chosen: if the risk is assumed to be low, instances near the decision boundary are strongly preferred; if the risk is equal to random guessing ($\breve{R} = 0.5$), active risk estimation falls back to uniform sampling. In the case that the empirical risk is close to zero, the estimate is strongly biased because some instances are not likely to be drawn from the resulting empirical sampling distribution. However, bounding $\breve{R}$ by a

constant, would ensure that all instances have positive probability to be chosen in practice. Furthermore, this approximation could be improved while evaluating by an estimate obtained from the labeled sample.

A pre-condition of Corollary 4.3 is that the observed labels $y \sim \mathcal{N}(y|\mu_x, \sigma_x^2)$ follow a normal distribution; the optimal sampling distribution depends on the unknown mean μ_x and variance σ_x^2 of the label distribution. If $p(y|x)$ is approximated by the model $p(y|x;\boldsymbol{\theta})$ and, thus, the first and second moment of the label distribution are replaced by $\mu_x \approx f_{\boldsymbol{\theta}}(x)$ and $\sigma_x^2 \approx \tau_x^2$, we do assume that $p(y|x;\boldsymbol{\theta})$ instead of the true distribution is Gaussian. A normal predictive distribution is a standard assumption for probabilistic regression models. The variance τ_x^2 at instance x would typically be available from a probabilistic predictor, such as Gaussian processes or Bayesian (kernelized) linear regression (see Section 2.2). The sampling distribution that minimizes σ_q^2 (see Corollary 4.3) in a pool-based setting, in which $p(x) \approx \hat{p}(x)$, can thus be approximated by

$$q^*(x) \propto \sqrt{(3\tau_x^2 - 2\check{R}_2)\tau_x^2 + \check{R}_2^2}, \text{ where } \check{R}_2 = \frac{1}{|D|}\sum_{x \in D} \tau_x^2. \tag{4.23}$$

Figure 4.2 depicts the sampling distribution given by Equation 4.23.

In the following, we study some alternative approximations of the conditional distribution. First, consider a classification setting in which the model assessment is based on the zero-one loss or on F_η-measures. Instead of using the predictive model to approximate the unknown conditionals, an uninformative approximation $p(y|x) \approx \frac{1}{|\mathcal{Y}|}$ can be used. In this case, optimal sampling according to Equation 4.14 and Equation 4.18 for $\eta < 1$ degenerates to uniform sampling with corresponding estimators given by Equation 3.9 for regular and Equation 3.28 for generalized risks. We denote this baseline as *passive* estimator. Furthermore, $p(y|x) \approx p(y)$ could be replaced by a multinomial distribution over y rather than an uniform distribution. Then, the resulting sampling distribution is also multinomial; the likelihood of choosing an instance depends only on the predicted label $f_{\boldsymbol{\theta}}(x)$. In the context of spam filtering, this could be more appropriate than uniform sampling. Non-spam emails incur higher misclassification costs than spam emails. If evaluating a loss function which is defined by a costs matrix, it is more efficient to identify the infrequent class by using an empirical estimate of the class prior $p(y)$. In a regression setting, the *passive* estimator is equivalent to assuming equal variances $\sigma_x^2 = \sigma^2$ for all instances $x \in \mathcal{X}$. This approximation yield also a uniform sampling distribution.

Algorithm 2: Active Estimation for Generalized Risks

input Model $f_{\boldsymbol{\theta}}$ with distribution $p(y|x;\boldsymbol{\theta})$, pool D_m , labeling budget n.

1: Compute sampling distribution q^* according to Corollary 4.2, 4.3 or 4.4 using the predictive distribution $p(y|x;\boldsymbol{\theta})$.
2: **for** $i = 1, \ldots, n$ **do**
3: Draw $x_i \sim q^*(x)$ from D_m with replacement.
4: Query label $y_i \sim p(y|x_i)$ from oracle.
5: **end for**
6: Compute $\hat{G}_{n,q^*}[f_{\boldsymbol{\theta}}]$ (see Equation 3.31).

output Risk estimate $\hat{G}_{n,q^*}[f_{\boldsymbol{\theta}}]$.

Algorithm 2 summarizes the active evaluation algorithm. It samples n instances with replacement from the pool according to the distribution prescribed by Corollary 4.2 (for zero-one loss), 4.3 (for squared loss), or 4.4 (for F-measures) using the predictive distribution $p(y|x;\boldsymbol{\theta})$. Labels are queried for these instances. An interesting special case occurs when the labeling process is deterministic. Since instances are sampled with replacement, elements may be drawn more than once. In this case, labels of previously drawn instances can be looked up rather than be queried from the deterministic labeling oracle repeatedly: hence, the actual labeling costs may stay below the sample size. In this case, the loop may be continued until the labeling budget is exhausted.

4.3 Active Evaluation under Instance-Specific Costs

Active evaluation processes choose instances such that the expected estimation error is minimal for fixed labeling costs. Up to this point, costs for acquiring labels were assumed to be identical for all instances $x \in \mathcal{X}$. In fact, time and effort required vary according to instance-specific features. Consider the following examples. The task in text classification domains is to predict some category as the topic of a given text. In order to obtain labeled instances, a human expert reads a text—at least partly—and assigns one of the predefined categories. Labeling costs may depend on the length of a document, its language, and its technical difficulty. Inhomogeneous labeling costs also arise when learning or evaluating ranking functions for web search and other information retrieval domains. Ranking functions are used to sort a set of items, such as text documents or websites,

by relevance according to an user-defined query. In practice, a representative set of test queries is acquired by manually assessing the relevance of all retrieved items for each query. The number of retrieved items and possibly other item-specific features influence the time to determine the corresponding relevance.

In this section, we generalize our problem setting studied in Section 4.1 by allowing instance-specific labeling costs and constraining overall costs $\Lambda \in \mathbb{R}$ rather than the number of test instances n that can be drawn. We denote labeling costs for an instance x by $\lambda(x)$, and assume that $\lambda(x)$ is bounded away from zero by $\lambda(x) \geq \epsilon > 0$. Our goal is to minimize the deviation of $\hat{G}_{n,q}$ from G under the constraint that the expected overall labeling costs stay below a budget Λ:

$$\begin{aligned}(q^*, n^*) = \arg\min_{q,n} \; & \mathrm{MSE}_{(x,y)\sim q(x)p(y|x)} \left[\hat{G}_{n,q}\right], \\ \text{s.t. } & \mathbb{E}_{x\sim q(x)} \left[\sum_{i=1}^{n} \lambda(x_i)\right] \leq \Lambda.\end{aligned} \tag{4.24}$$

Note that Equation 4.24 represents a trade-off between labeling costs and informativeness of a test instance: optimization over n implies that many inexpensive or few expensive instances could be chosen.

The following theorem extends the optimal sampling results from Theorem 4.1 to instance-specific labeling costs $\lambda(x)$ in the sense that it minimizes the estimation error by minimizing the variance as its dominating component.

Theorem 4.2 (Cost-sensitive Optimal Sampling Distribution)**.** *Let G be defined as in Equation 3.27 and $\sigma_q^2 = \lim_{n\to\infty} n \operatorname{Var}[\hat{G}_{n,q}]$. The instrumental distribution that minimizes the asymptotic variance σ_q^2 of the generalized risk estimator $\hat{G}_{n,q}$ for instance-specific labeling costs λ is given by*

$$\begin{aligned}q^*(x) &\propto \frac{p(x)}{\sqrt{\lambda(x)}} \sqrt{\int w(x, y, f_{\boldsymbol{\theta}})^2 \left(\ell(f_{\boldsymbol{\theta}}(x), y) - G\right)^2 p(y|x) \mathrm{d}y}, \\ n^* &= \frac{\Lambda}{\int \lambda(x) q(x) \mathrm{d}x}.\end{aligned} \tag{4.25}$$

Before we prove Theorem 4.2, we state the following Lemma:

Lemma 4.2. *Let $a : \mathcal{X} \to \mathbb{R}$ and $\lambda : \mathcal{X} \to \mathbb{R}$ denote functions on the instance space such that $\int \sqrt{a(x)}\mathrm{d}x$ exists and $\lambda(x) \geq \epsilon > 0$. The functional*

$$W[q] = \left(\int \frac{a(x)}{q(x)} \mathrm{d}x \right) \left(\int \lambda(x) q(x) \mathrm{d}x \right),$$

where $q(x)$ is a distribution over the instance space $\mathcal{X}$, is minimized over q by setting

$$q(x) \propto \sqrt{\frac{a(x)}{\lambda(x)}}.$$

Proof. We have to minimize the functional

$$\left(\int \frac{a(x)}{q(x)} \mathrm{d}x \right) \left(\int \lambda(x) q(x) \mathrm{d}x \right) \tag{4.26}$$

in terms of q under the constraints $\int q(x)\mathrm{d}x = 1$ and $q(x) > 0$. We first note that Objective 4.26 is invariant under multiplicative rescaling of $q(x)$, thus the constraint $\int q(x)\mathrm{d}x = 1$ can be dropped during optimization and enforced in the end by explicitly normalizing the unconstrained solution. We reformulate the problem as

$$\min_q C \int \frac{a(x)}{q(x)} \mathrm{d}x \quad \text{s.t.} \quad C = \int \lambda(x) q(x) \mathrm{d}x \tag{4.27}$$

which we solve using a Lagrange multiplier γ by

$$\min_q C \int \frac{a(x)}{q(x)} \mathrm{d}x + \gamma \left(\int \lambda(x) q(x) \mathrm{d}x - C \right).$$

The optimal point for the constrained problem satisfies the Euler-Lagrange equation

$$\gamma \lambda(x) = C \frac{a(x)}{q(x)^2}, \tag{4.28}$$

and therefore

$$q(x) = \sqrt{C \frac{a(x)}{\gamma \lambda(x)}}. \tag{4.29}$$

Resubstitution of Equation 4.29 into the constraint (see Equation 4.27) leads to

$$C = \int \sqrt{C\frac{a(x)}{\gamma\lambda(x)}}\lambda(x)\mathrm{d}x, \tag{4.30}$$

solving for γ we obtain

$$\gamma = \frac{1}{C^2}\left(\int \sqrt{Ca(x)\lambda(x)}\mathrm{d}x\right)^2. \tag{4.31}$$

Finally, back substitution of Equation 4.31 into Equation 4.29 yields

$$q(x) \propto \sqrt{\frac{a(x)}{\lambda(x)}}.$$

□

We now prove Theorem 4.2.

Proof of Theorem 4.2. We first study the minimization of $\frac{1}{n}\sigma_q^2$. Because

$$\mathbb{E}_{x\sim q(x)}\left[\sum_{i=1}^{n}\lambda(x_i)\right] = n\int \lambda(x)q(x)\mathrm{d}x,$$

the minimization problem can be reformulated as

$$\min_q \min_n \frac{1}{n}\sigma_q^2 \text{ s.t. } n \leq \frac{\Lambda}{\int \lambda(x)q(x)\mathrm{d}x}.$$

Clearly $n^* = \frac{\Lambda}{\int \lambda(x)q(x)\mathrm{d}x}$ solves the inner optimization. The remaining minimization over q is

$$q^* = \arg\min_q \ \sigma_q^2 \int \lambda(x)q(x)\mathrm{d}x.$$

Lemma 3.4 implies

$$\sigma_q^2 \propto \iint \frac{p(x)^2}{q(x)^2} w(x, y, f_{\boldsymbol{\theta}})^2 \left(\ell(f_{\boldsymbol{\theta}}(x), y) - R\right)^2 p(y|x)q(x)\mathrm{d}y\,\mathrm{d}x.$$

Setting

$$a(x) = p(x)^2 \int w(x, y, f_{\boldsymbol{\theta}})^2 \left(\ell(f_{\boldsymbol{\theta}}(x), y) - G\right)^2 p(y|x) \mathrm{d}y$$

and applying Lemma 4.2 implies Equation 4.25. □

Intuitively, the optimal sampling distribution given by Equation 4.25 constitutes a trade-off between labeling costs and informativeness. It gives preference to instances with low labeling costs and for which the expected loss deviates strongly from the expectation taken over all instances. In the case, in which the labeling costs $\lambda(x)$ of an instance x is exactly correlated with its expected (weighted) deviation

$$\mathbb{E}_{y \sim p(y|x)} \left[w(x, y, f_{\boldsymbol{\theta}})^2 \left(\ell(f_{\boldsymbol{\theta}}(x), y) - G\right)^2 \middle| x \right],$$

the optimal sampling distribution degenerates to sampling from $p(x)$.

In the next chapter, we study the active evaluation method with homogeneous and non-homogeneous labeling costs empirically.

4.4 Empirical Results

In this section we empirically study the estimation performance of active and passive evaluation methods. Estimating the risk of a model $f_{\boldsymbol{\theta}}$ on separately drawn test instances is motivated by scenarios, where we cannot obtain a risk estimate from the original training data (e.g., by cross-validation). Overall, we conduct experiments in the following application domains, in which the training data do not reflect the test distribution or are not available.

Spam Filtering Domain. In this domain (referred to as *EMAIL*), spammers impose a shift on the distribution of instances over time as they employ new strategies to generate spam messages. A classifier that has been trained in the past has to be evaluated with respect to the current distribution. We collected 169,612 emails from an email service provider between June 2007 and April 2010. Emails are represented using a binary bag-of-words, resulting in 541,713 distinct features; approximately 5% of all emails are spam.

Digit Recognition Domain. The digit recognition domain reflects an application scenario in which a classification system is procured and evaluated in an environment in which the input distribution may diverge from the training distribution. To realize this scenario, a digit recognition model is trained on the *USPS* data set and evaluated on the *MNIST* data set and vice versa. The *MNIST* database contains 70,000 images of scanned handwritten digits collected by the National Institute of Standards and Technology (NIST); we use a version of *MNIST* prepared by Sam Roweis. The *USPS* data set containing 11,000 instances, which were accrued in the context of a project sponsored by the US Postal Service (Hull, 1994). All digits occur roughly in the same proportion as in *MNIST*. We rescale the *MNIST* images from 28×28 to 16×16 to match the resolution of *USPS* and re-compute the bounding box. The rescaled *MNIST* images differ visually from the *USPS* images, the line strokes are generally thicker.

Text Classification Domain. The *REUTERS-21578* text classification task (Frank & Asuncion, 2010) allows us to study the effect of class skew, and serves as a prototypical domain for active learning. It contains 8,293 newswire articles represented as term-frequency-vectors and categorized into 65 topics. We experiment on the ten most frequently occurring topics, leaving 7,285 documents. Table A.1 in Appendix A.2.1 lists the class ratios.

Inverse Dynamics Domain. We use the *Sarcos* data set, containing 48,933 instances described by 21 features (Vijayakumar et al., 2005). In this regression problem, the task is to predict one of seven torques based on the motions of a seven degrees-of-freedom anthropomorphic robot arm.

Abalone Domain. We also use the *Abalone* benchmark data set (Frank & Asuncion, 2010), which includes 4,177 instances. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope. In this regression problem, the task is to predict the age of Abalone from ten physical measurements including length, diameter, and weight.

We employ kernelized logistic regression (see Section 2.1) for classification tasks. Hyperparameters are tuned a priori on the training portion of each data set by cross validation and then kept fixed. It is common to use an RBF kernel (see Equation 2.24) for digit recognition. The kernel width ς is tuned by maximizing the pairwise *generalized maximum mean discrepancy* (Sriperumbudur et al., 2009) between the class-conditional distributions $p(x|y)$ estimated by the respective

sample. For regression tasks, we employ Gaussian processes (see Section 2.2), using Bayesian model selection to determine the hyperparameters (Rasmussen & Williams, 2006). These predictive models provide us with an estimate of $p(y|x)$ and τ_x^2, respectively.

In each experiment, we train a model on the training data set and obtain an active estimate on the evaluation data set using Algorithm 2 (denoted *active*). As a baseline, we obtain a risk estimate using test instances drawn uniformly from the pool (denoted *passive*). For classification, we also study the online stratified sampling method proposed by Bennett & Carvalho (2010). The pool of instances is divided into disjoint strata based on the confidence of the classifier. In each iteration, we choose a stratum, draw an instance uniformly from that stratum, and query the label. The optimal strategy is to choose strata proportional to the standard deviation of the labels. The standard deviation is estimated iteratively by the queried labels. We split the data such that each stratum contains an equal portion of the instances, since it is more accurate than using strata of equal ranges (Bennett & Carvalho, 2010). This baseline is denoted by *strat*.

The evaluation process is repeated 1,000 times and results are averaged. In case one of the repetitions results in an undefined estimate, the entire experiment is discarded (*i.e.,* there is no data point for the method in the corresponding diagram). This can occur for example when measuring recall and only negative instances $y = 0$ were drawn. All methods operate on identical labeling budgets. In order to assess the estimation error $|\hat{G}_{n,q} - G|$ we determine the risk of the model on the entire evaluation data set D_m and use it as an approximation of the true risk; that is $\hat{G}_m \approx G$.

4.4.1 Estimating the Performance of a Model

In this section we study whether—and under which conditions—active risk estimation and active estimation of F_η-measures can lead to more accurate estimates than risk estimation based on a uniformly drawn sample. We now consider several regression and classification scenarios and determine the estimation error as a function of the labeling budget n, or Λ if labeling costs vary across the instances.

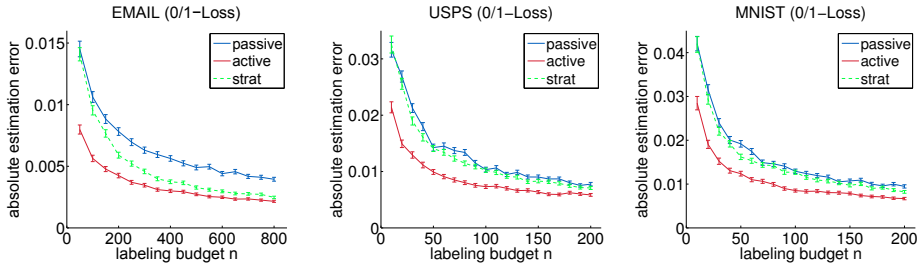


Figure 4.3: Absolute deviation from pool error over number of labeled data for spam filtering and digit recognition domain. Error bars indicate the standard error.

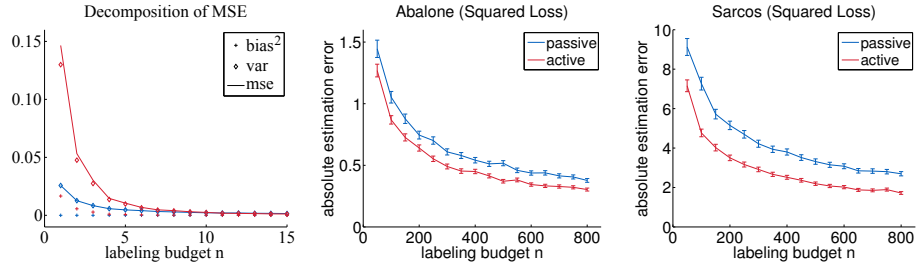


Figure 4.4: Empirical decomposition of the mean squared error in the spam filtering domain for passive (blue curve) and active (red curve) evaluation (left). Estimation error for squared loss over labeling costs for the regression tasks Abalone (center) and Sarcos (right). Error bars indicate the standard error.

Evaluation under Distribution Shift

Firstly, we study the estimation error for regular risks, precisely error rate and mean squared error as a function of the number of labeled test instances n. In the spam filtering domain, we use the first 42.165 emails received by February 2008 as training portion and the set of 33,296 emails received between February 2008 and October 2008 as evaluation portion. For digit recognition, we consider the popular problem of distinguishing between digits "4" and "9" which are easily confused; this results in 13,782 instances for the *MNIST* database, and 2,200 instances for the *USPS* database. Either data set are used as training and the other as test data.

Figure 4.3 shows the average absolute deviation between the risk estimate and the true risk for active and passive risk estimation for *EMAIL*, *MNIST-USPS*,

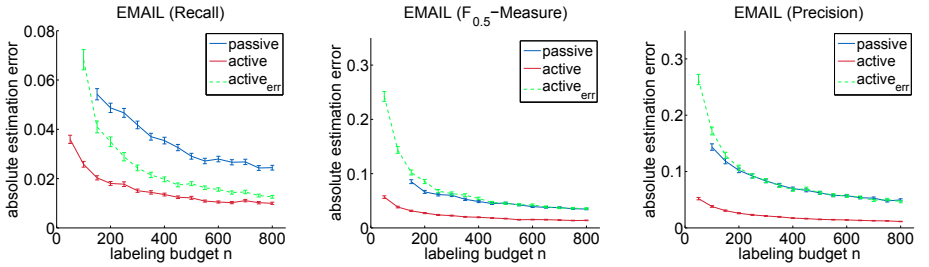


Figure 4.5: Estimation error for recall, $F_{0.5}$ and precision over labeling costs for spam filtering. Error bars indicate the standard error.

and *USPS-MNIST* problems. Error bars indicate the standard error; the zero-one risk on the entire pool of test instances is 0.0245 for *EMAIL*, 0.0205 for *USPS*, and 0.0280 for *MNIST*. Figure 4.4 (center, right) shows the estimation error for the regression tasks *Abalone* and *Sarcos*. Each model is trained on a randomly selected set of 500 instances and is evaluated by the risk on the remaining instance; the mean squared error on the entire pool is 5.00 and 27.12, respectively. In all learning problems, active risk estimates are significantly more accurate than passive risk estimates or, equivalently, a desired level of accuracy is achieved with significantly fewer test instances. For example, in the spam filtering domain, active evaluation with 300 test instances achieves approximately the same accuracy as passive evaluation with 800 instances. For *EMAIL* the online stratified sampling approach *strat* outperforms passive sampling for sufficiently many labeled instances; it relies on an estimate of the standard deviation within each strata. In the digit recognition domain, *strat* attains only a slightly lower estimation error than *passive*. This may be due to an inaccurate estimate of the stratas' standard deviation or an inadequate choice of the number of strata.

The optimal sampling distributions derived in Section 4.2.1 approximately minimize the estimation error by minimizing the asymptotic variance σ_q^2. This is motivated by the bias-variance decomposition. To study the error of this approximation we measure the deviation of the averaged estimates $\hat{R}_n$ and $\hat{R}_{n,q}$ over 1,000 repetitions from estimate over the entire pool $\hat{R}_{|D|}$ and the respective empirical variance. Figure 4.4 (left) shows the decomposition of the mean squared error in the spam filtering domain. In fact, except for very small sample size n, the error is clearly dominated by the variance of the estimator whereas the bias is negligible.

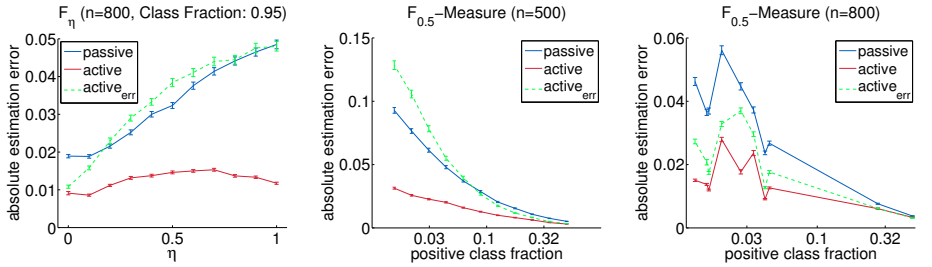


Figure 4.6: Absolute error for F_η-measure estimates for different values of η (left) and over class ratio on a logarithmic scale in the spam filtering domain (center) and text classification for all ten classes (right). Error bars indicate the standard error.

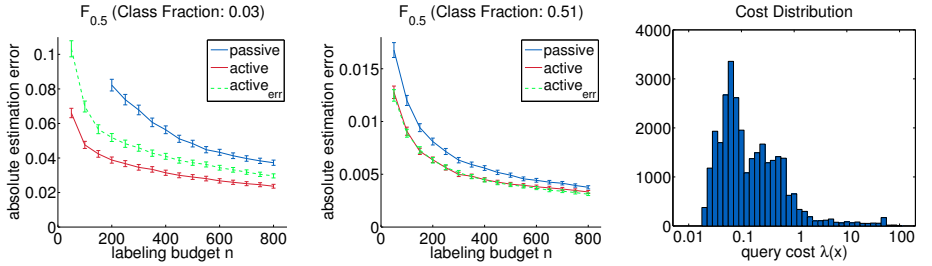


Figure 4.7: Text classification: Estimation error over number of labeled data for infrequent (left) and frequent (center) class. Distribution of query labeling costs $\lambda(x)$ in spam filtering (right).

We now compare active estimation of F_η-measures according to estimation based on a sample of instances drawn uniformly from the pool. As a baseline, we revisit the active estimator for zero-one loss. It might perform comparable even if evaluating F_η-measures, since both the optimal sampling distributions for the error rate as well as for $F_{0.5}$ (in particular if $F_0 \approx F_1$) prefer instances close to the decision boundary (see Figure 4.1). For this baseline (denoted $active_{err}$) instances are drawn according to the optimal sampling distribution $q^*_{0/1}$ for zero-one risk (see Equation 4.14), however the F_η-measure is computed according to Equation 3.26 using $q = q^*_{0/1}$. For the spam filtering domain, Figure 4.5 shows the average absolute estimation error for F_0 (recall), $F_{0.5}$, and F_1 (precision) estimates. True precision $\hat{G}_{|D|}$ and recall on the entire pool of test instances are 0.693 and 0.977, respectively. The active generalized risk estimate *active* significantly outperforms the passive estimate *passive* for all three measures. In order to reach the estimation accuracy of *passive* with a labeling budget of $n = 800$, *active* requires

fewer than 150 (recall), 180 ($F_{0.5}$), or 100 (precision) labeled test instances. Results obtained in the digit recognition domain are consistent with these findings (see Appendix A.2.2). Figure 4.6 (left) shows results for estimating several intermediate stages of F_η for a class ratio of 95% to 5%. Estimates obtained from *active* are at least as accurate as those of $active_{err}$, and more accurate for high η values.

F_η-measures are particularly suited for highly skewed prediction problems when measuring accuracy is not appropriate. This raises the question how strongly the benefit of the corresponding sampling distributions depend on the class distribution. In the spam filtering domain we artificially sub-sampled data to different ratios of spam and non-spam emails. Figure 4.6 (center) shows the performance of *active*, *passive*, and $active_{err}$ for $F_{0.5}$ estimation as a function of class skew. We observe that *active* outperforms *passive* consistently. Furthermore, *active* outperforms $active_{err}$ for imbalanced classes, while the approaches perform comparably when classes are balanced. This finding is consistent with the intuition that the values of accuracy and F-measure diverge more strongly for imbalanced classes.

In the text classification domain we estimate the $F_{0.5}$-measure for a ten-class classifiers. Class frequencies range from 0.012 for the smallest class to 0.51 for the majority class. We use the *active*, *passive*, and $active_{err}$ approaches to solve the correspondingly skewed per-class one-vs-rest $F_{0.5}$ estimation problem. Figure 4.7 shows the estimation error of *active*, *passive*, and $active_{err}$ for an infrequent class ("money-fx", left) and a frequent class ("earn", center). These results are representative for other frequent and infrequent classes; all results are included in the Appendix A.2.1. Figure 4.6 (right) shows the estimation error of *active*, *passive*, and $active_{err}$ on all ten one-versus-rest problems as a function of the problem's class skew (data points correspond to the ten different one-vs-rest estimation problems). We again observe that *active* outperforms *passive* consistently, and *active* outperforms $active_{err}$ for strongly skewed class distributions.

