Active Learning of Hyperparameters: An Expected Cross Entropy Criterion for Active Model Selection

Article · September 2014

Source: arXiv

CITATIONS
3

READS
112

3 authors:

Johannes Kulick
Universität Stuttgart
3 PUBLICATIONS **49** CITATIONS
SEE PROFILE

Robert Lieck
Universität Stuttgart
10 PUBLICATIONS **57** CITATIONS
SEE PROFILE

Marc Toussaint
Universität Stuttgart
170 PUBLICATIONS **2,762** CITATIONS
SEE PROFILE

Active Learning of Hyperparameters: An Expected Cross Entropy Criterion for Active Model Selection

Johannes Kulick Robert Lieck Marc Toussaint

September 26, 2014

Abstract

In standard active learning, the learner's goal is to reduce the predictive uncertainty with as little data as possible. We consider a slightly different problem: the learner's goal is to uncover latent properties of the model—e.g., which features are relevant ("active feature selection"), or the choice of hyper parameters—with as little data as possible. While the two goals are clearly related, we give examples where following the predictive uncertainty objective is suboptimal for uncovering latent parameters. We propose novel measures of information gain about the latent parameter, based on the divergence between the prior and expected posterior distribution over the latent parameter in question. Notably, this is different from applying Bayesian experimental design to latent variables: we give explicit examples showing that the latter objective is prone to get stuck in local minima, unlike its application the standard predictive uncertainty. Extensive evaluations show that active learning using our measures significantly accelerates the uncovering of latent model parameters, as compared to standard version space approaches (Query-by-committee) as well as predictive uncertainty measures.

1 Introduction

It is often the case that multiple statistical models are candidates for describing the data. Model selection tries to find the correct model class by applying various criteria or statistical methods like cross-validation (Akaike, 1974; Schwarz, 1978; Kohavi, 1995). However, existing methods to select the best model from a set of candidates rely on a given batch of data. While this is reasonable for a wide range of applications, where gathering labeled data is comparatively easy, traditional model selection methods do not directly transfer to an active learning scenario, where the labeling of data is very expensive and time consuming.

In active learning (Settles, 2012) the learning algorithm chooses its own data queries to optimize learning success. Active learning is very successful in reducing the number of training data. However, while the data acquired by active learning methods can also be used for model selection this is not the most efficient way to actively choose data for model selection.

In this paper we transfer active learning methods to the model selection context. This means that we want to actively choose samples to best discriminate between competing model classes or hypotheses. Generally, our problem setting is relevant whenever an early decision about latent parameters is more important than the reduction in predictive uncertainty itself—e.g. when the primary task is to infer the model class, the relevant features or hyperparameters. However, we will provide evidence that active model selection may also improve the learning performance w.r.t. predictive uncertainty.

This work is motivated by our observation that a straight-forward transfer of minimizing-predictive-uncertainty to minimizing-model-uncertainty—which is also utilized by Bayesian experimental design (discussed in detail below)—may fail in the iterative case: it is prone to get

stuck in local minima and exhibits the somewhat human behavior of always choosing queries that reinforce the current belief about the model—even if it is "wrong". We explicitly give an example for this behavior and the significant different behavior of our method.

We propose and investigate different measures for the information gain about latent variables, esp. a Kullback-Leibler divergence measure between the prior and expected posterior model belief, which literally captures an information gain about the models, but in a relative sense. We find this an interesting departure from absolute uncertainty (entropy) measures. Our evaluations show it to perform excellently compared to competitors that we discuss in the following related work section. The experiments will consider regression and classification tasks where the latent parameter refers to different kernels (including feature selection) of the Gaussian Process used to generate the data.

In the remainder of this paper we will first discuss related work. We then formally introduce active model selection, and discuss our method *MaxCE* in detail. Afterwards we evaluate our method in a synthetic regression and classification scenario and a high-dimensional real-world scenario and compare it to classical active learning, query-by-committee and random sampling. Finally we discuss the results and give an outlook to future work.

