Active Risk Estimation

Christoph Sawade[⋆] SAWADE@CS.UNI-POTSDAM.DE
Niels Landwehr[⋆] LANDWEHR@CS.UNI-POTSDAM.DE
Steffen Bickel[†] STEFFEN.BICKEL@NOKIA.COM
Tobias Scheffer[⋆] SCHEFFER@CS.UNI-POTSDAM.DE
[⋆] University of Potsdam, Department of Computer Science, August-Bebel-Strasse 89, 14482 Potsdam, Germany
[†] Nokia gate5 GmbH, Invalidenstrasse 117, 10115 Berlin, Germany

Abstract

We address the problem of evaluating the risk of a given model accurately at minimal labeling costs. This problem occurs in situations in which risk estimates cannot be obtained from held-out training data, because the training data are unavailable or do not reflect the desired test distribution. We study *active risk estimation* processes in which instances are actively selected by a sampling process from a pool of unlabeled test instances and their labels are queried. We derive the sampling distribution that minimizes the estimation error of the active risk estimator when used to select instances from the pool. 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.

1. Introduction

In order to make an informed decision about the deployment of a predictive model, it is crucial to know the model's approximate risk. In practice, however, it is not always possible to estimate the risk on held-out training data. Consider the following three scenarios.

Firstly, when a readily trained model is shipped and deployed, the training data 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. The supplier may be able to communicate an honest risk estimate. However, this estimate may still be biased because the confidential training data need not necessarily reflect the input distribution which the deployed model will be exposed to.

Secondly, the input distribution may change over time. In this case, one may wish to monitor the risk of the model in order to determine at which point an update becomes necessary. 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.

Thirdly, consider that a model may result from an active learning mechanism. In order to minimize the labeling effort, active learners query the class labels that they predict least confidently. Hence, the selected data are a biased sample of the input distribution.

In these three scenarios, no accurate risk estimates are readily available. Estimates that are communicated from the model provider or result from hold-out evaluation on outdated or biased samples can be arbitrarily inaccurate. In order to estimate the risk accurately, new test instances have to be drawn and labeled.

In many application scenarios, unlabeled test instances are readily available whereas the process of labeling instances is costly. We study an *active risk estimation process* that, in analogy to active learning, selects instances from a pool of unlabeled test instances and queries their labels. Instances are selected according to an instrumental distribution q. 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—that is, asymptotically unbiased—estimate. We analyze sources of estimation error of the empirical risk, and derive the sampling distribution q^* that minimizes

Appearing in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.

the estimation error.

This paper is organized as follows. Section 2 outlines the problem setting. Section 3 infers the instrumental distribution that minimizes the variance of a consistent risk estimator. Section 3.1 specializes our principal result for zero-one and squared loss. Section 3.2 details confidence intervals for active risk estimators. In Section 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. Section 5 discusses related work, Section 6 concludes.

2. Problem Setting

Let $\mathcal{X}$ denote the feature space and $\mathcal{Y}$ the label space; an unknown test distribution $p(\mathbf{x}, y)$ is defined over $\mathcal{X} \times \mathcal{Y}$. Let $p(y|\mathbf{x};\theta)$ be a given θ-parameterized model of $p(y|\mathbf{x})$ and let $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ with $f_\theta(\mathbf{x}) = \arg\max_y p(y|\mathbf{x};\theta)$ be the corresponding hypothesis.

We study the problem of estimating the risk

$$R = \iint \ell(f_\theta(\mathbf{x}), y) p(\mathbf{x}, y) dy d\mathbf{x} \tag{1}$$

of f_θ with respect to $p(\mathbf{x}, y)$. The loss function $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ measures the disagreement between prediction and the true label. For classification, the zero-one error $\ell_{0/1}$ is a widely-used choice; in this setting, the integral over $\mathcal{Y}$ reduces to a finite sum. For regression, quadratic loss ℓ_2 is a standard choice.