Evaluation of Actively Trained Models

Active learning can result in more accurate models than learning from uniformly sampled training examples (passive learning), but it has the disadvantage that risk estimates obtained on held-out training examples are severely biased (Schütze et al., 2006). In order to obtain an unbiased estimate of the risk of an actively learned model, additional test examples have to be labeled, which again increases

Table 4.1: Active vs. passive learning and active vs. passive risk estimation. Values in parenthesis indicate standard errors.

		EMAIL	MNIST	USPS
passive learning (3)	**model error**	0.1033 (0.00051)	0.0251 (0.00013)	0.0355 (0.00015)
	estimation error cross validation	0.0245 (0.00060)	0.0112 (0.00028)	0.0118 (0.00029)
active learning	**model error**	0.0492 (0.00018)	0.0070 (0.00004)	0.0272 (0.00007)
	estimation error active eval. (1)	0.0137 (0.00033)	0.0046 (0.00020)	0.0084 (0.00022)
	passive eval. (2)	0.0172 (0.00042)	0.0063 (0.00014)	0.0123 (0.00032)

the labeling costs. We will now study whether the combination of active learning and active risk estimation can outperform passive learning and risk estimation by cross validation on a uniformly drawn labeled sample.

We employ logistic regression as base learning algorithm; the active learner always selects the example with minimal functional margin

$$x = \arg\min_{\bar{x} \in D} \left(p(f_{\boldsymbol{\theta}}(\bar{x}) | \bar{x}; \boldsymbol{\theta}) - \max_{y \neq f_{\boldsymbol{\theta}}(\bar{x})} p(y | \bar{x}; \boldsymbol{\theta}) \right)$$

and updates the model. This is a straightforward multiclass adaption of the well-known greedy active learning strategy (see Chapter 2.3). We fix the labeling budget to $n = 220$ and compare the following three learning and evaluation protocols.

Protocol (1) draws 20 instances uniformly from the pool, trains an initial model, and then selects 100 additional training instances actively. The model is evaluated on further 100 test instances selected by the active risk estimation procedure.

Protocol (2) trains a model on an initial 20 uniformly drawn and an additional 100 actively selected training instances, and evaluates the model on 100 uniformly-drawn instances.

Protocol (3) draws 220 instances uniformly from the pool and runs 10-fold cross validation.

We consider these settings for *EMAIL*, *USPS* and *MNIST*. In this scenario, we are interested in both obtaining an accurate model and an accurate estimate of the risk of this model.

Table 4.1 shows the true risk over the entire pool (model error) and average absolute deviation of the estimated risk (estimation error) for strategies (1) to (3). Values in parenthesis indicate standard errors. Active learning consistently gives more accurate models than passive learning, even though models are trained on smaller samples. Moreover, we again observe that active risk estimation consistently outperforms passive risk estimation. Note that in all three domains the combination of active learning and active evaluation gives both the most accurate model and the most accurate risk estimate.

Evaluation with Instance-Specific Costs

In this section, we study active risk estimation processes under instance-specific costs in comparison to passive evaluation. To quantify the effect of modeling costs, we also consider active risk estimation according to Corollary 4.2 that assumes uniform costs $\lambda(x) = 1$ (labeled $\mathit{active}_{\mathit{uniC}}$) and a heuristic sampling distribution $q(x) \propto \lambda(x)^{1/2}$ only taking into account costs ($\mathit{active}_{\lambda}$).

In spam filtering, labels have been created by a human expert. The expert read—at least partially—the email and decides if it belongs to spam or non-spam. Furthermore, each email whose label is queried incurs storage costs. Since the actual labeling costs are unknown, we follow a semi-artificial setting to illustrate the effect of costs. We model the labeling effort as proportional to the corresponding file size; the cost unit is chosen such that average labeling costs for a query are one. Figure 4.7 (right) shows the distribution of labeling costs $\lambda(x)$.

Figure 4.8 (center) shows the average absolute deviation between the risk estimate and the true risk when the labeling budget depends on the storage costs of the labeled emails. Estimates obtained from *active* with a labeling budget of $\Lambda = 100$ are as accurate as *passive* for $\Lambda = 800$. We observe that modeling both the costs (indicated by $\mathit{active}_{\lambda}$) as well as the variance of the estimator (indicated by $\mathit{active}_{\mathit{uniC}}$) reduces the estimation error for fixed Λ. In the following, we study the influence of these two factors on the estimation error.

The benefit of the costs-sensitive sampling distribution over passive or active evaluation with uniform costs depends on the relation between costs $\lambda(x)$ and label uncertainty

$$u(x) = \mathbb{E}_{y \sim p(y|x;\boldsymbol{\theta})} \left[\left(\ell(f_{\boldsymbol{\theta}}(x), y) - R \right)^2 \middle| \, x \right]$$

namely the denominator and numerator of Equation 4.25; in the case in which expensive instances reduce the label uncertainty most, that is, $u(x)$ and $\lambda(x)$ are equal up to a constant factor, the optimal sampling distribution degenerates to random sampling, if $\lambda(x)^{-1} = u(x)$ *active* and $\textit{active}_{\textit{uniC}}$ coincide. We measure this dependence in terms of the Pearson product-moment correlation coefficient which is given by

$$\rho = \frac{\mathrm{Cov}_{x\sim p(x)}[u(x), \lambda(x)^{-1}]}{\sqrt{\mathrm{Var}_{x\sim p(x)}[u(x)]\,\mathrm{Var}_{x\sim p(x)}[\lambda(x)^{-1}]}}.$$

The correlation between costs and label uncertainty on the entire data set *EMAIL* is $\rho = 0.002$. In order to study different correlations, we reassign the labeling costs $\lambda(x)$ to the instances regardless of their actual file sizes. Resorting the labeling costs in a way such that the costs increase with the labeling uncertainty results in $\rho = 0.625$. The contrary case yields $\rho = -0.265$. Figure 4.8 shows the absolute estimation error for positive ($\rho = 0.625$), original ($\rho = 0.002$) and negative ($\rho = -0.265$) correlation between $u(x)$ and $\lambda(x)^{-1}$ as a function of the total labeling budget Λ. In the first case (left) we observe that $\textit{active}_{\textit{uniC}}$ and $\textit{active}_{\lambda}$ perform similarly which indicates that *active* profits equally from preferring cheap and informative instances. In the original setting file sizes and informativeness are slightly positively correlated. Here, $\textit{active}_{\lambda}$ attains a slightly lower absolute estimation error than $\textit{active}_{\textit{uniC}}$. Finally, in the case of a negative correlation, $\textit{active}_{\lambda}$ achieves significant better results than $\textit{active}_{\textit{uniC}}$, whereas the estimation accuracy of *active* and *passive* coincide. Figure 4.9 (left) shows the average of the used labeling budget when it is limited by $\Lambda = 200$. For all methods the budget is nearly exhausted. However, we can observe that $\textit{active}_{\lambda}$ comes closest to the stated limit independently of ρ, whereas budget used by *active* depends on the correlation between $u(x)$ and $\lambda(x)^{-1}$. Also the number of instances which are actually drawn depends on if the corresponding sampling distribution accounts for the costs $\lambda(x)$ or not; for *passive* and $\textit{active}_{\lambda}$ the number of drawn instances n is independent of the correlation ρ, whereas for *active* and $\textit{active}_{\textit{uniC}}$ n grows with the correlation ρ (see Figure 4.9, center).

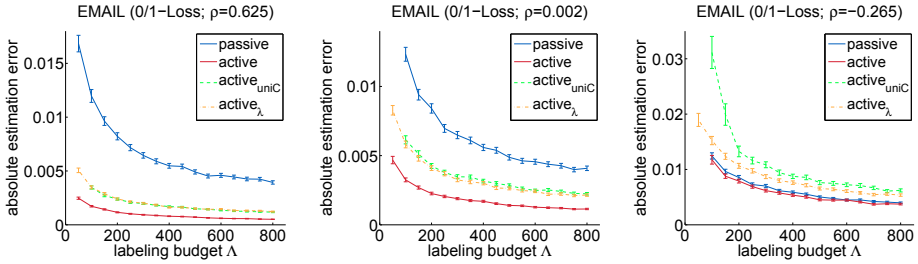


Figure 4.8: Estimation error for zero-one loss over instance-specific labeling costs for different correlation between costs and label uncertainties for spam filtering. Error bars indicate the standard error.

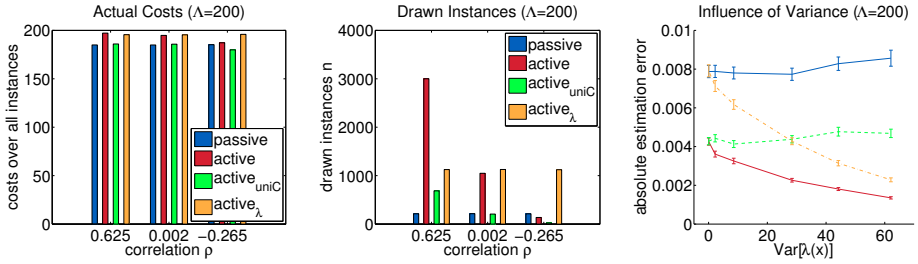


Figure 4.9: Actual used budget (left) and corresponding average number of drawn instances (center) for a fixed overall labeling budget of $\Lambda = 200$ in the spam filtering domain. Absolute estimation error over variances of the empirical cost distribution (right). Error bars indicate the standard error.

The impact of non-homogeneous instance-specific costs is also affected by the variance of the cost distribution. In particular, *active* reduces to $active_{uniC}$ if the variance goes to zero. In order to vary the variance of the empirical distribution of instance-specific costs, the original costs $\lambda'(x) \propto \lambda(x)^c$ are exponentiated with a positive number c and are normalized; a lower variance of the cost distribution can be obtained using a value $0 < c < 1$ between zero and one, a value above one increases the variance. Figure 4.9 (right) depicts the effect of cost variance. The benefit of *active* and $active_\lambda$ over *passive* and $active_{uniC}$ increases with higher variance. This is intuitive, since a high variance of the cost distribution yields instances with low labeling costs. Choosing these instances more likely reduces the variance of the estimator if $u(x)$ and $\lambda'(x)^{-1}$ are not strongly positive correlated.

4.4.2 Influence of the Predictive Distribution

Active evaluation relies on the model's estimate of the output probability in order to select uncertain instances from the pool. When training and test distributions differ, the approximation $p(y|x) \approx p(y|x;\boldsymbol{\theta})$ may become poor. In order to study the relation between the quality of $p(y|x;\boldsymbol{\theta})$ and the benefit of active (generalized) risk estimation, we consider two experimental setups.

First, in the spam filtering and digit recognition domain we let the size of the training sample vary over all powers of two, within the size of the available data sets. We evaluate each model 1,000 times actively and passively, and determine the average ratio $\frac{|\hat{G}_n - G|}{|\hat{G}_{n,q^*} - G|}$. A value of above one indicates that the active estimate is more accurate than a passive estimate. In order to probe the limitation of the active risk estimation model, we additionally train and evaluate a naïve Bayes classifier. Naïve Bayes delivers poorly calibrated probability estimates because the inaccuracy caused by its inherent independence assumption grows exponentially in the number of attributes. Figure 4.10 (left) shows the results in the spam filtering and digit recognition domain; each point corresponds to a model with fixed training set size. The horizontal axis quantifies the quality of $p(y|x;\boldsymbol{\theta})$ in terms of the theoretical label likelihood $\mathcal{L}(\boldsymbol{\theta})$ of the given model $\boldsymbol{\theta}$ (see Equation 2.8). It is estimated by the average per instance test likelihood on the entire evaluation data set D_m; that is $\hat{\mathcal{L}}_m(\boldsymbol{\theta}) \approx \mathcal{L}(\boldsymbol{\theta})$. For model likelihoods of 0.6 and above (corresponding to at least eight training instances), active evaluation outperforms passive evaluation, the advantage of active risk estimation grows with the model likelihood. The three leftmost points correspond to naïve Bayes: The likelihood of the naïve Bayesian model is close to zero as it misclassifies several test instances with extreme over-confidence. Active risk estimation rarely selects such over-confident misses; hence, for naïve Bayes, passive outperforms active risk estimation. To ensure that all instances have some probability to be chosen, the intrinsic risk could be bounded (see Section 4.2.2).

Additionally, we consider the gain of active over passive estimation in the spam filtering domain as a function of the discrepancy between training and testing data over time. To this end, we keep the training set of emails fixed and move the time interval from which test instances are drawn increasingly further away into the future, thereby creating a growing gap between training and test distribution. Specifically, we divide 127,447 emails from the *EMAIL* data set received between February 2008 and April 2010 into ten different test sets spanning approximately

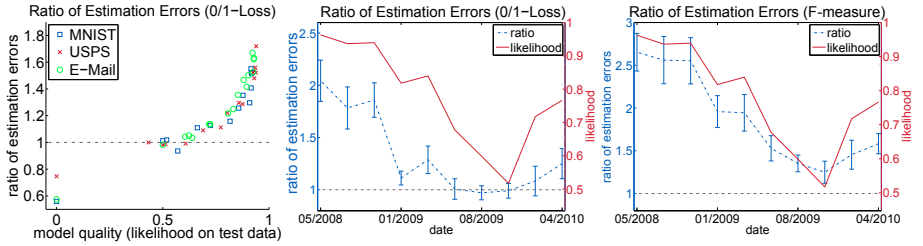


Figure 4.10: Ratio of estimation error of passive and active risk estimates for different models and learning tasks (left) and over time in the spam filtering domain for zero-one loss (center) and $F_{0.5}$-measure (right). The horizontal line indicates the break-even point. Error bars indicate standard errors

2.5 months each. The results for estimating error rate and $F_{0.5}$-measure are depicted in Figure 4.10 (center and right). The red curves show the discrepancy between training and test distribution measured in terms of the average per instance test likelihood. The likelihood at first continually decreases. It grows again for the two most recent batches; this coincides with a recent wave of text-based *vintage spam*. Blue curves also show the ratio of passive-to-active estimation errors $\frac{|\hat{G}_n - G|}{|\hat{G}_{n,q^*} - G|}$. The active estimate consistently outperforms the passive estimate; its advantage diminishes when training and test distributions diverge and the assumption of $p(y|x) \approx p(y|x;\boldsymbol{\theta})$ becomes less accurate. Particularly, as the average per-instance label likelihood falls towards 0.5, the optimal sampling distribution coincides with random guessing, and consequently the gain of active estimation vanishes.

4.4.3 Validation of Confidence Intervals

We have derived confidence intervals for active estimates in Section 3.1.2. These intervals are approximate for finite n. We now investigate how accurately the empirical coverage of the intervals matches the desired confidence level of $1 - \alpha$ and compare the width of the confidence intervals for active and passive evaluation. For the evaluation method *active* we use the Wald-interval (denoted $active_{Wald}$) based on the standard normal distribution for weighted test samples. For *passive*, that draws an unweighted sample of test instances, we compute intervals based on a two-sided t-test (denoted $passive_{t\text{-}Test}$). We do not include

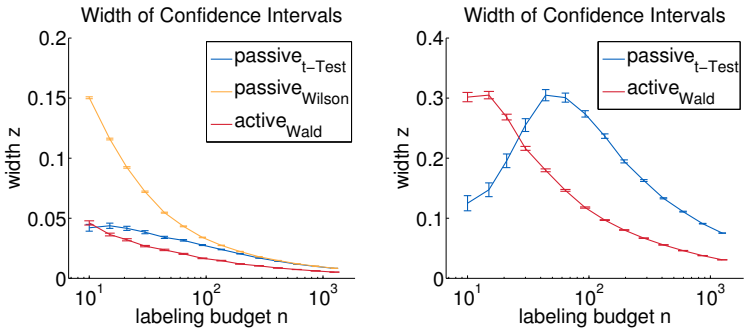


Figure 4.11: Width of confidence intervals for active and passive estimation of $R_{0/1}$ (left) and $F_{0.5}$ (right). Error bars indicate standard errors.

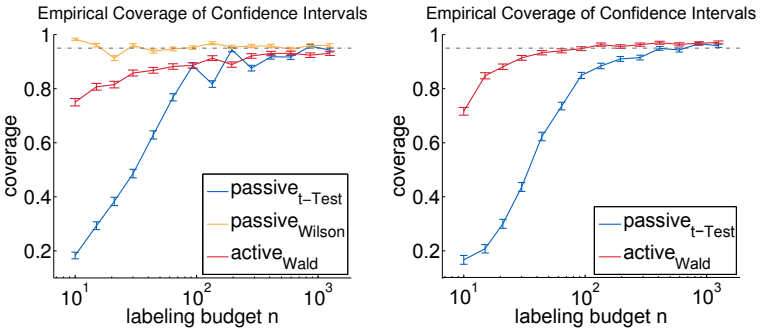


Figure 4.12: Empirical confidence levels for active and passive estimation of $R_{0/1}$ (left) and $F_{0.5}$ (right). Error bars indicate standard errors.

Wald intervals for passive evaluation, since the difference is negligible for the considered sample sizes (see Section 3.2.2). Additionally, we also consider the Wilson-interval (denoted $passive_{Wilson}$), which is more reasonable for error-rate estimates. Figure 4.11 depicts the width z_α of these confidence intervals for $R_{0/1}$ (left) and $F_{0.5}$ (right) estimates in the spam filtering domain and a confidence level of $1 - \alpha = 0.95$. Wald intervals obtained from active risk estimation are significantly tighter than those of passive risk estimation (standard confidence interval based on the t-test and Wilson interval).

We also investigate how accurately the empirical coverage of the intervals matches the desired confidence level of $1 - \alpha$. Figure 4.12 shows the fraction of iterations in which the true (generalized) risk lies within the confidence interval derived from active and passive risk estimation, determined over 1,000 repetitions of the

evaluation process. The empirical coverage of actively determined confidence intervals matches the desired confidence level more closely than the standard intervals ($passive_{t\text{-}Test}$). Still, at first glance it may appear surprising that empirical coverages are uniformly lower than the prescribed theoretical confidence levels. However, it is well-known that confidence intervals are only asymptotically correct (Wasserman, 2004, Section 6.3.2). On small test samples, empirical risks of zero occur regularly. An empirical risk of zero leads to an empirical variance of zero which in turn collapses the confidence interval into a single point (see Section 3.1.2). In contrast, the empirical coverage of Wilson intervals are reliable, however at the expense of considerably increased interval ranges (see Figure 3.1, bottom). Finally, note that the empirical coverage of $passive_{t\text{-}Test}$ and also $passive_{Wilson}$ oscillates due to the discrete lattice structure of the binomial distributed estimators, whereas it increases more smoothly with an increasing labeling budget for $\hat{R}_{n,q}$. For a detailed discussion see Section 3.1.2.

4.5 Summary and Related Work

This chapter has studied a setting in which a given model is to be evaluated at minimal labeling costs using test instances that can be selected from a large pool of unlabeled test data. We contribute an active evaluation procedure for generalized risks, including regular risks as well as F_η-measures, whose sampling distributions q^* minimize the asymptotic variance—and thus asymptotically the estimation error—of a self-normalized importance sampling estimator. We also extend our investigation to the case that instance-specific—not necessarily homogeneous—labeling costs are given. The derived instrumental distribution constitutes a trade-off between drawing informative and inexpensive instances. Active estimation can be applied immediately with a probabilistic classifier. Uncalibrated decision function values (such as an SVM would produce) have to be calibrated using, for instance, a one-dimensional logistic or isotonic regression on the decision function value.

Empirically, we observe that active evaluation outperforms passive evaluation when the model has a certain quality—a per-instance label likelihood of 0.6 or above. Active estimation performs poorly in combination with a naïve Bayesian classifier which delivers poorly calibrated class probabilities. In experiments with spam and handwriting recognition problems, we observed active estimates to be as accurate as passive estimates based on three times as many test examples.

Furthermore, we observed that a combination of active learning and active estimation produces more accurate models and more accurate risk estimates than cross validation on an equally large uniformly drawn sample. Active estimation for F_η-measures is preferable in particular for skewed classes. In order to reach the estimation accuracy based of an uniform drawn sample of size $n = 800$, the active estimation procedure requires fewer than 150 (recall), 200 ($F_{0.5}$-measure), or 100 (precision) labeled test instances in the spam filtering domain. Modeling instance-specific costs give an additional benefit if they are not anti-correlated with the label uncertainty of the model. Finally, we observe the confidence intervals of active risk estimates to be tighter and more reliable even for small test samples.

The presented approach can be seen as an application of the general technique of importance sampling (Hammersley & Handscomb, 1964) to the problem of estimating the performance of prediction models. Note that our approach exploits the predictive distribution defined by the model to be evaluated to derive the (approximately) optimal importance sampling distribution. In the context of sampling-based state inference in hidden Markov models, Cappé et al. (2005) quantify the variance of a self-normalized importance sampler.

Another approach to reduce the variance of an estimate is stratified sampling. In general, stratified sampling methods divide an inhomogeneous population into a number of disjoint and more homogeneous subpopulations called strata. Sampling uniformly from each stratum and combining the subpopulation estimates by a weighted average yields the final estimate. For an appropriate allocation of the strata, stratified sampling can be more accurate than uniform sampling. Bennett & Carvalho (2010) derive a procedure to evaluate binary classifiers based on an allocation proportional to the standard deviation of the true labels. The standard deviation per stratum is estimated iteratively from the labeled instances and the number of strata is fixed beforehand. Due to the update of the sampling distribution, resulting confidence intervals based on the Gaussian approximation are biased. In contrast to our optimal instrumental distribution, the sampling scheme does not depend on the calibration of the confidence scores. Finally, Druck & McCallum (2011) extend this approach and use the model-based predictive distribution to approximate the variance in each individual stratum.

Active evaluation can be considered to be a dual problem of active learning; in active learning, the goal of the selection process is to minimize the variance of the predictions or the variance of the model parameters, while in active evaluation the variance of the risk estimate is reduced (see Section 2.3). In analogy to our

approach, active learning algorithms use a current model to decide on instances whose class labels are queried. Specifically, Bach (2006) derives a sampling distribution under the assumption that the current model gives a good approximation to the conditional probability $p(y|x)$. Several active learning algorithms use importance weighting to compensate for the bias incurred by the instrumental distribution: for regression (Sugiyama, 2006), exponential family models (Bach, 2006), or SVMs (Beygelzimer et al., 2009). In many scenarios, labeling costs can vary over different instances. The optimal sampling distribution, which accounts for different labeling costs, extends the variance minimizing distribution by the reciprocal of the squared root of instance-specific costs. Costs-sensitive active learning strategies are proposed by Haertel et al. (2008) and Settles et al. (2008). Haertel et al. (2008) found that this sampling heuristic is effective for part-of-speech tagging.

Both, active learning and evaluation procedures collect labels and thus information about the true test distribution. Active learning algorithms exploit this knowledge to improve the predictive model and thus the model-based output probability $p(y|x;\boldsymbol{\theta})$, which serves as the basis on which the least confident instances are selected. It seems to be natural, to update progressively also the sampling distribution for risk estimations. In that protocol the probability of an instance to be selected depends on previously drawn instances. This is critical from a statistical point of view. Although the estimator $\hat{G}_{n,q}$ would still be consistent due to the resampling weights, the sampling distribution q^* is no longer optimal, since it is derived under the condition that instances are drawn independently from an identical distribution. Without this assumption, an appropriate sampling distribution is analytically intractable.

In the next chapter, we address situations, in which two candidate predictive models are given and we would like to identify the model with lower risk as label-efficiently as possible.

Chapter 5

Active Model Comparison

In machine learning, a properly conducted evaluation of predictive models plays a central role; time after time new algorithms are developed and are compared to hitherto state of the art methods. Typically, statements of the predictive performance are governed by comparing estimates of the corresponding risks obtained from held-out data. In the statistics literature a range of appropriate tests are proposed that allow us to make the decision to prefer the apparently best model confidently (see Section 3.2). However, when labeled data are not or only rarely available, new instances have to be drawn and labeled at a cost.

For example, in computer vision it is common to acquire pre-trained object or face recognizers from third parties. Such recognizers do not typically come with the image databases that have been used to train them. The suppliers of the models could provide risk estimates based on held-out training data; however, such estimates might be biased because the training data would not necessarily reflect the distribution of images the deployed models will be exposed to. Another example are domains where the input distribution changes over a period of time in which a baseline model, *e.g.,* a spam filter, has been employed. By the time a new predictive model is considered, a previous risk estimate of the baseline model may no longer be accurate. In this chapter, we transfer the idea of active risk estimation studied in the previous chapter to scenarios in which an informed choice between given predictive models has to be made. We study an *active model comparison process* that selects instances from a pool of unlabeled test instances according to an instrumental distribution and queries their labels. We derive an instrumental distribution that allows us to make the decision to prefer the superior model as confidently as possible given a fixed labeling budget, if one of the models is in fact superior. Equivalently, one may use the instrumental distribution to minimize the labeling costs required to reach a correct decision at a prescribed level of confidence.

Firstly, we study the case of comparing the risks of two predictive models—for instance, a baseline model and a challenger—as confidently as possible. The problem setting is laid out in Section 5.1. In Section 5.2, we analyze the statistic of the Wald test under the null and alternative hypothesis and develop the instrumental distribution which maximizes test power. Comparing multiple models is even harder. We discuss the case in which multiple models have to be compared in Section 5.3 and propose a heuristic sampling distribution. Section 5.4 explores active model comparison experimentally and Section 5.5 reviews related work and concludes. Results of this chapter has previously been published (Sawade et al., 2012b).

5.1 Problem Setting

Let $p(y|x;\boldsymbol{\theta}_1)$ and $p(y|x;\boldsymbol{\theta}_2)$ be given $\boldsymbol{\theta}$-parameterized models of $p(y|x)$, let $f_{\boldsymbol{\theta}_j} : \mathcal{X} \to \mathcal{Y}$ with

$$f_{\boldsymbol{\theta}_j}(x) = \arg\max_y p(y|x;\boldsymbol{\theta}_j)$$

be the corresponding predictive functions, and let $R[f_j]$ be their risks as defined by Equation 3.1. Our goal is to determine whether $R[f_{\boldsymbol{\theta}_1}] > R[f_{\boldsymbol{\theta}_2}]$ or $R[f_{\boldsymbol{\theta}_2}] > R[f_{\boldsymbol{\theta}_1}]$ using a sample of labeled test instances (x_i, y_i).

The standard approach to comparing two models would be to draw n test instances according to the test distribution which the models are exposed to in practice, label these data, and calculate the empirical risks $\hat{R}_n[f_{\boldsymbol{\theta}_1}]$ and $\hat{R}_n[f_{\boldsymbol{\theta}_2}]$. Then, the empirical difference

$$\begin{aligned}\hat{\Delta}_n &= \hat{R}_n[f_{\boldsymbol{\theta}_1}] - \hat{R}_n[f_{\boldsymbol{\theta}_2}] \\ &= \frac{1}{n}\sum_{i=1}^{n} \left(\ell(f_{\boldsymbol{\theta}_1}(x_i), y_i) - \ell(f_{\boldsymbol{\theta}_2}(x_i), y_i)\right)\end{aligned}$$

provides evidence which model is preferable; a positive sign of $\hat{\Delta}_n$ argues in favor of $f_{\boldsymbol{\theta}_2}$. Given the empirical variance S_n^2 of $\hat{\Delta}_n$ (see Equation 3.22), a (paired) Wald test can be performed (see Section 3.2); the corresponding p-value quantifies the likelihood that the empirical difference is due to chance, indicating how confidently the decision to prefer the apparently better model can be made.