2 Related Work

Active learning as general framework is widely used. Settles (2012) gives a wide overview over the different varieties. Its methods are successfully used in different fields of machine learning and a wide range of problems, as shown in a survey of projects using active learning (Tomanek & Olsson, 2009). However, it mainly focuses on speeding up the minimization of predictive error, or expected predictive entropy, in a fixed model class, whereas our method aims at actively selecting samples to speed up selecting the correct model class from a set of candidates.

Model selection techniques on the other side mainly focus on criteria to estimate the best model given a set of training data. Well known criteria are Akaikes Information Criterion (AIC) (Akaike, 1974; Burnham & Anderson, 2004) or the Bayesian Information Criterion (BIC) (Schwarz, 1978; Bhat & Kumar, 2010). Both are based on the likelihood ratios of models and are approximations of the model distribution. Since we will work with Gaussian Processes (GP) we can infer the model probabilities from the marginal likelihood of the GP and do not need to use an approximation for parametric models. Another approach to rate a model is cross-validation (see e.g. Kohavi (1995)), which statistically tests models with subsets of the training data for their generalization error.

Sugiyama & Rubens (2008) have shown that actively learning a whole set of predictors from different model classes and choosing among those models leads to what they call the "Active Learning with Model Selection Dilemma". That is, for actively choosing samples a model must be fixed while for choosing a model the training samples must be fixed. This, however, is true only for methods trying to reduce the generalization error and not for methods that actively search points for discriminating between models, as ours.

Query-by-committee (QBC) as introduced by Seung et al. (1992) tries to use active learning methods for version space reduction. The version space is the space of competing hypothesis. QBC finds new samples by evaluating the disagreement within a set of committee members (that is, models) concerning their predictions. These samples are then used to train the different models. In a binary classification scenario disagreement is easy to determine. In multi-class or regression scenarios it is harder to define. One approach, as suggested by McCallum & Nigam (1998), is to use as measure of disagreement the sum of Kullback-Leibler divergences (KLDs) from each committee member's predictive distribution to the mean of all committee member's predictive distributions. The QBC objective is

$$\frac{1}{\|M\|} \sum_{M} \mathrm{D}_{\mathrm{KL}} \left(p(y|M, D, x) \,\|\, \mathrm{E}_M[p(y|M, D, x)] \right), \tag{1}$$

with $\mathrm{E}_M[\cdot]$ being the mean over M. While aiming for model selection, QBC methods still focus on the prediction error. We will empirically compare our approach to QBC.

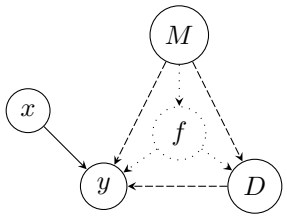


Figure 1: The graphical model of the assumed generative model; dotted arcs and variables describe dependencies before eliminating the predictive function f, dashed afterwards.

Our method is most closely related to *Bayesian experimental design*, where Bayesian techniques are used to optimally design a series of experiments (Chaloner & Verdinelli, 1995). These experiments can be seen as samples of a parameter space. They maximize the Shannon information of the posterior over models (which can be written as expected entropy as well as an expected KLD, as detailed below). Our method is closely related in that we also consider an expected KLD as a measure for information gain. However, while the traditional experimental design objective is not well designed for iterative methods as they may get stuck in local optima (more details in sec. 3.3) our method overcomes this flaw by using a different objective. Bayesian experimental design recently has regained interest due to an efficient implementation, the Bayesian Active Learning by Disagreement (BALD) (Houlsby et al., 2011), which exploits the equivalence of the experimental design criterion to a mutual information in order to make the computations tractable. Guestrin et al. (2005) also used this mutual information criterion for hyperparameter adjustment for positioning sensors for measuring data of a Gaussian Process model.

Another variant of active learning are expected model change methods such as the expected gradient length algorithm (Settles et al., 2008). Our method is in spirit related and might arguably be classified as a variant of expected model change since we are also interested in finding samples that contain a maximum amount of information with respect to the model. However, we apply the idea of greatest model change directly to the *distribution* of models by measuring the KLD between the distribution before and after new observations have been incorporated. In contrast, existing methods stay within one fixed model and can therefore not directly be applied to the model selection problem.

