Draft: Error Measures that make Outputs Probabilities

Barak A. Pearlmutter
Department of Computer Science and Engineering
Oregon Graduate Institute
19600 NW von Neumann Drive
Beaverton, OR 97006-1999
bap@cse.ogi.edu

DRAFT December 21, 1992 DRAFT

Abstract

When training a backpropagation network with binary labeled data, where the labels are generated stochastically from an underlying probability associated with each possible input, some error measures have the property that minimizing the error measure leads asymptotically to outputs which correctly estimate the involved probabilities. Below, we derive a necessary and sufficient condition for an error measure to have this property, solve the condition in general, and exhibit some families of such error measures.

1 The reasonable condition

Consider training a backpropagation network (Rumelhart et al., 1986) to perform a binary classification task using exemplars with binary labels, where the binary labels are generated stochastically from an underlying probability associated with each potential input. Let us further assume that this function is within the bias of the network, and that it is learnable.

We will denote by $E(y, d)$ the error measure by which we calculate the loss associated with an output of y when the desired output is d. We will assume that $0 \le y \le 1$, and that $E(y, d) = E(1 - y, 1 - d)$. Since the desired output d is binary, we will use $p = P(d = 1)$. Let us derive conditions under which determining y by the minimization of $\sum_i E(y, d_i)$ causes y to asymptotically converge to p. We will call an error measure with this property *reasonable* because it leads to outputs which can be reasonably interpreted as probabilities.

Define $f(y) = E(y, 0)$, and note that $f(1 - y) = E(y, 1)$. Let us compute the expected error

$$\begin{aligned} \mathcal{E}(y) &= (1-p)E(y,0) + pE(y,1) \\ &= (1-p)f(y) + pf(1-y) \end{aligned}$$

which is attained in the limit of many exemplars. Setting $\mathcal{E}'(y) = 0$ gives $(1 - p)f'(y) = pf'(1 - y)$ and if this holds exactly when $y = p$ then $\mathcal{E}$ has either a minimum or maximum at $y = p$. Since $p/(1 - p)$ is increasing in the range of interest, this condition is simply

$$\frac{f'(p)}{f'(1-p)} = \frac{p}{1-p}. \qquad (1)$$

To ensure that this is a minimum, it is necessary that

$$(1-p)f''(p) + pf''(1-p) > 0.$$

Taking the derivative of (1) gives $f''(1-p) = f'(p)/p^2 - (1-p)f''(p)/p$ and substituting this into the above inequality simplifies it to

$$f'(p) > 0 \text{ for } 0 < p < 1. \qquad (2)$$

If both (1) and (2) hold then f determines a reasonable error measure.

2 A general solution

If we choose a continuous function $g(y)$ for $0 < y \le \frac{1}{2}$ where $g(y) > 0$ then $f(y)$ defined by

$$f'(y) = \begin{cases} g(y) & \text{for } 0 < y \le \frac{1}{2} \\ \frac{y}{1-y} g(1-y) & \text{for } \frac{1}{2} \le y < 1 \end{cases} \qquad (3)$$

results in a reasonable error measure.

It is easy to see from (3) that if $f(y)$ satisfies the reasonable condition then $af(y)$ does as well, for $a > 0$, and that if two functions each satisfy it then so does their sum.

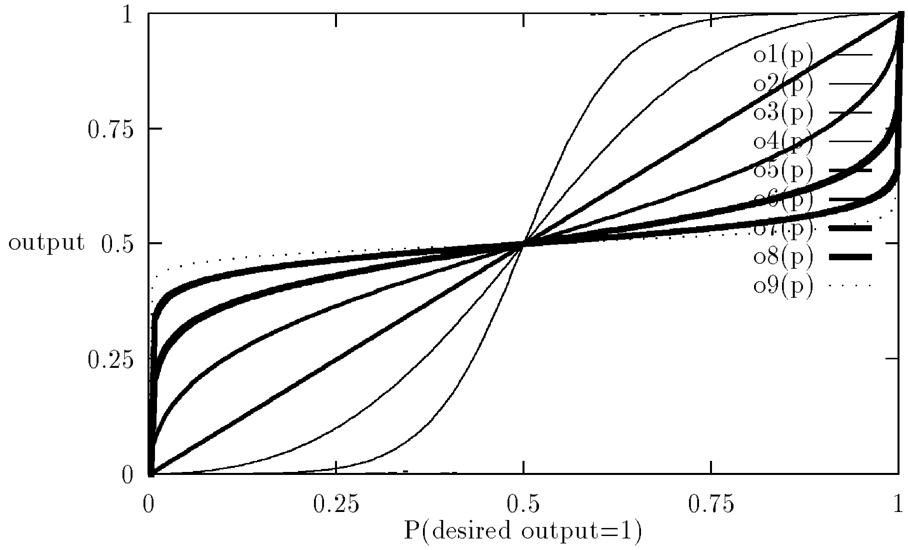

Figure 1: Expected minimal error output plotted as a function of the probability of a binary labeling when an L_R error metric is used, for $R = 1.0625, 1.125, 1.25, 1.5, 2, 3, 5, 9, 17$. Note how the output is biased towards certainty for $R < 2$ and away from it for $R > 2$.

3 Some specific reasonable error functions

One family of reasonable functions is given by

$$f(y) = \int y^r (1-y)^{r-1} dy. \tag{4}$$

This family has two special cases of particular importance.

3.1 $r = 0$: cross entropy

One function that satisfies the reasonable condition is $f(y) = \int (1-y)^{-1} dy = -\log(1-y)$, the well known cross entropy error measure (Hinton, 1987).