In analogy, when instances are drawn from an instrumental distribution $q(x)$ a

Wald test can be applied to quantify the significance of the observed difference

$$\hat{\Delta}_{n,q} = \hat{R}_{n,q}[f_{\boldsymbol{\theta}_1}] - \hat{R}_{n,q}[f_{\boldsymbol{\theta}_2}]$$

by discarding the null hypothesis that $\hat{\Delta}_{n,q}$ is observed by chance. The smaller the p-value the more confident the risks of $f_{\boldsymbol{\theta}_1}$ and $f_{\boldsymbol{\theta}_2}$ can be told apart by their empirical estimates. If the null hypothesis does not hold and the two models incur different risks, the distribution of the test statistic, and thus the p-value, depends on the chosen sampling distribution $q(x)$. Given a pre-specified confidence threshold, *e.g.,* $\alpha = 0.05$, the power $1 - \beta_{\alpha,q}$ is defined as the probability that the test will reject the null hypothesis. Our goal is to find the sampling distribution $q(x)$ that maximizes test power or, equivalently, minimizes the type II error

$$\begin{aligned} q^* &= \arg\max_q \; 1 - \beta_{\alpha,q} \\ &= \arg\min_q \; \beta_{\alpha,q} \end{aligned} \tag{5.1}$$

for a fixed labeling budget n.

We will derive an optimal sampling distribution q^* under the assumption that the models incur different risks. If the null hypothesis does hold, the test statistic is asymptotically normally distributed (see Equation 3.15), independently of the choice of q. Hence, in the case of identical risks the choice of $q(x)$ does not influence test results for large n.

Sackrowitz & Samuel-Cahn (1999) showed that the expected p-value equals the expected power $\mathbb{E}_{\alpha \sim \mathcal{U}(0,1)}[\beta_{\alpha,q}]$. Hence, minimizing the type II error of a test that compares $\hat{R}_{n,q}[f_{\boldsymbol{\theta}_1}]$ and $\hat{R}_{n,q}[f_{\boldsymbol{\theta}_2}]$ for any given confidence level α is equivalent to minimize the p-value in expectation over all samples of size n governed by $q(x)$.

5.2 Maximizing the Power of a Statistical Test

We now turn towards the problem of deriving an optimal sampling distribution q^* according to Equation 5.1. Our analysis considers the asymptotic case of $n \to \infty$, leading to a sampling distribution that is optimal if n is sufficiently large. Section 5.2.1 derives the asymptotically optimal sampling distribution. Section 5.2.2 discusses the empirical sampling distribution in a pool-based setting and presents the active model comparison algorithm.

5.2.1 Asymptotically Optimal Sampling Distribution

Let $\Delta = R[f_{\boldsymbol{\theta}_1}] - R[f_{\boldsymbol{\theta}_2}]$ denote the true risk difference, and assume $\Delta \neq 0$. Given a confidence threshold α, the test power of a Wald test (see Equation 3.25) equals the probability that the absolute value of the test statistic exceeds the critical value $z_\alpha = \Phi^{-1}(1 - \frac{\alpha}{2})$. Asymptotically, it holds that

$$\sqrt{n}\frac{\hat{\Delta}_{n,q} - \Delta}{\sigma_{n,q}} \sim \mathcal{N}(0,1).$$

Since $S_{n,q}$ consistently estimates $\sigma_{n,q}$, it follows that the test statistic $\sqrt{n}\frac{\hat{\Delta}_{n,q}}{S_{n,q}}$ is normally distributed with mean $\frac{\sqrt{n}\Delta}{\sigma_{n,q}}$ and unit variance,

$$\sqrt{n}\frac{\hat{\Delta}_{n,q}}{S_{n,q}} \sim \mathcal{N}\left(\frac{\sqrt{n}\Delta}{\sigma_{n,q}}, 1\right). \tag{5.2}$$

Equation 5.2 implies that the absolute value $\frac{\sqrt{n}|\hat{\Delta}_{n,q}|}{S_{n,q}}$ of the test statistic follows a *folded normal distribution* $\mathcal{N}_f(x|\mu,\sigma^2)$ with location parameter $\mu = \frac{\sqrt{n}\Delta}{\sigma_{n,q}}$ and scale parameter $\sigma^2 = 1$. Thus, the test power can be approximated in terms of the cumulative distribution of this folded normal distribution,

$$p\left(2 - 2\Phi\left(\sqrt{n}\frac{|\hat{\Delta}_{n,q}|}{S_{n,q}}\right) \leq \alpha\right) \approx 1 - \int_0^{z_\alpha} \mathcal{N}_f\left(T \middle| \frac{\sqrt{n}\Delta}{\sigma_{n,q}}, 1\right) \mathrm{d}T, \tag{5.3}$$

where

$$\mathcal{N}_f\left(T|\mu,1\right) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(T+\mu\right)^2\right) + \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(T-\mu\right)^2\right)$$

denotes the density of a folded normal distribution with location parameter μ and scale parameter one. We define the shorthand

$$\hat{\beta}_{n,q} = \int_0^{z_\alpha} \mathcal{N}_f\left(T \middle| \frac{\sqrt{n}\Delta}{\sigma_{n,q}}, 1\right) \mathrm{d}T \tag{5.4}$$

for the approximation of the type II error and the test power $1 - \hat{\beta}_{n,q}$ given by Equation 5.3. In the following, we derive a sampling distribution minimizing $\hat{\beta}_{n,q}$, and thereby approximately solving the optimization problem given by Equation 5.1.

Theorem 5.1 (Optimal Sampling Distribution)**.** *Let* $\Delta = R[f_1] - R[f_2]$ *with* $\Delta \neq 0$ *and* $\delta(x, y) = \ell(f_{\boldsymbol{\theta}_1}(x), y) - \ell(f_{\boldsymbol{\theta}_2}(x), y)$*. For all* $\alpha \in (0, 1)$*, the sampling distribution*

$$q^*(x) \propto p(x)\sqrt{\int (\delta(x, y) - \Delta)^2 \, p(y|x)\mathrm{d}y} \tag{5.5}$$

asymptotically minimizes $\hat{\beta}_{n,q}$*; that is, for any distribution* $q \neq q^*$ *it holds that* $\hat{\beta}_{n,q} > \hat{\beta}_{n,q^*}$ *for sufficiently large* n*.*

Before we prove Theorem 4.1, we show that a sampling distribution asymptotically minimizes $\hat{\beta}_{n,q}$ if and only if it minimizes the asymptotic variance of the estimator $\hat{\Delta}_{n,q}$.

Lemma 5.1 (Variance Optimality)**.** *Let* q*,* q' *denote two sampling distributions. Then it holds that* $\hat{\beta}_{n,q} < \hat{\beta}_{n,q'}$ *for sufficiently large* n *if and only if*

$$\lim_{n\to\infty} n \operatorname{Var}_{(x,y)\sim q(x)p(y|x)} \left[\hat{\Delta}_{n,q}\right] < \lim_{n\to\infty} n \operatorname{Var}_{(x,y)\sim q'(x)p(y|x)} \left[\hat{\Delta}_{n,q'}\right]. \tag{5.6}$$

Proof. Let $\alpha \in (0, 1)$ denote a confidence threshold, and let $\Delta = R[f_{\boldsymbol{\theta}_1}] - R[f_{\boldsymbol{\theta}_2}] \neq 0$ denote the true risk difference. The quantity $\hat{\beta}_{n,q}$ (see Equation 5.4) only depends on q through $\sigma_{n,q}$. For sufficiently large n, $\hat{\beta}_{n,q}$ is a monotonically increasing function of $\sigma_{n,q}$, because the partial derivative

$$\frac{\partial}{\partial \sigma_{n,q}} \hat{\beta}_{n,q} = \int_0^{z_\alpha} \frac{1}{2\pi} \left(\frac{n\Delta^2}{\sigma_{n,q}^3} - \frac{\sqrt{n}\Delta}{\sigma_{n,q}^2} T \right) \cdot \left(\exp\left(-\frac{1}{2} \left(T + \frac{\sqrt{n}\Delta}{\sigma_{n,q}} \right)^2 \right) + \exp\left(-\frac{1}{2} \left(T - \frac{\sqrt{n}\Delta}{\sigma_{n,q}} \right)^2 \right) \right) \mathrm{d}T$$

is positive for large n. Let q and q' denote two arbitrary sampling distributions. Since $\sigma_{n,q}^2 = n \operatorname{Var}_{(x,y)\sim q(x)p(y|x)}[\hat{\Delta}_{n,q}]$,

$$\lim_{n\to\infty} n \operatorname{Var}_{(x,y)\sim q(x)p(y|x)} \left[\hat{\Delta}_{n,q}\right] < \lim_{n\to\infty} n \operatorname{Var}_{(x,y)\sim q'(x)p(y|x)} \left[\hat{\Delta}_{n,q'}\right] \tag{5.7}$$

holds if and only if $\sigma_{n,q} < \sigma_{n,q'}$ for sufficiently large n. Condition 5.7 is thus equivalent to $\hat{\beta}_{n,q} < \hat{\beta}_{n,q'}$ for sufficiently large n. □

Lemma 5.1 shows that in order to solve the optimization problem given by Equation 5.1, we need to find the sampling distribution minimizing the asymptotic

variance of the estimator $\hat{\Delta}_{n,q}$. This asymptotic variance is characterized by the following Lemma.

Lemma 5.2 (Asymptotic Variance)**.** *Let* $\hat{\Delta}_{n,q} = \hat{R}_{n,q}[f_{\boldsymbol{\theta}_1}] - \hat{R}_{n,q}[f_{\boldsymbol{\theta}_2}]$*, and* $\sigma_q^2 = \lim_{n\to\infty} n \operatorname{Var}[\hat{\Delta}_{n,q}]$*. Then*

$$\sigma_q^2 = \iint \frac{p(x)^2}{q(x)^2} \left(\delta(x,y) - \Delta\right)^2 p(y|x) q(x) \mathrm{d}y \, \mathrm{d}x,$$

where $\delta(x,y) = \ell(f_{\boldsymbol{\theta}_1}(x), y) - \ell(f_{\boldsymbol{\theta}_2}(x), y)$.

The proof follows from Lemma 3.1 with loss function $\delta(x,y)$. We now prove Theorem 5.1 by deriving the distribution q^* minimizing the asymptotic variance as given by Lemma 5.2.

Proof of Theorem 5.1. According to Lemma 5.1 and Lemma 5.2, the distribution q^* asymptotically minimizing $\hat{\beta}_{n,q}$ can be derived by minimizing the functional σ_q^2 given by Lemma 5.2 in q under the constraint $\int q(x)\mathrm{d}x = 1$. Following the proof of Theorem 4.1 with

$$c(x) = p(x)^2 \int \left(\delta(x,y) - \Delta\right)^2 p(y|x) \mathrm{d}y$$

the optimal sampling distribution which asymptotically minimizes the type II error or, equivalently, maximizes the power of a two-sided Wald test that compares the risk of two predictive models $f_{\boldsymbol{\theta}_1}, f_{\boldsymbol{\theta}_2}$ is given by

$$q^*(x) = \frac{p(x)\sqrt{\int \left(\delta(x,y) - \Delta\right)^2 p(y|x)\mathrm{d}y}}{\int p(x)\sqrt{\int \left(\delta(x,y) - \Delta\right)^2 p(y|x)\mathrm{d}y}\mathrm{d}x}$$

This proves the claim. □

Intuitively, the sampling distribution $q^*(x)$ highlights disagreements between the models; it prefers instances for which the difference in performance $\delta(x,y)$ is expected to be high if Δ will be small. We will now derive the optimal sampling for two standard loss functions.

Corollary 5.1 (Optimal Sampling for Zero-One Loss)**.** *Let $\ell = \ell_{0/1}$ be the zero-one loss for a binary prediction problem with label space $\mathcal{Y} = \{0, 1\}$. The optimal sampling distribution that minimizes $\hat{\beta}_{n,q}$ resolves to*

$$q^*(x) \propto p(x) \begin{cases} \sqrt{1 - 2\Delta(1 - 2p(y = 1|x)) + \Delta^2}, & \text{if } f_{\boldsymbol{\theta}_1}(x) > f_{\boldsymbol{\theta}_2}(x) \\ \sqrt{1 + 2\Delta(1 - 2p(y = 1|x)) + \Delta^2}, & \text{if } f_{\boldsymbol{\theta}_1}(x) < f_{\boldsymbol{\theta}_2}(x) \\ |\Delta|, & \text{if } f_{\boldsymbol{\theta}_1}(x) = f_{\boldsymbol{\theta}_2}(x). \end{cases}$$

Proof. Rewriting the result of Theorem 5.1 in a classification setting, we obtain

$$\begin{aligned} q^*(x) &\propto p(x) \sqrt{\sum_{y \in \mathcal{Y}} \left(\ell(f_{\boldsymbol{\theta}_1}(x), y) - \ell(f_{\boldsymbol{\theta}_2}(x), y) - \Delta\right)^2 p(y|x) \mathrm{d}y} \\ &= p(x) \sqrt{\sum_{y \in \mathcal{Y}} \left((f_{\boldsymbol{\theta}_1}(x) - y)^2 - (f_{\boldsymbol{\theta}_1}(x) - y)^2 - \check{\Delta}\right)^2 p(y|x) \mathrm{d}y} \\ &= p(x) \Big(\left(f_{\boldsymbol{\theta}_1}(x) - f_{\boldsymbol{\theta}_2}(x)\right)^2 \\ &\qquad - 2\Delta \left(f_{\boldsymbol{\theta}_1}(x) - f_{\boldsymbol{\theta}_2}(x)\right) \left(1 - 2p(y = 1|x)\right) + \Delta^2 \Big)^{\frac{1}{2}}. \end{aligned} \tag{5.8}$$

Equation 5.8 expands the zero-one loss, exploiting $\ell(y, y') = (y - y')^2$ for $y, y' \in \{0, 1\}$. The claim follows by case differentiation according to the value of $f_{\boldsymbol{\theta}_1}(x)$ and $f_{\boldsymbol{\theta}_2}(x)$. □

In the following, we derive the optimal sampling distribution for regression problems and a squared loss function. In analogy to the task of evaluating a single model (see Chapter 4) we assume Gaussian distributed label noise:

Corollary 5.2 (Optimal Sampling for Squared Loss)**.** *Let $\ell = \ell_2$ be the squared loss and let the observed label y be normally distributed $p(y|x) = \mathcal{N}(y|\mu_x, \sigma_x^2)$ with label variance σ_x^2 at instance x. The sampling distribution that minimizes $\hat{\beta}_{n,q}$ resolves to*

$$q^*(x) \propto p(x) \sqrt{4\bar{f}^1(x)^2 \left(\mu_x^2 + \sigma_x^2\right) - 4\bar{f}^1(x)(\bar{f}^2(x) - \Delta)\mu_x + (\bar{f}^2(x) - \Delta)^2}, \tag{5.9}$$

where $\bar{f}^k(x) = f_{\boldsymbol{\theta}_1}(x)^k - f_{\boldsymbol{\theta}_2}(x)^k$.

Proof. Rewriting the result of Theorem 5.1 in a regression setting with squared

loss, we obtain

$$
\begin{aligned}
q^*(x) &\propto p(x)\sqrt{\int \left(\left(f_{\boldsymbol{\theta}_1}(x) - y\right)^2 - \left(f_{\boldsymbol{\theta}_2}(x) - y\right)^2 - \check{\Delta} \right)^2 p(y|x)\mathrm{d}y} \\
&= p(x) \left(4\bar{f}^1(x)^2 \int y^2 p(y|x)\mathrm{d}y \right. \\
&\qquad \left. -4\bar{f}^1(x)\left(\bar{f}^2(x) - \Delta\right) \int y p(y|x)\mathrm{d}y + \left(\bar{f}^2(x) - \Delta\right)^2 \right)^{\frac{1}{2}} \qquad (5.10)
\end{aligned}
$$

Equation 5.10 expands the loss function, orders terms by decreasing order of y, and makes use of the abbreviation $\bar{f}^k(x) = f_{\boldsymbol{\theta}_1}(x)^k - f_{\boldsymbol{\theta}_2}(x)^k$. Then, the claim follows by observing that the two integrals over $\mathcal{Y}$ are sums of the raw moments

$$
\begin{aligned}
&\int y \; p(y|x)\mathrm{d}y = \mu_x, \\
&\int y^2 p(y|x)\mathrm{d}y = \mu_x^2 + \sigma_x^2
\end{aligned}
$$

of the Gaussian distribution $p(y|x) = \mathcal{N}(y|\mu_x, \sigma_x^2)$. □

5.2.2 Empirical Sampling Distribution

The optimal sampling distribution prescribed by Theorem 5.1 depends on the unknown quantities $p(x)$ and $p(y|x)$. In this section, we focus again on pool-based settings in which instances can be sampled from a large pool D_m of m unlabeled test instances drawn from $p(x)$; we approximate $p(x)$ by the empirical sampling distribution $\hat{p}(x)$ defined over D_m (see Equation 2.29). In the following, we discuss different approximations of the true conditional distribution $p(y|x)$ for $\ell = \ell_{0/1}$ and $\ell = \ell_2$.

In Chapter 4, we have studied the problem of evaluating the risk of a single model as accurately as possible. We observed that approximating the conditional distribution by the predictive distribution of the model being evaluated, can lead to more accurate risk estimates than risk estimation based on a uniformly drawn sample. This approximation can be transferred in a natural way to the task of comparing two predictive models. We use the competing predictive models $p(y|x;\boldsymbol{\theta}_1)$ and $p(y|x;\boldsymbol{\theta}_2)$, and assume a mixture distribution $p(y|x;\boldsymbol{\theta})$

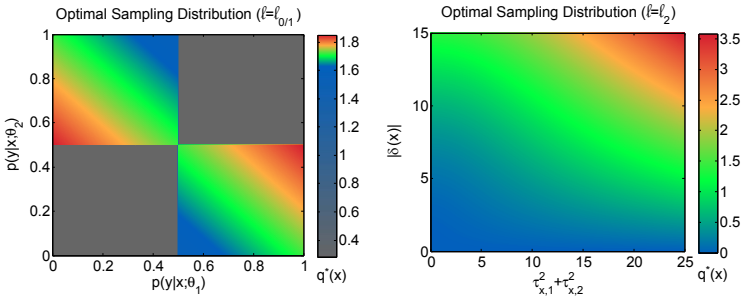


Figure 5.1: Heatmap of the sampling distribution with $\breve{\Delta} = 0.2$ when comparing the error rate of two models as a function of the predictive distributions of the competing models (left). Heatmap of the sampling distribution when comparing two regression models as a function of the variance of an instance τ_x^2 and the deviation of the models' predictions $|\delta(x)|$ (right).

giving equal weight to the models $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$:

$$\begin{aligned} p(y|x) &\approx p(y|x;\boldsymbol{\theta}) \\ &= \frac{1}{2}p(y|x;\boldsymbol{\theta}_1) + \frac{1}{2}p(y|x;\boldsymbol{\theta}_2). \end{aligned} \tag{5.11}$$

Note that any convex combination of the predictive distributions can be chosen. In particular, one can think of procedures to optimize the mixture weights during the evaluation process, as for example, by a maximum a posteriori estimate of the weights using a beta prior. However, see the discussion in Chapter 7 about difficulties arising from refining the instrumental distribution.

Finally, Theorem 5.1 depends on the true difference of risks $\Delta = R[f_{\boldsymbol{\theta}_1}] - R[f_{\boldsymbol{\theta}_2}]$. Straightforwardly, this risk difference is replaced by a difference $\breve{\Delta}$ of introspective risks calculated from Equation 3.1, where the integral over $\mathcal{X}$ is replaced by a sum over the pool, $p(x) \approx \hat{p}(x)$, and $p(y|x)$ is approximated by Equation 5.11:

$$\breve{\Delta} = \frac{1}{m}\sum_{x \in D} \int \delta(x,y) p(y|x;\boldsymbol{\theta}) \mathrm{d}y. \tag{5.12}$$

Recall that these approximations do not introduce any asymptotic bias: as long as $p(x) > 0$ implies $q(x) > 0$, the estimates given by Equation 3.11 and Equation 3.21 are consistent (see Section 3.1).

When $p(y|x)$ is approximated by $p(y|x;\boldsymbol{\theta})$ and Δ by the intrinsic difference $\breve{\Delta}$,

Corollary 5.1 gives immediately rise to a computable distribution for comparing models with respect to the zero-one loss:

$$q^*(x) \propto p(x) \begin{cases} \sqrt{1 - 2\check{\Delta}(1 - 2p(y = 1|x; \boldsymbol{\theta})) + \check{\Delta}^2}, & \text{if } f_{\boldsymbol{\theta}_1}(x) > f_{\boldsymbol{\theta}_2}(x) \\ \sqrt{1 + 2\check{\Delta}(1 - 2p(y = 1|x; \boldsymbol{\theta})) + \check{\Delta}^2}, & \text{if } f_{\boldsymbol{\theta}_1}(x) < f_{\boldsymbol{\theta}_2}(x) \\ |\check{\Delta}|, & \text{if } f_{\boldsymbol{\theta}_1}(x) = f_{\boldsymbol{\theta}_2}(x) \end{cases}$$

Figure 5.1 (left) shows q^* as a function of the predictive distributions $p(y|x; \boldsymbol{\theta}_i)$ for $\check{\Delta} = 0.2$, meaning that $f_{\boldsymbol{\theta}_2}$ is intrinsically more accurate. It gives highest preference to instances on which the apparently better model $p(y|x; \boldsymbol{\theta}_2) \approx 0.5$ is uncertain and on which the inferior model is certain about the corresponding label. Instances on which the models agree ($f_{\boldsymbol{\theta}_1}(x) = f_{\boldsymbol{\theta}_2}(x)$) are rarely chosen; they are drawn proportionally to the value of the intrinsic difference. Alternatively, an uninformative approximation $p(y = 1|x) \approx 0.5$ can be used. In this case, the intrinsic risk difference $\check{\Delta}$ is zero:

$$\begin{aligned} \check{\Delta} &= \frac{1}{m} \sum_{x \in D} \sum_{y \in \mathcal{Y}} \left(\ell_{0/1}(f_{\boldsymbol{\theta}_1}(x), y) - \ell_{0/1}(f_{\boldsymbol{\theta}_2}(x), y) \right) \frac{1}{|\mathcal{Y}|} \\ &= \frac{1}{m} \frac{1}{|\mathcal{Y}|} \sum_{x \in D} \left(\sum_{y \in \mathcal{Y}} [\![f_{\boldsymbol{\theta}_1}(x) \neq y]\!] - \sum_{y \in \mathcal{Y}} [\![f_{\boldsymbol{\theta}_2}(x) \neq y]\!] \right) \\ &= 0 \end{aligned}$$

and thus the sampling distribution given by Corollary 5.1 degenerates to uniform sampling from the subset of the pool where $f_{\boldsymbol{\theta}_1}(x) \neq f_{\boldsymbol{\theta}_2}(x)$. We denote this baseline as *active*$_{\neq}$. Note that *active*$_{\neq}$ yields an estimator $\hat{\Delta}_{n,q}$ that is consistent only on the subset of instances for which the models disagree. However, a consistent estimator for the overall difference in risks is obtained by multiplying $\hat{\Delta}_{n,q}$ with the fraction of pool instances for which $f_{\boldsymbol{\theta}_1}(x) \neq f_{\boldsymbol{\theta}_2}(x)$.

In order to apply the mixture model assumption (see Equation 5.11) to regression problems with a squared loss function (see Corollary 5.2), we assume that the predictive distributions $p(y|x; \boldsymbol{\theta}_1) = \mathcal{N}(y|f_{\boldsymbol{\theta}_1}(x), \tau^2_{x,1})$ and $p(y|x; \boldsymbol{\theta}_2) = \mathcal{N}(y|f_{\boldsymbol{\theta}_2}(x), \tau^2_{x,2})$ are Gaussian. Since the k-th raw moments of the mixture distribution (see Equation 5.11) are given by

$$\mathbb{E}_{y \sim p(y|x;\boldsymbol{\theta})} \left[y^k | x \right] = \frac{1}{2} \left(\mathbb{E}_{y \sim p(y|x;\boldsymbol{\theta}_1)} \left[y^k | x \right] + \mathbb{E}_{y \sim p(y|x;\boldsymbol{\theta}_2)} \left[y^k | x \right] \right),$$

the empirical sampling distribution can be obtained by substituting

$$\begin{aligned}\mathbb{E}_{y\sim p(y|x)}\left[y|x\right] = \mu_x &\approx \frac{1}{2}\left(f_{\boldsymbol{\theta}_1}(x) + f_{\boldsymbol{\theta}_2}(x)\right) \\ \mathbb{E}_{y\sim p(y|x)}\left[y^2|x\right] = \mu_x^2 + \sigma_x^2 &\approx \frac{1}{2}\left(f_{\boldsymbol{\theta}_1}(x)^2 + f_{\boldsymbol{\theta}_2}(x)^2 + \tau_{x,1}^2 + \tau_{x,2}^2\right)\end{aligned} \tag{5.13}$$

into Equation 5.9. Then, the introspective risk difference $\check{\Delta}$ defined by Equation 5.12 is zero: In Equation 5.14, we expand the squares and observe that the integral term reduces to the expectation of y. Finally, we insert the approximation given by Equation 5.13 (see Equation 5.15) and expand the product.

$$\begin{aligned}\check{\Delta} &= \frac{1}{m}\sum_{x\in D}\int (f_{\boldsymbol{\theta}_1}(x) - y)^2 - (f_{\boldsymbol{\theta}_2}(x) - y)^2 p(y|x;\boldsymbol{\theta})\mathrm{d}y \\ &= \frac{1}{m}\sum_{x\in D}\left(f_{\boldsymbol{\theta}_1}(x)^2 - f_{\boldsymbol{\theta}_2}(x)^2 - 2(f_{\boldsymbol{\theta}_1}(x) - f_{\boldsymbol{\theta}_2}(x))\mathbb{E}_{y\sim p(y|x;\boldsymbol{\theta})}\left[y\right]\right) \qquad (5.14) \\ &= \frac{1}{m}\sum_{x\in D}\left(f_{\boldsymbol{\theta}_1}(x)^2 - f_{\boldsymbol{\theta}_2}(x)^2 - (f_{\boldsymbol{\theta}_1}(x) - f_{\boldsymbol{\theta}_2}(x))(f_{\boldsymbol{\theta}_1}(x) + f_{\boldsymbol{\theta}_2}(x))\right) \qquad (5.15) \\ &= 0.\end{aligned}$$

Thus, the sampling distribution that asymptotically minimizes $\hat{\beta}_{n,q}$ in a pool based-setting resolves to

$$q^*(x) \propto \sqrt{\bar{f}^1(x)^2\left(\bar{f}^1(x)^2 + 2\left(\tau_{x,1}^2 + \tau_{x,2}^2\right)\right)}$$

for all $x \in D$, where again $\bar{f}^k(x) = f_{\boldsymbol{\theta}_1}(x)^k - f_{\boldsymbol{\theta}_2}(x)^k$. Figure 5.1 (right) shows $q^*(x)$. It prefers instances with high variance and instances on which the predictions of the models differ strongly.

