Supervised Learning: Information Theoretic Bounds on predictive errors

Manfred Opper
Institut für Theoretische Physik III,
Julius Maximilians Universität Würzburg,
Am Hubland, D–97074 Würzburg, Germany
Universität Würzburg, Germany
opper@physik.uni-wuerzburg.de

David Haussler
Department of Computer and Information Sciences,
UC Santa Cruz
Santa Cruz, California, U.S.A.
haussler@cse.ucsc.edu

Abstract

Within a Bayesian framework, we calculate general upper and lower bounds for a cumulative entropic error, which measures the success in the supervised learning of an unknown rule from examples.

1 Introduction

A standard task in supervised learning is the inference of an unknown rule from a set of examples. In the most simple approach, such a rule can be specified by a functional relation $y = f_\theta(x)$ that maps an input x onto an output y. For a classification problem, we may think of x as a vector of features of the object to be classified, and of y as the corresponding classification label. We will assume that f belongs to a parametric class of functions, and θ denotes a vector of parameters. E.g., f_θ may belong to the class of functions that are realizable by feedforward neural networks of a given architecture, where θ specifies the set of network couplings and thresholds. In a more realistic approach, the deterministic relation between x and y should be replaced by a stochastic rule that is described by a probability $P(y|\theta, x)$.

In this paper, we address the question of how much knowledge about the unknown rule is gained on average by observing t pairs of random input/output examples $(x^t, y^t) = (x_1, y_1), \ldots, (x_t, y_t)$ in a learning experiment. Our progress in learning about the true parameter will be measured by our ability to predict a new output y_{t+1} after having seen the previous observations $y^t = y_1, \ldots, y_t$ and $x^{t+1} = x_1, \ldots, x_{t+1}$.

Assume that our estimate for the unknown rule θ^* is given by the *predictive distribution* $\hat{P}_{t+1}(y_{t+1}|x_{t+1}) = \hat{P}(y_{t+1}|x_{t+1}, x^t, y^t)$. We adopt a Bayesian approach and assume that

nature has drawn the true parameter θ^* of the rule at random from a prior distribution. Hence we can measure the performance of any learning method that produces an estimated probability $\hat{P}_{t+1}(y_{t+1}|x_{t+1})$ for a new output through an entropic error which is defined by

$$\Delta I_{t+1} = \mathbb{E}_{X^{t+1},\Theta^*} \mathbb{E}_{Y^{t+1}|\theta^*,x^{t+1}} \ln \frac{P(y_{t+1}|\Theta^*, x_{t+1})}{\hat{P}_{t+1}(y_{t+1}|x_{t+1})}, \tag{1}$$

where the symbol $\mathbb{E}_{Y^{t+1}|\theta^*,x^{t+1}}$ denotes the expectation with respect to the conditional probability $P(y^{t+1}|\theta^*, x^{t+1})$. Similarly, $\mathbb{E}_{X^{t+1},\Theta^*}$ is the expectation with respect to the density of the inputs and with respect to the prior.

In coding theory ΔI_{t+1} is proportional to the expected extra number of bits needed to encode y_{t+1} by an encoder who uses the probability $\hat{P}$ relative to somebody who knows the true P, usually called *redundancy*. The x_t inputs are often called *side information*. A good choice of $\hat{P}$ with a small entropic error results in an efficient coding.

Using asymptotic methods of statistics, the scaling of the entropic error for large sample size t can be calculated for the regular case, where P depends smoothly on the parameter θ. This approach fails in the important nonregular cases, e.g., when the rule is given by a deterministic classification problem, or a nonsmooth noise model, where the outputs (classification labels) change discontinuously with the parameters. Using a different method, Amari [1] has performed an analysis of the asymptotics for deterministic binary classification problems. While this is a powerful new technique, it turns out to be more complicated and somewhat less general than the analysis of the regular models.

In the following, we present a new approach to this problem that covers both the regular and nonregular cases in a unified treatment. Our approach is not restricted to the asymptotics of the problem. We derive general upper and lower bounds for the cumulative entropic error, which hold for arbitrary numbers of examples. We find that both bounds match asymptotically. Details of the derivations, as well as references about related work can be found in [2, 3, 4].

