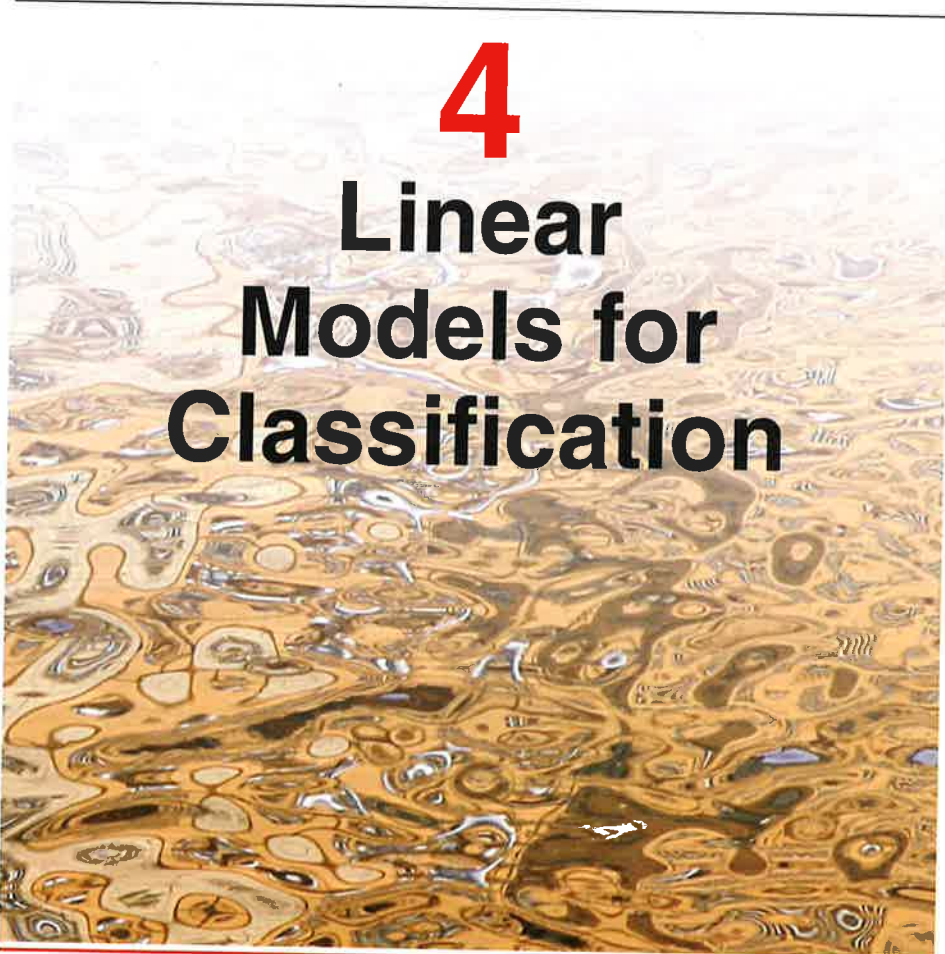


---

# 4

## Linear Models for Classification



In the previous chapter, we explored a class of regression models having particularly simple analytical and computational properties. We now discuss an analogous class of models for solving classification problems. The goal in classification is to take an input vector  $\mathbf{x}$  and to assign it to one of  $K$  discrete classes  $\mathcal{C}_k$  where  $k = 1, \dots, K$ . In the most common scenario, the classes are taken to be disjoint, so that each input is assigned to one and only one class. The input space is thereby divided into *decision regions* whose boundaries are called *decision boundaries* or *decision surfaces*. In this chapter, we consider linear 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*.

For regression problems, the target variable  $\mathbf{t}$  was simply the vector of real numbers whose values we wish to predict. In the case of classification, there are various

ways of using target values to represent class labels. For probabilistic models, the most convenient, in the case of two-class problems, is the binary representation in which there is a single target variable  $t \in \{0, 1\}$  such that  $t = 1$  represents class  $C_1$  and  $t = 0$  represents class  $C_2$ . We can interpret the value of  $t$  as the probability that the class is  $C_1$ , with the values of probability taking only the extreme values of 0 and 1. For  $K > 2$  classes, it is convenient to use a 1-of- $K$  coding scheme in which  $\mathbf{t}$  is a vector of length  $K$  such that if the class is  $C_j$ , then all elements  $t_k$  of  $\mathbf{t}$  are zero except element  $t_j$ , which takes the value 1. For instance, if we have  $K = 5$  classes, then a pattern from class 2 would be given the target vector

$$\mathbf{t} = (0, 1, 0, 0, 0)^T. \quad (4.1)$$

Again, we can interpret the value of  $t_k$  as the probability that the class is  $C_k$ . For nonprobabilistic models, alternative choices of target variable representation will sometimes prove convenient.

