Objective Functions for Probability Estimation

John W. Miller, Rod Goodman
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 paper 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.

Introduction

The results we present in this paper are discussed in the context of neural network models. The analysis however applies to non-network models also. Let $\underline{a}$ be a training input (vector) to the network and let $\underline{\theta}$ represent a particular set of network parameters, i.e., network weights and biases. Let $x(\underline{a},\underline{\theta})$ represent the network output for a particular input $\underline{a}$ and a parameter set $\underline{\theta}$ — consider for now a network with just a single such output unit. For the purposes of this paper we adopt the convention of dropping the explicit reference to $\underline{a}$ and $\underline{\theta}$ and instead refer to $x = x(\underline{a},\underline{\theta})$, for reasons which will become clear as we proceed.

Two well known objective functions (or loss functions) used to train a neural network are the mean square error:

$$L_{se}(x,t) = (x-t)^2$$

and the cross entropy [1, 2, 3]:

$$L_{ce}(x,t) = t \cdot \log\left(\frac{t}{x}\right) + (1-t)\log\left(\frac{1-t}{1-x}\right)$$

Here x is the output of the neural network and t is the target value. During learning, the network seeks a set of parameters $\underline{\theta}$ which minimize the expected value of the objective function. The two objective functions above have the important property that the minima of their expected value is achieved when x is equal to the expected value of t (as shown in [2]). The expected value is taken over all training samples, and the minima is taken over all possible mappings of the training sample inputs to an output x. A *possible mapping* is any mapping of input samples to outputs except for those mappings where two identical input patterns are mapped to two different outputs. In the context of neural networks this says that if a sufficiently powerful network is trained with a set of input-target pairs, the output during testing will be the average value of the target taken over all training samples which had the same input pattern as the test input. A ***sufficiently powerful network*** is one which can implement the actual mappings which minimize the expected error. If t is binary $\{0,1\}$, indicating the truth of some event, then the expected value of t is the probability that the event is true — in other words, stating that an objective function minimizes to a probability is equivalent to stating that the network output is an unbiased estimator of the posterior probability of the class given a particular input $\underline{a}$. More generally, if t is a sampled estimate of the probability of some event being true (with the number of samples used to calculate each t fixed), then (again) the expected value of t is the probability of the output being true. For this reason, these objective functions are said to minimize to a probability.

Note that for a sufficiently powerful network the expected error for all input samples is minimized when the expected error for each unique input sample is minimized. Since this analysis deals only with these ideal networks, all probabilities and function values can be understood to be conditioned on input $\underline{a}$, where input

0-7803-0164-1/91/0000-0881$01.00©1991 IEEE

$\underline{a}$ is some arbitrary fixed input pattern. The question arises, why study this ideal network? One reason is that the outputs of real networks in some sense approximate the outputs of this ideal network. For instance when the MSE object function is used, it has been shown the real network will minimize the average squared difference between its outputs and the outputs of the ideal network [4, 5]. Although the results presented in this paper deal with a network with a single output x, they also apply more generally to multiple output ideal networks. This is because these networks can minimize the error to each output independently, making the network equivalent to separate, single-output, ideal networks.

Minimization to a Probability

Necessary and Sufficient Conditions

Define $h(x) = L(x, 0)$ to be the value of the objective function when the target is zero. Let $p(t)$ be a probability function defined for $t \in [0, 1]$. The notation of a bar written over a variable indicates the expected value taken with respect to probability $p(t)$:

$$\bar{t} = \int_0^1 t \cdot p(t) dt = \text{ expected value of } t,$$

and

$$\overline{L}(x) = \int_0^1 p(t) \cdot L(x, t) dt.$$

An objective function is said to "minimize to a probability" if the following condition holds.

$$\begin{array}{c} \forall \; p(t) \;\; \text{s.t.} \;\; \int_0^1 p(t) \cdot dt = 1, \;\; \bar{t} \in (0, 1) \\ \min\limits_{0<x<1} \overline{L}(x) \; = \; \overline{L}(\bar{t}) \end{array} \qquad (c1)$$

In appendix A it is shown that minimization to a probability is equivalent to the restrictions ($r1$) and ($r2$).

$$L(x, t) = \int h'(x) \cdot \frac{x - t}{x} \cdot dx + C \qquad (r1)$$

$$h'(x) > 0 \quad \text{for} \;\; 0 < x < 1 \qquad (r2)$$

This result shows how the value of an objective function for all values of t may be determined by the function at $t = 0$.

