GENERAL BOUNDS FOR PREDICTIVE ERRORS IN SUPERVISED LEARNING

MANFRED OPPER
Institut für Theoretische Physik, Julius Maximilians Universität Würzburg, Am Hubland
D–97074 Würzburg, Germany
E-mail: opper@physik.uni-wuerzburg.de

and

DAVID HAUSSLER
Department of Computer and Information Sciences, UC Santa Cruz, Santa Cruz, California, U.S.A.
E-mail: 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. This performance measure is equivalent to the mutual information between the data and the parameter that specifies the rule to be learnt. Both bounds match asymptotically, when the number m of observed data grows large. Under mild conditions, we find that the information gain from observing a new example decreases universally like d/m. Here d is a dimension that is defined from the scaling of small volumes with respect to a suitable distance in the space of rules.

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 $y = f_w(x)$ that maps an input x onto an output y. For a classification problem, we may think of x as 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 w denotes a vector of parameters. E.g., f_w may belong to the class of functions that are realizable by feedforward neural networks of a given architecture, where w 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 $P(y|w, x)$.

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 $x^{t+1} = x_1, \ldots, x_{t+1}$. This problem is closely related to the calculation of the so called learning curves in the theory of neural networks. These give the probability of not predicting the

correct output on a new input x, sometimes called the 0-1 error, averaged over the distribution of training data, as a function of the size of the training set.

An important problem in the theory of learning curves is the question of universal asymptotic scaling properties. Rather than for the 0-1 error, such universality can be found for a different type of performance measure, the so called *entropic error*. Assume that our estimate for the unknown rule w^* is given by the *predictive distribution* $\hat{P}(y_{t+1}|x_{t+1})$. Then, we can define an entropic error by

$$\Delta I_{t+1} = -\Big\langle\Big\langle \ln \hat{P}(y_{t+1}|x_{t+1}) - \ln P(y_{t+1}|w^*, x_{t+1})\Big\rangle\Big\rangle \tag{1}$$

where the double brackets denote an average over the distribution of training examples and the new example (x_{t+1}, y_{t+1}). In a Bayesian framework, an additional average over w^* will be included. 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. 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 was calculated by Amari and Murata [1] for the regular case, where P depends smoothly on the parameter w. This approach fails in nonregular cases, e.g., when the rule is either entirely deterministic, i.e. when $y = f_w(x)$, or for nonsmooth noise models. An example for the latter case is a model for independent output noise, which can be formulated by the equation $y = \eta f_w(x)$, where η is a binary noise variable taking the values -1 or $+1$.

Using a different method, which is related to the idea of replicas, Amari [2] has also performed an analysis of the deterministic case. While this is a powerful new analysis 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. Within a Bayesian framework of learning, 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.

Some related work in this area is given in [7,12,13,14,15].

2. Cumulative Entropic Error, Mutual Information and the Free Energy

We adopt a Bayesian approach and assume that nature has drawn the true parameter w^* of the rule at random from a prior distribution. Hence we can measure the performance of any learning method that produces an estimated probability $\hat{P}(y_{t+1}|x_{t+1})$ for a new output by averaging its entropic error over the prior. In this sense, one can show that the Bayes optimal choice for the predictive distribution is

given by the posterior probability

$$P(y_{t+1}|x_{t+1}) = \frac{\int dw P(y^t|w,x^t) P(y_{t+1}|w,x_{t+1})}{\int dw P(y^t|w,x^t)}. \tag{2}$$

Here $\int dw$ denotes the expectation over the prior. If we assume that Bayes optimal method is applied sequentially each time a new example is observed, the *cumulative* entropic error on m examples is found to be

$$\begin{aligned} I_m = \sum_{t=1}^{m} \Delta I_t &= \left\langle \left\langle \int dw^* \sum_{y^m} P(y^m|w^*,x^m) \ln \frac{P(y^m|w^*,x^m)}{\int dw P(y^m|w,x^m)} \right\rangle \right\rangle_{x^m} \\ &= \left\langle \left\langle \ln \frac{P(y^m|w^*,x^m)}{\int dw P(y^m|w,x^m)} \right\rangle \right\rangle_{all}. \end{aligned} \tag{3}$$

Here, the the subscript *all* means an average taken over all random variables, i.e. the outputs y^m for fixed w^*, the parameter w^* and the inputs x^m. In the following, we will always assume that the input data $x^m = x_1, x_2, \ldots x_m$ are drawn independently from the same fixed density. A second interpretation of Eq.(3) comes from information theory. I_m is the (input) average of the so called mutual information [3] between the data y^m and the parameter w^*, given the inputs x^m. Hence, it tells us how much information we have gained about the unknown parameter w^* by observing the outputs y^m. Accordingly ΔI_t is the instantaneous information gain from the t–th example.