Since $p(\mathbf{x}, y)$ is unknown, the expected loss (Equation 1) is typically approximated by an empirical risk

$$\hat{R}_n = \frac{1}{n} \sum_{i=1}^{n} \ell(f_\theta(\mathbf{x}_i), y_i), \tag{2}$$

where n test instances $(\mathbf{x}_i, y_i)$ are drawn from $p(\mathbf{x}, y) = p(\mathbf{x})p(y|\mathbf{x})$. Drawing labels y for selected instances $\mathbf{x}$ according to $p(y|\mathbf{x})$ is a costly process that may involve a query to a human labeler.

Test instances $\mathbf{x}_i$ need not necessarily be drawn according to the distribution p. When instances $\mathbf{x}_i$ are drawn according to an instrumental distribution q, a risk estimator can be defined as

$$\hat{R}_{n,q} = \frac{1}{\sum_{i=1}^{n} \frac{p(\mathbf{x}_i)}{q(\mathbf{x}_i)}} \sum_{i=1}^{n} \frac{p(\mathbf{x}_i)}{q(\mathbf{x}_i)} \ell(f_\theta(\mathbf{x}_i), y_i), \tag{3}$$

where $(\mathbf{x}_i, y_i)$ are drawn from $q(\mathbf{x})p(y|\mathbf{x})$. Weighting factors $\frac{p(\mathbf{x}_i)}{q(\mathbf{x}_i)}$ compensate for the discrepancy between test and instrumental distributions, and the normalizer is the sum of weights. Because of the weighting factors, Equation 3 defines a consistent estimator (see *e.g.,* Liu, 2001, pg. 35). Consistency means asymptotical unbiasedness; that is, the expected value of the estimate $\hat{R}_{n,q}$ converges to the true risk R for $n \rightarrow \infty$. However, a precondition for $\hat{R}_{n,q}$ to be consistent is that $p(\mathbf{x}) > 0$ implies $q(\mathbf{x}) > 0$. Note that Equation 2 is a special case of Equation 3, using the instrumental distribution $q = p$.

In practice, Equation 3 cannot be evaluated because p is not known. Also, we might not be able to directly generate new test instances according to an arbitrary distribution q. Therefore, some of our results focus on the case in which a large pool D of m unlabeled test instances is available that has been drawn according to the test distribution p. Drawing instances from this pool then serves as an approximation of drawing from the test distribution; in this case, $p(\mathbf{x}) = \frac{1}{m}$ for all $\mathbf{x} \in D$ is the uniform distribution over the pool. Drawing test instances according to an instrumental distribution q is then implemented by sampling from the pool according to q. In this setting, Equation 3 simplifies to

$$\hat{R}_{n,q} = \frac{1}{\sum_{i=1}^{n} \frac{1}{q(\mathbf{x}_i)}} \sum_{i=1}^{n} \frac{1}{q(\mathbf{x}_i)} \ell(f_\theta(\mathbf{x}_i), y_i). \tag{4}$$

The estimate $\hat{R}_{n,q}$ given by Equations 3 and 4, respectively, depends on the selected instances $(\mathbf{x}_i, y_i)$, which are drawn according to the distribution $q(\mathbf{x})p(y|\mathbf{x})$. Thus, $\hat{R}_{n,q}$ is a random variable whose distribution depends on the distribution $q(\mathbf{x})p(y|\mathbf{x})$. Our goal is to find an instrumental distribution q such that the expected deviation from the true risk is minimal for fixed labeling costs n:

$$q^* = \arg\min_q \mathbb{E}\left[\left(\hat{R}_{n,q} - R\right)^2\right].$$

3. Active Risk Estimation

The bias-variance decomposition of Geman et al. (1992) constitutes the starting point of our investigation of the sources of estimation error:

$$\begin{aligned} & \mathbb{E}\left[(\hat{R}_{n,q} - R)^2\right] \\ &= \left(\mathbb{E}\left[\hat{R}_{n,q}\right] - R\right)^2 + \mathbb{E}\left[\left(\hat{R}_{n,q} - \mathbb{E}\left[\hat{R}_{n,q}\right]\right)^2\right] \\ &= \mathrm{Bias}^2[\hat{R}_{n,q}] + \mathrm{Var}[\hat{R}_{n,q}]. \end{aligned} \tag{5}$$

