May 28, 2004

Online Methods in Learning Theory

Jan Poland
IDSIA, Galleria 2 CH-6928 Manno (Lugano), Switzerland*
jan@idsia.ch

Abstract

We present theoretical results on the performance of Bayesian online learning, namely predictive error bounds for the Bayes mixture and MDL with respect to a countable model class. We briefly discuss how assertions and improvements might be obtained for active learning.

1 Online Learning

Proving loss bounds is an important issue in Statistical Learning Theory. Much work has been done in an *offline setup*: Given are a domain $\mathcal{X}$, a co-domain $\mathcal{Y}$, a class of models $\mathcal{C}$ which usually consists of functions from $\mathcal{X}$ to $\mathcal{Y}$, and a training data set $(x_i, y_i)_{1\leq i\leq n} \in \mathcal{X}\times\mathcal{Y}$ which is i.i.d. according to some distribution μ on $\mathcal{X}\times\mathcal{Y}$. Then one tries to obtain bounds on the expected prediction error when choosing some model from $\mathcal{C}$ in terms of the empirical error observed on the training data, thus identifying a model with good generalization capabilities [1, 2].

In contrast, we consider an *online setup*. No structural or probabilistic assumptions on the domain $\mathcal{X}$ are given. We concentrate on the task of classification, so the co-domain $\mathcal{Y}$ is a finite set (regression can be treated similarly but incurs additional technical problems). Sequences $x_{<\infty} = (x_1, x_2, \ldots) \subset \mathcal{X}$ are generated by an arbitrary mechanism (which may be also an adversary trying to maximize the prediction error). The corresponding target values $y_{<\infty} = (y_1, y_2, \ldots) \subset \mathcal{Y}$ are generated i.i.d.[1] according to a probability distribution $\mu(y|x)$ which depends on the actual input x. Let $\Delta(\mathcal{Y})$ denote the set of probability distributions on $\mathcal{Y}$, i.e. $\nu \in \Delta(\mathcal{Y})$ iff $\nu(y|x) \geq 0$ and $\sum_{y\in\mathcal{Y}} \nu(y|x) = 1$ for all $x \in \mathcal{X}$, then $\mu(\cdot|x) \in \Delta(\mathcal{Y})$. Note that the i.i.d. requirement is actually insignificant as long as there are no assumptions on $\mathcal{X}$: we can make the distribution dependent on the past by adding the past observations $x_{<t} = (x_1, \ldots, x_{t-1})$ and $y_{<t}$ to the current input x_t. The setup immediately generalizes to *binary sequence prediction*[2] by letting $x_t = y_1 \otimes y_2 \otimes \ldots \otimes y_{t-1}$ and $\mathcal{Y} = \{0, 1\}$. The model class $\mathcal{C}$ is assumed to be a *countable*[3] class of probability distributions $\mathcal{C} = \big\{\nu_i : \nu_i(\cdot|x) \in \Delta(\mathcal{Y}),\, i \geq 1\big\}$. A predictor φ is a distribution $\varphi(y|x_{1:t}, y_{<t})$ on $\mathcal{Y}$ given the past inputs $x_{1:t} = (x_1, \ldots, x_t)$ including the current and the past outputs $y_{<t}$. Let $d : \Delta(\mathcal{Y}) \times \Delta(\mathcal{Y}) \to \mathbb{R}^+$ be a distance between probability distributions. Then we are interested in bounding the *total expected distance*

$$D(\mu, \varphi|x_{<\infty}) = \sum_{t=1}^{\infty} \mathbf{E}\big[d\big(\mu(\cdot, x_t), \varphi(\cdot|x_{1:t}, y_{<t})\big)\big], \tag{1}$$

*This work was supported by SNF grant 2100-67712.02.

[1] Compare the prediction with expert setup [3] where even no assumptions on $y_{<\infty}$ are made.

[2] In fact, all our results generalize to semimeasures [4] and thus apply to universal sequence prediction [5].

[3] For results on continuously parameterized classes, see e.g. [6, 7].

where expectation is taken with respect to μ. This is an online error measure since in each time step t, the predictor may be chosen according to the complete set of past observations, i.e. the model is retrained after each step. Relevant distance measures are the Kullback-Leibler divergence (short KL divergence) and the square distance,

$$d_{KL}(\mu,\varphi|x_{1:t},y_{<t}) = \sum_{y_t\in\mathcal{Y}} \mu(y_t|x_t)\ln\frac{\mu(y_t|x_t)}{\varphi(y_t|x_{1:t},y_{<t})} \text{ and} \tag{2}$$

$$d_2^2(\mu,\varphi|x_{1:t},y_{<t}) = \sum_{y_t\in\mathcal{Y}} \big(\mu(y_t|x_t)-\varphi(y_t|x_{1:t},y_{<t})\big)^2, \tag{3}$$