For our experiments we use Gaussian Processes for regression and classification. See Rasmussen & Williams (2006) for an extensive introduction.

3 Active Model Selection

Let M, x, y, D, and f be random variables. M denotes a discrete latent random variable indicating the model class (or hyperparameter). Conditional to the model class we assume a distribution over functions f, for instance a Gaussian Process prior. (In the classification case these are discriminative functions.) D are the data observed so far consisting of (x_i, y_i) input-output pairs where $P(y_i|x_i, f)$ depends on f. x is the input that is to be chosen actively and y is the corresponding output received in response. The graphical model in Fig. 1, neglecting the dashed arcs, describes their dependence. For active model selection we will have to express the expected information gain about M depending on x and usually eliminate f. In the graphical model, after eliminating f, the dashed arcs describe the dependencies.

The scenario we describe here is iterative. The choice of a new query point x is guided by an objective function that scores all candidate points. The candidate with an optimal objective value is then used to generate a new data point (x, y) which, in the next iteration, is added to D.

3.1 Model Posterior and Predictive Posterior

As opposed to existing active learning methods we define the objective function directly in terms of the *model posterior* $P(M|D)$ instead of the *predictive posterior* $P(y|x, D)$. For comparison, query-by-committee (Seung et al., 1992) follows the classical approach by comparing the predictive posterior $P(y|x, M, D)$ for each model class M with the mean predictive posterior $\frac{1}{|M|}\sum_M P(y|x, M, D)$. We call $P(M|D, x, y)$ the *augmented* model posterior after we have seen an additional data point (x, y). The relation between the model posterior and predictive posterior is

$$\begin{aligned}\underbrace{p(M|D,x,y)}_{\text{model posterior}} &= \frac{p(x,y,D|M)p(M)}{p(x,y,D)} \\ &\propto \underbrace{p(y|x,M,D)}_{\text{predictive posterior}} \quad p(D|M)\ .\end{aligned} \tag{2}$$

3.2 Expected Model Entropy versus Cross-Entropy

Before introducing our active model selection criterion we summarize existing criteria for reference. Standard active learning chooses x to minimize expected predictive error or, identifying error with neg-log-likelihood, expected predictive entropy,

$$x_{PE} = \operatorname*{argmin}_x \int_y p(y|x,D)\left[\int_{x'} p(x')H\left[p(y'|x',D,x,y)\right]\right]. \tag{3}$$

Here, x is the point to be actively chosen, y the resulting observation we taken the expectation over, and $p(y'|x', D, x, y)$ the augmented predictive posterior (the predicted posterior for the augmented data $D \cup \{(x, y)\}$).

In GPs the *expected* predictive entropy is minimized by picking the point that *directly* maximizes the predictive entropy, a strategy called *uncertainty sampling*,

$$x_{US} = \operatorname*{argmax}_x H\left[\, p(y|x,D)\,\right]\ . \tag{4}$$

The idea of minimizing expected entropy can readily be transferred to the problem of model selection leading to the standard approach of Bayesian experimental design, maximizing the expected Shannon *information* (minimizing expected model entropy)

$$x_{ME} = \operatorname*{argmax}_x \int_y -p(y|x,D)H[p(M|D,x,y)]\ . \tag{5}$$

It is very instructive to rewrite this same criterion in various ways (we omit the data D for simplicity) by subtracting a constant $H[p(M)]$:

$$-\int_y p(y|x)\ H[p(M|y,x)] - H[p(M)] \tag{6}$$

$$= \int_{y,M} p(y|x)\ p(M|y,x)\ \log p(M|y,x) - H[p(M)] \tag{7}$$

$$= \int_{y,M} p(y,M|x)\ \log \frac{p(M|y,x)}{p(M)} \tag{8}$$

$$= \int_{y,M} p(y|x)\ p(M|y,x)\ \log \frac{p(M|y,x)}{p(M)} \tag{9}$$

$$= \int_y p(y|x)\ \mathrm{D}_{\mathrm{KL}}\left(p(M|y,x)\,\|\,p(M)\right)\ . \tag{10}$$