Equation 5 expresses the estimation error as a sum of a squared bias and a variance term. Because $\hat{R}_{n,q}$ is consistent, $\mathbb{E}[(\hat{R}_{n,q}-R)^2]$ vanishes for $n\rightarrow\infty$. Furthermore, Liu (2001, pg. 35) shows that $\text{Bias}^2[\hat{R}_{n,q}]$ is of order $\frac{1}{n^2}$; that is, there are $c_1>0$, $c_2>0$, and n_0 such that for all $n\geq n_0$

$$\frac{c_1}{n^2}\leq \text{Bias}^2[\hat{R}_{n,q}]\leq\frac{c_2}{n^2}. \tag{6}$$

Lemma 1 states that the active risk estimator $\hat{R}_{n,q}$ is asymptotically normally distributed, and characterizes its variance.

Lemma 1 (Asymptotic Variance of Estimator)**.** *Let $\hat{R}_{n,q}$ be as defined in Equation 3. Then,*

$$\sqrt{n}\left(\hat{R}_{n,q}-R\right)\overset{n\rightarrow\infty}{\longrightarrow}\mathcal{N}\left(0,\sigma_q^2\right) \tag{7}$$

with

$$\sigma_q^2=\int\frac{p(\mathbf{x})}{q(\mathbf{x})}\left(\int[\ell(f_\theta(\mathbf{x}),y)-R]^2p(y|\mathbf{x})dy\right)p(\mathbf{x})d\mathbf{x} \tag{8}$$

where $\overset{n\rightarrow\infty}{\longrightarrow}$ denotes convergence in distribution.

The proof can be found in the appendix. Taking the variance of both sides of Equation 7, we obtain

$$n\,\text{Var}\left[\hat{R}_{n,q}\right]\overset{n\rightarrow\infty}{\longrightarrow}\sigma_q^2. \tag{9}$$

Because $n\,\text{Var}[\hat{R}_{n,q}]$ converges to a constant, $\text{Var}[\hat{R}_{n,q}]$ 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 $\mathbb{E}[(\hat{R}_{n,q}-R)^2]$ will be dominated by the variance term $\text{Var}[\hat{R}_{n,q}]$. Dividing Equation 9 by n, we see that

$$\text{Var}\left[\hat{R}_{n,q}\right]\quad\approx\quad\frac{1}{n}\sigma_q^2 \tag{10}$$

for large n. In the following, we will use this approximation, and consequently derive a sampling distribution q^* that minimizes σ_q^2.

3.1. Optimal Sampling Distribution

The following theorem derives the sampling distribution that minimizes σ_q^2.

Theorem 1 (Optimal Sampling Distribution)**.** *The instrumental distribution that minimizes σ_q^2 is*

$$q^*(\mathbf{x})\propto p(\mathbf{x})\sqrt{\int[\ell(f_\theta(\mathbf{x}),y)-R]^2p(y|\mathbf{x})dy}. \tag{11}$$

The proof can be found in the appendix.

We will now focus on the pool-based setting in which $p(\mathbf{x})=\frac{1}{m}$ is the uniform distribution over a pool D. The active risk estimator samples n points according to a discrete instrumental distribution $q(\mathbf{x})$ and Equation 4 defines $\hat{R}_{n,q}$. Hence, σ_q^2 becomes a finite average over the pool and R a pool-based risk taken over the instances in D.

Note that the optimal sampling distribution q^* given by Equation 11 depends on the unknown risk R, and on the unknown true conditional $p(y|\mathbf{x})$. To compute q^* in practice, we approximate $p(y|\mathbf{x})$ by the given model $p(y|\mathbf{x};\theta)$, and R by the introspective risk

$$R_\theta=\frac{1}{m}\sum_{\mathbf{x}\in D}\int\ell(f_\theta(\mathbf{x}),y)p(y|\mathbf{x};\theta)dy$$

accordingly. This approximation constitutes an analogy to active learning. In active learning, the model-based output probability $p(y|\mathbf{x};\theta)$ serves as the basis on which the least confident instances are selected. Note that the introspective risk R_θ may approximate the true risk poorly, depending on the size of the training sample and other factors.

We will now derive the optimal sampling distribution for two standard loss functions.

