



Figure 5.14 The left-hand plot shows the class-conditional densities for three classes each having a Gaussian distribution, coloured red, green, and blue, in which the red and blue classes have the same covariance matrix. The right-hand plot shows the corresponding posterior probabilities, in which each point on the image is coloured using proportions of red, blue, and green ink corresponding to the posterior probabilities for the respective three classes. The decision boundaries are also shown. Notice that the boundary between the red and blue classes, which have the same covariance matrix, is linear, whereas those between the other pairs of classes are quadratic.

the prior class probabilities $p(\mathcal{C}_k)$, using maximum likelihood. This requires a data set comprising observations of \mathbf{x} along with their corresponding class labels.

First, suppose we have two classes, each having a Gaussian class-conditional density with a shared covariance matrix, and suppose we have a data set $\{\mathbf{x}_n, t_n\}$ where $n = 1, \dots, N$. Here $t_n = 1$ denotes class \mathcal{C}_1 and $t_n = 0$ denotes class \mathcal{C}_2 . We denote the prior class probability $p(\mathcal{C}_1) = \pi$, so that $p(\mathcal{C}_2) = 1 - \pi$. For a data point \mathbf{x}_n from class \mathcal{C}_1 , we have $t_n = 1$ and hence

$$p(\mathbf{x}_n, \mathcal{C}_1) = p(\mathcal{C}_1)p(\mathbf{x}_n|\mathcal{C}_1) = \pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}).$$

Similarly for class \mathcal{C}_2 , we have $t_n = 0$ and hence

$$p(\mathbf{x}_n, \mathcal{C}_2) = p(\mathcal{C}_2)p(\mathbf{x}_n|\mathcal{C}_2) = (1 - \pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}).$$

Thus, the likelihood function is given by

$$p(\mathbf{t}, \mathbf{X}|\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \prod_{n=1}^N [\pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})]^{t_n} [(1 - \pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})]^{1-t_n} \quad (5.54)$$

where $\mathbf{t} = (t_1, \dots, t_N)^T$. As usual, it is convenient to maximize the log of the likelihood function. Consider first the maximization with respect to π . The terms in