These rewritings of expected Shannon information establish the direct relation to Eq. (3) and (4) in (Chaloner & Verdinelli, 1995). We find that x_{ME} can also be interpreted as maximizing the expected change in neg-entropy (assuming a uniform prior $p(M)$),

$$\int_{y,M} p(y, M|x, D) \ \log p(M|x, y, D) \ , \tag{11}$$

as in Eq. (8), or to maximize the expected KL divergence

$$\int_{y} p(y|x, D) \ \mathrm{D}_{\mathrm{KL}} \left(p(M|D, y, x) \,\|\, p(M|D) \right) \ , \tag{12}$$

as in Eq. (10).

The more recent BALD algorithm computes the expected posterior entropy by leveraging the insight that Eq. (6) is equivalent to the mutual information between the unknown output and the model and thus transforms the model entropy to an entropy in y-space, which is efficiently computable (Houlsby et al., 2011).

Minimizing the expected model entropy is surely one way of maximizing information gain about M. However, in our interative setup we empirically show that this criterion can get stuck in local optima: Depending on the stochastic sample D, the model posterior $p(M|D)$ may be "mislead", that is, having low entropy while giving high probability to a wrong choice of M. As detailed below, the attempt to further minimize the entropy of $p(M|D, x, y)$ in such a situation may lead to suboptimal choices of x_{ME} that confirm the current belief instead of challenging it.

We therefore propose the *MaxCE* strategy which instead maximizes the expected model *cross entropy* between the (old) model posterior $p(M|D)$ and the augmented posterior $p(M|D, x, y)$,

$$x_{CE} = \operatorname*{argmax}_x \int_y p(y|x, D) \ H[p(M|D); p(M|D, x, y)] \tag{13}$$

$$= \operatorname*{argmax}_x \int_y p(y|x, D) \ \mathrm{D}_{\mathrm{KL}} \left(p(M|D) \,\|\, p(M|D, x, y) \right), \tag{14}$$

where $H[p(z), q(z)] = -\int_z p(z) \log q(z)$. The KLD $\mathrm{D}_{\mathrm{KL}} \left(p(M|D) \,\|\, p(M|D, x, y) \right)$ literally quantifies the additional information captured in $p(M|D, x, y)$ relative to the previous knowledge $p(M|D)$.

This does not require the entropy to decrease: the expected $\mathrm{D}_{\mathrm{KL}} \left(p(M|D) \,\|\, p(M|D, x, y) \right)$ can be high even if the expected entropy of $p(M|D, x, y)$ is higher than $H[p(M|D)]$—so our criteron is not the same as minimizing expected model entropy. In comparison to the KLD formulation Eq. (12) of expected model entropy the two arguments are switched. The following example and the later quantitative experiments will demonstrate the drastic effect of this difference.

3.3 An Example for Maximizing Cross Entropy versus Minimizing Entropy

Bayesian Optimal Design suggests to minimize the expected entropy of the model distribution Eq. (11). This is reasonable if one optimizes all samples (the whole experimental design) at once. However, in our iterative approach miminizing the expected entropy can get stuck in wrong hypotheses; intuitively it actually does not measure the *change* of distributions, but only the information of the augmented posterior $p(M|D, x, y)$.

We now explicitly show an example of such a situation. Assume a regression scenario where two GP hypotheses should approximate a ground truth function. Both GPs use a squared exponential kernel, but have a different length scale hyperparameter. One of these GPs is the correct underlying model.

Consider now a case where the first two observations by chance support the wrong hypothesis. This may happen due to the fact that the ground truth function that is actually sampled is itself only a random sample from the prior over all functions described by the underlying GP.