Derivation 1 (Optimal Sampling for zero-one loss)**.** *If $p(y|\mathbf{x})$ is approximated by the model $p(y|\mathbf{x};\theta)$, the sampling distribution that minimizes σ_q^2 for the zero-one loss $\ell_{0/1}$ in a pool-based setting resolves to*

$$q^*(\mathbf{x})\propto\sqrt{(1-2R_\theta)(1-p(f_\theta(\mathbf{x})|\mathbf{x};\theta))+R_\theta^2} \tag{12}$$

for all $\mathbf{x}\in D$.

Proof. Rewriting the result of Theorem 1 for $p(\mathbf{x})=\frac{1}{m}$ in a classification setting, we obtain

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

Equation 12 constructs $q^*(\mathbf{x})$ such that it gives preference to instances whose loss has a high variance according to $p(y|\mathbf{x};\theta)$. In binary classification, $q^*(\mathbf{x})$ gives a higher likelihood of selection to instances that are close to the decision boundary. If $R_\theta=\frac{1}{2}$, active risk estimation degenerates to uniform sampling.

In the following, we derive the optimal sampling distribution for regression problems and a squared loss function. Derivation 2 assumes that the model uses a Gaussian predictive distribution $p(y|\mathbf{x};\theta)$:

Derivation 2 (Optimal Sampling for squared loss). *If $p(y|\mathbf{x})$ is approximated by the model $p(y|\mathbf{x};\theta)$ and the model is Gaussian with $p(y|\mathbf{x};\theta) = \mathcal{N}(f_\theta(\mathbf{x}), \sigma^2_{\mathbf{x}})$, then the sampling distribution that minimizes σ^2_q for the squared loss ℓ_2 in a pool-based setting resolves to*

$$q^*(\mathbf{x}) \propto \sqrt{(3\sigma^2_{\mathbf{x}} - 2R_\theta)\sigma^2_{\mathbf{x}} + R^2_\theta}$$

for all $\mathbf{x} \in D$.

Proof. Rewriting Equation 11 for $\ell = \ell_2$ yields

$$\begin{aligned} q^*(\mathbf{x}) &\propto \sqrt{\int \left((f_\theta(\mathbf{x}) - y)^2 - R_\theta\right)^2 p(y|\mathbf{x};\theta)dy} \\ &= \left(R^2_\theta + \int (f_\theta(\mathbf{x}) - y)^4 p(y|\mathbf{x};\theta)dy \right. \\ &\qquad \left. -2R_\theta \int (f_\theta(\mathbf{x}) - y)^2 p(y|\mathbf{x};\theta)dy \right)^{\frac{1}{2}} \\ &= \sqrt{3\sigma^4_{\mathbf{x}} - 2R_\theta\sigma^2_{\mathbf{x}} + R^2_\theta}. \qquad (13) \end{aligned}$$

Equation 13 exploits that the two integrals over $\mathcal{Y}$ are central moments of the Gaussian predictive distribution, because mean and mode coincide. □

Note that the variance $\sigma^2_{\mathbf{x}}$ of the predictive distribution at instance $\mathbf{x} \in D$ would typically be available from a probabilistic predictor such as a Gaussian process (Williams & Rasmussen, 1996).

Algorithm 1 summarizes the active risk estimation algorithm. It samples n instances with replacement from the pool according to the distribution prescribed by Derivations 1 (for zero-one loss) and 2 (for squared loss), respectively. 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 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.

3.2. Confidence Intervals

This section derives confidence intervals for active risk estimators. According to Lemma 1, the estimator $\hat{R}_{n,q}$ is asymptotically normally distributed. The asymptotic variance is given by

$$\sigma^2_q = \iint \left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right)^2 \left[\ell(f_\theta(\mathbf{x}), y) - R\right]^2 p(y|\mathbf{x})q(\mathbf{x})dyd\mathbf{x},$$

where we have reformulated Equation 8 by changing the probability measure from p to q. A consistent estimate of σ^2_q is obtained from the labeled sample $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ by computing empirical variance

$$S^2_n = \frac{1}{\sum_{i=1}^n \frac{p(\mathbf{x}_i)}{q(\mathbf{x}_i)}} \sum_{i=1}^n \left(\frac{p(\mathbf{x}_i)}{q(\mathbf{x}_i)}\right)^2 [\ell(f_\theta(\mathbf{x}_i), y_i) - \hat{R}_{n,q}]^2.$$