respectively. These quantities induce total distances D_{KL} and D_2^2 according to (1). We associate a *prior weight* $w_\nu > 0$ with each $\mu \in \mathcal{C}$ and require that the Kraft inequality $\sum_\nu w_\nu \leq 1$ holds. Then for given observation $(x_{1:n}, y_{1:n})$, we define the *Bayes mixture* and the *MDL* = minimum description length (or MAP = maximum a posteriori) estimator with respect to the model class $\mathcal{C}$ as

$$\xi(y_{1:n}|x_{1:n}) = \sum_{\nu\in\mathcal{C}} w_\nu\nu(y_{1:n}|x_{1:n}) = \sum_{\nu\in\mathcal{C}} w_\nu \prod_{t=1}^{n}\nu(y_t|x_t) \tag{4}$$

$$\nu^* = \nu^*_{(x_{1:n},y_{1:n})} = \arg\max_{\nu\in\mathcal{C}}\{w_\nu\nu(y_{1:n}|x_{1:n})\}, \text{ and} \tag{5}$$

$$\varrho(y_{1:n}|x_{1:n}) = \max_{\nu\in\mathcal{C}}\{w_\nu\nu(y_{1:n}|x_{1:n})\} = w_{\nu^*}\nu^*(y_{1:n}|x_{1:n}). \tag{6}$$

These quantities induce predictors, namely the Bayes mixture, the static MDL, the dynamic MDL and the normalized dynamic MDL predictor, as follows:

$$\xi(y_n|x_{1:n},y_{1:n-1}) = \xi(y_{1:n}|x_{1:n})/\xi(y_{1:n-1}|x_{1:n-1}) \tag{7}$$

$$\varrho^{\text{static}}(y_n|x_{1:n},y_{1:n-1}) = \nu^*_{(x_{1:n-1},y_{1:n-1})}(y_n|x_n) \tag{8}$$

$$\varrho(y_n|x_{1:n},y_{1:n-1}) = \varrho(y_{1:n}|x_{1:n})/\varrho(y_{1:n-1}|x_{1:n-1}), \text{ and} \tag{9}$$

$$\bar\varrho(y_n|x_{1:n},y_{1:n-1}) = \varrho(y_{1:n}|x_{1:n})/\big[\textstyle\sum_{y_n}\varrho(y_{1:n}|x_{1:n})\big]. \tag{10}$$

Note that the static MDL predictor directly uses the MAP estimator and thus reflects a basic principle in machine learning: "choose the model which minimizes the error on the training data plus a regularization term". The following fundamental theorem which asserts excellent prediction properties of the Bayes mixture was found by Solomonoff [8] for universal sequence prediction:

Theorem 1 *Let $\mu \in \mathcal{C}$, then $D_2^2(\xi,\mu|x_{<\infty}) \leq \ln w_\mu^{-1}$ holds.*

That is, the only requirement is that the *true distribution* μ is contained in the model class. The assertion is strong: The total expected square error of the predictive probabilities is finitely bounded. This implies that they rapidly converge to the true probabilities *almost surely*. Moreover, one can derive good bounds for *any* bounded loss function [9]. In order to prove the theorem, we need that D_2^2 is bounded use by D_{KL} [10]. Use the *dominance* $\xi(\ldots) \geq w_\mu\mu(\ldots)$ which holds by definition to obtain

$$\sum_{t=1}^{n}\mathbf{E}\ln\frac{\mu(y_t|x_t)}{\varphi(y_t|x_{1:t},y_{<t})} = \mathbf{E}\ln\prod_{t=1}^{n}\frac{\mu(y_t|x_t)}{\varphi(y_t|x_{1:t},y_{<t})} = \mathbf{E}\ln\frac{\mu(y_{1:n}|x_{1:n})}{\xi(y_{1:n}|x_{1:n})} \leq \ln w_\mu^{-1}. \tag{11}$$

Taking the limit $n\to\infty$, the left hand side of (11) converges to D_{KL}, which implies the assertion. Note that the formal dependence on $x_{<\infty}$ is completely irrelevant for the proof.

Predicting according to the mixture ξ, also known as marginalization, is the optimal method from a Bayesian view point, but usually is computationally infeasible. Since a popular (and possibly less expensive) way is predicting according to the best model in the class, it is important to study the properties of the static MDL predictions. Here a corresponding result is more difficult to obtain. We achieve the goal using dynamic and normalized dynamic MDL as intermediate steps: the former quantity has the dominance property but is no (semi-)measure, hence it cannot be used with the KL divergence. On the other hand, the triangle inequality does not hold for the KL divergence.