Typically, the variances $\tau_{x,j}^2$ of the predictive distribution at instance x would be available from a probabilistic predictor such as a Gaussian process (see Section 2.2). If only predictive values $f_{\boldsymbol{\theta}_j}(x)$ but no predictive distribution is available, we cannot estimate the instance-specific label variance $\tau_x^2 = \tau_{x,1}^2 + \tau_{x,2}^2$. In this case, two simplified instances of Equation 5.9 can be considered. Firstly, we can assume peaked predictive distributions with $\tau_x^2 \to 0$. In this case, Equation 5.9 reduces to

$$q(x) \propto \left(f_{\boldsymbol{\theta}_1}(x) - f_{\boldsymbol{\theta}_2}(x)\right)^2.$$

Secondly, we can assume infinitely broad predictive distributions with $\tau_x^2 \to \infty$,

Algorithm 3: Active Model Comparison

input Models $f_{\boldsymbol{\theta}_1}, f_{\boldsymbol{\theta}_2}$ with distributions $p(y|x;\boldsymbol{\theta}_1)$, $p(y|x;\boldsymbol{\theta}_2)$; pool D_m, labeling budget n.

1: Compute sampling distribution q^* (Corollary 5.1 or 5.2) using the mixture distribution $p(y|x;\boldsymbol{\theta})$ given by Equation 5.11.
2: **for** $i = 1, \ldots, n$ **do**
3: Draw $x_i \sim q^*(x)$ from D_m with replacement.
4: Query label $y_i \sim p(y|x_i)$ from oracle.
5: **end for**
6: Compute $\hat{R}_{n,q}[f_{\boldsymbol{\theta}_1}]$ and $\hat{R}_{n,q}[f_{\boldsymbol{\theta}_2}]$ (see Equation 3.11).
7: Determine $f^* \leftarrow \arg\min_{f \in \{f_{\boldsymbol{\theta}_1}, f_{\boldsymbol{\theta}_2}\}} \hat{R}_{n,q}[f]$
8: Compute p-value for sample (see Equation 3.24)

output f^*, p-value.

leading to

$$\begin{aligned} q(x) &= \lim_{\tau \to \infty} \frac{\sqrt{\bar{f}^1(x)^2 \left(\bar{f}^1(x)^2 + 2\tau^2\right)}}{\sum_{x \in D} \sqrt{\bar{f}^1(x)^2 \left(\bar{f}^1(x)^2 + 2\tau^2\right)}} \\ &= \frac{\sqrt{\lim_{\tau \to \infty} \frac{\bar{f}^1(x)^4}{2\tau^2} + \bar{f}^1(x)^2}}{\sum_{x \in D} \sqrt{\lim_{\tau \to \infty} \frac{\bar{f}^1(x)^4}{2\tau^2} + \bar{f}^1(x)^2}} \\ &\propto |f_{\boldsymbol{\theta}_1}(x) - f_{\boldsymbol{\theta}_2}(x)|. \end{aligned}$$

We refer to these baselines as $active_0$ and $active_\infty$.

Algorithm 3 summarizes the active model comparison algorithm. It samples n instances with replacement from the pool according to the distribution prescribed by Corollary 5.1 (for zero-one loss) and 5.2 (for squared loss), respectively, using the predictive distribution $p(y|x;\boldsymbol{\theta})$. Labels are queried for these instances. In analogy to Algorithm 2, labels of previously drawn instances can be looked up rather than be queried repeatedly if the labeling process is deterministic.

5.3 Comparing Multiple Prediction Models

So far we have focused on the problem of comparing the risks of two predictive models. If multiple models are available, standard generalizations of the Wald

test such as ANOVA or the Tukey test (see Section 5.3) can be conducted. They try to reject the null hypothesis that the risks of all competing models are equal. This could occur if only one of the models performs clearly worst. Hence, rejection of the null hypothesis does not imply that all empirically observed differences are significant. Choosing a sampling distribution $q(x)$ that maximizes the power of such a test would thus not appropriately reflect the objectives of the empirical evaluation. However, researchers often resort to pairwise hypothesis testing when comparing multiple prediction models. Accordingly, we derive a heuristic sampling distribution for the comparison of multiple models $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k$ as a mixture of pairwise-optimal sampling distributions,

$$q^*(x) = \frac{1}{k(k-1)} \sum_{i \neq j} q^*_{i,j}(x), \tag{5.16}$$

where $q^*_{i,j}$ denotes the optimal distribution for comparing the models $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ given by Theorem 4.1. When comparing multiple models, we replace Equation 5.11 by a mixture over all models $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k$.

Learning a predictive model can be thought of as a search through the hypothesis space with the aim of finding the model with lowest risk. Intuitively, this can be done by comparing all hypothesis with respect to their risks estimated from a set of labeled instances. If no labeled instances are available, active learning strategies allow an label-efficient exploration; the underlying problem of active learning resembles active comparison with an infinite number of models. The heuristic sampling distribution given by Equation 5.16 is a mixture of asymptotically pairwise-optimal distributions giving equal weight to each model. In the case of an infinitive model space a posterior distribution $p(\boldsymbol{\theta}|T_n)$ can be used to weight the pairwise optimal sampling distribution by the probability of the model after observing a labeled data set T_n and, thus, give higher preference to instances that highlight differences between models with lower risk. A sampling distribution to comparing infinitely many models $\boldsymbol{\theta} \sim p(\boldsymbol{\theta}|T_n)$ can then be defined as

$$q^*(x) = \iint q^*_{\boldsymbol{\theta},\bar{\boldsymbol{\theta}}}(x) p(\boldsymbol{\theta}|T_n) \mathrm{d}\boldsymbol{\theta}\, p(\bar{\boldsymbol{\theta}}|T_n) \mathrm{d}\bar{\boldsymbol{\theta}}, \tag{5.17}$$

where $q^*_{\boldsymbol{\theta},\bar{\boldsymbol{\theta}}}$ denotes the optimal distribution for comparing the models $\boldsymbol{\theta}$ and $\bar{\boldsymbol{\theta}}$ given by Theorem 5.1.

To implement the sampling distribution we need to approximate the unknown quantities $p(x)$ and $p(y|x)$ (see Section 5.2.2). When the conditional distribution

$p(y|x) \approx \frac{1}{|\mathcal{Y}|}$ is approximated by an uniform distribution in the case of $\ell = \ell_{0/1}$, Equation 5.17 reduces to

$$q^*(x) \propto \iint [\![f_{\bar{\boldsymbol{\theta}}}(x) \neq f_{\boldsymbol{\theta}}(x)]\!] p(\boldsymbol{\theta}|T_n) \mathrm{d}\boldsymbol{\theta}\, p(\bar{\boldsymbol{\theta}}|T_n) \mathrm{d}\bar{\boldsymbol{\theta}} \tag{5.18}$$

in a pool-based setting. The integral in Equation 5.18 is generally intractable. One approach to approximate the sampling distribution is to sample a set of models from the posterior distribution $p(\boldsymbol{\theta}|T_n)$. The proportion of disagreements between the models on each instance $x \in D_m$ serves as selection criterion for the next instance to label. If the set of labeled instances T_n is progressively be used to update the posterior distribution, this procedure equals the query by committee algorithm discussed in Section 2.3.1.

For regression, consider the simplified instance of Theorem 4.1 which assumes an infinitely broad predictive distribution. In this case, the optimal sampling distribution given by Equation 5.17 reduces to

$$\begin{aligned} q^*(x) &\propto \iint \left(f_{\bar{\boldsymbol{\theta}}}(x) - f_{\boldsymbol{\theta}}(x)\right)^2 p(\boldsymbol{\theta}|T_n) \mathrm{d}\boldsymbol{\theta}\, p(\bar{\boldsymbol{\theta}}|T_n) \mathrm{d}\bar{\boldsymbol{\theta}} \\ &= 2 \int f_{\boldsymbol{\theta}}(x)^2 p(\boldsymbol{\theta}|T_n) \mathrm{d}\boldsymbol{\theta} - 2 \left(\int f_{\boldsymbol{\theta}}(x) p(\boldsymbol{\theta}|T_n) \mathrm{d}\boldsymbol{\theta} \right)^2 \\ &= 2 \operatorname{Var}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta}|T_n)} \left[f_{\boldsymbol{\theta}}(x) \right] \\ &\propto \tau_x^2 - \sigma_\varepsilon^2. \end{aligned} \tag{5.19}$$

Equation 5.19 prefers instances with high variance of the predictions and ignores the uncertainty which is caused by the label variance. This sampling distribution resembles the selection criterion of uncertainty sampling for regression (see Section 2.3.1).

5.4 Empirical Results

We study the empirical behavior of active comparison (Algorithm 3, labeled *active* in all diagrams) relative to a risk comparison based on a test sample drawn uniformly from the pool (labeled *passive*) and the baselines active_0, active_∞, and $\mathit{active}_{\neq}$ discussed in Section 5.2.2. We also include the active risk estimator presented in Chapter 4 in our study, which infers optimal sampling distributions q_1^* and q_2^* for individually estimating the risks of the models with

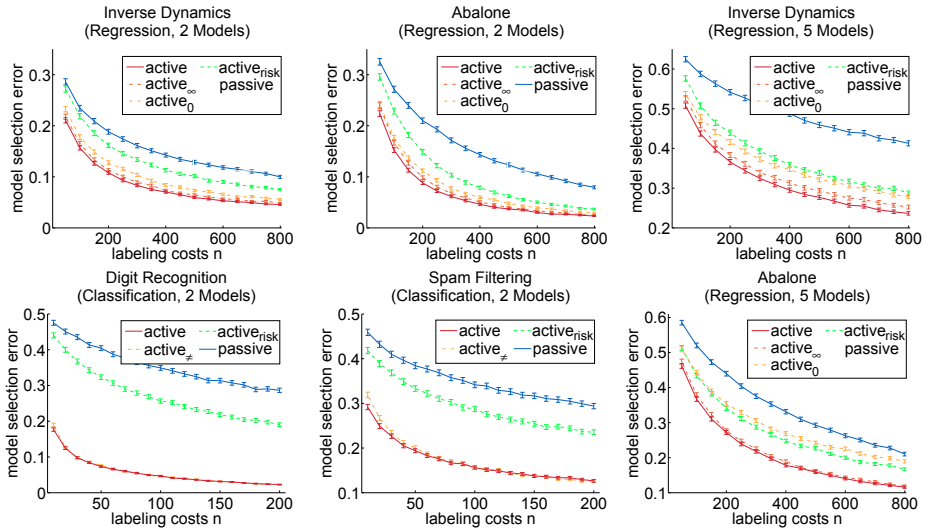


Figure 5.2: Model selection error over labeling costs for comparison of two prediction models (left and center) and comparison of multiple prediction models (right). Error bars indicate the standard error.

parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Test instances are sampled from a mixture distribution $q^*(x) = \frac{1}{2}q_1^*(x) + \frac{1}{2}q_2^*(x)$ (labeled *active*$_{risk}$). When studying classification, we also include the active learning algorithms A^2 (Balcan et al., 2006) and *IWAL* (Beygelzimer et al., 2009) as baselines by using them to sample test instances. Their model space is the set of predictive models that are to be compared. The confidence parameter of the active learning baselines A^2 and *IWAL* is set to $\delta = 0.05$, corresponding to a 95% confidence of the corresponding finite-sample error bound (for A^2 we use the *Chernoff bound*).

We again conduct experiments on the two classification domains and the two regression domains described in Section 4.4. Specifically, in the digit recognition problem logistic regression models with linear kernels against RBF kernels are compared; for *Sarcos* and *Abalone* we study whether a Gaussian process model with linear kernel or with Matérn kernel (see Equation 2.28) is preferable. In each domain, two differing models are trained on a randomly selected set of 500 instances; the remaining data serve as the pool of unlabeled test instances used to compare the models. Using non-linear kernels in spam filtering domain is uncommon, we instead compare models that differ in the recency of their training data. Specifically, we compare a logistic regression model trained on 5,000 randomly

sampled messages received between June 2007 and October 2007 to a logistic regression model trained on 5,000 randomly sampled messages received between December 2007 and April 2008. Emails received after April 2008 constitute the pool of test instances. In these four domains we average the evaluation results over ten pairs of models. Additionally, we study the following object recognition domain.

Object Recognition Domain. In this domain, the prediction task is to decide whether a given image contains a car (positive class) or not (negative class). Using Google Image Search, we built a corpus of 4,560 images; approximately 50% of the images belong to the positive class. For building the detection models, we follow a bag-of-visual-words approach. First, interest points are identified for all images, and SIFT features (Lowe, 2004) at the interest points are computed. Second, a visual vocabulary is built by clustering all SIFT features using k-means. Third, images are encoded as real vectors with one feature per cluster; a feature indicates how many interest points in the image fall into the corresponding cluster. Logistic regression models are trained on the resulting feature representation. We train 12 detection models that result from varying the interest point detection method (Harris & Stephens, 1988; Canny, 1986; Förstner & Gülch, 1987) and the size of the visual vocabulary $k \in \{50, 100, 500, 1000\}$. Additionally, we train a detection model based on SURF interest point detection (Bay et al., 2008) and a pyramid matching kernel, using the LIBPMK toolkit described by Lee (2008). The 13 models are trained on approximately 10% of the available images, the remaining images constitute the pool of unlabeled test examples on which the models are compared.

Our evaluation specifically addresses two main aspects. In Section 5.4.1 we evaluate how often the model with lower risk is correctly identified, as a function of the labeling budget n. In Section 5.4.2, we evaluate how often the corresponding paired test is unable to reject the null hypothesis although the models differ (Type II error), and how often a false-positive result is obtained under the null hypothesis (Type I error). Results are averaged over 5,000 repetitions of the evaluation process.

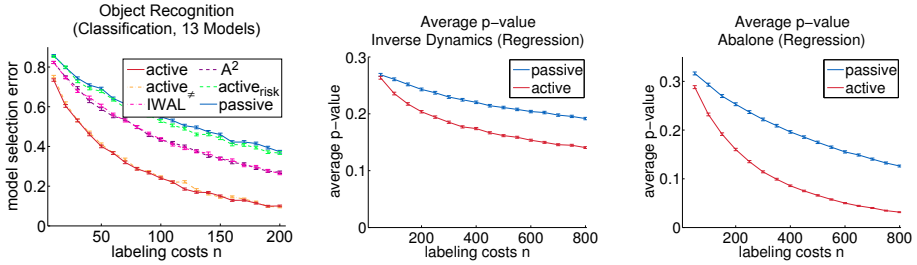


Figure 5.3: Model selection error over labeling costs in the object recognition domain (left). Average p-value over labeling costs n for *Sarcos* (center) and *Abalone* (right). Error bars indicate the standard error.

5.4.1 Identifying the Model With Lower Risk

We measure *model selection error*, defined as the fraction of experiments in which an evaluation method wrongly identifies the model with lower true risk. The true risk is taken to be the risk over all test instances in the pool. Figure 5.2 (left and center) shows that for the comparison of two models *active* results in significantly lower model selection error than *passive*, or, equivalently, saves between 70% and 90% of labeling effort. Differences between *active* and the simplified variants $active_0$, $active_\infty$, and $active_{\neq}$ are marginal. These variants do not require an estimate of $p(y|x)$ by the predictive distributions $p(y|x;\boldsymbol{\theta}_j)$, thus the method is applicable even if no predictive distributions are available or such an estimate would be very uncertain because of a shift between the training and test distributions. The active learning algorithms A^2 and *IWAL* applied to the model space $\{f_{\boldsymbol{\theta}_1}, f_{\boldsymbol{\theta}_1}\}$ coincide with $active_{\neq}$, as can be seen from inspection of Algorithm 1 in the paper by Balcan et al. (2006) and *IWAL* and Algorithms 1 and 2 in the paper by Beygelzimer et al. (2009); we omit these baselines for clarity reasons.

Figure 5.2 (right) shows results for the heuristic to compare multiple models given by Equation 5.16. Here, five differing models are trained using polynomial kernels of degree $d \in \{1, 2, 3, 4, 5\}$. We again observe that *active* outperforms *passive*, saving between 60% and 85% of labeling effort. In the object recognition domain, *active* saves approximately 70% of labeling effort compared to *passive*. A^2 and *IWAL* outperform *passive* but are less accurate than *active* (see Figure 5.3, left). In our experiments we observe that the finite-sample bounds of A^2 and *IWAL* are quite loose. Hence, in principal, both methods reduce to *passive*

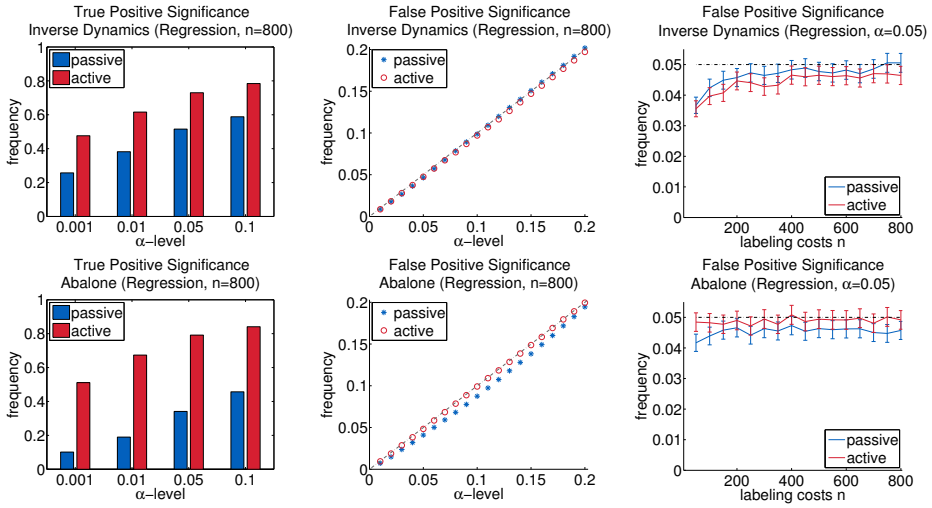


Figure 5.4: True-positive significance rate for different test levels α (left). False-positive significance rate over test level α (center) and labeling costs n (right). Dashed line indicates perfect calibration and error bars indicate the standard error.

applied to the subset $\{x \in D_m | \exists i, j : f_{\boldsymbol{\theta}_i}(x) \neq f_{\boldsymbol{\theta}_j}(x)\}$ of the pool D_m. In contrast, *active*$_{\neq}$ is based on a mixture distribution, from which instances are drawn proportional to the degree of disagreement among the models. The savings for *active* over *active*$_{\neq}$ are only slightly higher. This is intuitive, since the corresponding sampling distribution are similar (see Figure 5.1, left); they coincide for $\check{\Delta} \to 0$.

5.4.2 Significance Testing: Type I and II Errors

Each comparison method returns the model with lower empirical risk and the p-value of a paired two-sided test, which constitutes a measure of confidence for our decision to prefer the model with lower empirical risk. We now study how often a comparison method is able to reject the null hypothesis that two predictive models incur identical risks, and the calibration of the resulting p-values. For the evaluation method *active* p-values are computed according to the Wald test for weighted test samples discussed in Section 3.2 (see Equation 3.24). For the evaluation method *passive* that draws an unweighted sample of test instances we use the more standard t-test. For classification, the method *active*$_{\neq}$

Table 5.1: Incurred labeling costs, true-positive significance rate, and false decision rate for a protocol of drawing test instances until significance of $\alpha = 0.05$ is obtained or the labeling budget of $n = 800$ is exhausted.

		active	*active*$_\infty$	*active*$_0$	*active*$_{risk}$	*passive*
Abalone	labeling costs	362.35	390.24	395.36	484.79	544.74
	significance	82.56%	80.16%	76.57%	66.10%	52.39%
	false decisions	0.65%	0.73%	1.28%	1.78%	1.87%
Inverse Dynamics	labeling costs	354.73	357.28	380.56	428.59	444.56
	significance	79.51%	78.65%	75.10%	68.86%	66.10%
	false decisions	1.17%	1.35%	1.82%	1.85%	2.32%

is equivalent to *passive* applied to the subset $D_{\neq} = \{x \in D_m | f_{\boldsymbol{\theta}_1}(x) \neq f_{\boldsymbol{\theta}_2}(x)\}$ of the pool D_m (see Section 5.2.2). Labeling effort is thus simply reduced by a factor of $|D_{\neq}|/|D_m|$. For regression, the analysis is less straightforward as typically $D_{\neq} = D_m$. In this section, we therefore focus on regression problems.

The p-value of a paired two-sided test constitutes a measure of confidence for the decision to prefer the model with lower empirical risk. Figure 5.3 (center, right) shows the average p-value for *active* and *passive* as a function of the labeling budget n. Active comparison results in lower average p-values, in particular for large n. Figure 5.4 (left) shows how often the active and passive comparison methods are able to reject the null hypothesis that both models incur identical risks, for different α-levels of the test and a labeling budget of $n = 800$. The true risk incurred by the prediction models $f_{\boldsymbol{\theta}_1}$ and $f_{\boldsymbol{\theta}_2}$ is never equal in these experiments. We observe that *active* is able to reject the null hypothesis more often and with a higher confidence. In the Abalone domain, *active* rejects the null hypothesis at $\alpha = 0.001$ more often than *passive* is able to reject it at $\alpha = 0.1$.

In order to evaluate the Type I error of the tests based on active and passive sampling, we also conduct experiments under the null hypothesis. The experimental setup is changed such that whenever a new test instance x is chosen and the predictions $y = f_{\boldsymbol{\theta}_1}(x)$ and $y' = f_{\boldsymbol{\theta}_2}(x)$ are queried, we swap the predicted labels y and y' with probability 0.5. This protocol guarantees that the expected risks of $f_{\boldsymbol{\theta}_1}$ and $f_{\boldsymbol{\theta}_2}$ are identical. Figure 5.4 (center) shows the rate of false-positive test results as a function of the α-level of the test for a labeling budget of $n = 800$. We observe that Type I errors are well calibrated for both tests, as the false-positive rate stays below the (ideal) diagonal line. Figure 5.4 (right) shows the false-positive rate for $\alpha = 0.05$ as a function of n. Both tests are conservative for small n, and approach the expected false-positive rate of 0.05 as n grows larger.

We finally follow a protocol in which test instances are drawn and labeled until a paired two-sided test indicates a significant difference between the risks of $f_{\boldsymbol{\theta}_1}$ and $f_{\boldsymbol{\theta}_2}$ at $\alpha = 0.05$, or the labeling budget of $n = 800$ is exhausted. We do not enforce the null hypothesis by swapping prediction labels; the true risk incurred by the prediction models $f_{\boldsymbol{\theta}_1}$ and $f_{\boldsymbol{\theta}_2}$ is never equal. Note that due to the repeated statistical testing carried out in this protocol, the resulting p-values will not be correctly calibrated. Table 5.1 shows the average labeling costs incurred, fraction of experiments in which a significance result is obtained, and the fraction of experiments in which a significance result is obtained but the wrong model is chosen (*false decision rate*). In both domains, *active* incurs the lowest average labeling costs, obtains significance results most often, and has the lowest false decision rate.

5.5 Summary and Related Work

In this chapter, we studied a setting in which two predictive models have to be compared at minimal labeling costs using test instances that can be selected from a large pool of unlabeled test data. Typically, a statistical test is used to measure how confidently the decision to prefer the apparently better model can be made. We have derived an active comparison procedure for regression and classification tasks whose sampling distribution asymptotically maximizes the power or, equivalently, minimizes the Type II error of a two-sided Wald test. The sampling distribution intuitively gives preference to test instances on which the models disagree strongly. The proposed method is directly applicable with probabilistic models that provide a predictive distribution $p(y|x;\boldsymbol{\theta})$. We also proposed two simplified variants of the method that empirically perform almost as well and do not require a predictive distribution $p(y|x;\boldsymbol{\theta})$, and a heuristic generalization of the optimal sampling distribution for comparing multiple prediction models. In contrast to Chapter 4, in which we have studied active data acquisition strategies for the assessment of an individual existing model, in terms of generalized risks, the task was to assess the *relative* performance of two or more existing models, without necessarily determining absolute risks precisely.

The active comparison problem that we have studied can be seen as an extreme case of active learning, in which the model space contains only two (or, more generally, a small number of) models. For the special case of classification with zero-one loss and two models under study, a simplified version of the sampling

distribution we have derived coincides with the sampling distribution used in the A^2 and *IWAL* active learning algorithms proposed by Balcan et al. (2006) and Beygelzimer et al. (2009). For A^2 and *IWAL*, the derivation of this distribution is based on finite-sample complexity bounds, while in our approach, it is based on maximizing the power of a statistical test comparing the models under study. Furthermore, the latter approach has the advantage that it directly generalizes to regression problems. A further difference to active learning is that our goal is not only to choose the best model, but also to obtain a well-calibrated p-value indicating the confidence with which this decision can be made. We have discussed the relationship between existing active learning algorithms and our approach in the case of an infinite number of models in Section 5.3.

Madani et al. (2004) study *active model selection*, where the goal is also to identify a model with lowest risk. The main difference to our approach is that in their setting costs are associated with obtaining predictions $\hat{y} = f(x)$, while in our setting costs are associated with obtaining labels $y \sim p(y|x)$. *Hoeffding races* (Maron & Moore, 1993) and *sequential sampling algorithms* (Scheffer & Wrobel, 2003) perform efficient model selection by keeping track of risk bounds for candidate models and removing models that are clearly outperformed from consideration. The goal of these methods is to reduce computational complexity, not labeling effort.

Empirically, we observed that the proposed active comparison method consistently outperforms a traditional comparison based on a uniform sample of test instances. Active comparison identifies the model with lower true risk more often, and is able to detect significant differences between the risks of two given models more quickly. In the five experimental domains that we studied, performing active comparison resulted in a saved labeling effort of between 60% and over 90%. We also performed experiments under the null hypothesis that both models incur identical risks, and verified that active comparison does not lead to increased false-positive significance results.

In the next chapter, we study evaluation task in the ranking domain. In this domain, performance measures have to assess a list of returned items rather than a single label. Computing the optimal sampling strategy for evaluating and comparing ranking functions can thus be challenging.

Chapter 6

Active Evaluation of Ranking Functions

Evaluating the quality of ranking functions is a core task in web search and other information retrieval domains. Because query distributions and item relevance change over time, ranking models often cannot be evaluated accurately on held-out training data. Instead, considerable effort is spent on manually labeling the relevance of query results for test queries in order to track ranking performance.

The standard approach to evaluate a ranking function is to draw a random sample of test queries from $p(x)$, obtain relevance labels of all retrieved items for each query, and compute average empirical performance. However, the results in Chapter 4 and Chapter 5 indicate that estimation accuracy can be improved by drawing test examples from an appropriately engineered instrumental distribution $q(x)$ rather than the data distribution $p(x)$, and correcting for the discrepancy between $p(x)$ and $q(x)$ by importance weighting. In analogy to active learning, this is carried out by first sampling a large pool of unlabeled data from $p(x)$, and then actively sampling test instances from this pool. In this chapter, we apply the principle of active evaluation to the problem of estimating the performance of ranking functions. A crucial feature of ranking domains is that labeling costs vary according to the number of result items and item-specific attributes such as document length. This problem has been studied in Section 4.3.

Section 6.1 reviews ranking functions that are based on graded relevance and presents two commonly used performance measures, namely, *Discounted Cumulative Gain* (DCG) and *Expected Reciprocal Rank* (ERR). In Section 6.2, we derive cost-optimal sampling distributions for DCG and ERR. For these measures a naïve computation of the empirical sampling distribution is exponential in the number of the retrieved items. We derive polynomial-time solutions by dynamic programming. Section 6.3 presents empirical results and Section 6.4 concludes. Results of this chapter has previously been published (Sawade et al., 2012a).