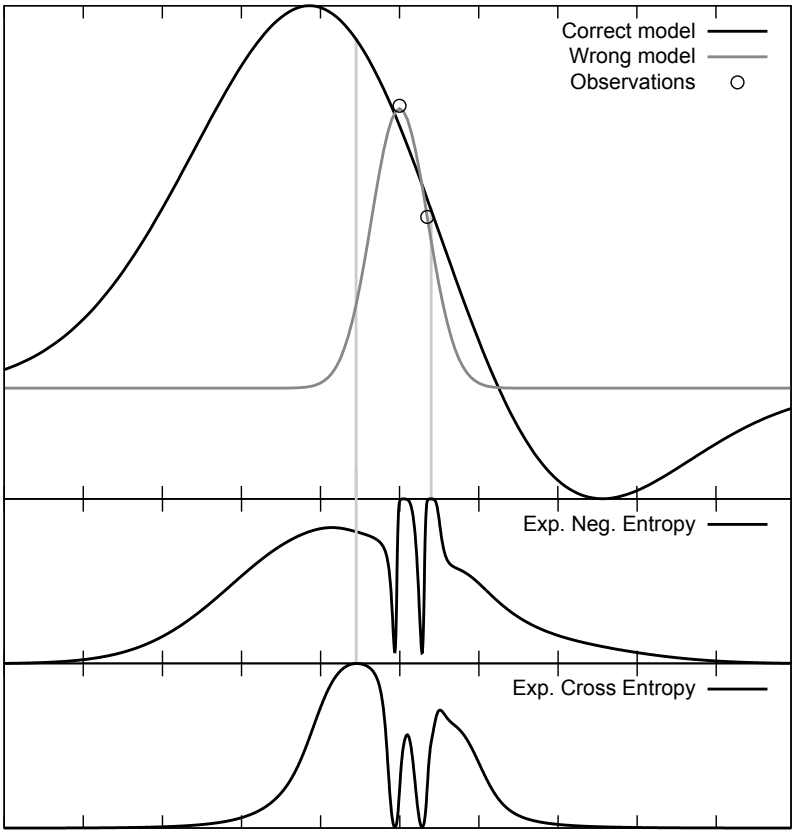


Figure 2: On top in two competing hypotheses are shown, one is correct the other one wrong. The two curves below correspond to two different objectives for choosing the next query by taking the maximum. The only difference between the two objectives is the direction of the KL divergence. The upper curve corresponds to the neg-entropy and KL divergence from Eqs. (11) and (12). The curve lower corresponds to the expected cross entropy and (reversed) KL divergence from Eqs. (13) and (14). The vertical lines show the next query for each of the objectives.

Furthermore observations may be noisy, which may lead to a similar effect. Such a scenario—one that actually occurred in our experiments—is shown in Fig. 2. The probability for the wrong model in this scenario is already around 90%. If we now compute the expected neg-entropy from Eq. (11) it has its maximum very close to the samples we already got. This is due to the fact, that samples possibly supporting the other—the correct—model would temporarily increase the entropy. The entropy would only decrease again if the augmented posterior actually flipped and the probability for the correct model got higher then 90%.

The MaxCE approach of maximizing cross entropy (see Eq. (13)) on the other hand favors *changes* of the model posterior in any direction, not only to lower entropy, and therefore recovers much faster from the misleading first samples. Fig. 2 shows both objectives for this explicit example.

3.4 Mixing Active Learning and Model Selection

While measuring the expected cross entropy is a good measure to find samples holding information about latent model parameters of competing hypothesis it might actually not query points that lead to minimal predictive error. For example, regions that are important for prediction but do not discriminate between hypothesis would not be sampled. Nevertheless, information about latent model parameters may help to increase the predictive performance as well. For our experiments

we therefore additionally tested a linear combination of the MaxCE measure f_{CE} from Eq. (13) with the uncertainty sampling measure f_{US} from Eq. (4) in a combined objective function f_{mix}

$$f_{mix} = \alpha \cdot f_{CE} + (1 - \alpha) \cdot f_{US} \ . \tag{15}$$

4 Experiments

4.1 Compared Methods

We compared six different strategies: our MaxCE, which maximizes the expected cross entropy (see Eq. (13)); classical Bayesian experimental design, which minimizes the expected entropy (see Eq. (11)); query-by-committee with Kullback-Leibler to the mean (see Eq. (1)); uncertainty sampling (see Eq. (4)), and random sampling, which randomly choses the next sample point. Additionally we tested a mixture of MaxCE and uncertainty sampling (see Eq. (15)). The mixing coefficient, which was found by a series of trial runs, was $\alpha = 0.5$ for both synthetic data sets and $\alpha = 0.3$ for the CT slices data set.

