We now use maximum likelihood to determine the parameters of the logistic regression model. To do this, we will make use of the derivative of the logistic sigmoid function, which can conveniently be expressed in terms of the sigmoid function itself:

Exercise 5.18

$$\frac{\mathrm{d}\sigma}{\mathrm{d}a} = \sigma(1-\sigma). \tag{5.72}$$

For a data set $\{\boldsymbol{\phi}_n, t_n\}$, where $\boldsymbol{\phi}_n = \boldsymbol{\phi}(\mathbf{x}_n)$ and $t_n \in \{0, 1\}$, with $n = 1, \ldots, N$, the likelihood function can be written

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n} \{1 - y_n\}^{1-t_n} \tag{5.73}$$

where $\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$ and $y_n = p(\mathcal{C}_1|\boldsymbol{\phi}_n)$. As usual, we can define an error function by taking the negative logarithm of the likelihood, which gives the *cross-entropy* error function:

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\} \tag{5.74}$$

where $y_n = \sigma(a_n)$ and $a_n = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n$. Taking the gradient of the error function with respect to $\mathbf{w}$, we obtain

Exercise 5.19

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n)\boldsymbol{\phi}_n \tag{5.75}$$

where we have made use of (5.72). We see that the factor involving the derivative of the logistic sigmoid has cancelled, leading to a simplified form for the gradient of the log likelihood. In particular, the contribution to the gradient from data point n is given by the 'error' $y_n - t_n$ between the target value and the prediction of the model times the basis function vector $\boldsymbol{\phi}_n$. Furthermore, comparison with (4.12) shows that this takes precisely the same form as the gradient of the sum-of-squares error function for the linear regression model.

Section 4.1.3

The maximum likelihood solution corresponds to $\nabla E(\mathbf{w}) = 0$. However, from (5.75) we see that this no longer corresponds to a set of linear equations, due to the nonlinearity in $y(\cdot)$, and so this equation does not have a closed-form solution. One approach to finding a maximum likelihood solution would be to use stochastic gradient descent, in which ∇E_n is the nth term on the right-hand side of (5.75).

Chapter 7

Stochastic gradient descent will be the principal approach to training the highly non-linear neural networks discussed in later chapters. However, the maximum likelihood equation is only 'slightly' nonlinear, and in fact the error function (5.74), in which the model is defined by (5.71), is a convex function of the parameters, which allows the error function to be minimized using a simple algorithm called *iterative reweighted least squares* or IRLS (Bishop, 2006). However, this does not easily generalize to more complex models such as deep neural networks.