6.1 Ranking Functions and Measures

Let $\mathcal{X}$ denote a space of queries, and $\mathcal{Z}$ denote a finite space of items. We study ranking functions

$$\mathbf{r} : x \mapsto \left(r_1(x), \ldots, r_{|\mathbf{r}(x)|}(x)\right)^{\mathsf{T}}$$

that, given a query $x \in \mathcal{X}$, return a list of $|\mathbf{r}(x)|$ items $r_i(x) \in \mathcal{Z}$ ordered by relevance. Ranking performance of $\mathbf{r}$ is defined in terms of graded relevance labels $y_z \in \mathcal{Y}$ that represent the relevance of an item $z \in \mathcal{Z}$ for the query x, where $\mathcal{Y} \subset \mathbb{R}$ is a finite space of relevance labels with minimum zero (irrelevant) and maximum y_{max} (perfectly relevant). We summarize the graded relevance of all $z \in \mathcal{Z}$ in a label vector $\mathbf{y} \in \mathcal{Y}^{\mathcal{Z}}$ with components y_z for $z \in \mathcal{Z}$.

In order to evaluate the quality of a ranking $\mathbf{r}(x)$ for a single query x, we consider two commonly used ranking performance measures: *discounted cumulative gain* (DCG), given by

$$\begin{aligned} \ell_{dcg}\left(\mathbf{r}(x), \mathbf{y}\right) &= \sum_{i=1}^{|\mathbf{r}(x)|} \kappa_{dcg}\left(y_{r_i(x)}, i\right) \\ \kappa_{dcg}\left(y, i\right) &= \frac{2^y - 1}{\log_2(i+1)}, \end{aligned} \tag{6.1}$$

and *expected reciprocal rank* (ERR), given by

$$\begin{aligned} \ell_{err}\left(\mathbf{r}(x), \mathbf{y}\right) &= \sum_{i=1}^{|\mathbf{r}(x)|} \frac{1}{i} \kappa_{err}\left(y_{r_i(x)}\right) \prod_{l=1}^{i-1} \left(1 - \kappa_{err}\left(y_{r_l(x)}\right)\right) \\ \kappa_{err}\left(y\right) &= \frac{2^y - 1}{2^{y_{max}}}, \end{aligned} \tag{6.2}$$

as introduced by Järvelin & Kekäläinen (2002) and Chapelle et al. (2009), respectively.

DCG scores a ranking by summing over the relevance of all documents discounted by their position in the ranking. ERR is based on a probabilistic user model: the user scans a list of documents in the order defined by $\mathbf{r}(x)$ and chooses the first document that appears sufficiently relevant; the likelihood of choosing a document z is a function of its graded relevance score y_z. If s denotes the position of the chosen document in $\mathbf{r}(x)$, then $\ell_{err}\left(\mathbf{r}(x), \mathbf{y}\right)$ is the expectation of the recipro-

cal rank $1/s$ under the probabilistic user model. Both DCG and ERR discount relevance with ranking position, ranking quality is thus most strongly influenced by documents that are ranked highly. Alternatively, evaluation measures can be truncated such that they only take into account the k most highly ranked documents (Järvelin & Kekäläinen, 2002). Truncated measures are popular in mobile applications where only a small number of documents can be presented to a user.

In contrast to the evaluation tasks studied in the previous chapters, the label space in ranking domains is typically structured. Let $p(x, \mathbf{y}) = p(x)p(\mathbf{y}|x)$ denote the joint distribution over queries $x \in \mathcal{X}$ and label vectors $\mathbf{y} \in \mathcal{Y}^{\mathcal{Z}}$ the model is exposed to. We assume that the individual relevance labels y_z for items z are drawn independently given a query x:

$$p(\mathbf{y}|x) = \prod_{z \in \mathcal{Z}} p(y_z|x, z). \tag{6.3}$$

This assumption is common in pointwise ranking approaches, *e.g.,* regression based ranking models (Cossock & Zhang, 2008; Mohan et al., 2011). The ranking performance of $\mathbf{r}$ with respect to $p(x, \mathbf{y})$ can then be expressed as a risk

$$R[\mathbf{r}] = \iint \ell\left(\mathbf{r}(x), \mathbf{y}\right) p(x, \mathbf{y}) \mathrm{d}x \, \mathrm{d}\mathbf{y}, \tag{6.4}$$

where $\ell \in \{\ell_{dcg}, \ell_{err}\}$ denotes the performance measure under study.

Recall that the joint distribution $p(x, \mathbf{y})$ is unknown; the ranking performance can be estimated either by an empirical average $\hat{R}_n$ (see Equation 3.8) over a set of test queries $x_1, ..., x_n$ and graded relevance labels $\mathbf{y}_1, ..., \mathbf{y}_n$ drawn i.i.d. from $p(x, \mathbf{y})$ or by a set of instances drawn according to an instrumental distribution $q(x)$. In the latter case, a consistent estimator $\hat{R}_{n,q}$ can be defined by Equation 3.11. For certain choices of the sampling distribution $q(x)$, $\hat{R}_{n,q}$ may be a more label-efficient estimator of the true performance R than $\hat{R}_n$.

We assume that drawing unlabeled data $x \sim p(x)$ from the true distribution of queries the model is exposed to is inexpensive, whereas obtaining relevance labels is costly. In the ranking domain, costs for acquiring each label are associated with the number of items $|\mathbf{r}(x)|$ returned by $\mathbf{r}$ and possibly item-specific features such as the length of a document whose relevance has to be determined; the labeling costs may vary over the queries $x \in \mathcal{X}$. The labeling costs for a query x are denoted by $\lambda(x)$. We assume that $\lambda(x)$ is bounded away from zero by $\lambda(x) \geq \epsilon > 0$. In this chapter, we are pursuing two objectives. Our first goal is to

minimize the deviation of $\hat{R}_{n,q}$ from R with respect to the ranking measures DCG and ERR under the constraint that expected overall labeling costs stay below a budget $\Lambda \in \mathbb{R}$:

$$(q^*, n^*) = \arg\min_{q,n} \; \mathbb{E}_{(x,\mathbf{y})\sim q(x)p(\mathbf{y}|x)} \left[\left(\hat{R}_{n,q} - R \right)^2 \right], \tag{6.5}$$
$$\text{s.t. } \mathbb{E}_{x\sim q(x)} \left[\sum_{i=1}^{n} \lambda(x_i) \right] \leq \Lambda.$$

A second task, which is of particular interest in practice, is to estimate the relative performance of two ranking models; for instance, in order to evaluate the result of an index update or the integration of novel sources of training data. To estimate *relative* performance of two ranking functions $\mathbf{r}_1$ and $\mathbf{r}_2$ as cost-efficiently as possible, Equation 6.5 can be replaced by

$$(q^*, n^*) = \arg\min_{q,n} \; \mathbb{E}_{(x,\mathbf{y})\sim q(x)p(\mathbf{y}|x)} \left[\left(\hat{\Delta}_{n,q} - \Delta \right)^2 \right], \tag{6.6}$$
$$\text{s.t. } \mathbb{E}_{x\sim p(x)} \left[\sum_{i=1}^{n} \lambda(x_i) \right] \leq \Lambda,$$

where $\hat{\Delta}_{n,q} = \hat{R}_{n,q}[\mathbf{r}_1] - \hat{R}_{n,q}[\mathbf{r}_2]$ and $\Delta = R[\mathbf{r}_1] - R[\mathbf{r}_2]$.

In the next section, we state sampling distributions q^* asymptotically solving Equations 6.5 and 6.6 for DCG and ERR and discusses the computation of empirical sampling distributions in a pool-based seeting.

6.2 Optimal Sampling Distributions

In Chapter 4 and Chapter 5, we derived the sampling distribution that asymptotically minimizes the estimation error. Since the expected deviation is dominated by the variance for large n (see Section 4.2.1), we have approximated

$$\mathbb{E}_{(x,\mathbf{y})\sim q(x)p(\mathbf{y}|x)} \left[\left(\hat{R}_{n,q} - R \right)^2 \right] \approx \frac{1}{n}\sigma_q^2 \text{ and}$$
$$\mathbb{E}_{(x,\mathbf{y})\sim q(x)p(\mathbf{y}|x)} \left[\left(\hat{\Delta}_{n,q} - \Delta \right)^2 \right] \approx \frac{1}{n}\tau_q^2,$$

where

$$\sigma_q^2 = \lim_{n\to\infty} n \operatorname{Var}_{(x,\mathbf{y})\sim q(x)p(\mathbf{y}|x)} \left[\hat{R}_{n,q}\right],$$
$$\tau_q^2 = \lim_{n\to\infty} n \operatorname{Var}_{(x,\mathbf{y})\sim q(x)p(\mathbf{y}|x)} \left[\hat{\Delta}_{n,q}\right].$$

Note that, Corollary 4.1 and Theorem 5.1 hold also in the case of a structured label space. The sampling distributions that minimize the quantities $\frac{1}{n}\sigma_q^2$ and $\frac{1}{n}\tau_q^2$, thereby approximately solving Problems 6.5 and 6.6, are thus given by

$$q^*(x) \propto \frac{p(x)}{\sqrt{\lambda(x)}} \sqrt{\int \left(\ell(\mathbf{r}(x), \mathbf{y}) - R\right)^2 p(\mathbf{y}|x) \mathrm{d}\mathbf{y}} \text{ and} \tag{6.7}$$

$$q^*(x) \propto \frac{p(x)}{\sqrt{\lambda(x)}} \sqrt{\int \left(\delta(x, y) - \Delta\right)^2 p(\mathbf{y}|x) \mathrm{d}\mathbf{y}}, \tag{6.8}$$

where

$$\delta(x, y) = \ell(\mathbf{r}_1(x), \mathbf{y}) - \ell(\mathbf{r}_2(x), \mathbf{y}) \tag{6.9}$$

denotes the performance difference of the two ranking models for a labeled test query $(x, \mathbf{y})$. The optimal number of drawn instances is

$$n^* = \frac{\Lambda}{\int \lambda(x) q(x) \mathrm{d}x}$$

in each case.

We now turn towards the problem of evaluating the sampling distributions in practice. The sampling distributions prescribed by Equation 6.7 and Equation 6.8 depend on the unknown test distribution $p(x)$, the true conditional distribution $p(\mathbf{y}|x) = \prod_{z\in\mathcal{Z}} p(y_z|x, z)$, and the true performance R and $\Delta = R[\mathbf{r}_1] - R[\mathbf{r}_2]$, respectively. In analogy to Section 4.2.2 and 5.2.2, we study a setting in which queries are sampled from a pool D_m of m unlabeled queries and approximate the distribution $p(x) \approx \hat{p}(x)$ by the empirical distribution (see Equation 2.29). For the large class of pointwise ranking methods—that is, methods that produce a ranking by predicting graded relevance scores for query-document pairs and then sorting documents according to their score—a model $p(y_z|x, z; \boldsymbol{\theta})$ can typically be derived from the graded relevance predictor. This model can be used to approximate the conditional $p(y_z|x, z)$. Likewise, $R[\mathbf{r}]$ is replaced by an introspective

performance $\check{R}[\mathbf{r}]$ calculated from Equation 3.1, where the integral over $\mathcal{X}$ is replaced by a sum over the pool, $p(x) \approx \hat{p}(x)$, and $p(\mathbf{y}|x) = \prod_{z \in \mathcal{Z}} p(y_z|x, z; \boldsymbol{\theta})$. The performance difference Δ is approximated by $\check{\Delta} = \check{R}[\mathbf{r}_1] - \check{R}[\mathbf{r}_2]$. Recall that as long as $p(x) > 0$ implies $q(x) > 0$, the weighting factors ensure that such approximations do not introduce an asymptotic bias in our estimator (see Equation 3.11). With these approximations, we arrive at the following empirical sampling distributions in a ranking setting.

Derivation 1. *When relevance labels for individual items are independent given the query (see Equation 6.3), and the conditional $p(y_z|x, z)$ is approximated by a model $p(y|x, z; \boldsymbol{\theta})$ of graded relevance, the sampling distributions minimizing $\frac{1}{n}\sigma_q^2$ and $\frac{1}{n}\tau_q^2$ in a pool-based setting resolve to*

$$q^*(x) \propto \frac{1}{\sqrt{\lambda(x)}} \sqrt{\mathbb{E}_{\mathbf{y} \sim p(\mathbf{y}|x)} \left[\left(\ell(\mathbf{r}(x), \mathbf{y}) - \check{R} \right)^2 \middle| \, x; \boldsymbol{\theta} \right]} \tag{6.10}$$

and

$$q^*(x) \propto \frac{1}{\sqrt{\lambda(x)}} \sqrt{\mathbb{E}_{\mathbf{y} \sim p(\mathbf{y}|x)} \left[\left(\delta(x, y) - \check{\Delta} \right)^2 \middle| \, x; \boldsymbol{\theta} \right]}, \tag{6.11}$$

respectively. Here, for any function $g(x, \mathbf{y})$ of a query x and label vector $\mathbf{y}$,

$$\mathbb{E}_{\mathbf{y} \sim p(\mathbf{y}|x)} \left[g(x, \mathbf{y}) | \, x; \boldsymbol{\theta} \right] = \sum_{\mathbf{y} \in \mathcal{Y}^{\mathcal{Z}}} g(x, \mathbf{y}) \prod_{z \in \mathcal{Z}} p(y_z|x, z; \boldsymbol{\theta}) \tag{6.12}$$

denotes expectation of $g(x, \mathbf{y})$ with respect to label vectors $\mathbf{y}$ generated according to $p(y_z|x, z, \boldsymbol{\theta})$.

For an intuition of the sampling distributions see Section 4.2.2, 4.3, and 5.2.1. Computation of the empirical sampling distributions given by Equations 6.10 and 6.11 requires the computation of $\mathbb{E}\left[g(x, \mathbf{y}) | \, x; \boldsymbol{\theta} \right]$, which is defined in terms of a sum over exponentially many relevance label vectors $\mathbf{y} \in \mathcal{Y}^{\mathcal{Z}}$; a mere application of the previous results would be intractable in practice. However, Theorem 6.1 states that the empirical sampling distributions can be computed in polynomial time.

Theorem 6.1 (Polynomial-time computation of empirical sampling distributions)**.** *The empirical sampling distribution given by Equation 6.10 can be computed in time*

$$\mathcal{O}\left(m|\mathcal{Y}|\max_x|\mathbf{r}(x)|\right) \quad \textit{for} \quad \ell \in \{\ell_{dcg}, \ell_{err}\}.$$

The empirical sampling distribution given by Equation 6.11 can be computed in time

$$\mathcal{O}\left(m|\mathcal{Y}|\max_x(|\mathbf{r}_1(x) \cup \mathbf{r}_2(x)|)\right) \quad \textit{for} \quad \ell = \ell_{dcg},$$
$$\mathcal{O}\left(m|\mathcal{Y}|\max_x(|\mathbf{r}_1(x)| \cdot |\mathbf{r}_2(x)|)\right) \quad \textit{for} \quad \ell = \ell_{err}.$$

where m is the number of unlabeled queries in the pool D_m.

Polynomial-time solutions are derived by dynamic programming. Specifically, after substituting Equations 6.1 and 6.2 into Equations 6.10 and 6.11 and exploiting the independence assumption given by Equation 6.3, Equations 6.10 and 6.11 decompose into cumulative sums and products of expectations over individual item labels $y \in \mathcal{Y}$. These sums and products can be computed in polynomial time. A detailed proof of Theorem 6.1 is included in the Appendix A.1.

The active estimation algorithm for DCG and ERR follows Algorithm 2 and 3; queries $x_1, ..., x_n$ are sampled with replacement from the pool according to the distribution prescribed by Derivation 1 and items included in $\mathbf{r}(x_i)$ or $\mathbf{r}_1(x_i) \cup \mathbf{r}_2(x_i)$ are annotated by a human labeler. Then, an estimate $\hat{R}_{n,q}$ of the true (relative) ranking performance can be computed with respect to the gain function ℓ_{dcg} and ℓ_{err}, respectively.

6.3 Empirical Results

We compare active estimation of ranking performance (Algorithm 2 and 3, labeled *active*), where instances are drawn according to Equation 6.10 or Equation 6.11, respectively, to estimation based on a test sample drawn uniformly from the pool (labeled *passive*). The computation of the optimal sampling distribution q^* from Derivation 1 requires a model $p(y_z|x, z; \boldsymbol{\theta})$ of graded relevance. If no such model is available, a uniform distribution $p(y_z|x, z; \boldsymbol{\theta}) = \frac{1}{|\mathcal{Y}|}$ can be used instead. In contrast to the measures studied in Chapter 4, the optimal sampling distributions for evaluating the performance in terms of DCG or ERR

do not degenerates to uniform sampling. We denote this baseline by $active_{uniD}$. In analogy to Section 4.4.1, we study two simplified sampling distributions in order to quantify the effect of modeling costs; $active_{uniC}$ assumes $\lambda(x) = 1$ for all $x \in \mathcal{X}$ in Equations 6.10 and 6.11 and the heuristic sampling distribution $q(x) \propto \lambda(x)^{-1/2}$ is based only on costs (labeled $active_{\lambda}$). We have shown how the resulting sampling distributions can be computed in polynomial time (see Derivation 1 and Theorem 6.1).

Experiments are performed on the Microsoft Learning to Rank MSLR-WEB30k data set (see Microsoft Research, 2010). It contains 31,531 queries, and a set of documents for each query whose relevance for the query has been determined by human labelers in the process of developing the Bing search engine. The resulting 3,771,125 query-document pairs are represented by 136 features widely used in the information retrieval community (such as query term statistics, page rank, and click counts). Relevance labels take values from 0 (irrelevant) to 4 (perfectly relevant).

The data are split into five folds. On one fold, we train ranking functions using different graded relevance models (details below). The remaining four folds serve as a pool of unlabeled test queries; we estimate (in Section 6.3.1) or compare (in Section 6.3.2) the performance of the ranking functions by drawing and labeling queries from this pool according to Algorithm 2 and 3 using the instrumental distribution discussed above. Test queries are drawn until a labeling budget Λ is exhausted. Labeling a query x involves rating the relevance of all documents in the associated list $\mathbf{r}(x)$ for the query (119 documents on average). To quantify the human effort realistically, we model the labeling costs $\lambda(x)$ as proportional to a sum of costs incurred for labeling individual documents $z \in \mathbf{r}(x)$; labeling costs for a single document z are assumed to be logarithmic in the document length.

All evaluation techniques, both active and passive, can approximate ℓ_{dcg} and ℓ_{err} for a query x by requesting labels only for the first k documents in the ranking. The number of documents for which the MSLR-WEB20k data set provides labels varies over the queries at an average of 119 documents per query. In our experiments, we use all documents for which labels are provided for each query and for all evaluation methods under investigation. Alternatively, one could choose to approximate ℓ_{dcg} and ℓ_{err} more coarsely by labeling fewer documents per query than suggested in the MSLR-WEB30k data set. Since the ranking measures ERR and DCG are monotonically increasing with the number of ranked items, this choice incur an additional bias regarding to the true ranking performance. The trade-off between bias and labeling costs per query applies equally to active

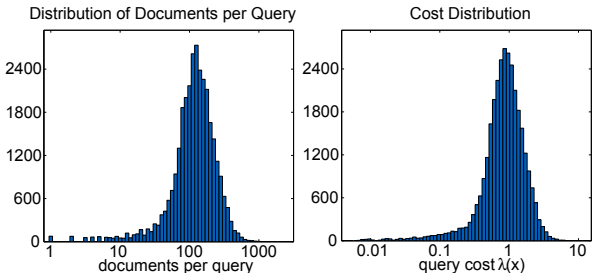


Figure 6.1: Distribution of the number of documents per query (left) and the query labeling costs $\lambda(x)$ (right) in the Microsoft Learning to Rank Dataset.

and passive estimation methods and is thus orthogonal to the query selection problem studied in this chapter.

Figure 6.1 (left) shows the distribution of the number of documents per query over the entire data set. The cost unit is chosen such that average labeling costs for a query are one. Figure 6.1 (right) shows the distribution of labeling costs $\lambda(x)$. All results are averaged over the five folds and 5,000 repetitions of the evaluation process. Error bars indicate the standard error.

6.3.1 Estimating the Performance of a Ranking Function

Based on the outcome of the 2010 Yahoo ranking challenge (Mohan et al., 2011; Chapelle & Chang, 2011), we choose a pointwise ranking approach and employ Random Forest regression (Breiman, 2001) to train graded relevance models on query-document pairs. The ranking function is obtained by returning all documents associated with a query sorted according to their predicted graded relevance. We apply the approach of Li et al. (2007) and Mohan et al. (2011) to obtain the probability estimates $p(y_z|x, z; \boldsymbol{\theta})$ required by *active* and $\textit{active}_{uniC}$ from the Random Forest model. As an alternative graded relevance model, we also study a MAP version of Ordered Logit (McCullagh, 1980); this model directly provides us with probability estimates $p(y_z|x, z; \boldsymbol{\theta})$. For both models, a ranking function is obtained by returning all documents associated with a query sorted according to their predicted graded relevance. Half of the available training fold is used for actual model training, the other half is used as a validation set to tune hyperparameters of the respective ranking model.

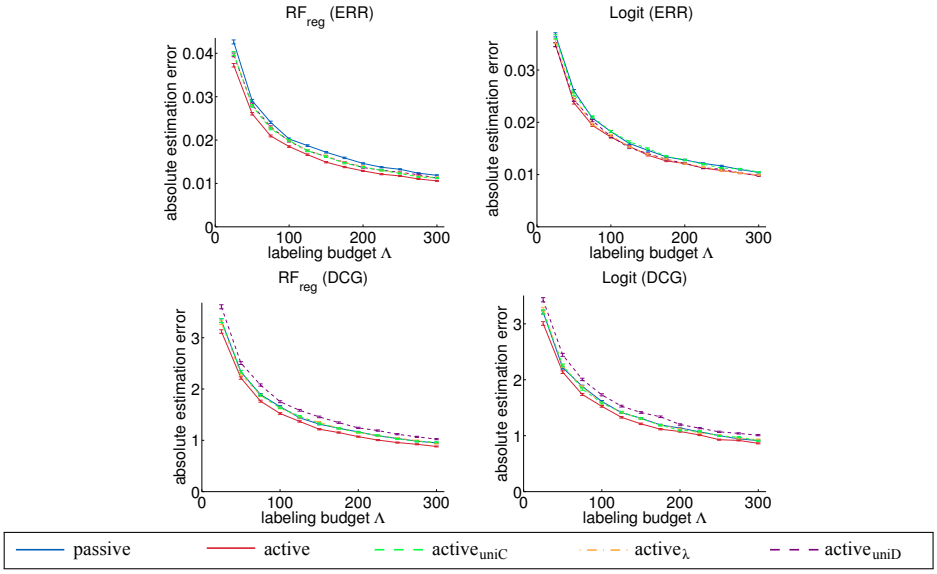


Figure 6.2: Estimation error over Λ when evaluating Random Forest regression (left column) and Ordered Logit (right column) with performance measure ERR (top) and DCG (bottom). Error bars indicate the standard errors.

Figure 6.2 shows absolute deviation between true ranking performance and estimated ranking performance as a function of the labeling budget Λ for the performance measures ERR and DCG. True performance is taken to be the performance over all test queries. We observe that active estimation is significantly more accurate than passive estimation; the labeling budget can be reduced from $\Lambda = 300$ by between 10% and 20% depending on the ranking method and performance measure under study.

6.3.2 Comparing the Performance of Ranking Functions

We additionally train linear Ranking SVM (Herbrich et al., 2000) and the ordinal classification extension to Random Forests (Li et al., 2007; Mohan et al., 2011), and compare the resulting ranking functions to those of the Ordered Logit and Random Forest regression models. For the comparison of Random Forest vs. Ordered Logit both models provide us with estimates $p(y_z|x, z; \boldsymbol{\theta})$; in this case a mixture model (see Equation 5.11) is employed as described in Section 5.2.2. We

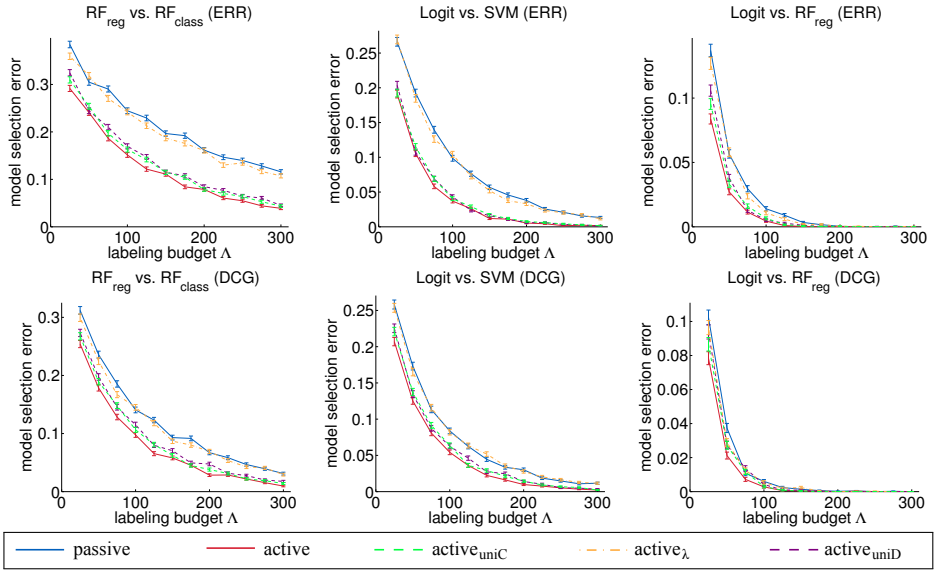


Figure 6.3: Model selection error over labeling budget Λ when comparing Random Forest regression vs. classification (left), Ordered Logit vs. Ranking SVM (center), and Ordered Logit vs. Random Forest regression (right). The performance measure is ERR (top) and DCG (bottom). Error bars indicate the standard error.

measure model selection error, defined as the fraction of experiments in which an evaluation method does not correctly identify the model with higher true performance. Figure 6.3 shows model selection error as a function of the available labeling budget for different pairwise comparisons and the performance measures ERR and DCG. Active estimation more reliably identifies the model with higher ranking performance, saving between 30% and 55% of labeling effort compared to passive estimation. We observe that the gains of *active* versus *passive* are not only due to differences in query costs; the baseline $active_{uniC}$, which does not take into account query costs for computing the sampling distribution, performs almost as well as *active*.