4.2 Measures

To measure progress in selecting a model we computed the entropy of the model posterior for each method. To measure progress in the predictive performance we plot the classification accuracy and the mean squared error for classification and regression, respectively. To compute an overall predictive performance for a method we took the weighted average over the different models, with the posterior probabilities as weights. This corresponds to the maximum a posteriori estimate of the marginal prediction

$$p(y|D, X) = \sum_M p(M|D) \ p(y|M, D, X) \ . \tag{16}$$

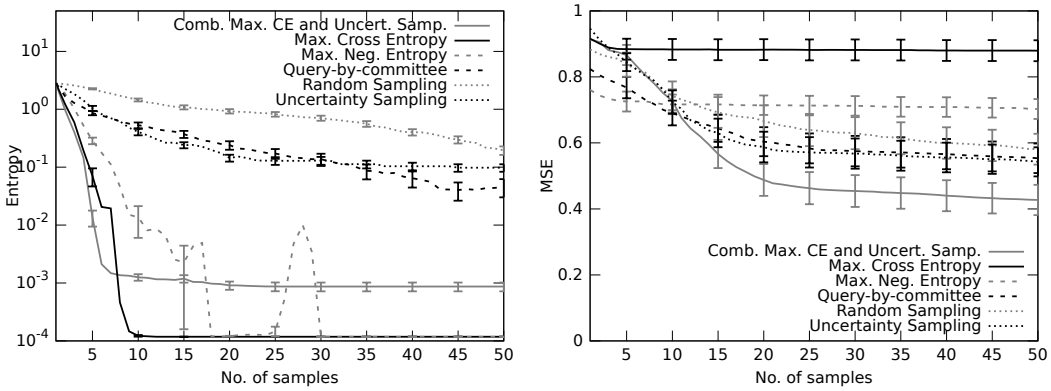
Figs. 3, 4 and 5 show these measures for all our experiments.

4.3 Synthetic Data

We tested our method in both a 3D-regression and a 3D-classification task. The setup for both experiments was essentially the same: A ground truth Gaussian Process (GP) was used to generate the data. The kernel of the ground truth GP was randomly chosen to depend either on all three dimensions (x, y, z), only a subset of two dimensions (x, y), (y, z) or (x, z), or on only one dimension (x), (y) or (z). Model selection in this case corresponds to a feature selection problem: uncovering on which features the unknown true GP depends on. The latent variable M, to be uncovered by the active learning strategies, enumerates exactly those seven possibilities. One run consisted of each method independently choosing fifty queries one-by-one from the same ground truth model. After each query the corresponding candidate GP was updated and the model posterior was computed.

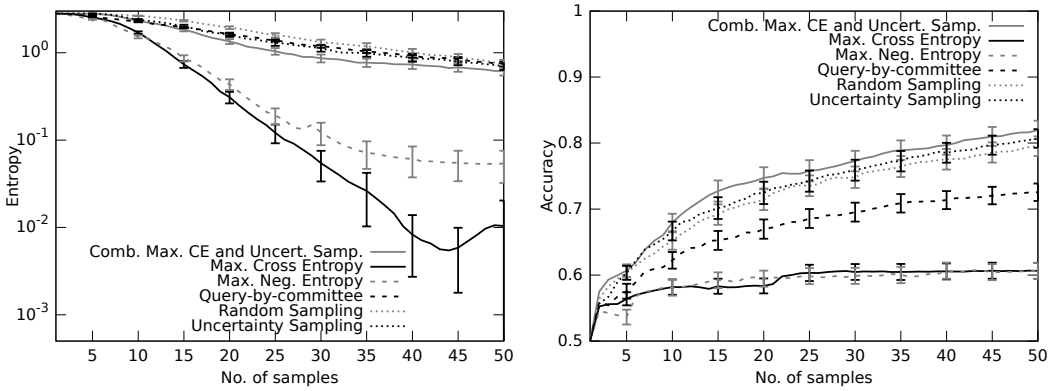
Figs. 3 and 4 show the mean performance over 100 runs of the synthetic classification and regression tasks, respectively. Since we average over 100 runs, the error bars of the mean estimators are very small. Both model entropy and accuracy/mean squared error are shown.