Algorithm 1 Active Risk Estimation

```
input Model parameters $\theta$, pool $D$, labeling costs $n$.
output Risk estimate $\hat{R}_{n,q^*}$.
Compute optimal sampling distribution $q^*$ according to Derivation 1 or 2, respectively.
for $i = 1, \ldots, n$ do
    Draw $\mathbf{x}_i \sim q^*(\mathbf{x})$ from $D$ with replacement.
    Query label $y_i \sim p(y|\mathbf{x}_i)$ from oracle.
end for
return $\frac{1}{\sum_{i=1}^n \frac{1}{q(\mathbf{x}_i)}} \sum_{i=1}^n \frac{1}{q(\mathbf{x}_i)} \ell(f_\theta(\mathbf{x}_i), y_i)$
```

A two-sided confidence interval $[\hat{R}_{n,q} - z, \hat{R}_{n,q} + z]$ with coverage $1 - \alpha$ is now given by

$$z = F^{-1}_n \left(1 - \frac{\alpha}{2}\right) \frac{S_n}{\sqrt{n}}$$

where F^{-1}_n is the inverse cumulative distribution function of the Student's t distribution. As in the standard case of drawing test instances $\mathbf{x}_i$ from the original distribution p, such confidence intervals are approximate for finite n, but become exact for $n \to \infty$.

4. Empirical Results

We study the empirical behavior of active risk estimation in comparison to risk estimation based on a sample drawn uniformly from the test distribution (passive evaluation). Our evaluation specifically addresses two scenarios: Section 4.1 focuses on evaluating models with respect to a test distribution that differs from the training distribution; Section 4.2 focuses on the evaluation of actively trained models. We conduct experiments in two application domains.

Spam Filtering Domain. In this domain, the distribution shifts over time and a classifier that has been trained in the past has to be evaluated with respect to the current distribution. We collected 84,330 emails from an email service provider between June 2007 and October 2008. We use emails received by February 2008 as training portion (referred to as EMAIL1) and the more recent emails as evaluation portion (EMAIL2). Emails are represented using a binary bag-of-words, resulting in 188,068 distinct features; approximately 75% of all emails are spam. We cannot use the standard Spam TREC benchmark data

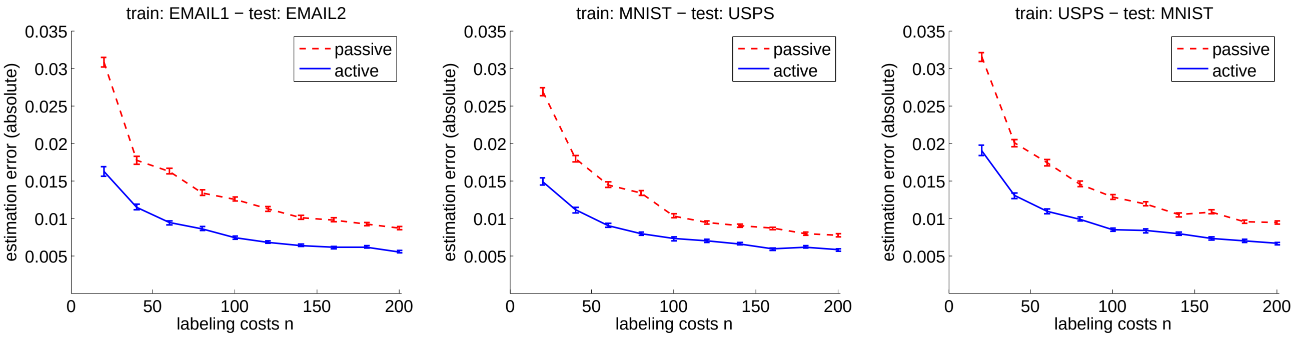

Figure 1. Absolute deviation from pool error over number of labeled data for spam filtering and digit recognition domain.

set because in this data set the original time stamps cannot be reconstructed, and messages cannot reliably be separated into old and new.

