A Derivation of Learning Theory from the Principle of Maximum Entropy

Griff L. Bilbro

Center for Communications and Signal Processing
Department Electrical and Computer Engineering
North Carolina State University

TR-91/4
February 1991

A Derivation of Learning Theory from the Principle of Maximum Entropy

Griff L. Bilbro

February 19, 1991

Abstract

An information-theoretic derivation is presented for the learning theory recently reported by Tishby, Levin, and Solla (TLS). The resulting learning theory is shown to apply to any problem of modeling data.

1 Introduction

A statistical theory which describes the learning of a relation from examples was reported in (Tishby, Levin, and Solla, 1989). It built on earlier work in (Schwartz, Samalan, Solla, and Denker, 1990) and has been carefully restated in (Levin, Tishby, and Solla, 1990). In that literature, statistical mechanics was used to relate the probability of independent input-output data pairs to a layered neural network.

2 Maximum entropy and modeling

In this section I will show that a TLS theory can be constructed for any problem in which parameters of a model are chosen to fit data. Consider a problem in which data $\{x_i\}_{i=1}^{i=N}$ drawn from an unknown density $\overline{p}(x)$ are to be fitted with some model by adjusting parameters w to minimize an additive error function during a training procedure. In the case of training a layered feedforward net, the data are input-output pairs; the parameters w are the

usual weights and biases; the error function and training procedure might be squared error and backpropagation. However the theory is much more general.

Except for the case of linear regression, there is little theoretical guidance for identifying when data is sufficient to determine the model. Often the total set of available data $\{x_i\}_{i=1}^{i=N}$ is divided into a training set $\{x_i\}_{i=1}^{i=m}$ and a remaining test set. The model is trained to a specified error ϵ_T on $\{x_i\}_{i=1}^{i=m}$ and then tested against the remaining data. In principle this procedure could be repeated for randomly selected test sets and initial values and the average results could be tabulated as functions of m and ϵ_T. Implicit in this hypothetical exercise is an average density $\langle \overline{\rho}^{(m)}(w) \rangle_x$ of nets trained to an error ϵ_T on m examples. The true $\langle \overline{\rho}^{(m)}(w) \rangle_x$ would be difficult to calculate in general, mostly because it involves the details of the training procedure. However the end result of training may depend more on the form of the model, the amount of the data, the noise in the data, and the training error ϵ_T, than on the details of the training procedure. In that case, the maximum entropy density $\langle \rho^{(m)}(w) \rangle_x$ resembles $\langle \overline{\rho}^{(m)}(w) \rangle_x$ and might be used to predict training behavior and generalization.

The *principle of maximum entropy* is a general inference tool which produces probabilities characterized by certain average values of specified functions(Jaynes, 1979). To the extent that entropy measures information, a maximum entropy estimate contains only the information implied by those average values and makes no other assumptions (*e.g.* about the training procedure). It is useful to also incorporate a prior estimate $\rho^{(0)}(w)$ of $\overline{\rho}^{(m)}(w)$ by considering the entropy of $\rho^{(m)}(w)$ relative to $\rho^{(0)}(w)$

$$R[\rho^{(m)}] = \int dw \rho^{(m)}(w) \ \ln \frac{\rho^{(m)}(w)}{\rho^{(0)}(w)}. \tag{1}$$

So little is known about $\overline{\rho}^{(m)}(w)$ that $\rho^{(0)}(w)$ is chosen merely as a restriction to reasonable portions of w space. For example, in back propagation it is unreasonable to expect the weights as large as $O(10^6)$ and perhaps $O(1)$ is a little too small. In any case the first test of this theory must be for sensitivity to $\rho^{(0)}(w)$.

