models for classification, by which we mean that the decision surfaces are linear functions of the input vector $\mathbf{x}$ and, hence, are defined by $(D-1)$-dimensional hyperplanes within the D-dimensional input space. Data sets whose classes can be separated exactly by linear decision surfaces are said to be *linearly separable*. Linear classification models can be applied to data sets that are not linearly separable, although not all inputs will be correctly classified.

We can broadly identify three distinct approaches to solving classification problems. The simplest involves constructing a *discriminant function* that directly assigns each vector $\mathbf{x}$ to a specific class. A more powerful approach, however, models the conditional probability distributions $p(\mathcal{C}_k|\mathbf{x})$ in an *inference* stage and subsequently uses these distributions to make optimal *decisions*. Separating inference and decision brings numerous benefits. There are two different approaches to determining the conditional probabilities $p(\mathcal{C}_k|\mathbf{x})$. One technique is to model them directly, for example by representing them as parametric models and then optimizing the parameters using a training set. This will be called a *discriminative probabilistic model*. Alternatively, we can model the class-conditional densities $p(\mathbf{x}|\mathcal{C}_k)$, together with the prior probabilities $p(\mathcal{C}_k)$ for the classes, and then compute the required posterior probabilities using Bayes' theorem:

Section 5.2.4

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}. \tag{5.1}$$

This will be called a *generative probabilistic model* because it offers the opportunity to generate samples from each of the class-conditional densities $p(\mathbf{x}|\mathcal{C}_k)$. In this chapter, we will discuss examples of all three approaches: discriminant functions, generative probabilistic models, and discriminative probabilistic models.

5.1. Discriminant Functions

A discriminant is a function that takes an input vector $\mathbf{x}$ and assigns it to one of K classes, denoted $\mathcal{C}_k$. In this chapter, we will restrict attention to *linear discriminants*, namely those for which the decision surfaces are hyperplanes. To simplify the discussion, we consider first two classes and then investigate the extension to $K > 2$ classes.

5.1.1 Two classes

The simplest representation of a linear discriminant function is obtained by taking a linear function of the input vector so that

$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0 \tag{5.2}$$

where $\mathbf{w}$ is called a *weight vector*, and w_0 is a *bias* (not to be confused with bias in the statistical sense). An input vector $\mathbf{x}$ is assigned to class $\mathcal{C}_1$ if $y(\mathbf{x}) \geqslant 0$ and to class $\mathcal{C}_2$ otherwise. The corresponding decision boundary is therefore defined by the relation $y(\mathbf{x}) = 0$, which corresponds to a $(D-1)$-dimensional hyperplane within