In Chapter 1, we identified three distinct approaches to the classification problem. 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 distribution  $p(C_k|\mathbf{x})$  in an inference stage, and then subsequently uses this distribution to make optimal decisions. By separating inference and decision, we gain numerous benefits, as discussed in Section 1.5.4. There are two different approaches to determining the conditional probabilities  $p(C_k|\mathbf{x})$ . One technique is to model them directly, for example by representing them as parametric models and then optimizing the parameters using a training set. Alternatively, we can adopt a generative approach in which we model the class-conditional densities given by  $p(\mathbf{x}|C_k)$ , together with the prior probabilities  $p(C_k)$  for the classes, and then we compute the required posterior probabilities using Bayes' theorem

$$p(C_k|\mathbf{x}) = \frac{p(\mathbf{x}|C_k)p(C_k)}{p(\mathbf{x})}. \quad (4.2)$$

We shall discuss examples of all three approaches in this chapter.

In the linear regression models considered in Chapter 3, the model prediction  $y(\mathbf{x}, \mathbf{w})$  was given by a linear function of the parameters  $\mathbf{w}$ . In the simplest case, the model is also linear in the input variables and therefore takes the form  $y(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_0$ , so that  $y$  is a real number. For classification problems, however, we wish to predict discrete class labels, or more generally posterior probabilities that lie in the range  $(0, 1)$ . To achieve this, we consider a generalization of this model in which we transform the linear function of  $\mathbf{w}$  using a nonlinear function  $f(\cdot)$  so that

$$y(\mathbf{x}) = f(\mathbf{w}^T \mathbf{x} + w_0). \quad (4.3)$$

In the machine learning literature  $f(\cdot)$  is known as an *activation function*, whereas its inverse is called a *link function* in the statistics literature. The decision surfaces correspond to  $y(\mathbf{x}) = \text{constant}$ , so that  $\mathbf{w}^T \mathbf{x} + w_0 = \text{constant}$  and hence the decision surfaces are linear functions of  $\mathbf{x}$ , even if the function  $f(\cdot)$  is nonlinear. For this reason, the class of models described by (4.3) are called *generalized linear models*

(McCullagh and Nelder, 1989). Note, however, that in contrast to the models used for regression, they are no longer linear in the parameters due to the presence of the nonlinear function  $f(\cdot)$ . This will lead to more complex analytical and computational properties than for linear regression models. Nevertheless, these models are still relatively simple compared to the more general nonlinear models that will be studied in subsequent chapters.

The algorithms discussed in this chapter will be equally applicable if we first make a fixed nonlinear transformation of the input variables using a vector of basis functions  $\phi(\mathbf{x})$  as we did for regression models in Chapter 3. We begin by considering classification directly in the original input space  $\mathbf{x}$ , while in Section 4.3 we shall find it convenient to switch to a notation involving basis functions for consistency with later chapters.

## 4.1. Discriminant Functions

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

### 4.1.1 Two classes

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

$$y(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_0 \quad (4.4)$$

where  $\mathbf{w}$  is called a *weight vector*, and  $w_0$  is a *bias* (not to be confused with bias in the statistical sense). The negative of the bias is sometimes called a *threshold*. An input vector  $\mathbf{x}$  is assigned to class  $C_1$  if  $y(\mathbf{x}) \geq 0$  and to class  $C_2$  otherwise. The corresponding decision boundary is therefore defined by the relation  $y(\mathbf{x}) = 0$ , which corresponds to a  $(D - 1)$ -dimensional hyperplane within the  $D$ -dimensional input space. Consider two points  $\mathbf{x}_A$  and  $\mathbf{x}_B$  both of which lie on the decision surface. Because  $y(\mathbf{x}_A) = y(\mathbf{x}_B) = 0$ , we have  $\mathbf{w}^T(\mathbf{x}_A - \mathbf{x}_B) = 0$  and hence the vector  $\mathbf{w}$  is orthogonal to every vector lying within the decision surface, and so  $\mathbf{w}$  determines the orientation of the decision surface. Similarly, if  $\mathbf{x}$  is a point on the decision surface, then  $y(\mathbf{x}) = 0$ , and so the normal distance from the origin to the decision surface is given by

$$\frac{\mathbf{w}^T \mathbf{x}}{\|\mathbf{w}\|} = -\frac{w_0}{\|\mathbf{w}\|}. \quad (4.5)$$

We therefore see that the bias parameter  $w_0$  determines the location of the decision surface. These properties are illustrated for the case of  $D = 2$  in Figure 4.1.

Furthermore, we note that the value of  $y(\mathbf{x})$  gives a signed measure of the perpendicular distance  $r$  of the point  $\mathbf{x}$  from the decision surface. To see this, consider