To motivate the bounds on the mutual information that are derived in the following sections, let us finally discuss a third interpretation of (3) that is related to statistical mechanics. Let us assume for the moment, that the likelihood $P(y^m|w,x^m)$ of the data can be written in the form of a Boltzmann factor

$$P(y^m|w,x^m) \propto e^{-\beta E(w,y^m,x^m)},$$

where $E(w,y^m,x^m)$ may be interpreted as suitable measure for the goodness of fit of the data using a deterministic model and the temperature β^{-1} measures the strength of the noise in the data. In the context of learning, E is called the training energy. Then the mutual information can be written

$$I_m = -\left\langle \left\langle \ln \int dw\, e^{-\beta E(w,y^m,x^m)} \right\rangle \right\rangle_{all} - \beta \left\langle \langle E(w^*,y^m,x^m) \rangle \right\rangle_{all},$$

Apart from the trivial second term, a calculation of the mutual information becomes essentially equivalent to a calculation of the quenched average of a free energy in statistical mechanics.

Besides the nonrigorous replica trick, which enables an exact evaluation of the quenched average for certain models,[4,5,6] mainly two other approximate methods have been used in statistical mechanics to estimate the free energy. These are the *annealed*

[4] and the *high temperature* [4] approximations. Both are summarised in the rigorous inequalities

$$-\ln\int dw\Big\langle\Big\langle e^{-\beta E}\Big\rangle\Big\rangle \leq -\Big\langle\Big\langle \ln\int dw\, e^{-\beta E}\Big\rangle\Big\rangle \leq -\ln\int dw\, e^{-\beta\langle\langle E\rangle\rangle}. \tag{4}$$

The lower bound gives the annealed free energy, which is derived from Jensen's inequality. While this bound gives qualitatively good results in the case of noise-free and exactly learnable rules, it is known to fail for noisy or nonrealizable rules.[4] On the other hand, the expression on the righthand side has been used to estimate the free energy in the high temperature limit. Both bounds have been recently applied to the stochastic complexity of learning a rule.[7]

The fact that the righthand expression is a rigorous upper bound [8] on the quenched average appears not to be widely known. Thus, we will briefly sketch one of the possible proofs. Setting $E = E_0 + \delta E$ with $E_0 = \langle\langle E\rangle\rangle$ and applying the well known thermodynamic variational principle

$$-\ln\int dw e^{-\beta E} \leq -\ln\int dw e^{-\beta E_0} + \beta\,\frac{\int dw\, e^{-\beta E_0}\,\delta E}{\int dw\, e^{-\beta E_0}}, \tag{5}$$

the bound is immediately obtained by taking the average. While the lower bound fails for noisy rules, the upper bound applied to the deterministic case ($\beta = \infty$) yields the trivial result $I_m \leq \infty$. In order to derive expressions for I_m that are useful in general, we will derive new bounds in the following sections.

3. An Improved Upper Bound

While the bounds (4) hold for general distributions of data y^m and x^m, and are also true for fixed parameter w^*, the special type of expectation involved in the definition of (3) allows for a better type of bound. Note that the same probabilities (both the prior and $P(y^m|w,x^m)$) that define the average also appear inside the logarithm. This allows us to utilize the following simple inequality. For an arbitrary probability $Q(y^m|w,x^m)$ one has

$$\int dw^* \sum_{y^m} P(y^m|w^*,x^m)\ln\frac{P(y^m|w^*,x^m)}{\int dw\, P(y^m|w,x^m)} \leq \tag{6}$$
$$\int dw^* \sum_{y^m} P(y^m|w^*,x^m)\ln\frac{P(y^m|w^*,x^m)}{\int dw\, Q(y^m|w,x^m)}$$

This follows from the fact that the difference between the right and left expressions forms a relative entropy,[3] and thus is always nonnegative. If

$$Q(y^m|w,x^m) = \prod_{i=1}^{m} Q(y_i|w,x_i)$$

then combining (6) with the upper bound of (4) we get

$$I_m \leq -\int dw^* \ln \int dw \, \exp\left[-m\left\langle\left\langle \sum_y P(y|w^*,x) \ln \frac{P(y|w^*,x)}{Q(y|w,x)}\right\rangle\right\rangle_x\right] \tag{7}$$

Here, we have also used the independence of the inputs x^m together with

$$P(y^m|w^*,x^m) = \prod_{i=1}^{m} P(y_i|w^*,x_i).$$

Optimising with respect to Q, inequality (7) defines a variational method to bound I_m.

