Following the same argument as for a single target variable, we see that maximizing the likelihood function with respect to the weights is equivalent to minimizing the sum-of-squares error function:

Exercise 6.9

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} \|\mathbf{y}(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n\|^2. \tag{6.29}$$

The noise variance is then given by

$$\sigma^{2\star} = \frac{1}{NK}\sum_{n=1}^{N} \|\mathbf{y}(\mathbf{x}_n, \mathbf{w}^\star) - \mathbf{t}_n\|^2 \tag{6.30}$$

where K is the dimensionality of the target variable. The assumption of conditional independence of the target variables can be dropped at the expense of a slightly more complex optimization problem.

Exercise 6.10

Recall that there is a natural pairing of the error function (given by the negative log likelihood) and the output-unit activation function. In regression, we can view the network as having an output activation function that is the identity, so that $y_k = a_k$. The corresponding sum-of-squares error function then has the property

Section 5.4.6

$$\frac{\partial E}{\partial a_k} = y_k - t_k. \tag{6.31}$$

6.4.2 Binary classification

Now consider binary classification in which we have a single target variable t such that $t = 1$ denotes class $\mathcal{C}_1$ and $t = 0$ denotes class $\mathcal{C}_2$. Following the discussion of canonical link functions, we consider a network having a single output whose activation function is a logistic sigmoid (6.13) so that $0 \leqslant y(\mathbf{x}, \mathbf{w}) \leqslant 1$. We can interpret $y(\mathbf{x}, \mathbf{w})$ as the conditional probability $p(\mathcal{C}_1|\mathbf{x})$, with $p(\mathcal{C}_2|\mathbf{x})$ given by $1 - y(\mathbf{x}, \mathbf{w})$. The conditional distribution of targets given inputs is then a Bernoulli distribution of the form

Section 5.4.6

$$p(t|\mathbf{x}, \mathbf{w}) = y(\mathbf{x}, \mathbf{w})^t \left\{1 - y(\mathbf{x}, \mathbf{w})\right\}^{1-t}. \tag{6.32}$$

If we consider a training set of independent observations, then the error function, which is given by the negative log likelihood, is then a *cross-entropy* error of the form

$$E(\mathbf{w}) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\} \tag{6.33}$$

where y_n denotes $y(\mathbf{x}_n, \mathbf{w})$. Simard, Steinkraus, and Platt (2003) found that using the cross-entropy error function instead of the sum-of-squares for a classification problem leads to faster training as well as improved generalization.

Note that there is no analogue of the noise variance σ^2 in (6.32) because the target values are assumed to be correctly labelled. However, the model is easily extended to allow for labelling errors by introducing a probability ϵ that the target

Exercise 6.11