OBJECTIVE FUNCTIONS FOR NEURAL NETWORK CLASSIFIER DESIGN

Rod Goodman, John W. Miller
Department of Electrical Engineering
California Institute of Technology 116-81
Pasadena, CA 91125, USA

Padhraic Smyth
Communications Systems Research
Jet Propulsion Laboratory 238-420
Pasadena, CA 91109, USA

Abstract

Backpropagation was originally derived in the context of minimizing a mean-squared error (MSE) objective function. More recently there has been interest in objective functions that provide accurate class probability estimates. In this talk we derive necessary and sufficient conditions on the required form of an objective function to provide probability estimates. This leads to the definition of a general class of functions which includes MSE and cross entropy (CE) as two of the simplest cases. We establish the equivalence of these functions to Maximum Likelihood estimation and the more general principle of Minimum Description Length models. Empirical results are used to demonstrate the tradeoffs associated with the choice of objective functions which minimize to a probability.

Summary

The results we present are discussed in the context of feed forward neural network models. The results, however, are dependent on the training scheme (such as BEP), not on the actual modeling scheme (such as a neural network, a decision tree, or a truth table).

We have a set of N samples where each sample consists of a vector of feature measurements and a class label (say there are K features and m possible class labels, $m \geq 2$). Define the class to be a discrete m-ary variable C, and let us refer to the K-dimensional feature variable as $\underline{x}$. From the training data we seek to infer a classifier, where a classifier takes as input an unlabeled feature vector and produces as output posterior probability estimates of the classes, i.e., an estimate of the conditional probability of each class given a particular feature value $\underline{x}$, $\hat{p}(c_i|\underline{x}), 1 \leq i \leq m$. We will find the following notation convenient: for each of the i training samples let $c_i(j)$ be the true class, i.e., the given class label.

We will refer to the estimated network parameters (weights and biases) collectively by $\underline{\theta}$. The most widely used error function is the Mean-Squared Error (MSE) function defined as

$$E_{MSE} = \sum_{i=1}^{N}\sum_{k=1}^{m}(t_i(k) - o_i(k))^2 \tag{1}$$

where $t_i(k)$ is the "target" value for node k $o_i(k)$ is the network's output at node k ($o_i(k)$ is actually a function of the input features $\underline{x}(i)$ and the network parameters, but we ignore this dependence for notational convenience). Note that for labelled class data that $t_i(j) = 1, t_i(k) = 0, k \neq j, 1 \leq k \leq m$.

Two other objective functions have been proposed in the literature. The so-called cross-entropy (CE) measure (Hinton [1], Baum and Wilczek [2], Solla et al. [3]) is defined as

$$E_{CE} = \sum_{i=1}^{N}\left(\sum_{k=1}^{m} t_i(k)\log\frac{t_i(k)}{1-o_i(k)} + (1-t_i(k))\log\frac{1-t_i(k)}{1-o_i(k)}\right) \tag{2}$$

This definition is motivated by a desire to minimize the cross-entropy between the target distribution and the network estimate of the distribution for each class. In effect, it is a sum of binary cross-entropy measures for each node, rather than a true cross-entropy.

Consider now objective functions of the form

$$E = \sum_{i=1}^{N}\sum_{k=1}^{m} L(o_i(k), t_i(k))$$

The function $L(y,t)$ is said to minimize to a probability when the its minimum is achieved when y is equal to $\bar{t}$, the average value of t taken over the training samples. In this talk the conditions on $L(y,t)$ necessary for this to occur are developed. It is found that

$$L(y,t) = \int h'(y)\cdot\frac{y-t}{y}\cdot dy + C(t) \tag{3}$$

$$h'(y) > 0 \quad \text{for } 0 < y < 1 \tag{4}$$

where $h(y)$ is any smooth function which satisfies (4). If the symmetry condition $L(y,t) = L(1-y, 1-t)$ is imposed then $h(y)$ must also satisfy (5):

$$\frac{1-y}{y} = \frac{h'(1-y)}{h'(y)} \tag{5}$$

Equation (3) ensures the existence of a local extremum at $y = \bar{t}$, while (4) forces this to be the unique minimum. It follows from (5) that at least one of the following cases must be true:

$h'(y)$ has a zero at $y = 0$ or $h'(y)$ has a pole at $y = 1$

The simplest functions satisfying the above restrictions are:

$$h'(y) = y \quad \Rightarrow \text{MSE}$$

$$h'(y) = \frac{1}{1-y} \quad \Rightarrow \text{CE}$$

Using these equations one can propose a variety of other more complicated objective functions which minimize to a probability. Each of the MSE and CE loss functions for probability estimation have advantages and drawbacks in their own right — we discuss the general conditions under which their use is appropriate.

For the case $m = 2$ using the CE objective function is equivalent to using maximum likelyhood (ML) as shown by Baum and Wilczek. But for non-binary classes, with non-probabilistic training labels, using the CE criterion amounts to independent ML estimation of each class, ignoring the other class information. The more direct ML procedure in this case is equivalent to an objective function defined as

$$E_{MMI} = \sum_{i=1}^{N}\log\frac{1}{o_i(j)} \tag{6}$$

which has been shown by Bridle to be the the neural equivalent of the Maximum Mutual Information (MMI) criterion used for Hidden Markov Model parameter estimation [4].

The research described in this talk was carried out in part by the Jet Propulsion Laboratories, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. In addition this work was supported in part by the Air Force Office of Scientific Research under grant number AFOSR-90-0199.

1. G. E. Hinton, 'Connectionist learning procedures,' *Connectionist Learning Procedures*, Technical Report CMU-CS-87-115, Carnegie Mellon University, 1987.
2. E. Baum and F. Wilczek, 'Supervised learning of probability distributions by neural networks,' in *Neural Information Processing Systems*, pages 52–61, American Institute of Physics, 1988.
3. S. Solla, E. Levin, and M. Fleisher, 'Accelerated learning in layered neural networks,' *Complex Systems*, January 1989.
4. J. Bridle, 'Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters,' in *Advances in Neural Information Processing II* (ed. D. Touretzky), Morgan Kaufmann, pp.211–217, 1990.
5. A. Barron and R. Barron, 'Statistical learning networks: a unifying view,' presented at the *1988 Symposium on the Interface: Statistics and Computing Science*, Virginia.