Digit Recognition Domain. This setting 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. We use the MNIST (in a version prepared by Sam Roweis) and USPS image recognition data sets; we use either data set as training and the other as test data. 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. We rescale the MNIST images from 28×28 to 16×16 to match the resolution of USPS and recompute the bounding box. The rescaled MNIST images differ visually from the USPS images, the line strokes are generally thicker.

For most experiments, we train a regularized logistic regression classifier that provides us with an estimate of $p(y|\mathbf{x};\theta)$. The regularization parameter is tuned a priori on the training portion of each data set by cross validation and then kept fixed. In each experiment, we train a model on the training data set and obtain an active risk estimate on the evaluation data set using Algorithm 1. As a baseline, we obtain a risk estimate using test instances drawn uniformly from the pool (passive evaluation). Both methods operate on identical labeling budgets n. The evaluation process is repeated $1,000$ times and results are averaged. In order to assess the estimation error we determine the risk of the model on the entire evaluation data set and use it as an approximation of the true risk.

4.1. Evaluation under Distribution Shift

This section studies whether—and under which conditions—active risk estimation can lead to more accurate estimates than risk estimation based on a uniformly drawn sample.

Figure 1 shows the average absolute deviation between the risk estimate and the true risk for active and passive risk estimation as a function of the labeling costs n for the EMAIL1-EMAIL2, MNIST-USPS, 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 EMAIL2, 0.0205 for USPS, and 0.0280 for MNIST. In all three 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 labeled test instances. For example, in the spam filtering domain, active evaluation with 70 test instances achieves approximately the same accuracy as passive evaluation with 200 instances.

Active evaluation relies on the model's estimate of the output probability in order to select uncertain instances from the pool. In order to study the relation between the quality of $p(y|\mathbf{x};\theta)$ and the benefit of active risk estimation, 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 $\frac{|\hat{R}_n - R|}{|\hat{R}_{n,q^*} - R|}$. A ratio of above one indicates that active evaluation incurs a smaller estimation error than passive learning. 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 2 (left) shows the results; the horizontal axis quantifies the quality of $p(y|\mathbf{x};\theta)$ in terms of the exponentiated average log-likelihood per test example which grows with the size of the training sample. For model likelihoods of 0.6

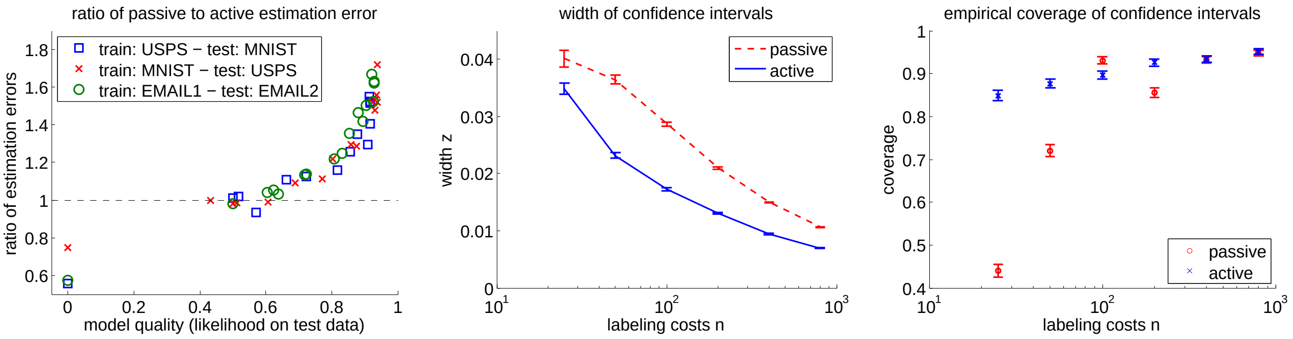

Figure 2. Ratio of estimation error of passive and active risk estimates (left). Width of confidence intervals for active and passive risk estimates (center). Empirical confidence levels for active and passive risk estimates (right).

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.

