Active Comparison of Prediction Models (Online Appendix)

Christoph Sawade, Niels Landwehr, and Tobias Scheffer
University of Potsdam
Department of Computer Science
August-Bebel-Strasse 89, 14482 Potsdam, Germany
{sawade, landwehr, scheffer}@cs.uni-potsdam.de

A. Proofs

Proof of Lemma 2

Let $\alpha \in (0,1)$ denote a confidence threshold, and let $\Delta = R[f_1] - R[f_2] \neq 0$ denote the true risk difference. The quantity

$$\beta_{n,q} = 1 - \int_0^{z_\alpha} f\left(T; \frac{\sqrt{n}\Delta}{\sigma_{n,q}}, 1\right) \mathrm{d}T,$$

where

$$f\left(T; \frac{\sqrt{n}\Delta}{\sigma_{n,q}}, 1\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(T + \frac{\sqrt{n}\Delta}{\sigma_{n,q}}\right)^2\right) + \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(T - \frac{\sqrt{n}\Delta}{\sigma_{n,q}}\right)^2\right),$$

only depends on q through $\sigma_{n,q}$. For sufficiently large n, $\beta_{n,q}$ is a monotonically decreasing function of $\sigma_{n,q}$, because the partial derivative

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

is negative for large n.

Let q, q' denote two sampling distributions. Since $\sigma_{n,q} = \sqrt{n \operatorname{Var}[\hat{\Delta}_{n,q}]}$,

$$\lim_{n\to\infty} n \operatorname{Var}[\hat{\Delta}_{n,q}] < \lim_{n\to\infty} n \operatorname{Var}[\hat{\Delta}_{n,q'}] \tag{16}$$

holds if and only if $\sigma_{n,q} < \sigma_{n,q'}$ for sufficiently large n. Condition 16 is thus equivalent to $\beta_{n,q} > \beta_{n,q'}$ for sufficiently large n. □

Proof of Lemma 3

Let $\hat{\Delta}^0_{n,q} = \sum_{i=1}^n w_i \ell_i$ and $W_n = \sum_{i=1}^n w_i$ with $w_i = \frac{p(\mathbf{x}_i)}{q(\mathbf{x}_i)}$ and $\ell_i = \ell(f_1(\mathbf{x}_i), y_i) - \ell(f_2(\mathbf{x}_i), y_i)$. We note that for examples drawn according to $q(\mathbf{x})$, $\mathbb{E}[\hat{\Delta}^0_{n,q}] = n\Delta$ 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 $\frac{1}{n}\hat{\Delta}^0_{n,q}$ and $\frac{1}{n}W_n$ are asymptotically normally distributed with

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

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

We employ the multivariate *delta method* (see, *e.g.,* [10], Chapter 5) to extend the convergence results for $\hat{\Delta}^0_{n,q}$ and W_n to a convergence result for the normalized estimator $\hat{\Delta}_{n,q}$. 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 $x = \frac{1}{n}\hat{\Delta}^0_{n,q}$ and $y = \frac{1}{n}W_n$ yields

$$\sqrt{n}\left(\begin{array}{c}\frac{1}{n}\hat{\Delta}^0_{n,q}\\ \frac{1}{n}W_n\end{array} - \Delta\right) \overset{n\to\infty}{\longrightarrow} \mathcal{N}(0, \nabla f(\Delta,1)^\mathsf{T}\Sigma\nabla f(\Delta,1)), \tag{17}$$

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

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

Taking the variance of both sides of Equation 17, we observe that

$$\lim_{n\to\infty} n \operatorname{Var}\left[\hat{\Delta}_{n,q}\right] = \nabla f(\Delta,1)^\mathsf{T}\Sigma\nabla f(\Delta,1).$$