In the important case where the ratio $\frac{P(y|w,x)}{P(y|w^*,x)}$ is not bounded (this holds for deterministic rules where $P(y|w^*,x) \in \{0,1\}$), and the output variable assumes only $K < \infty$ different values, the following simple ansatz for Q yields a good bound for I_m.

$$Q_m(y|w,x) = (1 - \frac{K}{(K-1)(m+1)})P(y|w,x) + \frac{1}{(K-1)(m+1)}. \tag{8}$$

Note that for $m \to \infty$, Q_m approaches P. If on the other hand $\frac{P(y|w,x)}{P(y|w^*,x)}$ is bounded,[9] we can chose $Q = P$. With the above choices, it is possible to obtain the general upper bounds

$$I_m \leq -\int dw^* \ln \int dw \, e^{-m[const \cdot \Delta(w,w^*)]}, \tag{9}$$

for the case when $\frac{P(y|w,x)}{P(y|w^*,x)}$ is bounded, and for the general case

$$I_m \leq -\int dw^* \ln \int dw \, e^{-m[c(m)\Delta(w,w^*)+o(1)]}, \tag{10}$$

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

$$\Delta(w,w^*) = \frac{1}{2}\left\langle\left\langle \sum_y \left(\sqrt{P(y|w,x)} - \sqrt{P(y|w^*,x)}\right)^2\right\rangle\right\rangle_x. \tag{11}$$

It is the average of the so-called *Hellinger distance* between $P(y|w,x)$ and the true probability $P(y|w^*,x)$.

For deterministic rules, $\Delta(w,w^*)$ reduces to the standard 0-1 generalization error, i.e. the probability that w and w^* disagree on a new input. On the other hand, for a noisy rule, when the probability P depends smoothly on w, the Hellinger distance is related to the *Fisher information*,[1] which plays an important role for the asymptotics of regular models. One finds the expansion

$$\Delta(w,w^*) \simeq \frac{1}{4}\sum_{i,j} \langle\langle J_{ij}\rangle\rangle (w_i - w_i^*)(w_j - w_j^*) \tag{12}$$

which holds for w close to w^*, where J_{ij} is the Fisher information.

4. A Lower Bound

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

$$J_m(\lambda) = -\int dw^* \sum_{y^m} P(y^m|w^*, x^m) \ln \int dw \left(\frac{P(y^m|w, x^m)}{P(y^m|w^*, x^m)} \right)^{\lambda}. \tag{13}$$

($J_m(\lambda)$ depends implicitly on x^m.) One can show that $J_m(\lambda)$ is maximal at $\lambda = 1$, where it equals the mutual information. An application of this result to $\lambda = \frac{1}{2}$ yields

$$\begin{aligned} I_m = \langle\langle J_m(1)\rangle\rangle_{x^m} \geq \left\langle \left\langle J_m(\frac{1}{2}) \right\rangle \right\rangle_{x^m} = \\ -\int dw^* \left\langle \left\langle \sum_{y^m} P(y^m|w^*, x^m) \ln \int dw \left(\frac{P(y^m|w, x^m)}{P(y^m|w^*, x^m)} \right)^{\frac{1}{2}} \right\rangle \right\rangle_{x^m} \geq \\ -\int dw^* \ln \int dw\, e^{m \ln(1-\Delta(w,w^*))}. \end{aligned} \tag{14}$$

For a deterministic rule, the rightmost expression coincides with the result of the annealed average.

5. Asymptotic Scaling

The bounds (9,10) and (14) become similar when the number m of examples grows large. The behaviour of the integrals over w are then determined from small neighbourhoods of w^* in the large m limit. Under some mild conditions explained in ref. 9, we can show that

$$\lim_{m\to\infty} \frac{-\ln \int dw\, e^{-m\Delta(w,w^*)}}{\ln m} = d(w^*) \tag{15}$$

where the *local dimension* $d(w^*)$ is defined by

$$d(w^*) = \lim_{\epsilon\to 0} \frac{\ln V_\epsilon(w^*)}{\ln \epsilon} \tag{16}$$

and $V_\epsilon(w^*)$ is the volume (probability) of the set of all parameters w which satisfy $\Delta(w, w^*) \leq \epsilon$, measured with respect to the prior.