Parameterization of h(x)

Since $L(x, t)$ is used to minimize the error to a probability, one natural restriction on the class of all error functions is to require a symmetry between $t = 1$ ("True") and $t = 0$ ("False"). When the choice of labels "True" and "False" are arbitrary, the following symmetry condition ensures that probability estimates are not affected by the choice of labels:

$$L(x, t) = L(1 - x, 1 - t) \qquad (c2)$$

A final condition which will be useful is to require smoothness:

$$L_x(x, t) \;\; \text{and all its partial derivatives w.r.t. } x \text{ exist for } x \in (0, 1) \qquad (c3)$$

We show in Appendix B that minimization to a probability restricts the form of $h(x)$, specifically $h(x)$ must satisfy (r3),(r4) and (r5).

$$\frac{1-x}{x} = \frac{h'(1-x)}{h'(x)}, \qquad (r3)$$

Hampshire and Pearlmutter [6] independently arrived at equation (r3) for the case where the targets are binary $\{0,1\}$. In this paper we show that this result applies to objective function analysis for more general distributions $p(t)$. The other restrictions on the form of $h(x)$ are:

$$h^{(n+1)}(.5) = 2n \cdot h^{(n)}(.5) \quad \text{(for } n \text{ odd)}, \qquad (r4)$$

(here $h^{(n)}$ represents the n^{th} derivative of h),
and

$$\exists\, C_1, C_3, \ldots \;\; s.t. \;\; h(x) = \sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} C_n\, (x-.5)^n\, (2nx+1). \qquad (r5)$$

The logical relation between these results and the stated conditions is:

$$((c1)\cdot(c3)) \Rightarrow ((c2) \Leftrightarrow (r3) \Leftrightarrow (r4) \Leftrightarrow (r5))$$

Combined with the conditions for minimization to a probability, this result shows that all smooth objective functions which obey the logical symmetry condition can be generated by choosing a $h(x)$ function which satisfies $(r2)$ and any one of $(r3)$, $(r4)$, or $(r5)$. This $h(x)$ is then substituted into equation $(r1)$ to get the objective function $L(x,t)$.

It follows from $(r3)$ that at least one of the following cases must be true:
$h'(x)$ has a zero at $x=0$ or $h'(x)$ has a pole at $x=1$

The simplest functions satisfying the above restriction are:

$$h_1'(x) = x$$
$$h_2'(x) = \frac{1}{1-x}$$

By substitution into $(r1)$, it is seen that h_1 defines an objective function

$$L_1(x,t) = .5x^2 - xt$$

Since additive and multiplicative values independent of x do not change the minimization, it is seen that $L_1(x,t)$ is equivalent to $L_{se} = (x-t)^2$. Similarly h_2 may be substituted into $(r3)$ to generate the cross entropy objective function.

A result of $(r4)$ is that the *only* function $L(x,t)$ for which $\overline{L}(x)$ is symmetric about $\overline{t}$ is the squared error objective function. However, depending on the particular problem, a symmetric error measure need not be the most appropriate. As discussed by El-Jaroudi and Makhoul [2], and Gish [5], in applications such as speech it is the *relative error* which is most critical for success, in which case objective functions such as the cross-entropy function will yield better results than the squared error function.

Since the derived conditions are *sufficient*, equations $(r1)$, $(r3)$, $(r4)$ and $(r5)$ may be used to generate equations for other objective functions which minimize to a probability. For example the following three:

$$h_3'(x) = C \cdot x - x^2 + x^3 \quad \text{for} \quad C > 1.5$$
$$h_4'(x) = \sqrt{\frac{(1-x)}{x}}$$
$$h_5'(x) = x + \frac{C}{1-x}$$

were found by equations ($r3$) and ($r5$) to be members of this general class.

Likewise, since the conditions derived in this paper are *necessary* for minimization to a probability, functions which do *not* satisfy the requirements will not produce accurate probability estimates if used as objective functions. For example, the class of L_p norm objective functions as proposed by Burrascano [7], do not minimize to a probability except for $p = 2$. In particular there is a distinction between *class discrimination* and *class probability estimation*. In the former it is sufficient to approximate the optimal Bayes disriminant by simply identifying the most likely class given any input $\underline{a}$. In the latter case one wishes not only to discriminate among the classes but also to accurately estimate their individual posterior probabilities. Such information is necessary in the general statistical decision framework such as, for example, in a medical decision problem where there are tangible costs associated with various actions and their outcomes.