2 Mutual Information

One can show that the Bayes optimal choice for the predictive distribution is given by the posterior probability

$$\hat{P}_{t+1}(y_{t+1}|x_{t+1}) = \frac{P(y^{t+1}|x^{t+1})}{P(y^t|x^t)}, \tag{2}$$

where $P(y^t|x^t) = \mathbb{E}_\Theta\ P(y^t|\theta, x^t)$. Here and in the sequel we assume that the inputs x_t are independent and the outputs y_t are independent given θ^* and x_t. If we assume that Bayes optimal method is applied sequentially each time a new example is observed, the *cumulative* entropic error on n examples is found to be

$$I_n = \sum_{t=1}^{n} \Delta I_t = \mathbb{E}_{X^n,\Theta^*} \mathbb{E}_{Y^n|\theta^*,x^n} \ln \frac{P(y^n|\theta^*, x^n)}{P(y^n|x^n)}. \tag{3}$$

A second interpretation of Eq.(3) comes from information theory. I_n is the (input) average of the mutual information between the data y^n and the parameter θ^*, given the inputs x^n. Hence, it tells us how much information we have gained about the unknown parameter θ^* by observing the outputs y^n. Accordingly ΔI_t is the instantaneous information gain from the t–th example.

3 Upper bound on I_n

Our upper bound on I_n uses two main inequalities. The first one states that for any random variables W and V and real-valued function $u(w, v)$,

$$-\mathbf{E}_V \ln \mathbf{E}_W e^{u(w,v)} \leq -\ln \mathbf{E}_W e^{\mathbf{E}_V u(w,v)}.$$

This is easily proved by applying Jensen's inequality, utilizing that $\ln \mathbf{E}_W e^{u(w,v)}$ is convex in u.

Second, for any probability $Q(y^n|\theta, x^n)$ one can show that

$$\mathbf{E}_{\Theta^*} \mathbf{E}_{Y^n|\theta^*,x^n} \ln \frac{P(y^n|\theta^*, x^n)}{P(y^n|x^n)} \leq \mathbf{E}_{\Theta^*} \mathbf{E}_{Y^n|\theta^*,x^n} \ln \frac{P(y^n|\theta^*, x^n)}{Q(y^n|x^n)}.$$

This follows from the fact that the difference between the right and left expressions forms a relative entropy, and thus is always nonnegative.

Using these inequalities, and setting $Q(y^n|\theta, x^n) = \prod_{t=1}^n Q(y_t|\theta^*, x_t)$ we can prove that for every Q

$$I_m \leq -\mathbf{E}_{\Theta^*} \ln \mathbf{E}_\Theta \exp\left[-n\mathbf{E}_X D_K(P_{y|\theta^*,x}||Q_{y|\theta,x})\right], \tag{4}$$

where

$$D_K(P_{y|\theta^*,x}||Q_{y|\theta,x}) = \sum_y P(y|\theta^*, x) \ln \frac{P(y|\theta^*, x)}{Q(y|\theta, x)}$$

denotes the *Kullback-Leibler (KL) divergence* (or *relative entropy distance*).

Optimising with respect to Q, inequality (4) defines a variational method to bound I_n. In the important case where the ratio $\frac{P(y|\theta,x)}{P(y|\theta^*,x)}$ is not bounded (this holds for deterministic classification rules where $P(y|w^*, x) \in \{0, 1\}$), and the output variable assumes only $K < \infty$ different values, a simple ansatz for Q yields the upper bound

$$I_m \leq -\mathbf{E}_{\theta^*} \ln \mathbf{E}_\theta \, e^{-n[c(n)\Delta(\theta,\theta^*)+o(1)]}, \tag{5}$$

where $c(n) \leq const \cdot \ln(n)$ (the $const$ will depend on K) and $o(1) \to 0$ as $n \to \infty$. The function Δ in the exponent is defined as

$$\Delta(\theta, \theta^*) = \frac{1}{2}\mathbf{E}_X \sum_y \left(\sqrt{P(y|\theta, x)} - \sqrt{P(y|\theta^*, x)}\right)^2. \tag{6}$$

This is the average of the so-called *Hellinger distance* between $P(y|\theta, x)$ and the true probability $P(y|\theta^*, x)$. For deterministic rules, $\Delta(\theta, \theta^*)$ reduces to the standard 0-1 generalization error, i.e. the probability that θ and θ^* disagree on a new input. If on the other hand $\frac{P(y|\theta,x)}{P(y|\theta^*,x)}$ is bounded, we can chose $Q = P$ and an analogous bound to (5) is obtained, where now $c(n) \leq const$.

