

For a 1-of- $c$  coding scheme this minimum value is 0. However, the error function (6.148) is still valid, as we shall see, when  $t_k^n$  is a continuous variable in the range  $(0, 1)$  representing the probability that input  $\mathbf{x}^n$  belongs to class  $C_k$ . In this case the minimum of the error function need not vanish (it represents the entropy of the distribution of target variables, as will be discussed shortly). It is then convenient to subtract off this minimum value, and hence obtain the error function in the form

$$E = - \sum_n \sum_{k=1}^c t_k^n \ln \left( \frac{y_k^n}{t_k^n} \right) \quad (6.150)$$

which is non-negative, and which equals zero when  $y_k^n = t_k^n$  for all  $k$  and  $n$ .

We now consider the corresponding activation function which should be used for the network output units. If the output values are to be interpreted as probabilities they must lie in the range  $(0, 1)$ , and they must sum to unity. This can be achieved by using a generalization of the logistic sigmoid activation function which takes the form

$$y_k = \frac{\exp(a_k)}{\sum_{k'} \exp(a_{k'})} \quad (6.151)$$

which is known as the normalized exponential, or *softmax* activation function (Bridle, 1990). The term softmax is used because this activation function represents a smooth version of the *winner-takes-all* activation model in which the unit with the largest input has output +1 while all other units have output 0. If the exponentials in (6.151) are modified to have the form  $\exp(\beta a_k)$ , then the winner-takes-all activation is recovered in the limit  $\beta \rightarrow \infty$ . The softmax activation function can be regarded as a generalization of the logistic function, since it can be written in the form

$$y_k = \frac{1}{1 + \exp(-A_k)} \quad (6.152)$$

where  $A_k$  is given by

$$A_k = a_k - \ln \left\{ \sum_{k' \neq k} \exp(a_{k'}) \right\}. \quad (6.153)$$

As with the logistic sigmoid, we can give a very general motivation for the softmax activation function by considering the posterior probability that a hidden unit activation vector  $\mathbf{z}$  belongs to class  $C_k$ , in which the class-conditional densities are assumed to belong to the family of exponential distributions of the general form

$$p(\mathbf{z}|C_k) = \exp \left\{ A(\theta_k) + B(\mathbf{z}, \phi) + \theta_k^T \mathbf{z} \right\}. \quad (6.154)$$

From Bayes' theorem, the posterior probability of class  $C_k$  is given by

$$p(C_k|\mathbf{z}) = \frac{p(\mathbf{z}|C_k)P(C_k)}{\sum_{k'} p(\mathbf{z}|C_{k'})P(C_{k'})}. \quad (6.155)$$

Substituting (6.154) into (6.155) and re-arranging we obtain

$$p(C_k|\mathbf{z}) = \frac{\exp(a_k)}{\sum_{k'} \exp(a_{k'})} \quad (6.156)$$

where

$$a_k = \mathbf{w}_k^T \mathbf{z} + w_{k0} \quad (6.157)$$

and we have defined

$$\mathbf{w}_k = \theta_k \quad (6.158)$$

$$w_{k0} = A(\theta_k) + \ln P(C_k). \quad (6.159)$$

The result (6.156) represents the final layer of a network with softmax activation functions, and shows that (provided the distribution (6.154) is appropriate) the outputs can be interpreted as probabilities of class membership, conditioned on the outputs of the hidden units.

In evaluating the derivatives of the softmax error function we need to consider the inputs to all output units, and so we have (for pattern  $n$ )

$$\frac{\partial E^n}{\partial a_k} = \sum_{k'} \frac{\partial E^n}{\partial y_{k'}} \frac{\partial y_{k'}}{\partial a_k}. \quad (6.160)$$

From (6.151) we have

$$\frac{\partial y_{k'}}{\partial a_k} = y_{k'} \delta_{kk'} - y_{k'} y_k \quad (6.161)$$

while from (6.150) we have

$$\frac{\partial E^n}{\partial y_{k'}} = -\frac{t_{k'}}{y_{k'}}. \quad (6.162)$$

Substituting (6.161) and (6.162) into (6.160) we find

$$\frac{\partial E^n}{\partial a_k} = y_k - t_k \quad (6.163)$$

which is the same result as found for both the sum-of-squares error (with a linear activation function) and the two-class cross-entropy error (with a logistic activation function). Again, we see that there is a natural pairing of error function and activation function.

### 6.10 Entropy

The concept of entropy was originally developed by physicists in the context of equilibrium thermodynamics and later extended through the development of statistical mechanics. It was introduced into information theory by Shannon (1948). An understanding of basic information theory leads to further insights into the entropy-based error measures discussed in this section. It also paves the way for an introduction to the minimum description length framework in Section 10.10. Here we consider two distinct but related interpretations of entropy, the first based on *degree of disorder* and the second based on *information content*.

