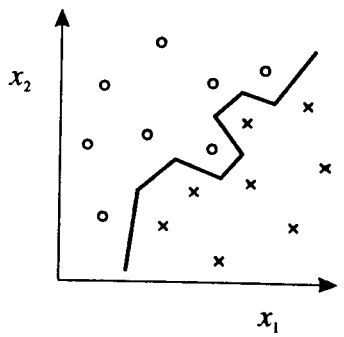

Figure 2.11. Example of the decision boundary produced by the nearest-neighbour classification rule. Note that the boundary is piecewise linear, with each segment corresponding to the perpendicular bisector between two data points belonging to different classes.

high, so that the model density is noisy and very sensitive to the individual data points. (Bias and variance are defined more precisely in Section 9.1). The choice of a suitable value for the smoothing parameter is analogous to the problem of choosing the number of terms in a polynomial used in curve fitting, discussed in Section 1.5. Similar smoothing parameters will appear in our discussions of neural networks. For instance, the number of hidden units in a layered feed-forward network can play a similar role to the number of terms in a polynomial.

It is important to realize that we cannot simply pick the value of the smoothing parameter which gives the largest value for the likelihood, as the likelihood can always be increased indefinitely by choosing ever smaller values for the smoothing parameter. Consider for instance the case of kernel estimators. The likelihood function can be written as

$$\mathcal{L}(h) = \prod_{n=1}^{N} p(\mathbf{x}^n | h; \mathbf{x}^1, \ldots \mathbf{x}^N) \tag{2.66}$$

where $p(\mathbf{x}|\ldots)$ is given by (2.58) for the case of Gaussian kernels. It is easily verified that unconstrained maximization of $\mathcal{L}(h)$ leads to $h \to 0$ so that the resulting density estimate consists of a delta function at each data point, with zero density elsewhere.

The goal in selecting smoothing parameters is to produce a model for the probability density which is as close as possible to the (unknown) true density $p(\mathbf{x})$. It is often convenient to have a formal measure of the difference, or 'distance', between two density functions. If $\widetilde{p}(\mathbf{x})$ is our model of the density function, then the average negative log-likelihood per data point, in the limit as

the number of data points goes to infinity, can be written as an expectation in the form

$$\mathcal{E}\left[-\ln \mathcal{L}\right] = -\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \ln \widetilde{p}(\mathbf{x}^n) \tag{2.67}$$

$$= -\int p(\mathbf{x}) \ln \widetilde{p}(\mathbf{x})\, d\mathbf{x} \tag{2.68}$$

which can be regarded as a measure of the extent to which the model density and the true density agree. When $\widetilde{p}(\mathbf{x}) = p(\mathbf{x})$ this measure has a residual value given by

$$-\int p(\mathbf{x}) \ln p(\mathbf{x})\, d\mathbf{x} \tag{2.69}$$

which is known as the *entropy* of $p(\mathbf{x})$ (Section 6.10). It is convenient to subtract off this residual value to give a measure of the 'distance' between $\widetilde{p}(\mathbf{x})$ and $p(\mathbf{x})$ in the form

$$L = -\int p(\mathbf{x}) \ln \frac{\widetilde{p}(\mathbf{x})}{p(\mathbf{x})}\, d\mathbf{x} \tag{2.70}$$

which is known as the *Kullback–Leibler distance* or *asymmetric divergence* (Kullback and Leibler, 1951; Kullback, 1959). It is easily shown (Exercise 2.10) that $L \geq 0$ with equality if, and only if, the two density functions are equal. Note that L is not symmetric with respect to the two probability distributions. This is reasonable since it is more important for the model distribution $\widetilde{p}(\mathbf{x})$ to be close to the true distribution $p(\mathbf{x})$ in regions where data is more likely to be found. Thus the integral in (2.70) is weighted by the true distribution.

In a practical density estimation problem we are therefore faced with the difficulty of deciding a suitable value for the smoothing parameter. This is an example of a very general, and very important, issue which is concerned with choosing the optimal level of complexity, or flexibility, of a model for a given data set. Rather than consider this problem in the framework of density estimation, we defer further discussion until Chapters 9 and 10, where we consider the analogous issue in the context of neural network models. There we shall discuss two general approaches for dealing with model complexity, based respectively on cross-validation and Bayesian inference.

2.6 Mixture models

So far in this chapter we have considered two general approaches to density estimation, parametric and non-parametric, each of which has its merits and limitations. In particular, the parametric approach assumes a specific form for