As a further comparative evaluation we simulate an index update. An outdated index with lower coverage is simulated by randomly removing 10% of all query-document pairs from each result list $\mathbf{r}(x)$ for all queries. Random Forest regression is employed for graded relevance prediction. Active and passive estimation methods are applied to estimate the difference in performance between

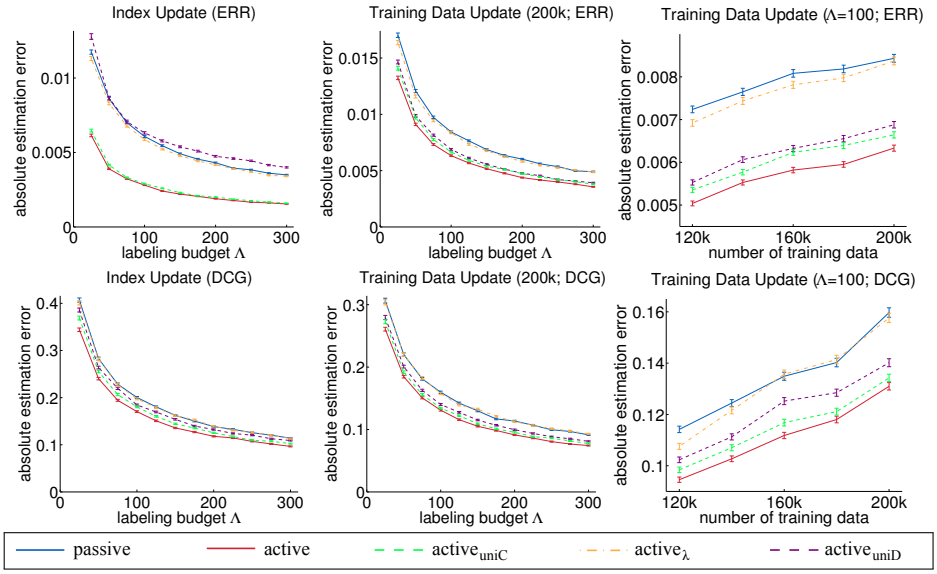


Figure 6.4: Absolute estimation error over labeling costs Λ for a simulated index update affecting 10% of items for each query (left). Absolute estimation error comparing ranking functions trained on 100,000 vs. 200,000 query-document pairs over Λ (center), and over training set size of second model at $\Lambda = 100$ (right). The performance measure is ERR (top) and DCG (bottom). Error bars indicate the standard error.

models based on the outdated and current index. Figure 6.4 (left) shows absolute deviation of estimated from true performance difference over labeling budget Λ for the performance measures ERR and DCG. We observe that active estimation quantifies the impact of the index update more accurately than passive estimation, saving approximately 75% of labeling effort for the performance measure ERR and about 30% labeling effort for DCG.

We finally simulate the incorporation of novel sources of training data by comparing a Random Forest model trained on 100,000 query-document pairs ($\mathbf{r}_1$) to a Random Forest model trained on between 120,000 and 200,000 query-document pairs ($\mathbf{r}_2$). The difference in performance between $\mathbf{r}_1$ and $\mathbf{r}_2$ is estimated using active and passive methods. Figure 6.4 (center) shows absolute deviation of estimated from true performance difference for models trained on 100,000 and 200,000 instances as a function of Λ. Active estimation quantifies the performance gain resulting from additional training data more accurately, reducing labeling costs by approximately 45% (ERR) and 30% (DCG). Figure 6.4 (right)

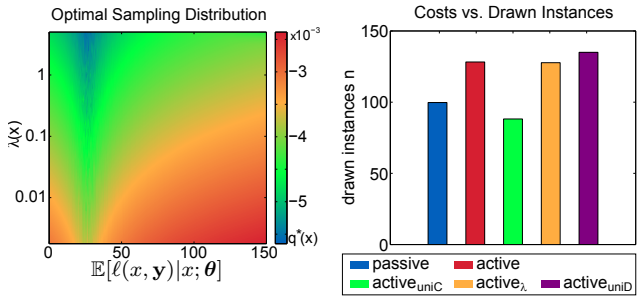


Figure 6.5: Heatmap of the sampling distribution $q^*(x)$ when evaluating a Random Forest regression model in terms of DCG, plotted into a two-dimensional space with axes $\lambda(x)$ and $\mathbb{E}[\ell(x, \mathbf{y})|x; \theta]$ (left). Number of queries drawn by passive and active estimation methods for a labeling budget of $\Lambda = 100$ when evaluating a Random Forest regression model in terms of DCG (right).

shows estimation error as a function of the number of query-document pairs the model $\mathbf{r}_2$ is trained on for $\Lambda = 100$. Since the model $\mathbf{r}_1$, which is trained on 100,000 instances, is kept fixed, the performance difference between the models and thus the estimation error increases with an increasing number of training data. Active estimation significantly reduces the estimation error compared to passive estimation for all training set sizes.

6.3.3 Influence of Query Costs on the Sampling Distribution

The empirical sampling distributions prescribed by Derivation 1 select queries $x \in D_m$ based on their cost $\lambda(x)$ and intrinsic expectations of ranking performance given by $\breve{R}$ (Equation 6.10) and $\breve{\Delta}$ (Equation 6.11), respectively. Figure 6.5 (left) shows a representative example of the sampling distribution $q^*(x)$ given by Equation 6.10 plotted into a two-dimensional space with axes $\lambda(x)$ and intrinsic expected ranking performance $\mathbb{E}[\ell(x, \mathbf{y})|x; \boldsymbol{\theta}]$ for the query x. We observe that low-cost queries are preferred over high-cost queries, and queries for which the intrinsic ranking performance $\mathbb{E}[\ell(x, \mathbf{y})|x; \boldsymbol{\theta}]$ for query x is far away from the intrinsic overall performance $\breve{R}$ are more likely to be chosen (in the example, approximately $\breve{R} = 26$).

Optimization problems 6.5 and 6.6 constitute a trade-off between labeling costs

and informativeness of a test query: optimization over n implies that many inexpensive or few expensive queries could be chosen. On average, active estimation prefers to draw more (but cheaper) queries than passive estimation (Equations 6.10 and 6.11). Figure 6.5 (right) shows the number of queries actually drawn by the different evaluation methods for a labeling budget of $\Lambda = 100$ when estimating absolute DCG for Random Forest regression. We observe that the active estimation methods that take into account costs in the computation of the optimal sampling distribution (*active* and $active_{uniD}$) draw more instances than *passive* and $active_{uniC}$.

6.4 Summary and Related Work

There has been significant interest in learning ranking functions from data in order to improve search result relevance in information retrieval (Li et al., 2007; Zheng et al.; Burges, 2010; Mohan et al., 2011). This has partly been driven by the recent release of large-scale data sets derived from commercial search engines, such as the Microsoft Learning to Rank data sets (see Section 6.3) and the Yahoo Learning to Rank Challenge data sets (Chapelle & Chang, 2011). These data sets serve as realistic benchmarks for evaluating and comparing the performance of different ranking algorithms.

In this chapter, we have applied ideas from active risk estimation (see Chapter 4) and active comparison (see Chapter 5) to the problem of estimating the performance of ranking functions as accurately as possible for a fixed labeling budget. We explicitly model instance-specific labeling costs as proportional to a sum of logarithmic document lengths and constrain overall costs rather than the number of test instances that can be drawn. In a ranking setting, optimal sampling distributions derived from Theorem 4.1 and Theorem 5.1 involve sums over an exponential number of joint relevance label assignments (see Derivation 1). We have shown that they can be computed in polynomial time using dynamic programming (see Theorem 6.1).

Besides sampling queries, it is also possible to sample subsets of documents to be labeled for a given query. Carterette et al. (2006) use document sampling to decide which of two ranking functions achieves higher *precision at k*. Aslam et al. (2006) use document sampling to obtain unbiased estimates of mean average precision and mean R-precision. Carterette & Smucker (2007) study statistical significance testing from reduced document sets.

Empirically, we observed that active estimates of ranking performance are more accurate than passive estimates. In different experimental settings—estimation of the performance of a single ranking model, comparison of different types of ranking models, simulated index updates—performing active estimation resulted in saved labeling efforts of between 10% and 75%.

Chapter 7

Conclusion

Evaluating the performance of predictive models becomes challenging if labeled instances that represent the desired test distribution are unavailable. The overall goal of this thesis was to devise evaluation procedures that enable us to estimate the predictive performance of a given model as accurately as possible at minimal labeling costs. Chapter 3 has presented commonly used performance measures and estimators which are the fundamental tools for model evaluation and comparison. In particular, we have generalized the regular risk functional and have shown that F-measures, which are defined as empirical estimates, consistently estimate a quantity that falls into the class of generalized risks (see Section 3.3). An analysis of the distributions that governs the estimators gives rise to confidence intervals and statistical tests which quantify the remaining uncertainty of the estimate with respect to the true quantity. All estimates require a set of labeled instances drawn either directly from the test distribution or from a known instrumental distribution.

When labeled instances are not available in advance or do not represent the test distribution, new instances have to be drawn and labeled. In many practical applications, drawing unlabeled instances from the distribution the model is exposed to is inexpensive, whereas obtaining the labels is a costly process and typically involves a human expert. A standard approach is to draw instances directly from the test distribution and query their labels at a cost until some pre-defined labeling budget is exhausted. However, the set of test instances—for which the relevance labels have to be determined—does not necessarily have to be a random set sampled from the test distribution; drawing instances independently of their expected label outcome may not be label-efficient. In Chapter 4, we introduced a new, active evaluation process. This process consists of an instrumental distribution and a corresponding estimator. The instrumental distribution is used to select more label-efficient set of instances. The choice determines the effectiveness of the evaluation process in terms of the labeling costs needed to achieve a certain level of accuracy or, equivalently, the expected estimation error

for a fixed labeling budget. We have analyzed the asymptotic distribution of the performance estimator and have derived the instrumental distribution that asymptotically minimizes the estimation error of the active estimator when used to select instances.

The active estimator exploits properties of the considered performance measure and the model to be evaluated. It can be immediately applied to evaluate probabilistic predictive models with respect to any generalized risk. We studied different performance measures for regression and classification problems in several domains. Compared to estimates based on sampling directly from the test distribution, estimates from active evaluation are more accurate when the model has a certain quality; the active evaluation procedure outperforms the standard approach for a per-instance label likelihood of 0.6 or above. The advantage of active estimates of F-measures are particularly strong for skewed classes. Moreover, we observed the confidence intervals of active estimates to be tighter and more reliable even for small test samples. These results indicate that estimation accuracy can be improved by drawing test examples from an appropriately engineered instrumental distribution. However, performance measures, which assess the predictive performance independently of the discrimination threshold cannot be expresses as a generalized risk. This holds for the area under the receiver operating characteristic curve (AUC), the precision-recall break-even point, as well as the maximal F-score. Devising an optimal sampling procedure for these measures is an interesting opportunity for future research.

The task of comparing two predictive models and identifying the model with higher performance as confidently as possible on a fixed labeling budget was studied in Chapter 5. A statistical test is typically used to reject the hypothesis that observed performance differences are due to chance. We have derived the instrumental distribution that asymptotically maximizes the power of a two-sided paired Wald test, and thereby minimizes the likelihood of choosing the inferior model for a fixed labeling budget. The instrumental distribution intuitively gives preference to test instances on which the models disagree strongly. Proper test procedures for multiple models consist of several sequent steps and are thus hard to be analyzed. We have derived a heuristic instrumental distribution for comparing multiple models as a mixture of pairwise-optimal sampling distributions. Empirically, we observed that the active comparison method identifies the model with lower risk more often. Furthermore, significant risk differences are more quickly detected than by a traditional comparison, which is based on a uniform sample of test instances. We also have verified that active comparisons do not

lead to increased false-positive significance results. Simplified variants of the active comparison method have been identified which do not rely on the predictive distribution. They do not depend on the model quality and are empirically still competitive.

In Chapter 4 and 5, we have derived active evaluation and comparison processes which are applicable for performance measures that can be decomposed into instance-specific loss functions. However, this decomposition can be arbitrarily complex and computing the corresponding optimal distribution efficiently may be challenging. In Chapter 6, we used active estimators to evaluate the quality of ranking functions. For the commonly used performance measures discounted cumulative gain (DCG) and expected reciprocal rank (ERR) a naïve computation of the empirical sampling distribution is exponential in the number of the retrieved items. We have derived a polynomial-time solution using dynamic programming and have studied the benefit of our method using a real word data set for web search. In addition of drawing queries, it is also reasonable in practice to sample subsets of documents to be labeled for a given query. However, this approach leads to serious problems, since the considered performance measures depend on the relevance labels of all individual documents for a given query. Substituting missing entries with an estimate or a default value yields an estimate that is strongly biased.

Active evaluation is reminiscent of active learning (see Section 2.3) in many ways. Firstly, the active comparison problem can be seen as an extreme case of active learning, in which the model space contains only a finite number of models. We have discussed extensions to comparing infinitely many models and the relationship to existing active learning methods in Section 5.3. Secondly, the unknown conditional distribution is of particular interest and has to be estimated accurately in order to derive predictive models and to evaluate their performance. Unfortunately, the theoretical optimal sampling distribution in each case depends on this unknown distribution. In analogy to active learning algorithms, our approach uses the current model to decide on instances whose class labels are queried. Active learning algorithms improve the estimate of the true conditional distribution with increasing number of labeled instances. It seems to be natural to extend also the active evaluation process to iteratively refine the instrumental distribution. However, the instrumental distribution that minimizes the estimation error of the predictive performance, does not yield an optimal estimate of the model parameters. Therefore the labeled instances are not optimal to improve the active evaluation process. Aside from this, a sampling distribution

that varies as the instances are drawn is in conflict with the i.i.d.-assumption. However, this assumption enables us to analyze the estimation error. Developing dual techniques which trade off between sampling instances according to the given model and improving the predictive distribution provides interesting research opportunities beyond the scope of this thesis. The active evaluation procedure proposed in this thesis uses a constant sampling distribution, which is proven to be asymptotically optimal.

Appendix A

Appendix

A.1 Proof of Theorem 6.1

In order to show that the empirical sampling distributions, which are given by Equations 6.10 and 6.11 can be computed efficiently, we have to show that Equation 6.12 can be computed efficiently. This can be done by suitable algebraic manipulation, exploiting the independence assumption given by Equation 6.3. In the following proofs, we denote by m the number of unlabeled queries of the pool D_m. All expectations and variances are over the distribution $q(x)p(\mathbf{y}|x)$. We omit the underlying distribution to keep the notation uncluttered.

Empirical Sampling Distribution for Absolute Estimation (see Equation 6.10) with $\ell = \ell_{dcg}$

It suffices to show that the intrinsic performance $\check{R}$ (see Equation 4.22) can be computed in time $\mathcal{O}(m|\mathcal{Y}| \max_x |\mathbf{r}(x)|)$, and that for any $x \in \mathcal{X}$ the quantity $\mathbb{E}[(\ell(\mathbf{r}(x), \mathbf{y}) - \check{R})^2 | x, \boldsymbol{\theta}]$ can be computed in time $\mathcal{O}(|\mathcal{Y}||\mathbf{r}(x)|)$ given $\check{R}$.

We first note that for any $z \in \mathcal{Z}$, it holds that

$$
\begin{aligned}
&\mathbb{E}\left[\kappa_{dcg}\left(y_z, i\right) | \, x; \boldsymbol{\theta}\right] \\
&= \sum_{\mathbf{y} \in \mathcal{Y}^{\mathcal{Z}}} \kappa_{dcg}\left(y_z, i\right) \prod_{z' \in \mathcal{Z}} p(y_{z'} | x, z'; \boldsymbol{\theta}) \\
&= \sum_{y_z} \sum_{\mathbf{y} \in \mathcal{Y}^{\mathcal{Z} \setminus \{z\}}} \kappa_{dcg}\left(y_z, i\right) \prod_{z' \in \mathcal{Z}} p(y_{z'} | x, z'; \boldsymbol{\theta}) \\
&= \sum_{y_z} \kappa_{dcg}\left(y_z, i\right) p(y_z | x, z; \boldsymbol{\theta}) \sum_{\mathbf{y} \in \mathcal{Y}^{\mathcal{Z} \setminus \{z\}}} \prod_{z' \in \mathcal{Z} \setminus \{z\}} p(y_{z'} | x, z'; \boldsymbol{\theta}) \\
&= \sum_{y_z} \kappa_{dcg}\left(y_z, i\right) p(y_z | x, z; \boldsymbol{\theta}), \qquad \text{(A.1)}
\end{aligned}
$$

where $\mathbf{y} \in \mathcal{Y}^{\mathcal{Z}\setminus\{z\}}$ is a vector of relevance labels $y_{z'}$ for all $z' \in \mathcal{Z} \setminus \{z\}$. The expected value $\mathbb{E}\left[\kappa_{dcg}\left(y_z, i\right)\middle| x; \boldsymbol{\theta}\right]$ can thus be computed in time $\mathcal{O}(|\mathcal{Y}|)$.

Let $\kappa_i = \kappa_{dcg}\left(y_{r_i(x)}, i\right)$, such that the DCG can be expressed as $\ell_{dcg}\left(\mathbf{r}(x), \mathbf{y}\right) = \sum_{i=1}^{|\mathbf{r}(x)|} \kappa_i$. Then, for $\ell = \ell_{dcg}$ it holds that

$$\begin{aligned}
\check{R} &= \frac{1}{m} \sum_{x \in D_m} \mathbb{E}\left[\ell\left(\mathbf{r}(x), \mathbf{y}\right)\middle| x; \boldsymbol{\theta}\right] \\
&= \frac{1}{m} \sum_{x \in D_m} \sum_{\mathbf{y} \in \mathcal{Y}^{\mathcal{Z}}} \sum_{i=1}^{|\mathbf{r}(x)|} \kappa_i \prod_{z \in \mathcal{Z}} p(y_z | x, z; \boldsymbol{\theta}) \\
&= \frac{1}{m} \sum_{x \in D_m} \sum_{i=1}^{|\mathbf{r}(x)|} \mathbb{E}\left[\kappa_i\middle| x; \boldsymbol{\theta}\right],
\end{aligned}$$

therefore $\check{R}$ can be computed in time $\mathcal{O}(m|\mathcal{Y}| \max_x |\mathbf{r}(x)|)$. We furthermore derive

$$\begin{aligned}
&\mathbb{E}\left[\left(\ell\left(\mathbf{r}(x), \mathbf{y}\right) - \check{R}\right)^2 \middle| x; \boldsymbol{\theta}\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{|\mathbf{r}(x)|} \kappa_i - \check{R}\right)^2 \middle| x; \boldsymbol{\theta}\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{|\mathbf{r}(x)|} \kappa_i^2 + 2 \sum_{i=1}^{|\mathbf{r}(x)|} \sum_{l=i+1}^{|\mathbf{r}(x)|} \kappa_i \kappa_l - 2\check{R} \sum_{i=1}^{|\mathbf{r}(x)|} \kappa_i + \check{R}^2 \middle| x; \boldsymbol{\theta}\right] \qquad \text{(A.2)} \\
&= \sum_{i=1}^{|\mathbf{r}(x)|} \left(\mathbb{E}\left[\kappa_i^2\middle| x; \boldsymbol{\theta}\right] + 2\mathbb{E}\left[\kappa_i\middle| x; \boldsymbol{\theta}\right] \sum_{l=i+1}^{|\mathbf{r}(x)|} \mathbb{E}\left[\kappa_l\middle| x; \boldsymbol{\theta}\right] \right. \\
&\qquad \left. - 2\check{R}\mathbb{E}\left[\kappa_i\middle| x; \boldsymbol{\theta}\right] \right) + \check{R}^2 \qquad \text{(A.3)}
\end{aligned}$$

Equation A.2 expands the square of sums twice and in Equation A.3 we make use of the independence assumption for item relevance (see Equation 6.3) from which follows that $\mathbb{E}\left[\kappa_i \kappa_l\middle| x; \boldsymbol{\theta}\right] = \mathbb{E}\left[\kappa_i\middle| x; \boldsymbol{\theta}\right] \mathbb{E}\left[\kappa_l\middle| x; \boldsymbol{\theta}\right]$. Equation A.3 can now be evaluated in time $\mathcal{O}(|\mathcal{Y}||\mathbf{r}(x)|)$: cumulative sums over l and all terms of the form $\mathbb{E}\left[\kappa_i\middle| x; \boldsymbol{\theta}\right]$ as well as $\mathbb{E}\left[\kappa_i^2\middle| x; \boldsymbol{\theta}\right]$ can be precomputed in time $\mathcal{O}(|\mathcal{Y}||\mathbf{r}(x)|)$ (see Equation A.1), and the second summand of Equation A.3 can then be computed in time $\mathcal{O}(|\mathcal{Y}||\mathbf{r}(x)|)$. Thus, the empirical sampling distribution can overall be computed in time $\mathcal{O}(m|\mathcal{Y}| \max_x |\mathbf{r}(x)|)$. □

Empirical Sampling Distribution for Absolute Estimation (see Equation 6.10) with $\ell = \ell_{err}$

It suffices to show that the intrinsic performance $\check{R}$ (see Equation 4.22) can be computed in time $\mathcal{O}(m|\mathcal{Y}| \max_x |\mathbf{r}(x)|)$, and that for any $x \in \mathcal{X}$ the quantity $\mathbb{E}[(\ell(\mathbf{r}(x), \mathbf{y}) - \check{R})^2 | x; \boldsymbol{\theta}]$ can be computed in time $\mathcal{O}(|\mathcal{Y}||\mathbf{r}(x)|)$ given $\check{R}$. Let $\kappa_i = \kappa_{err}\left(y_{r_i(x)}\right)$ and $\kappa_{1:i} = \frac{1}{i}\kappa_i \prod_{l=1}^{i-1}(1-\kappa_l)$, such that the ERR can be expressed as $\ell_{err}(\mathbf{r}(x), \mathbf{y}) = \sum_{i=1}^{|\mathbf{r}(x)|} \kappa_{1:i}$. From the independence assumption for item relevance (Equation 6.3), it follows that

$$\mathbb{E}\left[\kappa_{1:i} \middle| x; \boldsymbol{\theta}\right] = \frac{1}{i}\mathbb{E}\left[\kappa_{err}\left(y_{r_i(x)}\right) \middle| x; \boldsymbol{\theta}\right] \prod_{l=1}^{i-1}\left(1 - \mathbb{E}\left[\kappa_{err}\left(y_{r_l(x)}\right) \middle| x; \boldsymbol{\theta}\right]\right)$$

In analogy to $\ell = \ell_{dcg}$ (see Equation A.1) it follows that

$$\mathbb{E}\left[\kappa_{err}(y_z) \middle| x; \boldsymbol{\theta}\right] = \sum_{y_z} \kappa_{err}(y_z)\, p(y_z | x, z; \theta) \tag{A.4}$$

for any item $z \in \mathcal{Z}$ and can thus be computed in time $\mathcal{O}(|\mathcal{Y}|)$. Thus, all terms $\mathbb{E}\left[\kappa_{1:i} \middle| x; \boldsymbol{\theta}\right]$ for $i = 1, ..., |\mathbf{r}(x)|$ can together be computed in time $\mathcal{O}(|\mathcal{Y}||\mathbf{r}(x)|)$, since all terms of the form $\mathbb{E}\left[\kappa_{err}(y_z) \middle| x; \boldsymbol{\theta}\right]$ (see Equation A.4) can be precomputed in time $\mathcal{O}(|\mathcal{Y}||\mathbf{r}(x)|)$ and subsequently the cumulative products over l can be computed in time $\mathcal{O}(|\mathcal{Y}||\mathbf{r}(x)|)$. Therefore, $\check{R}$ for $\ell = \ell_{err}$ can be computed in time $\mathcal{O}(|\mathcal{Y}||D| \max_x |\mathbf{r}(x)|)$, since it holds that

$$\begin{aligned}
\check{R} &= \frac{1}{m} \sum_{x \in D_m} \mathbb{E}\left[\ell(\mathbf{r}(x), \mathbf{y}) \middle| x; \boldsymbol{\theta}\right] \\
&= \frac{1}{m} \sum_{x \in D_m} \sum_{\mathbf{y} \in \mathcal{Y}^{\mathcal{Z}}} \sum_{i=1}^{|\mathbf{r}(x)|} \kappa_{1:i} \prod_{z \in \mathcal{Z}} p(y_z | x, z; \theta) \\
&= \frac{1}{m} \sum_{x \in D_m} \sum_{i=1}^{|\mathbf{r}(x)|} \mathbb{E}\left[\kappa_{1:i} \middle| x; \boldsymbol{\theta}\right].
\end{aligned}$$

We now turn towards the quantity

$$\mathbb{E}\left[(\ell(\mathbf{r}(x), \mathbf{y}) - \check{R})^2 \middle| x; \boldsymbol{\theta}\right].$$

We expand the square of sums twice in Equation A.5. Equation A.6 follows from the independence assumption (Equation 6.3):

$$\begin{aligned}
&\mathbb{E}\left[\left(\ell\left(\mathbf{r}(x),\mathbf{y}\right)-\check{R}\right)^2\middle| x;\boldsymbol{\theta}\right]\\
&=\mathbb{E}\left[\sum_{i=1}^{|\mathbf{r}(x)|}\kappa_{1:i}^2+2\sum_{i=1}^{|\mathbf{r}(x)|}\sum_{l=i+1}^{|\mathbf{r}(x)|}\kappa_{1:i}\kappa_{1:l}-2\check{R}\sum_{i=1}^{|\mathbf{r}(x)|}\kappa_{1:i}+\check{R}^2\middle| x;\boldsymbol{\theta}\right] &\text{(A.5)}\\
&=\sum_{i=1}^{|\mathbf{r}(x)|}\left(\mathbb{E}\left[\kappa_{1:i}^2\middle| x;\boldsymbol{\theta}\right]+2\sum_{l=i+1}^{|\mathbf{r}(x)|}\mathbb{E}\left[\kappa_{1:i}\kappa_{1:l}\middle| x;\boldsymbol{\theta}\right]-2\check{R}\mathbb{E}\left[\kappa_{1:i}\middle| x;\boldsymbol{\theta}\right]\right)+\check{R}^2. &\text{(A.6)}
\end{aligned}$$

In contrast to Equation A.3 in the proof of $\ell=\ell_{dcg}$, items used to calculate $\kappa_{1:i}$ and $\kappa_{1:l}$ in the second summand overlap. However, we note that for $l>i$ the following decomposition holds:

$$\kappa_{1:l}=\frac{1}{l}\kappa_l\left(\prod_{k=1}^{i-1}(1-\kappa_k)\right)(1-\kappa_i)\left(\prod_{k=i+1}^{l-1}(1-\kappa_k)\right). \tag{A.7}$$

Thus $\sum_{l=i+1}^{|\mathbf{r}(x)|}\mathbb{E}\left[\kappa_{1:i}\kappa_{1:l}\middle| x;\boldsymbol{\theta}\right]$ can be expressed as follows. In Equation A.8, we insert the definition of $\kappa_{1:i}$ and the decomposition given by Equation A.7 for $\kappa_{1:l}$. The rest follows from the independence assumption (see Equation 6.3) and reordering terms:

$$\begin{aligned}
&\sum_{l=i+1}^{|\mathbf{r}(x)|}\mathbb{E}\left[\frac{1}{i}\kappa_i\left(\prod_{k=1}^{i-1}(1-\kappa_k)\right)\cdot\right.\\
&\qquad\left.\frac{1}{l}\kappa_l\left(\prod_{k=1}^{i-1}(1-\kappa_k)\right)(1-\kappa_i)\left(\prod_{k=i+1}^{l-1}(1-\kappa_k)\right)\middle| x;\boldsymbol{\theta}\right] &\text{(A.8)}\\
&=\frac{1}{i}\sum_{l=i+1}^{|\mathbf{r}(x)|}\mathbb{E}\left[\kappa_i\left(1-\kappa_i\right)\middle| x;\boldsymbol{\theta}\right]\mathbb{E}\left[\prod_{k=1}^{i-1}(1-\kappa_k)^2\middle| x;\boldsymbol{\theta}\right]\cdot\\
&\qquad\mathbb{E}\left[\frac{1}{l}\kappa_l\prod_{k=i+1}^{l-1}(1-\kappa_k)\middle| x;\boldsymbol{\theta}\right]
\end{aligned}$$

$$= \frac{1}{i}\mathbb{E}\left[\kappa_i\left(1-\kappa_i\right)\middle|\, x;\boldsymbol{\theta}\right]\left(\prod_{k=1}^{i-1}\mathbb{E}\left[\left(1-\kappa_k\right)^2\middle|\, x;\boldsymbol{\theta}\right]\right)\cdot$$
$$\sum_{l=i+1}^{|\mathbf{r}(x)|}\frac{1}{l}\mathbb{E}\left[\kappa_l\middle|\, x;\boldsymbol{\theta}\right]\frac{\prod_{k=i+1}^{|\mathbf{r}(x)|}\mathbb{E}\left[\left(1-\kappa_k\right)\middle|\, x;\boldsymbol{\theta}\right]}{\prod_{k=l}^{|\mathbf{r}(x)|}\mathbb{E}\left[\left(1-\kappa_k\right)\middle|\, x;\boldsymbol{\theta}\right]}$$
$$= \frac{1}{i}\mathbb{E}\left[\kappa_i\left(1-\kappa_i\right)\middle|\, x;\boldsymbol{\theta}\right]\left(\prod_{k=1}^{i-1}\mathbb{E}\left[\left(1-\kappa_k\right)^2\middle|\, x;\boldsymbol{\theta}\right]\right)\cdot$$
$$\left(\prod_{k=i+1}^{|\mathbf{r}(x)|}\mathbb{E}\left[\left(1-\kappa_k\right)\middle|\, x;\boldsymbol{\theta}\right]\right)\sum_{l=i+1}^{|\mathbf{r}(x)|}\frac{\mathbb{E}\left[\kappa_l\middle|\, x;\boldsymbol{\theta}\right]}{l\prod_{k=l}^{|\mathbf{r}(x)|}\mathbb{E}\left[\left(1-\kappa_k\right)\middle|\, x;\boldsymbol{\theta}\right]}. \tag{A.9}$$

Equation A.9 and thus Equation A.6 can be evaluated in time $\mathcal{O}(|\mathcal{Y}||\mathbf{r}(x)|)$, since all cumulative products over k, cumulative sums over l, and expected values can together be precomputed in time $\mathcal{O}(|\mathcal{Y}||\mathbf{r}(x)|)$. Thus, the empirical sampling distribution can overall be computed in time $\mathcal{O}(|\mathcal{Y}||D|\max_x|\mathbf{r}(x)|)$. □

Empirical Sampling Distribution for Comparative Estimation (Equation 6.11) with $\ell = \ell_{dcg}$

Since the intrinsic risks $\check{R}[\mathbf{r}_i]$ can be computed in time $\mathcal{O}(|\mathcal{Y}||D|\max_x|\mathbf{r}_i(x)|)$ and hence the difference $\check{\Delta} = \check{R}[\mathbf{r}_1] - \check{R}[\mathbf{r}_2]$ can be computed in time

$$\mathcal{O}(|\mathcal{Y}||D|\max_x(|\mathbf{r}_1(x)\cup\mathbf{r}_2(x)|))$$

it is sufficient to show that for any $x\in\mathcal{X}$ the quantity

$$\mathbb{E}\left[\left(\ell(\mathbf{r}_1(x),\mathbf{y}) - \ell(\mathbf{r}_2(x),\mathbf{y}) - \check{\Delta}\right)^2\middle|\, x;\boldsymbol{\theta}\right]$$

can be computed in time $\mathcal{O}(|\mathcal{Y}||D||\mathbf{r}_1(x)\cup\mathbf{r}_2(x)|)$. In order to simplify notation we define an inverse ranking function

$$\mathbf{r}^{-1}:\mathcal{X}\times\mathcal{Z}\to\{1,\ldots,|\mathbf{r}(x)|\}\cup\{\infty\}. \tag{A.10}$$

Given a query $x\in\mathcal{X}$ and an item $z\in\mathcal{Z}$ it returns the position in the list of results $\mathbf{r}^{-1}:(x,r_i(x))\mapsto i$ if item z is contained in the results of ranking func-

tion $\mathbf{r}$ and infinity otherwise. Then, ℓ_{dcg} (see Equation 6.1) can be equivalently written as

$$\ell_{dcg}\left(\mathbf{r}(x), \mathbf{y}\right) = \sum_{z \in \mathcal{Z}} \kappa_{dcg}\left(y_z, \mathbf{r}^{-1}(x, z)\right),$$

since $\kappa_{dcg}\left(y, \infty\right) = 0$.