On this synthetic data MaxCE significantly outperforms all other tested methods in terms of model selection, followed by Bayesian experimental design, and the mixture of MaxCE and uncertainty sampling (Fig. 3a and 4a). As expected, In terms of classification accuracy and predictive error both MaxCE and Bayesian experimental design perform poorly. This is because their objectives are not designed for prediction but for model selection. However, the mixture of MaxCE and uncertainty sampling, performs best (Fig. 3b and 4b), which is presumably due to its capability to uncover the correct model quickly.



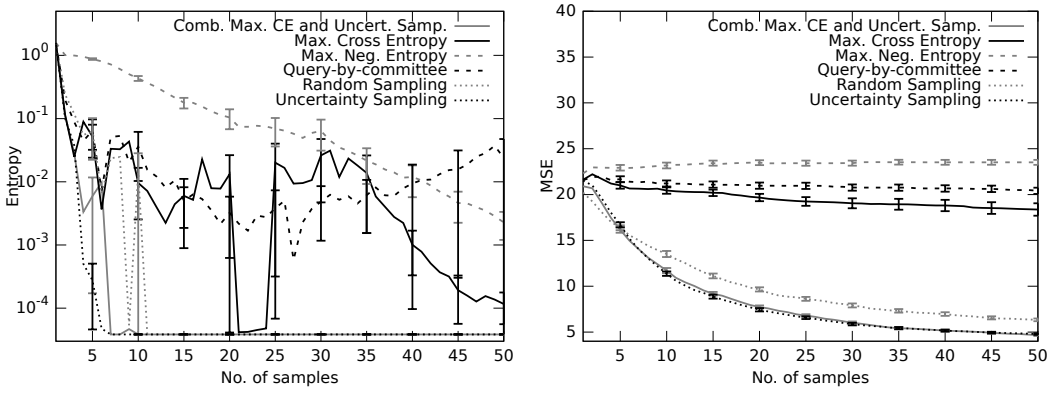
(a) Regression Model Entropy (b) Regression MSE

Figure 3: The mean performance of the different methods for synthetic regression task.



(a) Classification Model Entropy (b) Classification Accuracy

Figure 4: The mean performance of the different methods for a synthetic classification task.



(a) CT Slices Model Entropy (b) CT Slices Error

Figure 5: The mean performance of the different methods for the CT slice regression task.

4.4 CT-Slice Data

We also tested our methods on a high dimensional (384 dimensions) real world data set from the machine learning repository of the University of California, Irvine (Bache & Lichman, 2013). The task on this set is to find the relative position of a computer tomography (CT) slice in the human body based on two histograms measuring the position of bone (240 dimensions) and gas (144 dimensions). We used three GPs with three different kernels: a γ-exponential kernel with $\gamma = 0.4$, an exponential kernel, and a squared exponential kernel. Although obviously none of these processes generated the data we try to find the best matching process alongside with a good regression result. Fig. 5 shows the mean performance over 40 runs on the CT slice data set.

In the CT slice data set neither MaxCE nor Bayesian experimental design minimize the entropy quickly (Fig. 5a). This may be a consequence of the true model *not* being among the available alternatives. As a consequence both methods continuously challenge the belief thereby preventing it from converging. QBC may be subject to the same struggle, here even resulting in an increase of entropy after the first 25 samples. In contrast for uncertainty sampling, our mixture method, and random sampling the entropy converges reliably. Concerning the predictive performance MaxCE, Bayesian experimental design, and QBC do not improve noticeably over time (cf. explanation above). Again uncertainty sampling and our mixture method perform much better, while here the difference between them is not significant.