Theorem 2 *Let $\mu \in \mathcal{C}$, then we have $D_2^2(\bar{\varrho}, \mu|x_{<\infty}) \leq \ln w_\mu^{-1} + w_\mu^{-1}$, $D_2^2(\varrho, \bar{\varrho}|x_{<\infty}) \leq 2w_\mu^{-1}$, and $D_2^2(\varrho^{\text{static}}, \varrho|x_{<\infty}) \leq 3w_\mu^{-1}$.*

The proofs can be found in [4]. Since the triangle inequality holds for $\sqrt{D_2^2}$, this implies bounds on the total expected distances for the static MDL predictor. Observe that the bounds are exponentially worse than for the Bayes mixture, namely $O(w_\mu^{-1})$ instead of $\ln w_\mu^{-1}$. This is no artifact of the proof, as the following example shows.

Example 3 Let $N \geq 1$, $\mathcal{X} = \{1, \ldots, N-1\}$, $\mathcal{Y} = \{0, 1\}$, $\mathcal{C} = \{\nu_1, \ldots, \nu_N\}$, where $\nu_i(1|x) = 1$ iff $x \geq i$. Let $w_\nu = \frac{1}{n}$ for all ν and assume that the true distribution is $\mu = \nu_N$, i.e. it generates 0 on all inputs almost surely. Let the input sequence be $x_{1:N-1} = 1, 2, \ldots, N-1$, then $\bar{\rho}(1|x_{1:t}, y_{<t}) = \frac{1}{2}$ for all $t \in \{1, \ldots, N-1\}$, thus the total error of the normalized dynamic MDL predictions is $\frac{N-1}{2}$. Assume that in case of a tie, the static MDL predictor always chooses the element of $\mathcal{C}$ with the smaller index (or make the weights slightly non-uniform), then the total static MDL prediction error is $N-1$.

It might be surprising on a first glance that even if all elements of $\mathcal{C}$ do not depend on the input (prediction of Bernoulli sequences), examples can be found where the bound $O(w_\mu^{-1})$ is sharp [11]. However under additional mild conditions a bound of $O(\ln w_\mu^{-1})$ holds here.

2 Active Learning

So far, results have presented for a *passive* learner. If in contrast the learner has influence on the generation or selection of training data, on could hope that either the expected prediction error decreases or the amount of (labeled) training examples is reduced, or both. As far as we know, no general results of this type have been proven yet in our setup. Thus there are some interesting open questions in active learning.

In the special case of Example 3, it is immediate how active learning improves the performance: If the learner chooses the inputs by an iterative bisection of the set $\{1, \ldots, N-1\}$, then the true distribution is identified after $\log_2 N$ steps, thus the total prediction error is reduced to $\log_2 N$. This gives hope that a more general result might hold: If the learner is allowed to construct queries, then the MDL prediction error reduces to $\ln w_\mu^{-1}$.

In order to construct queries, the input space $\mathcal{X}$ must be known to the learner and hence must fulfil structural assumptions. Without such assumptions, one might hope to draw benefits from a *selective sampling* method in the spirit of [12]. There unlabeled inputs are presented to the learner that may pick the "interesting" ones for labeling in the sense of an expected information gain criterion. In this case, one might hope to reduce the total request of labeled examples as in [12], while the error does not increase significantly.

Another issue is greedy vs. non-greedy construction of queries. Empirical evidence indicates that a sophisticated multi-step lookahead strategy of query construction does not gain much over a greedy method. We give an informal argument in favor of this claim in our framework, provided that all models are independent on the history (i.e. no information about the history or the current

step is integrated in the input): Given the bound on the *total* error, we should choose inputs first where our predictive performance is poor, in order to maximize the quality of future predictions. Of course, this argument should be made formal.

References

[1] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, second edition, 1999.

[2] O. Bousquet. New approaches to statistical learning theory. *Annals of the Institute of Statistical Mathematics*, 55(2):371–389, 2003.

[3] P. Auer, N. Cesa-Bianchi, and C. Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64:48–75, 2002.

[4] J. Poland and M. Hutter. Convergence of discrete MDL for sequential prediction. 17th Annual Conference on Learning Theory (COLT), to appear, 2004.

[5] M. Li and P. M. B. Vitányi. *An introduction to Kolmogorov complexity and its applications*. Springer, 2nd edition, 1997.

[6] B. S. Clarke and A. R. Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory*, 36:453–471, 1990.

[7] A. R. Barron, J. J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760, 1998.

[8] R. J. Solomonoff. Complexity-based induction systems: comparisons and convergence theorems. *IEEE Trans. Inform. Theory*, IT-24:422–432, 1978.

[9] M. Hutter. Convergence and loss bounds for Bayesian sequence prediction. *IEEE Transactions on Information Theory*, 49(8):2061–2067, 2003.

[10] M. Hutter. Convergence and error bounds for universal prediction of nonbinary sequences. *Proceedings of the 12th Eurpean Conference on Machine Learning (ECML-2001)*, pages 239–250, December 2001.

[11] J. Poland and M. Hutter. On the convergence speed of MDL predictions for Bernoulli sequences. Technical Report, 2004.

[12] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28:133, 1997.