Consider a probability density function  $p(x)$  for a single random variable  $x$ . It is convenient to represent the density function as a histogram in which the  $x$ -axis has been divided into bins labelled by the integer  $i$ . Imagine constructing the histogram by putting a total of  $N$  identical discrete objects into the bins, such that the  $i$ th bin contains  $N_i$  objects. We wish to count the number of distinct ways in which objects can be arranged, while still giving rise to the same histogram. Since there are  $N$  ways of choosing the first object,  $(N-1)$  ways of choosing the second object, and so on, there are a total of  $N!$  ways to select the  $N$  objects. However, we do not wish to count rearrangements of objects within a single bin. For the  $i$ th bin there are  $N_i!$  such rearrangements and so the total number of distinct ways to arrange the objects, known as the multiplicity, is given by

$$W = \frac{N!}{\prod_i N_i!}. \quad (6.164)$$

The entropy is defined as (a constant times) the negative logarithm of the multiplicity

$$S = -\frac{1}{N} \ln W = -\frac{1}{N} \{\ln N! - \sum_i \ln N_i!\}. \quad (6.165)$$

We now consider the limit  $N \rightarrow \infty$ , and make use of Stirling's approximation  $\ln N! \simeq N \ln N - N$  together with the relation  $\sum_i N_i = N$ , to give

$$S = -\lim_{N \rightarrow \infty} \sum_i \left( \frac{N_i}{N} \right) \ln \left( \frac{N_i}{N} \right) = -\sum_i p_i \ln p_i \quad (6.166)$$

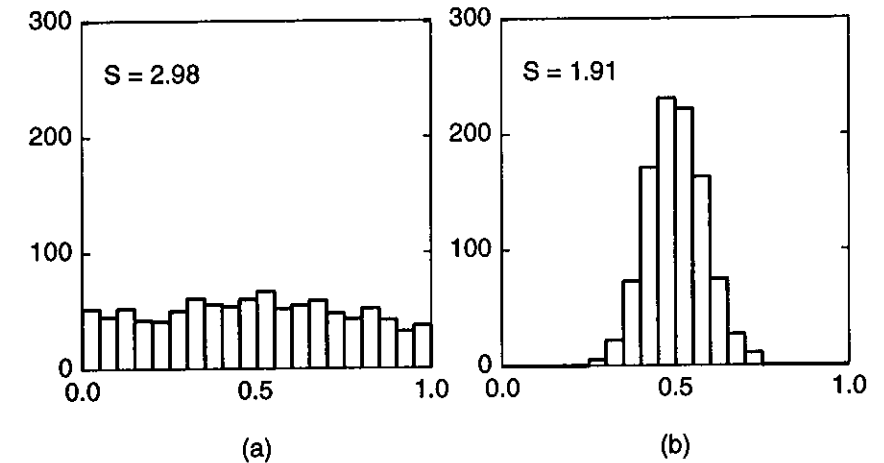


Figure 6.13. Examples of two histograms, together with their entropy values defined by (6.166). The histograms were generated by sampling two Gaussian functions with variance parameters  $\sigma = 0.4$  and  $\sigma = 0.08$ , and each contain 1000 points. Note that the more compact distribution has a lower entropy.

where  $p_i = N_i/N$  (as  $N \rightarrow \infty$ ) represents the probability corresponding to the  $i$ th bin. The entropy therefore gives a measure of the number of different *microstates* (arrangements of objects in the bins) which can give rise to a given *macrostate* (i.e. a given set of probabilities  $p_i$ ). A very sharply peaked distribution has a very low entropy, whereas if the objects are spread out over many bins the entropy is much higher. The smallest value for the entropy is 0 and occurs when all of the probability mass is concentrated in one bin (so that one of the  $p_i$  is 1 and all the rest are 0). Conversely the largest entropy arises when all of the bins contain equal probability mass, so that  $p_i = 1/M$  where  $M$  is the total number of bins. This is easily seen by maximizing (6.166) subject to the constraint  $\sum_i p_i = 1$  using a Lagrange multiplier (Appendix C). An example of two histograms, with their respective entropies, is shown in Figure 6.13.

For continuous distributions (rather than histograms) we can take the limit in which the number  $M$  of bins goes to infinity. If  $\Delta$  is the width of each bin, then the probability mass in the  $i$ th bin is  $p_i = p(x_i)\Delta$ , and so the entropy can be written in the form

$$S = \lim_{M \rightarrow \infty} \sum_{i=1}^M p(x_i)\Delta \ln \{p(x_i)\Delta\} \quad (6.167)$$

$$= \int p(x) \ln p(x) dx + \lim_{M \rightarrow \infty} \ln \Delta \quad (6.168)$$