In the maximum entropy sense, the density $\rho^{(m)}(w)$ that contains the least information beyond $\rho^{(0)}(w)$ but is nevertheless a normalized density

with integral

$$\int dw\rho^{(m)}(w) \ = 1 \tag{2}$$

and with specific average training error

$$\langle \sum_{i=1}^{i=m} \epsilon(x_i, w) \rangle_w \ \equiv \int dw\rho^{(m)}(w) \ \sum_{i=1}^{i=m} \epsilon(x_i, w) \ = \epsilon_T, \tag{3}$$

can be found with calculus of variations and two Lagrange multipliers α and β in the usual way. The extremum satistifies

$$\ln \frac{\rho^{(m)}(w)}{\rho^{(0)}(w)} - 1 + \alpha + \beta \sum_{i=1}^{i=m} \epsilon(x_i, w) \ = 0 \tag{4}$$

which can be solved for $\rho^{(m)}(w)$ to get

$$\rho^{(m)}(w) \ = \frac{\rho^{(0)}(w)}{Z^{(m)}} \exp\left(-\beta \sum_{i=1}^{i=m} \epsilon(x_i, w)\right) \tag{5}$$

where β has been left, but $\exp(1-\alpha)$ has been evaluated in terms of the more conventional normalization

$$Z^{(m)} \ = \int dw\rho^{(0)}(w) \ \exp\left(-\beta \sum_{i=1}^{i=m} \epsilon(x_i, w)\right) \tag{6}$$

that depends on the particular set of examples as well as their number.

Equation 5 is significant. It is an estimate of the probability density of models of the form (*i.e.* architecture for a neural net) defined by the functional relation between w and x in $\epsilon(x, w)$ after being trained from random initial values of w to an average error Equation 3 on the particular data set $\{x_i\}_{i=1}^{i=m}$. However Equation 5 is still not useful as a predictor of average training behavior because it does depend on the particular examples in the set $\{x_i\}_{i=1}^{i=m}$. In order to remove that effect, I average Equation 5 over all possible m examples

$$\langle \rho^{(m)}(w) \rangle_x \ = \int dx^{(m)} \overline{p}(x_1)\overline{p}(x_2)...\overline{p}(x_m)\rho^{(m)}(w). \tag{7}$$

Equation 7 can be used to define a performance criterion, the *average prediction fraction*

$$\phi^{(m)} \ = \int dw \int dx \overline{p}(x_{m+1}) \exp(-\beta\epsilon(x_{m+1}, w)) \langle \rho^{(m)}(w) \rangle_x \tag{8}$$

which is the fraction of models distributed according to $\rho^{(m)}(w)$ which have an error function ϵ within about $1/\beta$ of the next unseen example x_{m+1} on average.

Equations 7 and 8 are inconvenient to evaluate exactly because of the $Z^{(m)}$ term of Equation 6 in their denominators. TLS propose an "annealed approximation" which in the present context is equivalent to replacing $Z^{(m)}$ in Equation 7 by its average over ways of choosing $\{x_i\}_{i=1}^{i=m}$

$$\langle Z^{(m)} \rangle_x = \int dw \int dx^{(m)} \overline{p}(x_1)\overline{p}(x_2)...\overline{p}(x_m)\rho^{(0)}(w) \ \exp(-\beta \textstyle\sum_{i=1}^{i=m} \epsilon(x_i, w) \ ) \tag{9}$$

which can be written

$$\langle Z^{(m)} \rangle_x = \int dw \rho^{(0)}(w) f^m(w) \tag{10}$$

where

$$f(w) \equiv \int dx \overline{p}(x) \exp(-\beta \epsilon(x, w)). \tag{11}$$

With this the average prediction fraction of Equation 8 becomes

$$\phi^{(m)} \ = \frac{\int dw \rho^{(0)}(w) f^{m+1}(w)}{\int dw \rho^{(0)}(w) f^m(w)}. \tag{12}$$

Equation 12 predicts generalization behavior by drawing a maximum entropy inference of the average consistency between the model represented by $\epsilon(x, w)$ and m examples drawn from $\overline{p}(x)$.

I will show that Equation 12 is is well suited for theoretical analysis and is also convenient in practical numerical calculations for small problems. This is because it is easy to produce Monte Carlo estimates for the averages over the $\{x_i\}_{i=1}^{i=m}$ by using the entire set of available data $\{x_i\}_{i=1}^{i=N}$.