3.2 $r = 1$: squared error

The standard squared error measure, $f(y) = \int y dy = y^2/2$, is also a special case of (4).

3.3 Combinations of the two

Both squared error and cross entropy have information theoretic justifications as the description length of the residuals, with squared error resulting from an assumption of Gaussian perturbations while cross entropy assumes binary targets. Squared error has the pleasant property that the output converges to the average of the desired outputs even for non-binary data, while cross entropy performs a proper combination of independently sampled probability estimates of noiseless binary data. It might be desirable to blend the two, and if so the most general reasonable combination of squared error and cross entropy is

$$f(y) = ay^2 - b \log(1-y)$$

where $b \geq 0$ and $a > -2b$.

3.4 L_R metrics

A number of researchers, e.g. (Hanson and Burr, 1988), have advocated the use of L_R metrics. Burrascano (Burrascano, 1991) has shown that L_R metrics are optimal in the MaxEnt sense for real desired outputs perturbed by generalized Gaussian noise, where the characteristics of the noise determine the proper R; however, after this careful derivation, he applies L_R metrics to problems with binary labels, which do not satisfy the conditions under which he showed their optimality.

For an L_R metric, we have $f(y) = y^R$, so the reasonable condition is $p^{R-2} = (1-p)^{R-2}$ or $R = 2$. Another way of seeing this is that $\mathcal{E}$ is minimized when

$$y = \frac{\sqrt[R-1]{p}}{\sqrt[R-1]{1-p} + \sqrt[R-1]{p}} \tag{5}$$

which simplifies to the above. Since an L_R metric is reasonable only when $R = 2$, this argues against using L_R error metrics (other than L_2) when the output is being interpreted as a probability.

4 Three categories of reasonable functions

Using the above it is simple to construct new reasonable error functions. For instance, the functions $\sin^{-1}(2y-1)-2\sqrt{y(1-y)}$, $\sqrt{y/(1-y)}$, and $1/(1-y)+\ln y/(1-y)$ are all reasonable. Such reasonable error measures can be classified depending on whether (a) $\lim_{y\to 0^+} f(y) = -\infty$ and (b) $\lim_{y\to 1^-} f(y) = \infty$. It can be easily shown that this is determined by the order of the pole at 1, which we call o. Only three of the four possible combinations of (a) and (b) can occur.

- $o = 0$. Neither (a) nor (b) hold, in which case $f(y)$ is bounded, as in squared error.
- $0 < o \leq 1$. (b) holds but not (a), so the error is unbounded above, but bounded below. Cross entropy is such an error measure.
- $1 < o$. Both (a) and (b) hold, so the error is unbounded above, and also unbounded below. This class of reasonable error measure, typified by something like $3y - 2/y(1-y)^2$, is not found in practice, and would appear impractical, as it would seem to encourage overtraining.

5 Conclusion

In order to integrate neural networks into larger system under a uniform probabilistic framework, it is typically necessary that the network's outputs be interpretable as probabilities. If an extra layer of interpretation, as in (Denker and le Cun, 1991), is not desired, and local error measures are to be used, this can only be done with reasonable error measures. However, there are many such reasonable error measures, and a principled way to choose one as opposed to another, in the absence of strong prior knowledge of the noise and a-priori distributions, still awaits us.

Hampshire has shown (Hampshire and Pearlmutter, 1990; Hampshire, 1991) that if the network is in a forced classification setting, reasonable error functions are sufficient but not necessary to achieve Bayesian decision performance, as the intermediate step of interpreting the outputs as probabilities before forcing the classification is not in general needed.

6 Acknowledgments

I would like to thank John B. Hampshire II for our many conversations, for his dogged determination, and for the rare pleasure of occasionally discovering that he has not anticipated me. Some of this work was carried out while the author was a Hertz fellow. An abbreviated version of this document appeared as a section of a workshop paper (Hampshire and Pearlmutter, 1990).

References

Burrascano, P. (1991). A norm selection criterion for the generalized delta rule. *IEEE Transactions on Neural Networks*, 2(1).

Denker, J. S. and le Cun, Y. (1991). Transforming neural-net output levels to probability distributions. In Lippmann, R. P., Moody, J. E., and Touretzky, D. S., editors, *Advances in Neural Information Processing Systems 3*, pages 839–845. Morgan Kaufmann.

Hampshire, II, J. B. (1991). Multi-layer perceptron classifiers and the Bayesian discriminant function. Unpublished Ph.D. thesis proposal.

Hampshire, II, J. B. and Pearlmutter, B. A. (1990). Equivalence proofs for multi-layer perceptron classifiers and the Bayesian discriminant function. In Touretzky, D. S., Elman, J. L., Sejnowski, T. J., and Hinton, G. E., editors, *Connectionist Models: Proceedings of the 1990 Summer School*, pages 159–172. Morgan Kaufmann.

Hanson, S. J. and Burr, D. J. (1988). Minkowski-r back-propagation: Learning in connectionist model with non-euclidean error signals. In Anderson, D. Z., editor, *Neural Information Processing Systems*, pages 348–357, New York, New York. American Institute of Physics.

Hinton, G. E. (1987). Connectionist learning procedures. Technical Report CMU-CS-87-115, Carnegie Mellon University, Pittsburgh, PA 15213.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning internal representations by error propagation. In Rumelhart, D. E., McClelland, J. L., and the PDP research group., editors, *Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundations*. MIT Press.