Let $\kappa_z = \kappa_{dcg}\left(y_z, \mathbf{r}_1^{-1}(x, z)\right) - \kappa_{dcg}\left(y_z, \mathbf{r}_2^{-1}(x, z)\right)$ be the difference of the loss functions for two ranking functions. Then, we derive

$$\begin{aligned}
&\mathbb{E}\left[\left(\ell(\mathbf{r}_1(x), \mathbf{y}) - \ell(\mathbf{r}_2(x), \mathbf{y}) - \check{\Delta}\right)^2 \middle| \, x; \boldsymbol{\theta}\right] \\
&= \mathbb{E}\left[\left(\sum_{z \in \mathcal{Z}} \kappa_z - \check{\Delta}\right)^2 \middle| \, x; \boldsymbol{\theta}\right] \\
&= \mathbb{E}\left[\sum_{z \in \mathcal{Z}} \kappa_z^2 + \sum_{z \in \mathcal{Z}} \sum_{\substack{\bar{z} \in \mathcal{Z} \\ \bar{z} \neq z}} \kappa_z \kappa_{\bar{z}} - 2 \sum_{z \in \mathcal{Z}} \kappa_z \check{\Delta} + \check{\Delta}^2 \middle| \, x; \boldsymbol{\theta}\right] \qquad \text{(A.11)} \\
&= \sum_{z \in \mathcal{Z}} \left(\mathbb{E}\left[\kappa_z^2 \middle| \, x; \boldsymbol{\theta}\right] + \mathbb{E}\left[\kappa_z \middle| \, x; \boldsymbol{\theta}\right] \left(\sum_{\substack{\bar{z} \in \mathcal{Z} \\ \bar{z} \neq z}} \mathbb{E}\left[\kappa_{\bar{z}} \middle| \, x; \boldsymbol{\theta}\right] - 2\check{\Delta} \right) \right) + \check{\Delta}^2 \qquad \text{(A.12)}
\end{aligned}$$

Equation A.11 expands the square of sums twice and in Equation A.12 we make use of the independence between different items (see Equation 6.3).

A single quantity of the form $\mathbb{E}\left[\kappa_{\bar{z}} \middle| \, x; \boldsymbol{\theta}\right]$ can be computed in time $\mathcal{O}(|\mathcal{Y}|)$ (see Equation A.1). Since the expectation $\mathbb{E}\left[\kappa_{dcg}\left(y_z, \mathbf{r}_i^{-1}(x, z)\right) \middle| \, x; \boldsymbol{\theta}\right]$ equals zero for all $z \in \mathcal{Z}$ not contained in the results of $\mathbf{r}_i(x)$, the sums over $z \in \mathcal{Z}$ can be computed in time $\mathcal{O}(|\mathcal{Y}| \max_x(|\mathbf{r}_1(x) \cup \mathbf{r}_2(x)|))$. Thus, the empirical sampling distribution can overall be computed in time $\mathcal{O}(|\mathcal{Y}||D| \max_x(|\mathbf{r}_1(x) \cup \mathbf{r}_2(x)|))$. □

Empirical Sampling Distribution for Comparative Estimation (Equation 6.11) with $\ell = \ell_{err}$

Since the intrinsic risks $\check{R}[\mathbf{r}_i]$ can be computed in time $\mathcal{O}(|\mathcal{Y}||D| \max_x |\mathbf{r}_i(x)|)$ and hence the difference $\check{\Delta} = \check{R}[\mathbf{r}_1] - \check{R}[\mathbf{r}_2]$ can be computed in

$$\mathcal{O}(|\mathcal{Y}||D| \max_x(|\mathbf{r}_1(x) \cup \mathbf{r}_2(x)|))$$

it is sufficient to show that for any $x \in \mathcal{X}$ the quantity

$$\mathbb{E}\left[(\ell(\mathbf{r}_1(x), \mathbf{y}) - \ell(\mathbf{r}_2(x), \mathbf{y}) - \check{\Delta})^2 \middle| x; \boldsymbol{\theta}\right]$$

can be computed in time $\mathcal{O}(|\mathcal{Y}||\mathbf{r}_1(x)||\mathbf{r}_2(x)|)$. From the independence assumption (Equation 6.3) it follows that

$$\begin{aligned}
&\mathbb{E}\left[(\ell(\mathbf{r}_1(x), \mathbf{y}) - \ell(\mathbf{r}_2(x), \mathbf{y}) - \check{\Delta})^2 \middle| x; \boldsymbol{\theta}\right] \\
&= \sum_{k=1}^{2}\left(\mathbb{E}\left[\ell_{err}\left(\mathbf{r}_k(x), \mathbf{y}\right)^2 \middle| x; \boldsymbol{\theta}\right] + 2\check{\Delta}(-1)^k \mathbb{E}\left[\ell_{err}\left(\mathbf{r}_k(x), \mathbf{y}\right) \middle| x; \boldsymbol{\theta}\right]\right) + \check{\Delta}^2 \\
&\quad - 2\mathbb{E}\left[\ell_{err}\left(\mathbf{r}_1(x), \mathbf{y}\right) \ell_{err}\left(\mathbf{r}_2(x), \mathbf{y}\right) \middle| x; \boldsymbol{\theta}\right]. \qquad \text{(A.13)}
\end{aligned}$$

In analogy to Equation A.6 in the proof of Theorem 6.1, it can be shown that all terms except $\mathbb{E}\left[\ell_{err}\left(\mathbf{r}_1(x), \mathbf{y}\right) \ell_{err}\left(\mathbf{r}_2(x), \mathbf{y}\right) \middle| x; \boldsymbol{\theta}\right]$ in Equation A.13 can be computed in linear time. In the following we show, that

$$\mathbb{E}\left[\ell_{err}\left(\mathbf{r}_1(x), \mathbf{y}\right) \ell_{err}\left(\mathbf{r}_2(x), \mathbf{y}\right) \middle| x; \boldsymbol{\theta}\right]$$

can be computed in $\mathcal{O}(|\mathcal{Y}||\mathbf{r}_1(x)||\mathbf{r}_2(x)|)$. Let $\kappa_z = \kappa_{err}(y_z)$ and the inverse ranking function $\mathbf{r}^{-1}$ be defined as in proof for $\ell = \ell_{dcg}$ (see Equation A.10). Furthermore, we define

$$\bar{\kappa}_{z'<z,t} = (1 - \kappa_{z'})^{[\![\mathbf{r}_t^{-1}(x,z') < \mathbf{r}_t^{-1}(x,z)]\!]},$$

where $[\![\cdot]\!] \to \{0, 1\}$ denotes the indicator function. Then, $\ell_{err}\left(\mathbf{r}(x), \mathbf{y}\right)$ can be expressed as sum over the pool of items $\mathcal{Z}$:

$$\ell_{err}\left(\mathbf{r}(x), \mathbf{y}\right) = \sum_{z \in \mathcal{Z}} \frac{1}{\mathbf{r}^{-1}(x, z)} \kappa_z \prod_{z' \in \mathcal{Z}} \bar{\kappa}_{z'<z,t}.$$

Then we derive

$$\begin{aligned}
&\mathbb{E}\left[\ell_{err}\left(\mathbf{r}_1(x), \mathbf{y}\right) \ell_{err}\left(\mathbf{r}_2(x), \mathbf{y}\right) \middle| x; \boldsymbol{\theta}\right] \\
&= \sum_{z \in \mathcal{Z}} \sum_{\bar{z} \in \mathcal{Z}} \frac{1}{\mathbf{r}_1^{-1}(x, z)} \frac{1}{\mathbf{r}_2^{-1}(x, \bar{z})} \cdot \\
&\qquad \mathbb{E}\left[\left(\kappa_z \prod_{z' \in \mathcal{Z}} \bar{\kappa}_{z'<z,1}\right)\left(\kappa_{\bar{z}} \prod_{z' \in \mathcal{Z}} \bar{\kappa}_{z'<\bar{z},2}\right) \middle| x; \boldsymbol{\theta}\right]. \qquad \text{(A.14)}
\end{aligned}$$

The expectation in Equation A.14 can be decomposed as follows. In analogy to the proof of Theorem 6.1 for absolute estimation with $\ell = \ell_{err}$ (see Equation A.7), we decompose the cumulative products into disjoint item sets: In Equation A.15, we exclude factors depending on $\bar{z}$ and z, respectively. Equation A.16 and A.17 follow from the independence assumption (Equation 6.3). Furthermore, we summarize factors over the same items in Equation A.17. Finally, we again make use of the independence assumption for item relevance and abbreviate the products over the disjoint factors (Equation A.18):

$$
\begin{aligned}
&\mathbb{E}\left[\left(\kappa_z \prod_{z' \in \mathcal{Z}} \bar{\kappa}_{z'<z,1}\right)\left(\kappa_{\bar{z}} \prod_{z' \in \mathcal{Z}} \bar{\kappa}_{z'<\bar{z},2}\right) \middle| x; \boldsymbol{\theta}\right] \\
&= \mathbb{E}\Bigg[\kappa_z \bar{\kappa}_{\bar{z}<z,1}\left(\prod_{z' \neq \bar{z}} \bar{\kappa}_{z'<z,1}\right) \\
&\qquad \kappa_{\bar{z}} \bar{\kappa}_{z<\bar{z},2}\left(\prod_{\substack{z' \neq \bar{z} \\ \mathbf{r}_1^{-1}(x,z')<\mathbf{r}_1^{-1}(x,z)}} \bar{\kappa}_{z'<\bar{z},2}\right)\left(\prod_{\substack{z' \neq \bar{z} \\ \mathbf{r}_1^{-1}(x,z')>\mathbf{r}_1^{-1}(x,z)}} \bar{\kappa}_{z'<\bar{z},2}\right) \middle| x; \theta\Bigg] \qquad \text{(A.15)} \\
&= \mathbb{E}\left[\kappa_z \bar{\kappa}_{\bar{z}<z,1} \kappa_{\bar{z}} \bar{\kappa}_{z<\bar{z},2} \middle| x; \boldsymbol{\theta}\right] \\
&\qquad \mathbb{E}\left[\left(\prod_{\substack{z' \neq \bar{z} \\ \mathbf{r}_1^{-1}(x,z')<\mathbf{r}_1^{-1}(x,z)}} \bar{\kappa}_{z'<z,1}\right)\left(\prod_{\substack{z' \neq \bar{z} \\ \mathbf{r}_1^{-1}(x,z')<\mathbf{r}_1^{-1}(x,z)}} \bar{\kappa}_{z'<\bar{z},2}\right) \middle| x; \boldsymbol{\theta}\right] \\
&\qquad \mathbb{E}\left[\left(\prod_{\substack{z' \neq \bar{z} \\ \mathbf{r}_1^{-1}(x,z')>\mathbf{r}_1^{-1}(x,z)}} \bar{\kappa}_{z'<\bar{z},2}\right) \middle| x; \boldsymbol{\theta}\right] \qquad \text{(A.16)} \\
&= \underline{\ell}_x^{=}(z,\bar{z}) \left(\prod_{\substack{z' \neq \bar{z} \\ \mathbf{r}_1^{-1}(x,z')<\mathbf{r}_1^{-1}(x,z)}} \mathbb{E}\left[\bar{\kappa}_{z'<z,1} \bar{\kappa}_{z'<\bar{z},2} \middle| x; \boldsymbol{\theta}\right]\right) \cdot \\
&\qquad \left(\prod_{\substack{z' \neq \bar{z} \\ \mathbf{r}_1^{-1}(x,z')>\mathbf{r}_1^{-1}(x,z)}} \mathbb{E}\left[\bar{\kappa}_{z'<\bar{z},2} \middle| x; \boldsymbol{\theta}\right]\right) \qquad \text{(A.17)} \\
&= \underline{\ell}_x^{=}(z,\bar{z})\, \underline{\ell}_x^{<}(z,\bar{z})\, \underline{\ell}_x^{>}(z,\bar{z}), \qquad \text{(A.18)}
\end{aligned}
$$

where

$$\underline{\ell}_x^{=}(z,\bar{z}) = \begin{cases} \mathbb{E}\left[\kappa_z^2 \middle| x;\boldsymbol{\theta}\right], & \text{if } z=\bar{z} \\ \mathbb{E}\left[\kappa_z(1-\kappa_z)^{[\![\mathbf{r}_2^{-1}(x,z)<\mathbf{r}_2^{-1}(x,\bar{z})]\!]} \middle| x;\boldsymbol{\theta}\right] & \\ \quad \cdot\mathbb{E}\left[\kappa_{\bar{z}}(1-\kappa_{\bar{z}})^{[\![\mathbf{r}_1^{-1}(x,\bar{z})<\mathbf{r}_1^{-1}(x,z)]\!]} \middle| x;\boldsymbol{\theta}\right], & \text{if } z\neq\bar{z}, \end{cases} \tag{A.19}$$

$$\underline{\ell}_x^{<}(z,\bar{z}) = \prod_{\substack{z'\neq\bar{z} \\ \mathbf{r}_1^{-1}(x,z')<\mathbf{r}_1^{-1}(x,z)}} \mathbb{E}\left[(1-\kappa_{z'})^{[\![\mathbf{r}_2^{-1}(x,z')<\mathbf{r}_2^{-1}(x,\bar{z})]\!]+1} \middle| x;\boldsymbol{\theta}\right], \tag{A.20}$$

$$\underline{\ell}_x^{>}(z,\bar{z}) = \prod_{\substack{z'\neq\bar{z} \\ \mathbf{r}_1^{-1}(x,z')>\mathbf{r}_1^{-1}(x,z)}} \mathbb{E}\left[(1-\kappa_{z'})^{[\![\mathbf{r}_2^{-1}(x,z')<\mathbf{r}_2^{-1}(x,\bar{z})]\!]} \middle| x;\boldsymbol{\theta}\right]. \tag{A.21}$$

Individual expected values in Equations A.19–A.21 can be computed in $\mathcal{O}(|\mathcal{Y}|)$ (see Equation A.4). Cumulative products in Equation A.20 and A.21 can be precomputed incrementally for all items $z \in \mathcal{Z}$ using the order given by $\mathbf{r}_1^{-1}$ in time $\mathcal{O}(|\mathcal{Y}||\mathcal{Z}|)$. Since $\frac{1}{\mathbf{r}_i^{-1}(x,z)}$ equals zero for all $z \in \mathcal{Z}$ not contained in the results of $\mathbf{r}_i(x)$, Equation A.14 can be computed in $\mathcal{O}(|\mathcal{Y}||\mathbf{r}_1(x)||\mathbf{r}_2(x)|)$. Thus, the sampling distribution can overall be computed in $\mathcal{O}(|\mathcal{Y}||D|\max_x(|\mathbf{r}_1(x)||\mathbf{r}_2(x)|))$. □

A.2 Comprehensive Empirical Results

In this section, we present additional empirical results. We study the task of estimating F-measures of text classification models in Section A.2.1 and of digit recognition models in Section A.2.2.

A.2.1 Text Classification Domain

Table A.1 (top) lists the true one-vs-rest precision, $F_{0.5}$-measure, recall, and accuracy of the actively trained model for the ten classes in the *Reuters-21578* text classification domain. Figure A.1 shows the estimation error of *active*, *passive*, and $\textit{active}_{\textit{err}}$ over number of labeled data, for precision, $F_{0.5}$ and recall estimates and all ten classes in the text classification domain.

Table A.1: Class ratios and model quality for active learned model for ten most frequent occurring topics in Reuters corpus (top). Class ratios and model quality of digit recognition model on *MNIST* (bottom).

	class	class ratio	accuracy	precision	recall	$F_{0.5}$
Reuters	earn	0.510	0.978	0.976	0.980	0.978
	acq	0.280	0.974	0.942	0.966	0.954
	crude	0.044	0.994	0.972	0.879	0.923
	trade	0.041	0.995	0.937	0.940	0.938
	money-fx	0.034	0.991	0.872	0.861	0.867
	interest	0.027	0.992	0.851	0.843	0.847
	ship	0.020	0.994	0.886	0.820	0.850
	sugar	0.016	0.998	0.972	0.921	0.946
	coffee	0.015	0.999	1.000	0.900	0.947
	gold	0.012	0.998	0.954	0.911	0.932
MNIST	0	0.097	0.995	0.966	0.978	0.972
	1	0.113	0.978	0.955	0.844	0.896
	2	0.099	0.986	0.937	0.918	0.927
	3	0.102	0.980	0.885	0.920	0.902
	4	0.098	0.987	0.930	0.937	0.933
	5	0.092	0.980	0.932	0.840	0.884
	6	0.098	0.991	0.951	0.962	0.956
	7	0.104	0.978	0.972	0.810	0.883
	8	0.096	0.962	0.744	0.926	0.825
	9	0.099	0.977	0.849	0.940	0.891

A.2.2 Digit Recognition Domain

We consider a digit recognition domain in which training and testing distributions diverge because the data originate from different sources. To realize this scenario, a digit recognition model is trained on the *USPS* data set and evaluated on the *MNIST* data set. We use a version of *MNIST* prepared by Sam Roweis. *MNIST* images were rescaled from 28×28 to 16×16 pixels to match the resolution of *USPS* and the bounding box was recomputed. Images are represented by their 256 numeric pixel values. There are 10 classes, 11,000 training and 70,000 test instances. We train a single multi-class model using an RBF kernel.

We evaluate one-versus-rest F-measures for each class, resulting in ten different evaluation tasks. Table A.1 (bottom) lists the true one-vs-rest precision, $F_{0.5}$-measure, recall, and accuracy of the trained model on the ten different estimation problems corresponding to digits 0 to 9. Figures A.2 and A.3 show the estimation error of *active*, *passive*, and $active_{err}$ over number of labeled data, for precision, $F_{0.5}$ and recall estimates and the ten different one-vs-rest estimation problems corresponding to digits 0 to 9 in the digit recognition domain.

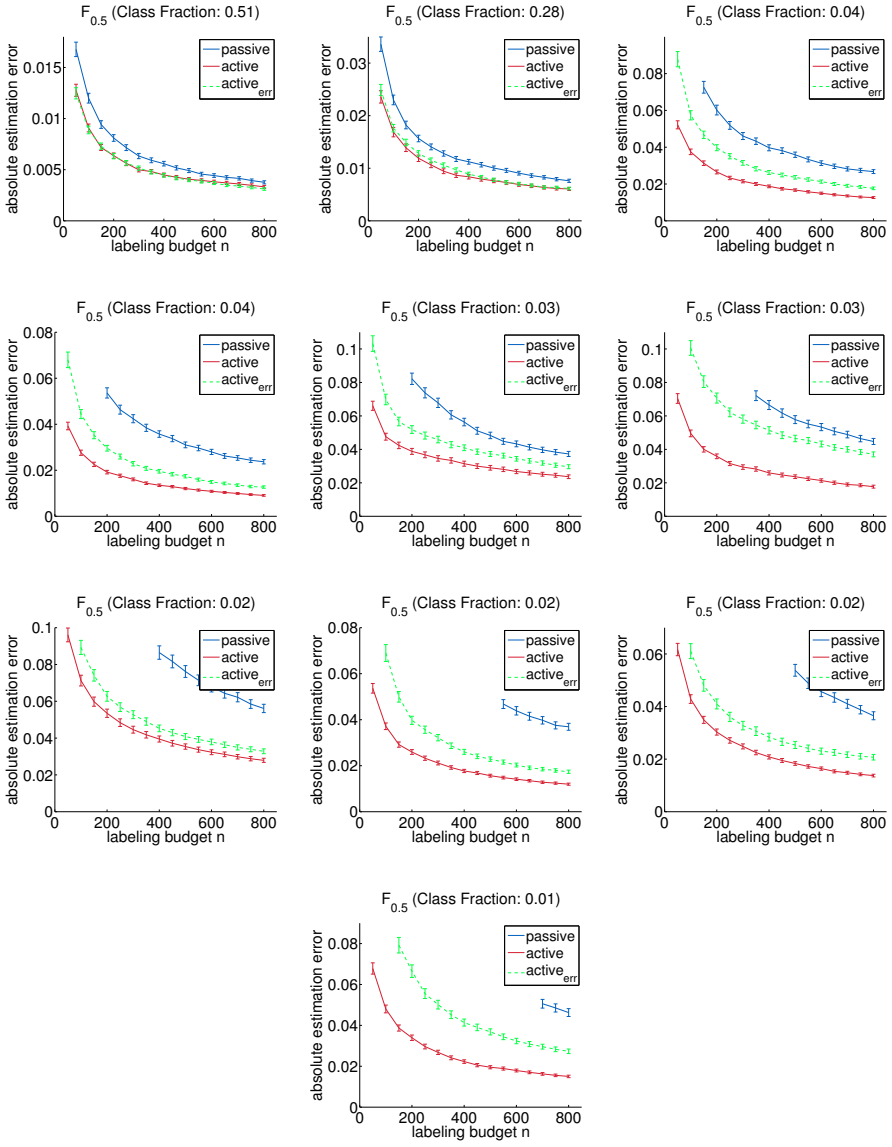


Figure A.1: Text Classification: Estimation error over number of labeled data, for recall, $F_{0.5}$ and precision estimates. Classes (from left to right and top to bottom): "earn", "acq", "crude", "trade", "money-fx", "interest", "ship", "sugar", "coffee" and "gold". Error bars indicate the standard error.

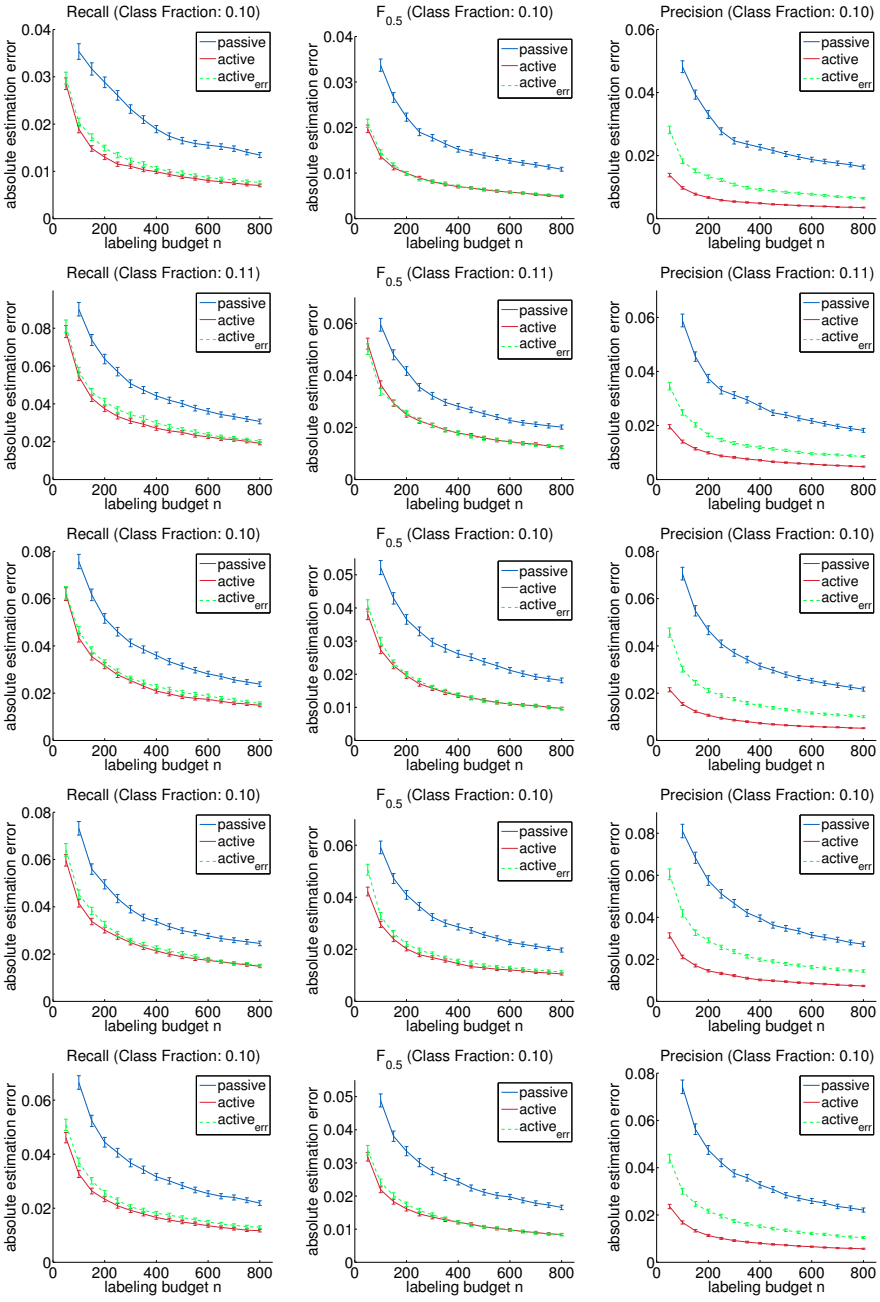


Figure A.2: Digit Recognition: Estimation error over number of labeled data for recall, $F_{0.5}$, and precision estimates for digits 0 to 4 (from top to bottom). Error bars indicate the standard error.

