Chapter 7

data points, or the individual gradients for each data point can be used directly in gradient-based optimization algorithms.

It is convenient to introduce the following variables:

$$\gamma_{nk} = \gamma_k(\mathbf{t}_n|\mathbf{x}_n) = \frac{\pi_k \mathcal{N}_{nk}}{\sum_{l=1}^{K} \pi_l \mathcal{N}_{nl}} \tag{6.44}$$

Exercise 6.17

where $\mathcal{N}_{nk}$ denotes $\mathcal{N}\left(\mathbf{t}_n|\boldsymbol{\mu}_k(\mathbf{x}_n), \sigma_k^2(\mathbf{x}_n)\right)$. These quantities have a natural interpretation as posterior probabilities for the components of the mixture in which the mixing coefficients $\pi_k(\mathbf{x})$ are viewed as $\mathbf{x}$-dependent prior probabilities.

Exercise 6.18

The derivatives of the error function with respect to the network output pre-activations governing the mixing coefficients are given by

$$\frac{\partial E_n}{\partial a_k^{\pi}} = \pi_k - \gamma_{nk}. \tag{6.45}$$

Exercise 6.19

Similarly, the derivatives with respect to the output pre-activations controlling the component means are given by

$$\frac{\partial E_n}{\partial a_{kl}^{\mu}} = \gamma_{nk} \left\{ \frac{\mu_{kl} - t_{nl}}{\sigma_k^2} \right\}. \tag{6.46}$$

Exercise 6.20

Finally, the derivatives with respect to the output pre-activations controlling the component variances are given by

$$\frac{\partial E_n}{\partial a_k^{\sigma}} = \gamma_{nk} \left\{ L - \frac{\|\mathbf{t}_n - \boldsymbol{\mu}_k\|^2}{\sigma_k^2} \right\}. \tag{6.47}$$

6.5.4 Predictive distribution

We illustrate the use of a mixture density network by returning to the toy example of an inverse problem shown in Figure 6.17. Plots of the mixing coefficients $\pi_k(x)$, the means $\mu_k(x)$, and the conditional density contours corresponding to $p(t|x)$, are shown in Figure 6.19. The outputs of the neural network, and hence the parameters in the mixture model, are necessarily continuous single-valued functions of the input variables. However, we see from Figure 6.19(c) that the model is able to produce a conditional density that is unimodal for some values of x and trimodal for other values by modulating the amplitudes of the mixing components $\pi_k(\mathbf{x})$.

Once a mixture density network has been trained, it can predict the conditional density function of the target data for any given value of the input vector. This conditional density represents a complete description of the generator of the data, so far as the problem of predicting the value of the output vector is concerned. From this density function, we can calculate more specific quantities that may be of interest in different applications. One of the simplest of these is the mean, corresponding to the conditional average of the target data, and is given by

$$\mathbb{E}\left[\mathbf{t}|\mathbf{x}\right] = \int \mathbf{t} p(\mathbf{t}|\mathbf{x})\,\mathrm{d}\mathbf{t} = \sum_{k=1}^{K} \pi_k(\mathbf{x}) \boldsymbol{\mu}_k(\mathbf{x}) \tag{6.48}$$