3 Relation to TLS

In this subsection I will show that under the assumptions of Tishby, Levin, and Solla, my average predition fraction is proportional to their *Average Prediction Probability* $\langle\langle p^{(m)} \rangle\rangle$

$$\phi^{(m)} \ = z(\beta)\langle\langle p^{(m)} \rangle\rangle \tag{13}$$

where

$$z(\beta) = \int dx \exp(-\beta\epsilon(x,w)) \tag{14}$$

normalizes the probability density for the conditional probability

$$p(x|w) \equiv \frac{1}{z(\beta)} \exp(-\beta\epsilon(x,w)) \tag{15}$$

which TLS use to describe the behavior of a single certain net w. Equation 15 can itself be obtained from a maximum entropy argument in x space but I will not do so here.

Now z is a function of β but *is assumed in TLS to be independent of* w. TLS show this to be rigorously true for layered nets with real outputs if ϵ is the usual squared error between data and output. It is true in that case because the area under a Gaussian is independent of the mean of the Gaussian.

My derivation here is more direct than existing derivations of TLS, who apply Bayes' rule to statistical mechanics. In their treatment, the extra factor of z appears naturally. I can demonstrate the equivalence Equation 13 as follows. I solve Equation 15 for the exponential and substitute it into Equation 11

$$f(w) = \int dx\overline{p}(x)\ z(\beta)p(x|w) = z(\beta)g(w) \tag{16}$$

where $g(w)$ was defined as "the average generalization of the network" in (Tishby, Levin, Solla, 1989) "the sample average of the likelihood" of the network (Levin, Tishby, Solla, 1990) to be

$$g(w) \equiv \int dx\overline{p}(x)p(x|w). \tag{17}$$

I now substitute this into Equation 12 to get

$$\phi^{(m)} = \frac{z^{m+1}(\beta) \int dw\rho^{(0)}(w)g^{m+1}(w)}{z^{m}(\beta) \int dw\rho^{(0)}(w)g^{m}(w)} \tag{18}$$

which yields Equation 13 since Levin, Tishby, and Solla define APP as

$$\langle\langle p^{(m)}\rangle\rangle \equiv \frac{\langle\gamma^{m+1}\rangle_{\rho^{(0)}}}{\langle\gamma^{m}\rangle_{\rho^{(0)}}} \tag{19}$$

where I have used γ to distinguish their varable g from the function $g(w)$. They define the denominator of this expression as

$$\langle \gamma^m \rangle_{\rho^{(0)}} \int dw \rho^{(0)}(w) \int d\gamma \ \gamma^m \delta(g(w) - \gamma). \tag{20}$$

which can be evaluated with the Dirac delta function to get

$$\langle \gamma^m \rangle_{\rho^{(0)}} = \int dw \rho^{(0)}(w) \ g^m(w) \tag{21}$$

with a similar expression for the numerator, so that

$$\langle \langle p^{(m)} \rangle \rangle \ = \frac{\int dw \rho^{(0)}(w) g^{m+1}(w)}{\int dw \rho^{(0)}(w) g^m(w)}. \tag{22}$$

Equation 13 follows immediately from Equations 22 and 12. Therefore the average prediction fraction is identical (except for a scale factor $z(\beta)$) to the TLS average prediction probability under the conditions that TLS assume.

References

E. T. Jaynes. (1979) Where Do We Stand on Maximum Entropy?. In R. D. Leven and M. Tribus (Eds.), *Maximum Entropy Formalism*, M. I. T. Press, Cambridge, pages 17-118.

E. Levin, N. Tishby, and S. A. Solla. (1990) A Statistical approach to learning and generalization in layered neural networks. *Proceedings of the IEEE* , Vol. 78, No. 10, pages 1568-1574.

D. B. Schwartz, V. K. Samalan, S. A. Solla & J. S. Denker. (1990) Exhaustive Learning. *Neural Computation.*

N. Tishby, E. Levin, and S. A. Solla. (1989) Consistent inference of probabilities in layered networks: Predictions and generalization. *IJCNN*, IEEE, New York, pages II:403-410.