We have derived confidence intervals for active risk estimates in Section 3.2. Figure 2 (center) depicts their width z in comparison to confidence intervals of passive risk estimates, for the spam filtering domain and $\alpha = 0.05$. Intervals obtained from active risk estimation are significantly tighter than those of passive risk estimation. We also investigate how accurately the empirical coverage of the intervals matches the desired confidence level of $1 - \alpha$. Figure 2 (right) shows the fraction of iterations in which the true 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. 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.

4.2. 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 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 that is closest to the current decision boundary (Lewis & Gale, 1994) and updates the model. 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.

Table 1 shows the true risk (model error) and average absolute deviation of the estimated risk (estimation error) for strategies (1) to (3). EMAIL indicates the dataset consisting of all $84,330$ emails collected in the email domain. 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

Table 1. Active vs. passive learning and active vs. passive risk estimation. Values in parenthesis indicate standard errors.

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

and the most accurate risk estimate.

5. Related Work

Active risk estimation 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. In analogy to our approach, active learning algorithms use a current model to decide on instances whose class labels are queried. Specifically, Bach (2007) derives a sampling distribution under the assumption that the current model gives a good approximation to the conditional probability $p(y|\mathbf{x})$. 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, 2007), or SVMs (Beygelzimer et al., 2009).

Drawing instances from an instrumental distribution instead of the test distribution implies covariate shift. It is a standard approach to compensate for the covariate shift by reweighting instance-specific losses according to the density ratio of the original and auxiliary distribution (see, *e.g.,* Shimodaira, 2000).

Finally, 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 risk of prediction models. 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.

6. Conclusion

We have 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. Our analysis of the sources of estimation error has lead to an instrumental distribution q^* that, when used to select instances to be labeled from the pool, minimizes the variance term of the error. The active risk estimator is a consistent estimator of the true risk. It intuitively gives preference to uncertain test instances, using the model $p(y|\mathbf{x};\theta)$ to quantify this uncertainty. Active risk 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 risk evaluation outperforms passive evaluation when the model has a certain quality—an exponentiated per-instance log-likelihood of 0.6 or above—which in our experiments was the case with eight or more training examples. Active risk 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 risk estimates to be as accurate as passive estimates based on three times as many test examples. 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. We observe the confidence intervals of active risk estimates to be tighter and more reliable even for small test samples.

Appendix

Proof of Lemma 1

Let $\hat{R}^0_{n,q} = \sum_{i=1}^n w_i \ell_i$ and $W_n = \sum_{i=1}^n w_i$ with $w_i = \frac{p(\mathbf{x}_i)}{q(\mathbf{x}_i)}$ and $\ell_i = \ell(f_\theta(\mathbf{x}_i), y_i)$. We note that for examples drawn according to $q(\mathbf{x})$, $\mathbb{E}[\hat{R}^0_{n,q}] = nR$ and $\mathbb{E}[W_n] = n$. The random variables $w_1, \ldots, w_n$ and $w_1\ell_1, \ldots, w_n\ell_n$ are *iid*, therefore the central limit theorem implies that $\frac{1}{n}\hat{R}^0_{n,q}$ and $\frac{1}{n}W_n$ are asymptotically normally distributed with

$$\sqrt{n}\left(\frac{1}{n}\hat{R}^0_{n,q} - R\right) \overset{n\to\infty}{\longrightarrow} \mathcal{N}(0, \mathrm{Var}[w_i\ell_i]) \quad (14)$$

$$\sqrt{n}\left(\frac{1}{n}W_n - 1\right) \overset{n\to\infty}{\longrightarrow} \mathcal{N}(0, \mathrm{Var}[w_i]) \quad (15)$$

where $\overset{n\to\infty}{\longrightarrow}$ denotes convergence in distribution.