Again, under conditions given in ref. 9, from (9,10) and (14) we obtain the asymptotic scaling

$$I_m \simeq d \ln(m) \tag{17}$$

for $m \to \infty$, where $d = \int dw^* \, d(w^*)$. As a consequence, it can be shown that whenever the limit $\lim_{m\to\infty} m\, I_m$ exists, the asymptotic value of the information gain is given by

$$\Delta I_m \simeq \frac{d}{m} \tag{18}$$

The values for the dimension d can be calculated for a variety of models: For regular probabilistic rules, one can use the expansion (12) to show that $d = \frac{1}{2}N$. On the other hand, for a deterministic linearly separable (perceptron) rule with N independent parameters and a spherical distribution of inputs, one finds that $d = N$. These special cases agree with the results of refs. 1 and 2.

A different scaling is obtained for a special limit of the committee machine [10] with tree architecture, N weights, h hidden units and spherical distribution of inputs. When $N \to \infty$, $h \to \infty$, but $\frac{h}{N} \to 0$, the dimension scales like $d \simeq 2N$.

6. Error of the Gibbs–Learning

Using our results, it is possible to get bounds on the error of a different type of learning strategy, the so called *Gibbs algorithm*. This algorithm, which has been extensively studied using methods of statistical mechanics,[4,5,6] chooses an estimate $\hat{w}$ for the unknown parameter w^* at random from the posterior distribution $p(w|y^m, x^m)$. We will be interested in the question of how close the estimate $\hat{w}$ comes to the true parameter w^* on average. Thus we define the Gibbs–error by

$$\varepsilon_m = \left\langle \left\langle \frac{\int dw P(y^m|w, x^m)\Delta(w, w^*)}{\int dw P(y^m|w, x^m)} \right\rangle \right\rangle_{all} . \tag{19}$$

ε_m reduces to the average 0-1 generalization error of the Gibbs algorithm after m training examples in the case of a deterministic rule. It can be shown that we always have

$$\varepsilon_m \leq 2\Delta I_m. \tag{20}$$

A better bound was found to hold in the deterministic case in ref. 11:

$$\varepsilon_m \leq \frac{1}{2\ln 2}\Delta I_m. \tag{21}$$

Hence, the last inequality combined with (18) shows that the generalization error for the Gibbs–algorithm is (again under some mild conditions) asymptotically upper bounded by $\simeq 0.721\frac{d}{m}$ in the deterministic case.

7. Summary and Outlook

We have developed new general upper and lower bounds for the entropic error of Bayes method in learning a rule from examples. Using these bounds a universal asymptotic scaling for the error could be obtained.

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 (see recent work of Merhav and Feder [12] and Feder, Freund and Mansour [13]). 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. Finally we hope to find further applications of our bounds in other areas of statistical mechanics.

Acknowledgments

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

References

1. S. Amari and N. Murata, *Neural Computation* **5** (1993) 140.
2. S. Amari, *Neural Networks* **6** (1993) 161.
3. T. Cover and J. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
4. H. Seung, H. Sompolinsky, and N. Tishby, *Physical Review* **A 45** (1992) 6056.
5. T. L. H. Watkin, A. Rau and M. Biehl, *Rev. Mod. Phys.* **65** (1993) 499.
6. M. Opper and W. Kinzel, *Statistical Mechanics of Generalization*, to appear in: *Physics of Neural Networks*, ed. J. L. van Hemmen, E. Domany and K. Schulten, to be published by Springer Verlag.
7. R. Meir and N. Merhav, *Stochastic Complexity of Learning Realizable and Unrealizable Rules*, to be published in *Machine Learning*.
8. K. Symanzik, *J.Math. Phys.* **6** (1965) 1155.
9. D. Haussler and M. Opper, *Mutual Information and Bayes Methods for Learning a Distribution*, in this volume.
10. H. Schwarze, J. Hertz, *Europhys. Lett.* **20** (1992) 375.
11. D. Haussler, M. Kearns, and R. Schapire, *Machine Learning* **14** (1994) 84.
12. N. Merhav and Meir Feder, *A Strong Version of the Redundancy-Capacity Theorem of Universal Coding*, submitted to *IEEE Trans. Inform. Theory.*
13. M. Feder, Yoav Freund and Yishay Mansour, *Optimal Universal Learning and Prediction of Probabilistic Concepts*, preprint.
14. K. Yamanishi, *A learning criterion for stochastic rules*, *Machine Learning* (1992), Special Issue on the Proceedings of the 3nd Workshop on Computational Learning Theory.
15. D. Haussler and A. Barron, *How well do Bayes methods work for on-line*

prediction of $\{+1, -1\}$ *values?*, in *Proceedings of the Third NEC Symposium on Computation and Cognition*. SIAM (1992).