Finally,

$$\begin{aligned}&\nabla f(\Delta,1)^\mathsf{T}\Sigma\nabla f(\Delta,1)\\ =&\mathbb{E}\left[w_i\right]^{-2}\left(\operatorname{Var}[w_i\ell_i] - 2\Delta\operatorname{Cov}[w_i, w_i\ell_i] + \Delta^2\operatorname{Var}[w_i]\right)\\ =&\mathbb{E}\left[w_i^2\ell_i^2\right] - 2\Delta\mathbb{E}\left[w_i^2\ell_i\right] + \Delta^2\mathbb{E}\left[w_i^2\right]\\ =&\iint \frac{p(\mathbf{x})^2}{q(\mathbf{x})^2}\left(\ell(f_1(\mathbf{x}), y - \ell(f_2(\mathbf{x}), y)) - \Delta\right)^2 p(y|\mathbf{x})q(\mathbf{x})\mathrm{d}y\,\mathrm{d}\mathbf{x}.\end{aligned}$$

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

Derivation 4

Rewriting the result of Theorem 2 for $p(\mathbf{x}) = \frac{1}{m}$ in a classification setting, we obtain

$$\begin{aligned}q^*(\mathbf{x}) &\propto \sqrt{\sum_{y\in\mathcal{Y}}\left(\ell(f_1(\mathbf{x}), y) - \ell(f_2(\mathbf{x}), y) - \Delta_\theta\right)^2 p(y|\mathbf{x};\theta)}\\ &= \sqrt{\left(f_1(\mathbf{x}) - f_2(\mathbf{x})\right)^2 - 2\Delta_\theta\left(f_1(\mathbf{x}) - f_2(\mathbf{x})\right)\left(1 - 2p(y=1|\mathbf{x})\right) + {\Delta_\theta}^2}.\end{aligned} \tag{18}$$

Equation 18 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 $f_j(\mathbf{x})$. □

Derivation 5

Rewriting the result of Theorem 2 for $p(\mathbf{x}) = \frac{1}{m}$ in a regression setting, we obtain

$$\begin{aligned}q^*(\mathbf{x}) &\propto \sqrt{\int\left(\ell(f_1(\mathbf{x}), y) - \ell(f_2(\mathbf{x}), y) - \Delta_\theta\right)^2 p(y|\mathbf{x};\theta)\mathrm{d}y}\\ &= \sqrt{c_1\int y^2 p(y|\mathbf{x};\theta)\mathrm{d}y + c_2\int y p(y|\mathbf{x};\theta)\mathrm{d}y + c_3} &(19)\\ &= \sqrt{\frac{c_1}{2}\left(f_1^2(\mathbf{x}) + f_2^2(\mathbf{x}) + \tau^2_{1,\mathbf{x}} + \tau^2_{2,\mathbf{x}}\right) + \frac{c_2}{2}\left(f_1(\mathbf{x}) + f_2(\mathbf{x})\right) + c_3}. &(20)\end{aligned}$$

Equation 19 expands the loss function, orders terms by decreasing order of y, and makes use of the abbreviations

$$\begin{aligned}c_1 &= 4(f_1(\mathbf{x}) - f_2(\mathbf{x}))^2,\\ c_2 &= 4(f_1(\mathbf{x}) - f_2(\mathbf{x}))(\Delta_\theta - (f_1^2(\mathbf{x}) - f_2^2(\mathbf{x}))), \text{ and}\\ c_3 &= ((f_1^2(\mathbf{x}) - f_2^2(\mathbf{x})) - \Delta_\theta)^2.\end{aligned}$$

Equation 20 exploits that the two integrals over $\mathcal{Y}$ are sums of raw moments of the Gaussian predictive distribution under the mixture model assumption (Equation 13). Furthermore, a straightforward calculation shows that the introspective risk difference Δ_θ equals zero. The claim follows by reinserting c_i. □

B. Detailed Experimental Setup

For the classification tasks, we train logistic regression models. The regularization parameter is tuned *a priori* on the training portion of each data set by cross validation and then kept fixed. For the regression tasks, we employ Gaussian processes [8], using Bayesian model selection to determine the hyperparameters. All prediction models provide us with an estimate of $p(y|\mathbf{x})$.

For the evaluation methods *active*, *active*$_0$, *active*$_\infty$, *active*$_\neq$ and *ARE* we compute p-values according to the Wald test for weighted test samples discussed in Section 2 (Equation 5). For the passive evaluation method *passive* that draws an unweighted sample of test instances we use the more standard t-test. 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.

