where σ^2 is the variance of the Gaussian noise. Of course this is a somewhat restrictive assumption, and in some applications we will need to extend this approach to allow for more general distributions. For the conditional distribution given by (6.23), it is sufficient to take the output-unit activation function to be the identity, because such a network can approximate any continuous function from $\mathbf{x}$ to y. Given a data set of N i.i.d. observations $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, along with corresponding target values $\mathbf{t} = \{t_1, \ldots, t_N\}$, we can construct the corresponding likelihood function:

Section 6.5

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n|y(\mathbf{x}_n, \mathbf{w}), \sigma^2). \tag{6.24}$$

Note that in the machine learning literature, it is usual to consider the minimization of an error function rather than the maximization of the likelihood, and so here we will follow this convention. Taking the negative logarithm of the likelihood function (6.24), we obtain the error function

$$\frac{1}{2\sigma^2} \sum_{n=1}^{N} \{y(\mathbf{x}_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2} \ln \sigma^2 + \frac{N}{2} \ln(2\pi), \tag{6.25}$$

which can be used to learn the parameters $\mathbf{w}$ and σ^2. Consider first the determination of $\mathbf{w}$. Maximizing the likelihood function is equivalent to minimizing the sum-of-squares error function given by

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(\mathbf{x}_n, \mathbf{w}) - t_n\}^2 \tag{6.26}$$

where we have discarded additive and multiplicative constants. The value of $\mathbf{w}$ found by minimizing $E(\mathbf{w})$ will be denoted $\mathbf{w}^\star$. Note that this will typically not correspond to the global maximum of the likelihood function because the nonlinearity of the network function $y(\mathbf{x}_n, \mathbf{w})$ causes the error $E(\mathbf{w})$ to be non-convex, and so finding the global optimum is generally infeasible. Moreover, regularization terms may be added to the error function and other modifications may be made to the training process, so that the resulting solution for the network parameters may differ significantly from the maximum likelihood solution.

Chapter 9

Having found $\mathbf{w}^\star$, the value of σ^2 can be found by minimizing the error function (6.25) to give

Exercise 6.8

$$\sigma^{2\star} = \frac{1}{N} \sum_{n=1}^{N} \{y(\mathbf{x}_n, \mathbf{w}^\star) - t_n\}^2. \tag{6.27}$$

Note that this can be evaluated once the iterative optimization required to find $\mathbf{w}^\star$ is completed.

If we have multiple target variables, and we assume that they are independent, conditional on $\mathbf{x}$ and $\mathbf{w}$, with shared noise variance σ^2, then the conditional distribution of the target values is given by

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}) = \mathcal{N}\left(\mathbf{t}|\mathbf{y}(\mathbf{x}, \mathbf{w}), \sigma^2 \mathbf{I}\right). \tag{6.28}$$