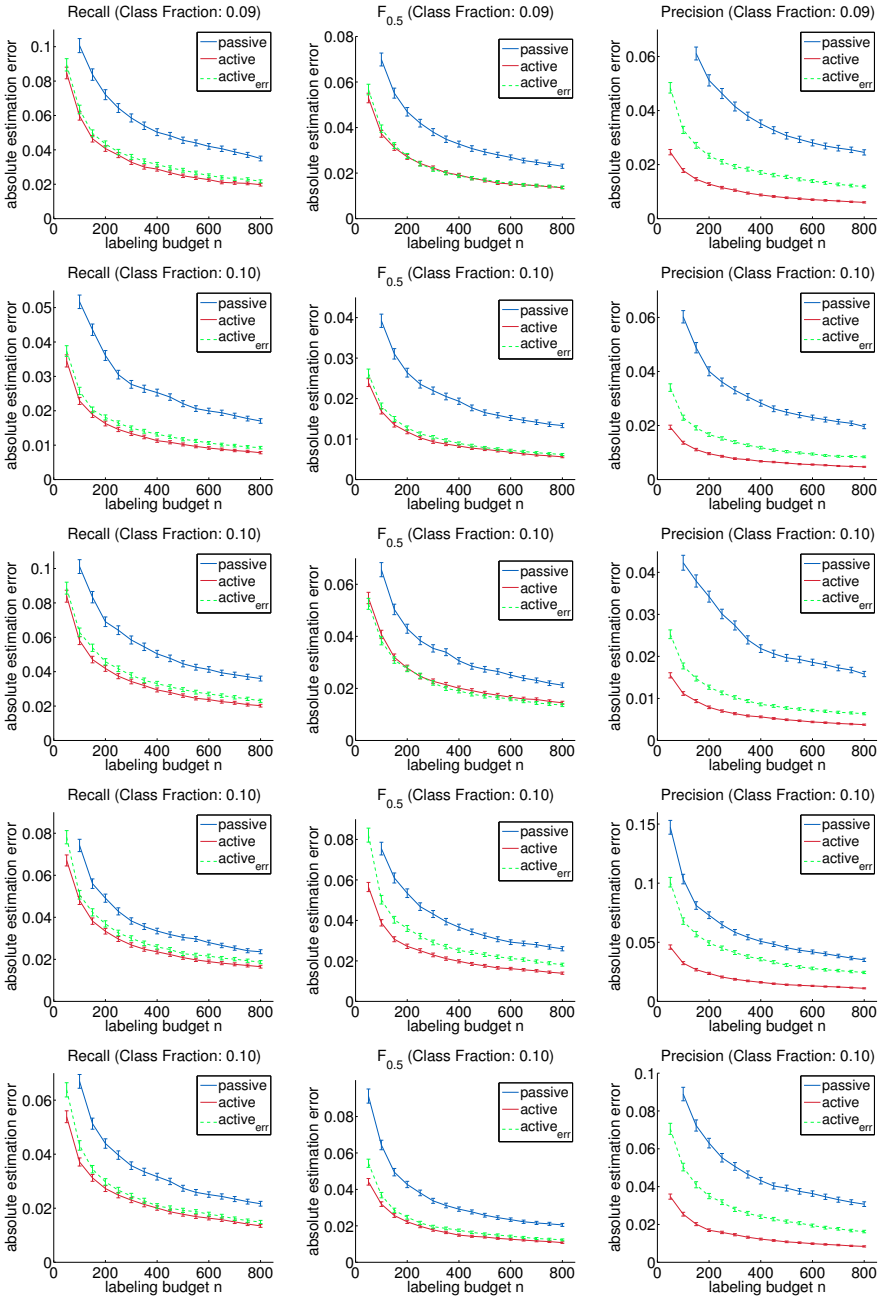


Figure A.3: Digit Recognition: Estimation error over number of labeled data for recall, $F_{0.5}$, and precision estimates for digits 5 to 9 (from top to bottom). Error bars indicate the standard error.

Notation

This section summarizes the notation and symbols used throughout the thesis. In general, a vector $\mathbf{x} = (x_i)_{i=1,\ldots,n}$ with components $x_1, \ldots, x_n$ is denoted by a lower case bold letter. If the components are the elements of a finite set X, we also write $x = (x)_{x \in X}$. Upper case bold letters such as $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\mathsf{T} = (x_{ij})_{i,j=1,\ldots,n}$ refer to matrices; the element in the i-th row and j-th column is denoted by x_{ij}. Greek letters are used to emphasize scalar values, vectors, and matrices as parameters of a model or a distribution. The notations $x \sim p(x)$ and $x \sim p$ denote that a random variable x has density $p(x)$. Estimates of a quantity are denoted by a hat such as $\hat{R}$. If the estimate is based on a model, we highlight this intrinsic approximation by a check such as $\check{R}$. The following list gives an overview over frequently used terms.

Model and Parameters

x	Observed data instance, page 7.
$\mathbf{x}$	Numerical Euclidean vector representation of instance x, page 11.
$\phi(\mathbf{x})$	Feature mapping (or sufficient statistic) of a vector $\mathbf{x}$, page 11.
$k(x, x')$	Kernel function, *i.e.,* inner product of instances x and x' in a Hilbert space induced by a mapping ϕ, page 15.
$\mathcal{X}$	Instance space such as $\mathbb{R}^d$, page 7.
y	Target label, page 7.
$\mathcal{Y}$	Label space. Labels can be categorical (classification problem), continuous (regression problem), or complex (structured prediction problem), page 7.
$\mathcal{Z}$	Space of items which can be retrieved by a ranking function, page 112.
T_n	Training set of n instance-label pairs (x_i, y_i) which is used to infer model parameters, page 7.
D_m	Pool of m unlabeled instances that are drawn independently from the distribution $p(x)$, page 20.

$\boldsymbol{\theta}$	Vector of model parameters. The data are assumed to be generated by a model with unknown parameters (denoted $\boldsymbol{\theta}^*$), page 7.
Θ	Model space. Set of all possible parameter vectors such as $\mathbb{R}^e$, page 7.
$f_{\boldsymbol{\theta}}$	Predictive model parameterized with a model parameter $\boldsymbol{\theta}$. Maps an instance to a label, page 10.
$\lambda(x)$	Instance-specific labeling costs, *i.e.,* the costs resulting from labeling an instance x with the true target label, page 68.
Λ	Available budget to label instances, page 68.

Probability Notation

$p(x, y)$	Probability density function of test data, page 7.
$q(x, y)$	Probability density function of an instrumental distribution used to highlight instance-label pairs, page 31.
$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$	Probability density function of the multivariate Gaussian (or normal) distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, page 12.
$\mathcal{N}_f(x \mid \mu, \sigma^2)$	Probability density function of the folded normal distribution with location parameter μ and scale parameter σ^2, page 92.
$U(x \mid a, b)$	Probability density function of the (continuous) uniform distribution with boundaries a and b, page 91.
$\chi(x \mid \nu)$	Probability density function of the χ^2-distribution with ν degrees of freedom, page 44.
$\Phi(x)$	Cumulative distribution function $\Phi : \mathbb{R} \to [0, 1]$ of the standard normal distribution, page 36.
$F_\nu(x)$	Cumulative distribution function $F_\nu : \mathbb{R} \to [0, 1]$ of Student's t-distribution with ν degrees of freedom, page 44.
$\Omega_\nu(x)$	Cumulative distribution function $\Omega_\nu : \mathbb{R} \to [0, 1]$ of the studentized range distribution with ν degrees of freedom, page 47.

$\mathrm{KL}[p \parallel q]$	Kullback-Leibler divergence. It measures the difference between two distributions p and q, page 9.
$\mathbb{E}_{x\sim p(x)}[u(x)]$	Expectation of a real-valued function $u(x)$ w.r.t. the underlying distribution $p(x)$ of the argument x, page 9.
$\mathrm{Var}_{x\sim p(x)}[u(x)]$	Variance of a real-valued function $u(x)$ w.r.t. the underlying distribution $p(x)$ of the argument x, page 29.
$\mathrm{Cov}_{x\sim p(x)}[u(x), v(x)]$	Covariance of the real-valued functions $u(x)$ and $v(x)$ w.r.t. the underlying distribution $p(x)$ of their argument x, page 52.

Risks and Estimators

$\mathcal{L}[f_{\boldsymbol{\theta}}]$	Theoretical label likelihood of model $f_{\boldsymbol{\theta}}$, page 9.
$\ell(y, \bar{y})$	Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ measures the disagreement between two labels, *e.g.,* the prediction $\bar{y} = f_{\boldsymbol{\theta}}(x)$ and the true label y, page 28.
$R[f_{\boldsymbol{\theta}}]$	Risk of model $f_{\boldsymbol{\theta}}$, *i.e.,* the expected value of a loss function ℓ w.r.t the test distribution $p(x, y)$, page 28.
$G[f_{\boldsymbol{\theta}}]$	Generalized risk of model $f_{\boldsymbol{\theta}}$. It is parameterized with a loss function ℓ and a further function w that assigns a weight $w(x, y, f_{\boldsymbol{\theta}})$ to each instance x, page 48.
$F_\eta[f_{\boldsymbol{\theta}}]$	F-measure of classifier $f_{\boldsymbol{\theta}}$, page 48.
$\delta(x, y)$	Difference of the loss between two models $f_{\boldsymbol{\theta}_1}$ and $f_{\boldsymbol{\theta}_2}$ for a data point (x, y), page 94.
Δ	Risk difference between two models $f_{\boldsymbol{\theta}_1}$ and $f_{\boldsymbol{\theta}_2}$, page 41.
α	Type I error of a statistical test or of a confidence interval, page 34.
$\beta_{\alpha,q}$	Type II error of a statistical test, page 44.
$\mathfrak{p}_q$	p-value of a statistical test, page 43.
$\mathrm{MSE}_{x\sim p(x)}\left[\hat{R}\right]$	Mean squared error of estimate $\hat{R}$ w.r.t. the distribution $p(x)$. It is a measure for the estimation error of $\hat{R}$, page 28.

$\mathrm{Bias}_{x\sim p(x)}\left[\hat{R}\right]$	Bias of estimate $\hat{R}$ w.r.t. the distribution $p(x)$. It quantifies the systematic deviation of $\hat{R}$ from the value being estimated, page 29.
$\mathrm{Var}_{x\sim p(x)}\left[\hat{R}\right]$	Variance of estimate $\hat{R}$ w.r.t. the distribution $p(x)$. It quantifies the amount of variation of $\hat{R}$, page 29.

Miscellaneous

$\lvert S\rvert$	Cardinality of a finite set S, page 63.
$[\![\psi]\!]$	Indicator function. It returns 1 if statement ψ is satisfied and 0 otherwise, page 11.
$\mathbf{u}\otimes\mathbf{v}$	Kronecker product. It multiplies each component of $\mathbf{u}$ by each component of $\mathbf{v}$, page 12.
$\mathrm{vec}(\mathbf{U})$	Operator that stacks the column vectors of a matrix $\mathbf{U}$ below another, page 12.
$\lvert\mathbf{X}\rvert$	Determinant of a matrix $\mathbf{X}$, page 12.
$\mathcal{O}(g(n))$	Landau notation. A function $f(n)$ is a member of the class $\mathcal{O}(g(n))$, if the quotient $\lvert f(x)/g(x)\rvert$ is bounded as x goes to infinity, page 15.

Bibliography

Abe, N. and Mamitsuka, H. Query learning strategies using boosting and bagging. In *Proceedings of the 15th International Conference on Machine Learning*, 1998.

Abramowitz, M. and Stegun, I. *Handbook of Mathematical Functions*. Dover Publications, 1964.

Agresti, A. and Coull, B.A. Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, pp. 119–126, 1998.

Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

Aslam, J., Pavlu, V., and Yilmaz, E. A statistical method for system evaluation using incomplete judgments. In *Proceedings of the 29th SIGIR Conference on Research and Development on Information Retrieval*, 2006.

Bach, F.R. Active learning for misspecified generalized linear models. In *Proceedings of the 19th Annual Conference on Neural Information Processing Systems*, 2006.

Balcan, M., Beygelzimer, A., and Langford, J. Agnostic active learning. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

Banerjee, A. An analysis of logistic models: exponential family connections and online performance. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, 2007.

Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. Surf: speeded up robust features. *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

Bennett, P.N. and Carvalho, V.R. Online stratified sampling: evaluating classifiers at web-scale. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management*, 2010.

Beygelzimer, A., Dasgupta, S., and Langford, J. Importance weighted active learning. In *Proceedings of the 26th International Conference on Machine Learning*, 2009.

Bishop, C.M. *Pattern Recognition and Machine Learning*, volume 4. Springer New York, 2006.

Breiman, L. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Brown, L.D., Cai, T.T., and DasGupta, A. Interval estimation for a binomial proportion. *Statistical Science*, 16(2):101–117, 2001.

Brown, L.D., Cai, T.T., and DasGupta, A. Confidence intervals for a binomial proportion and asymptotic expansions. *The Annals of Statistics*, 30(1):160–201, 2002.

Burges, C. Ranknet to lambdarank to lambdamart: an overview. Technical Report MSR-TR-2010-82, Microsoft Research, 2010.

Canny, J. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.

Cappé, O., Moulines, E., and Rydén, T. *Inference in Hidden Markov Models*. Springer, 2005.

Carterette, B. and Smucker, M. Hypothesis testing with incomplete relevance judgments. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, 2007.

Carterette, B., Allan, J., and Sitaraman, R. Minimal test collections for retrieval evaluation. In *Proceedings of the 29th SIGIR Conference on Research and Development in Information Retrieval*, 2006.

Castro, R.M. and Nowak, R.D. Minimax bounds for active learning. In *Proceedings of the 20th Annual Conference on Learning Theory*, 2007.

Chapelle, O. Active learning for parzen window classifier. In *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, 2005.

Chapelle, O. and Chang, Y. Yahoo! learning to rank challenge overview. *Journal of Machine Learning Research, Workshop and Conference Proceedings*, 14:1–24, 2011.

Chapelle, O., Metzler, D., Zhang, Y., and Grinspan, P. Expected reciprocal rank for graded relevance. In *Proceeding of the 18th ACM Conference on Information and Knowledge Management*, 2009.

Clopper, C.J. and Pearson, E.S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404, 1934.

Cohn, D.A. Neural network exploration using optimal experiment design. *Neural Networks*, 9(6):1071–1083, 1996.

Cossock, D. and Zhang, T. Statistical analysis of Bayes optimal subset ranking. *IEEE Transactions on Information Theory*, 54(11):5140–5154, 2008.

Cramér, H. *Mathematical Methods of Statistics*. Princeton University Press, 1946.

Dagan, I. and Engelson, S.P. Committee-based sampling for training probabilistic classifiers. In *Proceedings of the 12th International Conference on Machine Learning*, 1995.

Dasgupta, S. Coarse sample complexity bounds for active learning. In *Proceedings of the 19th Annual Conference on Neural Information Processing Systems*, 2006.

Dasgupta, S. and Hsu, D. Hierarchical sampling for active learning. In *Proceedings of the 25th International Conference on Machine Learning*, 2008.

David, I.P. and Sukhatme, B.V. On the bias and mean square error of the ratio estimator. *Journal of the American Statistical Association*, 69(346):464–466, 1974.

Davidson, I. and Fan, W. When efficient model averaging out-performs boosting and bagging. *Knowledge Discovery in Databases*, pp. 478–486, 2006.

Demšar, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Domingos, P. Bayesian averaging of classifiers and the overfitting problem. In *Proceedings of the 17th International Conference on Machine Learning*, 2000.

Donmez, P., Carbonell, J., and Bennett, P. Dual strategy active learning. *Machine Learning: ECML 2007*, pp. 116–127, 2007.

Druck, G. and McCallum, A. Toward interactive training and evaluation. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management*, 2011.

Fedorov, V.V. Theory of optimal experiments. 1972.

Flueck, J.A. and Holland, B.S. Ratio estimators and some inherent problems in their utilization. *Journal of Applied Meteorology*, 15:535–543, 1976.

Förstner, W. and Gülch, E. A fast operator for detection and precise location of distinct points, corners and centers of circular features. In *Proceedings of the Intercommission Workshop on Fast Processing of Photogrammetric Data*, 1987.

Frank, A. and Asuncion, A. Uci machine learning repository. Technical report, University of California, Irvine, School of Information and Computer Sciences, 2010.

Geman, S., Bienenstock, E., and Doursat, R. Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58, 1992.

Geweke, J. Bayesian inference in econometric models using monte carlo integration. *Econometrica*, 57(6):1317–1339, 1989.

Haertel, R., Seppi, K.D., Ringger, E.K., and Carroll, J.L. Return on investment for active learning. In *Proceedings of the NIPS Workshop on Cost-Sensitive Learning*, 2008.

Hammersley, J.M. and Handscomb, D.C. *Monte Carlo Methods*. Taylor & Francis, 1964.

Hanneke, S. Rates of convergence in active learning. *The Annals of Statistics*, 39(1):333–361, 2011.

Harris, C. and Stephens, M. A combined corner and edge detector. In *Proceedings of the 4th Alvey Vision Conference*, 1988.

Henderson, M. and Meyer, M.C. Exploring the confidence interval for a binomial parameter in a first course in statistical computing. *The American Statistician*, 55(4):337–344, 2001.

Herbrich, R., Graepel, T., and Obermayer, K. Large margin rank boundaries for ordinal regression. *Advances in Large-Margin Classifiers*, pp. 115–132, 2000.

Hull, J.J. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994.

Järvelin, K. and Kekäläinen, J. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

Jordan, M.I. Why the logistic function? a tutorial discussion on probabilities and neural networks. Technical report, Computational Cognitive Science, MIT, Cambridge, MA, 1995.

Kanamori, T. and Shimodaira, H. Active learning algorithm using the maximum weighted log-likelihood estimator. *Journal of Statistical Planning and Inference*, 116:149–162, 2003.

Kapoor, A., Grauman, K., Urtasun, R., and Darrell, T. Active learning with gaussian processes for object categorization. In *Proceedings of the 11st International Conference on Computer Vision*, 2007.

Karsmakers, P., Pelckmans, K., and Suykens, J.A.K. Multi-class kernel logistic regression: a fixed-size implementation. In *International Joint Conference on Neural Networks*, 2007.

Kimeldorf, G. and Wahba, G. Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95, 1971.

Lee, J.J. Libpmk: a pyramid match toolkit. Technical Report MIT-CSAIL-TR-2008-17, MIT Computer Science and Artificial Intelligence Laboratory, 2008.

Lewis, D.D. and Gale, W.A. A sequential algorithm for training text classifiers. In *Proceedings of the 17th SIGIR Conference on Research and Development in Information Retrieval*, 1994.

Li, P., Burges, C., and Wu, Q. Mcrank: Learning to rank using multiple classification and gradient boosting. In *Proceedings of the 20th Annual Conference on Neural Information Processing Systems*, 2007.

Liu, J.S. *Monte Carlo Strategies in Scientific Computing*. Springer, 2001.

Long, B., Chapelle, O., Zhang, Y., Chang, Y., Zheng, Z., and Tseng, B. Active learning for ranking through expected loss optimization. In *Proceedings of the 33rd SIGIR Conference on Research and Development in Information Retrieval*, 2010.

Lowe, D. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

MacKay, D.J.C. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

Madani, O., Lizotte, D.J., and Greiner, R. Active model selection. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, 2004.

Mardia, K.V. and Marshall, R.J. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71(1):135–146, 1984.

Maron, O. and Moore, A.W. Hoeffding races: Accelerating model selection search for classification and function approximation. In *Proceedings of the 6th Annual Conference on Neural Information Processing Systems*, 1993.

McCallum, A. and Nigam, K. Employing em in pool-based active learning for text classification. In *Proceedings of the 15th International Conference on Machine Learning*, 1998.

McCullagh, P. Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):109–142, 1980.

McCullagh, P. and Nelder, J.A. *Generalized Linear Models*. Chapman & Hall/CRC, 1989.

Microsoft Research. Microsoft learning to rank datasets. http://research.microsoft.com/en-us/projects/mslr/, 2010. Data sets released on June 16, 2010.

Mohan, A., Chen, Z., and Weinberger, K. Web-search ranking with initialized gradient boosted regression trees. *Journal of Machine Learning Research, Workshop and Conference Proceedings*, 14:77–89, 2011.

Nguyen, H.T. and Smeulders, A. Active learning using pre-clustering. In *Proceedings of the 21st International Conference on Machine Learning*, 2004.

O'Hagan, A. Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–42, 1978.

Osugi, T., Kun, D., and Scott, S. Balancing exploration and exploitation: A new algorithm for active machine learning. 2005. ISSN 1550-4786.

Pandey, G., Gupta, H., and Mitra, P. Stochastic scheduling of active support vector learning algorithms. In *Proceedings of the 20th Annual Symposium on Applied Computing*, 2005.

Rasmussen, C.E. and Williams, C. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Roy, N. and McCallum, A. Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of the 18th International Conference on Machine Learning*, 2001.

Sackrowitz, H. and Samuel-Cahn, E. P values as random variables-expected p values. *The American Statistician*, pp. 326–331, 1999.

Salzberg, S.L. On comparing classifiers: pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1(3):317–328, 1997.

Sawade, C., Landwehr, N., Bickel, S., and Scheffer, T. Active risk estimation. In *Proceedings of the 27th International Conference on Machine Learning*, 2010a.

Sawade, C., Landwehr, N., and Scheffer, T. Active estimation of f-measures. In *Proceedings of the 24th Annual Conference on Neural Information Processing Systems*, 2010b.

Sawade, C., Bickel, S., von Oertzen, T., Scheffer, T., and Landwehr, N. Active evaluation of ranking functions based on graded relevance. In *Proceedings of the 22nd European Conference on Machine Learning*, 2012a.

Sawade, C., Landwehr, N., and Scheffer, T. Active comparison of prediction models. In *Proceedings of the 26th Annual Conference on Neural Information Processing Systems*, 2012b.

Scheffer, T. and Wrobel, S. Finding the most interesting patterns in a database quickly by using sequential sampling. *Journal of Machine Learning Research*, 3:833–862, 2003.

Scheffer, T., Decomain, C., and Wrobel, S. Active hidden markov models for information extraction. *Advances in Intelligent Data Analysis*, pp. 309–318, 2001.

Schohn, G. and Cohn, D. Less is more: active learning with support vector machines. In *Proceedings of the 17th International Conference on Machine Learning*, 2000.

Schölkopf, B. and Smola, A.J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, volume 64. MIT Press, Cambridge, MA, USA, 2002.

Schölkopf, B., Herbrich, R., and Smola, A. A generalized representer theorem. In *Proceedings of the 14th Annual Conference on Computational Learning Theory*, 2001.

Schütze, H., Velipasaoglu, E., and Pedersen, J.O. Performance thresholding in practical text classification. In *Proceedings of the 15th ACM Conference on Information and Knowledge Management*, 2006.

Schwarz, G. Estimating the dimension of a model. *The Annals of Statistics*, 6 (2):461–464, 1978.

Settles, B. Active learning literature survey. Technical report, University of Wisconsin–Madison, 2009.

Settles, B., Craven, M., and Friedland, L. Active learning with real annotation costs. In *Proceedings of the NIPS Workshop on Cost-Sensitive Learning*, 2008.

Seung, H.S., Opper, M., and Sompolinsky, H. Query by committee. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, 1992.

Sheskin, D. *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall, 2004. ISBN 1584884401.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90: 227–244, 2000.

Singh, R., Palmer, N., Gifford, D., Berger, B., and Bar-Joseph, Z. Active learning for sampling in time-series experiments with application to gene expression analysis. In *Proceedings of the 22nd International Conference on Machine Learning*, 2005.

Spall, J.C. *Introduction to Stochastic Search and Optimization*, volume 64. Wiley and Sons, 2003.

Sriperumbudur, B.K., Fukumizu, K., Gretton, A., Lanckriet, G.R.G., and Schölkopf, B. Kernel choice and classifiability for rkhs embeddings of probability distributions. In *Proceedings of the 22nd Annual Conference on Neural Information Processing Systems*, 2009.

Stein, M.L. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, 1999.

Sugiyama, M. Active learning in approximately linear regression based on conditional expectation of generalization error. *Journal of Machine Learning Research*, 7:141–166, 2006.

Tikhonov, A.N. and Arsenin, V.I.A. *Solutions of Ill-Posed Problems*. V. H. Winston and Sons, 1977.

Tong, S. and Koller, D. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2, 2002.

Van der Vaart, A.W. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

van Rijsbergen, C. *Information Retrieval*. Butterworths, 2nd edition, 1979.

Vijayakumar, S., D'souza, A., and Schaal, S. Incremental online learning in high dimensions. *Neural Computation*, 17:2602–2634, 2005.

Wasserman, L. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2004.

Weiss, S.M. and Kulikowski, C.A. *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. Morgan Kaufmann, 1990.

Wiens, D.P. Robust weights and designs for biased regression models: least squares and generalized m-estimation. *Journal of Statistical Planning and Inference*, 83(2):395–412, 2000.

Wilson, E.B. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Zheng, Z., Zha, H., Zhang, T., Chapelle, O., Chen, K., and Sun, G. A general boosting method and its application to learning ranking functions for web search. In *Proceedings of the 20th Annual Conference on Neural Information Processing Systems*.

The field of machine learning studies algorithms that infer predictive models from data. Predictive models are applicable for, e.g., spam filtering, face and handwritten digit recognition, and personalized product recommendation. To estimate a model's performance, a set of labeled test instances is required that is sampled from the same distribution to which the model will be applied. In many practical scenarios, unlabeled test instances are readily available, but the process of labeling them is a time- and cost-intensive task that usually involves human experts.

This thesis addresses the problem of accurately evaluating and comparing given predictive models with minimal labeling effort. We study *active model evaluation processes* that select, according to an instrumental sampling distribution, instances of the data to be labeled. We derive optimal sampling distributions that minimize estimation error with respect to several performance measures. For the related problem of efficiently comparing the performance of predictive models, we devise an active comparison method that maximizes the likelihood of identifying the superior model.

Empirically, we investigate model evaluation and comparison problems in several domains and show under which conditions the active evaluation processes are more accurate than standard estimates given equally many test instances.

ISBN 978-3-86956-255-1