4 A Lower Bound on I_n

It is possible to obtain a lower bound that has a structure similar to the upper bound. This is, to our knowledge, new and is based on the following extremal property of the mutual information. Define

$$J_n(\lambda) = -\mathbf{E}_{\Theta^*} \mathbf{E}_{Y^n|\theta^*,x^n} \ln \mathbf{E}_\Theta \left(\frac{P(y^n|w, x^n)}{P(y^n|w^*, x^n)}\right)^\lambda. \tag{7}$$

One can show that $J_n(\lambda)$ is *maximal* at $\lambda = 1$, where it equals the mutual information. A combination of this result for $\lambda = \frac{1}{2}$ with Jensen's inequality enables us to show that

$$I_n \geq -\mathbf{E}_{\Theta^*} \ln \mathbf{E}_\Theta e^{-n\Delta_H(\theta,\theta^*)}. \tag{8}$$

5 Asymptotic Scaling

The bounds (5) and (8) become similar when the number n of examples grows large. The behaviour of the expectations over θ are then determined from small neighbourhoods of θ^* in the large n limit. Under some mild conditions explained in [3], we can show that

$$\lim_{n\to\infty} \frac{-\ln \mathbf{E}_\theta\, e^{-n\Delta(\theta,\theta^*)}}{\ln n} = d(\theta^*) \tag{9}$$

where the *local dimension* $d(\theta^*)$ is defined by $d(\theta^*) = \lim_{\epsilon\to 0} \ln V_\epsilon(\theta^*)/\ln\epsilon$ and $V_\epsilon(\theta^*)$ is the volume (probability) of the set of all parameters θ that satisfy $\Delta(\theta,\theta^*) \le \epsilon$, measured with respect to the prior. Again, under some mild conditions from (5,8) and (9) we obtain the asymptotic scaling $I_n \simeq d\ln(n)$ for $n\to\infty$, where $d = \mathbf{E}_{\theta^*}\, d(\theta^*)$. As a consequence, it can be shown that whenever the limit $\lim_{n\to\infty} n\, I_n$ exists, the asymptotic value of the information gain is given by

$$\Delta I_n \simeq \frac{d}{n} \tag{10}$$

(see [2]). The values for the dimension d will depend on the type of learning problem and can be calculated for several models. For a regular noise model, when the probability P depends smoothly on an N component parameter vector θ, a quadratic expansion $\Delta(\theta,\theta^*) \simeq \sum_{i,j} C_{ij}(w_i - w_i^*)(w_j - w_j^*)$ is valid for w close to w^*. Inserted into (9), assuming a well behaved prior, this shows that $d = \frac{1}{2}N$. On the other hand, for deterministic, binary classification rules, using polar coordinates, other types of expansions for Δ around θ^* are often possible [1]. Here, for fixed angle variables the scaling $\Delta(\theta,\theta^*) \simeq ||\theta - \theta^*||^\gamma$ with $\gamma = 1$ is a typical case, resulting in $d = N$. Further work has to show in which cases other types of universality classes will appear.

So far our results hold on *average* over the prior distribution on the rules. It will be interesting to see under which conditions these results will also hold pointwise for almost all rules. It is also a challenge to extend our methods to the important case of learning unrealizable rules. A further line of our future research considers the evaluation and optimization of our bounds in the so-called *thermodynamic limit*, where the dimension of the rule space grows with the size of the set of examples.

Acknowledgements

D. Haussler was supported by NSF grant IRI-9123692 and M. Opper by a Heisenberg fellowship of the DFG.

References

[1] S. Amari, "A universl theorem on learning curves", *Neural Networks*, vol 6, pp. 161-166, 1993

[2] D. Haussler and M. Opper, "General bounds on the mutual information between a parameter and n conditionally independent observations", to appear in the *8th ACM Conference on Computational Learning Theory (COLT95)* (Santa Cruz 1995); published by ACM Press.

[3] D. Haussler and M. Opper,"Mutual information and Bayes methods for learning a distribution",*Proc. Workshop on the Theory of Neural Networks: The Statistical Mechanics Perspective*., World Scientific, 1995, to appear.

[4] M. Opper and D. Haussler, "General bounds for predictive errors in supervised learning", *Proc. Workshop on the Theory of Neural Networks: The Statistical Mechanics Perspective*. World Scientific, 1995, to appear