Conclusion

We have generalized and extended previously known results on the topic of obtaining probability estimates from neural network classifiers. In particular, we derived necessary and sufficient conditions for an objective function which minimizes to a probability. The objective function $L(x,t)$ was found to be uniquely specified by the function $L(x,0)$. This function $L(x,0)$ was found to satisfy further restrictions when a condition of logical symmetry is required. These restrictions and the relation between $L(x,t)$ and $L(x,0)$ define the class of all objective functions which minimize to a probability. The two simplest functions in this class were found to be the well-known MSE and CE objective functions.

Acknowledgments

The research described in this paper was supported in part by DARPA under grant number AFOSR–90–0199 and by NSF grant ENG–8711673. In addition this work was carried out in part by the Jet Propulsion Laboratories, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

References

1. 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.
2. A. El-Jaroudi and J. Makhoul, "A New Error Criterion For Posterior Probability Estimation With Neural Nets", in *Proc. 1990 Int. Joint Conf. Nerual Networks*, pp. III-185-192, San Diego, June 1990.
3. S. Solla, E. Levin, and M. Fleisher, "Accelerated learning in layered neural networks," *Complex Systems*, January 1989.
4. H. White, "Learning in artificial neural networks: a statistical perspective," *Neural Computation*, 1(4), pp.425–464, 1990.
5. H. Gish, "A probabilistic approach to the understanding and training of neural network classifiers," in *Proceeding of the 1990 IEEE Conference on Acoustics, Speech and Signal Processing*, pp.1361–1364, 1990.
6. J. Hampshire, B. Pearlmutter, "Equivalence proofs for multi-layer perceptron classifiers and the Bayesian discriminant function," in *Proceedings of the 1990 Connectionist Models Summer School*, Morgan Kaufmann, 1990.
7. P. Burrascano, "A norm selection criterion for the generalized delta rule," *IEEE Trans. on Neural Networks*, vol.2, no.1, pp.125–130, 1991.

Appendix A

Conditions for Minimization to a Probability

Proof: $(c1) \Leftrightarrow \big((r1)\cdot(r2)\big)$
(Please refer to the body of the paper for appropriate equation references.)

First it will be shown that $\big((r1)\cdot(r2)\big) \Rightarrow (c1)$. Substitute $(r1)$ into the formula for the expected value of the objective function:

$$\overline{L}(x) = \int_0^1 p(t)\Big[\int h'(x)\frac{x-t}{x}dx + C\Big]\ dt.$$

Take the derivative with respect to x:

$$\overline{L}_x(x) = \int_0^1 p(t)h'(x)\Big(\frac{x-t}{x}\Big)dt. \qquad (a1)$$

A local minimum can occur for $x \in (0,1)$ if and only if:

$$\overline{L}_x(x) = 0 \quad \text{and} \quad \overline{L}_{xx}(x) > 0.$$

From $(a1)$:

$$\begin{aligned}\overline{L}_x(x) &= h'(x)\int_0^1 p(t)dt - \frac{h'(x)}{x}\int_0^1 t\ p(t)dt \\ &= h'(x)\Big(1-\frac{\overline{t}}{x}\Big).\end{aligned} \qquad (a2)$$

Given $(r2)$ it is clear that $\overline{L}_x(x) = 0$ for $x \in (0,1)$ if and only if $x = \overline{t}$. Taking derivatives again shows that this extremum represents a minimum:

$$\overline{L}_{xx}(x) = h''(x) - h''(x)\ \frac{\overline{t}}{x} + \frac{\overline{t}}{x^2} + h'(x)\ \frac{\overline{t}}{x^2}$$
$$\overline{L}_{xx}(\overline{t}) = \frac{h'(\overline{t})}{\overline{t}} > 0.$$

Thus it has been shown $\big((r1)\cdot(r2)\big) \Rightarrow (c1)$. In order to show equivalence it remains to be proven that $(c1) \Rightarrow \big((r1)\cdot(r2)\big)$. Condition $(c1)$ certainly requires:

$$\int_0^1 p(t)L_x(x,t)dt = 0 \ \text{ at } \ x = \overline{t}. \qquad (a3)$$

Without loss of generality let

$$L_x(x,t) = g(x,t)\cdot(x-t). \qquad (a4)$$

Substituting this into $(a3)$:

$$\int_0^1 p(t)g(x,t)(x-t)dt = 0 \ \text{ at } \ x = \overline{t}. \qquad (a5)$$

Now consider the distribution

$$p(t) = \begin{cases} \rho, & \text{if } t = t_1; \\ (1-\rho), & \text{if } t = t_2; \quad (0 < t_1 < t_2 < 1) \\ 0, & \text{otherwise.}\end{cases} \qquad (a6)$$