Spam Filtering Domain. We collected 169,612 emails described by 541,713 binary bag-of-words features from an email service provider between June 2007 and April 2010; approximately 75% of all emails are spam. In this domain, spammers impose a shift on the distribution of instances over time as they employ new strategies to generate spam messages. We 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 2000 and October 2007 to a logistic regression model trained on 5,000 randomly sampled messages received between December 2007 and April 2008. Both models employ a linear kernel. Emails received after April 2008 constitute the pool of test instances. Experimental results are averaged over 10 different sets of models and 5,000 repetitions of the evaluation process. Note that we could not use the standard Spam TREC benchmark data set because in this data set the original time stamps cannot be reconstructed, and messages therefore cannot be chronologically stratified.

Object Recognition Domain. We study the problem of detecting 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 [7] features 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 operator [5], Canny edge detector [2], Förstner operator [3]) and the size of the visual vocabulary $k \in \{50, 100, 500, 1000\}$. Additionally, we train a detection model based on SURF [1] interest point detection and a pyramid matching kernel, using the LIBPMK toolkit described in [6]. 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. Experimental results are averaged over 5,000 repetitions of the evaluation process.

Inverse Dynamics Domain. 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. We use the *Sarcos* dataset, containing 48,933 instances described by 21 features [9]. As a two-model comparison problem, we evaluate whether Gaussian process models with linear kernel or Matern kernels are preferable (Figure 1, top; Figure 2; Figure 3). As a multi-model comparison problem, we study the relative performance of Gaussian process models using polynomial kernels of degree $d \in \{1, 2, 3, 4, 5\}$ (Figure 1, bottom). Models are trained on a randomly selected set of 500 training instances, the remaining data constitute the unlabeled pool of test instances. Experimental results are averaged over 10 different sets of models and 5,000 repetitions of the evaluation process.

Abalone Domain. In this regression problem, the task is to predict the age of Abalone from ten physical measurements including length, diameter, and weight. We use the *Abalone* benchmark dataset, which includes 4,177 instances [4]. We again study the comparison of Gaussian process

Table 1: Average labeling costs, true-positive significance rate, and false decision rate when drawing test instances until significance at $\alpha = 0.05$ is obtained or a labeling budget of $n = 800$ is exhausted.

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

models with linear kernels and Matern kernels (two-model comparison) and with polynomial kernels of degree $d \in \{1, 2, 3, 4, 5\}$ (multi-model comparison). Models are trained on a randomly selected set of 500 training instances, the remaining data constitute the unlabeled pool of test instances. Experimental results are averaged over 10 different sets of models and 5,000 repetitions of the evaluation process.

C. Additional Experimental Results

For the comparison tasks discussed in Section 4.2 (two-model comparison, inverse dynamics and Abalone domains) we have also studied a protocol in which test instances are drawn and labeled until the null hypothesis is rejected at $\alpha = 0.05$, or a 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 is never equal. Note that due to the repeated statistical testing in this protocol, the resulting p-values will not be correctly calibrated. Table 1 shows average labeling costs incurred, fraction of experiments in which a significant 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.

References

[1] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[2] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.

[3] W. Förstner and E. Gülch. A fast operator for detection and precise location of distinct points, corners and centers of circular features. In *Proceedings of the ISPRS Intercommission Workshop on Fast Processing of Photogrammetric Data*, 1987.

[4] A. Frank and A. Asuncion. Uci machine learning repository. Technical report, University of California, Irvine, School of Information and Computer Sciences, 2010.

[5] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of the 4th Alvey Vision Conference*, 1988.

[6] John J. Lee. Libpmk: a pyramid match toolkit. Technical Report MIT-CSAIL-TR-2008-17, MIT Computer Science and Artificial Intelligence Laboratory, 2008.

[7] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[8] Carl Edward Rasmussen and Christopher Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[9] S. Vijayakumar, A. D'souza, and S. Schaal. Incremental online learning in high dimensions. *Neural Computation*, 17(12):2602–2634, 2005.

[10] L. Wasserman. *All of Statistics: a Concise Course in Statistical Inference*. Springer, 2004.