We employ the multivariate *delta method* (see *e.g.,* Wasserman, 2004, pg. 79) to extend the convergence results for $\hat{R}^0_{n,q}$ and W_n to a convergence result for the normalized estimator $\hat{R}_{n,q}$. The delta method allows to derive the asymptotic distribution of a differentiable function f whose input variables are asymptotically normally distributed. Applying it to the function $f(x,y) = \frac{x}{y}$ with $x = \frac{1}{n}\hat{R}^0_{n,q}$ and $y = \frac{1}{n}W_n$ yields

$$\sqrt{n}\left(\frac{\frac{1}{n}\hat{R}^0_{n,q}}{\frac{1}{n}W_n} - R\right) \overset{n\to\infty}{\longrightarrow} \mathcal{N}(0, \nabla f(R,1)^{\mathsf{T}}\Sigma\nabla f(R,1))$$

where ∇f denotes the gradient of f and Σ is the (asymptotic) covariance matrix of the input arguments

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

Furthermore,

$$\begin{aligned} &\nabla f(R,1)^{\mathsf{T}}\Sigma\nabla f(R,1) \\ &= \mathrm{Var}[w_i\ell_i] - 2R\,\mathrm{Cov}[w_i, w_i\ell_i] + R^2\,\mathrm{Var}[w_i] \\ &= \mathbb{E}[w_i^2\ell_i^2] - 2R\,\mathbb{E}[w_i^2\ell_i] + R^2\,\mathbb{E}[w_i^2] \\ &= \iint \left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right)^2 \left[\ell(f_\theta(\mathbf{x}), y) - R\right]^2 p(y|\mathbf{x})q(\mathbf{x})dyd\mathbf{x}. \end{aligned}$$

From this, the claim follows by canceling $q(\mathbf{x})$. □

Proof of Theorem 1

We minimize the functional σ_q^2 in terms of q under the constraint $\int q(\mathbf{x})d\mathbf{x} = 1$ using a Lagrange multiplier β.

$$\begin{aligned} \mathcal{L}[q,\beta] &= \sigma_q^2 + \beta\left(\int q(\mathbf{x})d\mathbf{x} - 1\right) \\ &= \int \underbrace{\frac{c(\mathbf{x})}{q(\mathbf{x})} + \beta\left(q(\mathbf{x}) - 1\right)}_{=G(q(x),x)} d\mathbf{x} \end{aligned}$$

where $c(\mathbf{x}) = p(\mathbf{x})^2 \int \left[\ell(f_\theta(\mathbf{x}), y) - R\right]^2 p(y|\mathbf{x})dy$.

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

$$\frac{\partial G}{\partial q(\mathbf{x})} = -\frac{c(\mathbf{x})}{q(\mathbf{x})^2} + \beta = 0 \qquad (16)$$

A solution for Equations 16 with respect to the normalization constraint is given by

$$q^*(\mathbf{x}) = \frac{\sqrt{c(\mathbf{x})}}{\int \sqrt{c(\mathbf{x})}d\mathbf{x}}. \qquad (17)$$

Note that we dismiss the negative solution, since $q(\mathbf{x})$ is a probability. Resubstitution of c in Equation 17 implies the theorem. □

Acknowledgments

We gratefully acknowledge that this work was supported by a Google Research Award. We wish to thank Michael Brückner for his help with the experiments on spam data.

References

Bach, F.R. Active learning for misspecified generalized linear models. In *Advances in Neural Information Processing Systems*, 2007.

Beygelzimer, A., Dasgupta, S., and Langford, J. Importance weighted active learning. In *Proceedings of the International Conference on Machine Learning*, 2009.

Cappé, O., Moulines, E., and Rydén, T. *Inference in hidden Markov models*. Springer, 2005.

Geman, S., Bienenstock, E., and Doursat, R. Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58, 1992.

Hammersley, J.M. and Handscomb, D.C. *Monte carlo methods*. Taylor & Francis, 1964.

Lewis, D.D. and Gale, W.A. A sequential algorithm for training text classifiers. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994.

Liu, J.S. *Monte carlo strategies in scientific computing*. Springer, 2001.

Schütze, H., Velipasaoglu, E., and Pedersen, J.O. Performance thresholding in practical text classification. In *Proceedings of the ACM Conference on Information and Knowledge Management*, 2006.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Statistical Planning and Inference*, 90:227–244, 2000.

Sugiyama, M. Active learning in approximately linear regression based on conditional expectation of generalization error. *Journal of Machine Learning Research*, 7:141–166, 2006.

Wasserman, L. *All of statistics: a concise course in statistical inference*. Springer, 2004.

Williams, C. K. I. and Rasmussen, C. E. Gaussian processes for regression. In *Advances in Neural Information Processing Systems*, 1996.