Condition $(c1)$ requires that the minima be at $\overline{t}$ for all $p(t)$, so it must be true for the particular distribution given in $(a6)$. Notice $\overline{t} = \rho t_1 + (1-\rho)t_2$. Evaluating $(a5)$ with this distribution gives:

$$\rho g(x,t_1)(\overline{t}-t_1) + (1-\rho)g(x,t_2)(\overline{t}-t_2) = 0 \ \text{ at } \ x = \overline{t},$$

$$\rho g(\bar{t}, t_1)[\rho t_1 + (1-\rho)t_2 - t_1] + (1-\rho)g(\bar{t}, t_2)[\rho t_1 + (1-\rho)t_2 - t_2] = 0,$$
$$\rho(1-\rho)(t_2 - t_1)(g(\bar{t}, t_1) - g(\bar{t}, t_2)) = 0.$$

Therefore $g(\bar{t}, t_1) = g(\bar{t}, t_2)$, where ρ can be chosen to set $\bar{t}$ arbitrarily in (t_1, t_2). Thus $g(x, t_1) = g(x, t_2))$ for any $t_1 < x < t_2$. Therefore $g(x,t)$ is independent of t, so we may write:

$$L_x(x,t) = g(x) \cdot (x - t).$$

Evaluating at $t = 0$ and using the definition of $h(x)$ shows:

$$g(x) = \frac{h'(x)}{x}.$$

Integrating and substituting into (a4) gives the desired result (r1). The result $(c1) \Rightarrow (r2)$ follows easily from the requirement that the unique extremum found by $(r1)$, be a minimum rather than a maximum. Since $(c1) \Rightarrow ((r1) \cdot (r2))$ and $((r1) \cdot (r2)) \Rightarrow (c1)$, it has been shown $(c1) \Leftrightarrow ((r1) \cdot (r2))$. *Q.E.D.*

Appendix B

Restrictions on $h(x)$

Claim: $((c1) \cdot (c3)) \Rightarrow ((c2) \Leftrightarrow (r3) \Leftrightarrow (r4) \Leftrightarrow (r5))$

Proof: $((c1) \cdot (c2)) \Rightarrow (r3)$

By $(c1)$ the equilibrium equation $(a3)$ must hold for distribution $(a6)$:

$$\rho L_x(x,1) + (1-\rho)L_x(x,0) = 0 \text{ at } x = \rho.$$

Using $(c2)$:

$$\rho(-1)L_x(1-x,0) + (1-\rho)L_x(x,0) = 0 \text{ at } x = \rho.$$
$$\frac{1-\rho}{\rho} = \frac{h'(1-\rho)}{h'(\rho)}$$

Since ρ is arbitrary $\in (0,1)$, equation (r3) is proven. *Q.E.D.*

Proof: $((r3) \cdot (c3)) \Rightarrow (r4)$

Rewrite $(r3)$:

$$xh'(1-x) = (1-x)h'(x) \qquad (a7)$$

Now examine the successive derivatives of the left hand side (LHS) of this equation:

$$LHS(x) = xh'(1-x)$$
$$LHS'(x) = h'(1-x) + xh''(1-x)$$
$$\vdots$$
$$LHS^{(n)}(x) = (-1)^n \left[xh^{(n+1)}(1-x) - nh^{(n)}(1-x)\right]. \qquad (a8)$$

Similarly, for the right hand side of the equation:

$$RHS^{(n)}(x) = -nh^{(n)}(x) + (1-x)h^{(n+1)}(x). \qquad (a9)$$

Equating $(a8)$ with $(a9)$ and evaluating at $x = .5$ gives $(r4)$. *Q.E.D.*

Proof: $((c3) \cdot (r4)) \Rightarrow (r5)$

This follows by expanding $h(x)$ in a Taylor series about .5:

$$h(x) = \sum_{i=1}^{\infty} h^{(i)}(.5)(\frac{1}{i!})(x - .5)^i$$

Using $(r4)$

$$h(x) = \sum_{i \text{ odd}} h^{(i)}(.5)(\frac{1}{i!})(x - .5)^i \left(1 + 2i(\frac{x - .5}{i+1})\right)$$

Let $C_i = h^{(i)}(.5)\frac{1}{(i+1)!}$. Substitute to find (r5). *Q.E.D.*

The remaining proofs required to establish the claim follow easily by substituting $(r3)$, $(r4)$, or $(r5)$ into the equation given by $(r1)$ to establish $(c2)$.