5 Conclusion and Outlook

The presented results strongly suggest that our newly developed strategy of maximizing the expected cross entropy is superior to classical Bayesian experimental design for uncovering latent parameters in an iterative setting.

The results on predictive performance additionally demonstrate a successful application of MaxCE for prediction by mixing it with an uncertainty sampling objective. The resulting objective at the same time actively learns the latent parameters and accurate predictions. This initially goes at the expense of accurate predictions but at some point more than compensates this fall-back. This might be the case, because this way areas which are important for false model hypothesis can be ignored and thus the right model is trained better.

So far our mixing strategy is rather simple. But the results suggest, that the mixing helps. Investigating better mixing strategies might lead to more improvements.

So far we only investigated the discrete case of k distinct models. The same techniques described in this paper may be useful to find samples to optimize continuous hyperparameter. In this case the sum over models will become an integral and efficient integration techniques need to be applied to the method to keep it computationally tractable. It also might be applicable to leverage the insight of Ko et al. (1995) that the entropy is submodular to implement efficient approximations of the optimization.

Another direction of research would involve finding better optimization techniques to find the actual maxima to up the process. When using GPs all involved distributions are Gaussian (or approximated by Gaussians for the classification case). As such they are infinitely differentiable, so higher order methods might prove useful.

Acknowledgments

This work was partly funded by the German Research Foundation (DFG) within the priority programm “Autonomous learning” (SPP1597) and the German National Academic Foundation. The CT slices database was kindly provided by the UCI machine learning repository (Bache & Lichman, 2013).

References

Akaike, Hirotugo. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, pp. 716–723, 1974.

Bache, K. and Lichman, M. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

Bhat, Harish S. and Kumar, Nitesh. On the Deviation of the Bayesian Information Criterion. Technical report, University of California, Merced, 2010.

Burnham, Kenneth P. and Anderson, David R. Multimodel Inference - Understanding AIC and BIC in Model Selection. *Sociological Methods and Research*, 33:261–304, 2004.

Chaloner, Kathryn and Verdinelli, Isabella. Bayesian experimental design: A review. *Statistical Science*, pp. 273–304, 1995.

Guestrin, Carlos, Krause, Andreas, and Singh, Ajit Paul. Near-Optimal Sensor Placements in Gaussian Processes. In *Proc. of the Int. Conf. on Machine Learning (ICML)*, pp. 265–272, 2005.

Houlsby, Neil, Huszár, Ferenc, Ghahramani, Zuobin, and Lengyel, Máté. Bayesian Active Learning for Classification and Preference Learning. *arXiv*, 1112.5745 (stat.ML), 2011.

Ko, C., Lee, J., and Queyranne, M. An exact algorithm for maximum entropy sampling. *Ops Research*, 43:684–691, 1995.

Kohavi, Ron. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proc. of the Int. Conf. on Artificial Intelligence (IJCAI)*, 1995.

McCallum, Andrew and Nigam, Kamal. Employing EM in pool-based active learning for text classification. In *Proc. of the Int. Conf. on Machine Learning (ICML)*, pp. 359–367, 1998.

Rasmussen, Carl and Williams, Christopher. *Gaussian processes for machine learning*. MIT Press, 2006.

Schwarz, Gideon E. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

Settles, Burr. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan and Claypool, 2012.

Settles, Burr, Craven, Mark, and Ray, Soumya. Multiple-Instance Active Learning. In *Proc. of the Conf. on Neural Information Processing Systems (NIPS)*, pp. 1289–1296, 2008.

Seung, H. Sebastian, Opper, Manfred, and Sompolinsky, Haim. Query by committee. In *Proc. of the Annual Conf. on Computational Learning Theory*, pp. 287–294, 1992.

Sugiyama, Masashi and Rubens, Neil. Active Learning with Model Selelction in Linear Regression. In *Proc. of the Int. Conf. of Data Mining*, pp. 518–529, 2008.

Tomanek, Katrin and Olsson, Fredrik. A web survey on the use of active learning to support annotation of text data. In *Proceedings of the Workshop an Active Learning for Natural Language Processing*, pp. 45–48, 2009.