the computer. The sections related to estimation of the number of clusters and neural network implementations are bypassed.

Chapter 13 deals with hierarchical clustering algorithms. In a first course, only the general agglomerative scheme is considered with an emphasis on single link and complete link algorithms, based on matrix theory. Agglomerative algorithms based on graph theory concepts as well as the divisive schemes are bypassed.

Chapter 14 deals with clustering algorithms based on cost function optimization, using tools from differential calculus. Hard clustering and fuzzy and possibilistic schemes are considered, based on various types of cluster representatives, including point representatives, hyperplane representatives, and shell-shaped representatives. In a first course, most of these algorithms are bypassed, and emphasis is given to the isodata algorithm.

Chapter 15 features a high degree of modularity. It deals with clustering algorithms based on different ideas, which cannot be grouped under a single philosophy. Spectral clustering, competitive learning, branch and bound, simulated annealing, and genetic algorithms are some of the schemes treated in this chapter. These are bypassed in a first course.

Chapter 16 deals with the clustering validity stage of a clustering procedure. It contains rather advanced concepts and is omitted in a first course. Emphasis is given to the definitions of internal, external, and relative criteria and the random hypotheses used in each case. Indices, adopted in the framework of external and internal criteria, are presented, and examples are provided showing the use of these indices.

Syntactic pattern recognition methods are not treated in this book. Syntactic pattern recognition methods differ in philosophy from the methods discussed in this book and, in general, are applicable to different types of problems. In syntactic pattern recognition, the structure of the patterns is of paramount importance, and pattern recognition is performed on the basis of a set of pattern *primitives*, a set of rules in the form of a *grammar*, and a recognizer called *automaton*. Thus, we were faced with a dilemma: either to increase the size of the book substantially, or to provide a short overview (which, however, exists in a number of other books), or to omit it. The last option seemed to be the most sensible choice.

CHAPTER

2 Classifiers Based on Bayes Decision Theory

2.1 INTRODUCTION

This is the first chapter, out of three, dealing with the design of the classifier in a pattern recognition system. The approach to be followed builds upon probabilistic arguments stemming from the statistical nature of the generated features. As has already been pointed out in the introductory chapter, this is due to the statistical variation of the patterns as well as to the noise in the measuring sensors. Adopting this reasoning as our kickoff point, we will design classifiers that classify an unknown pattern in the most probable of the classes. Thus, our task now becomes that of defining what "most probable" means.

Given a classification task of M classes, $\omega_1, \omega_2, \ldots, \omega_M$, and an unknown pattern, which is represented by a feature vector $\boldsymbol{x}$, we form the M conditional probabilities $P(\omega_i|\boldsymbol{x}), i = 1, 2, \ldots, M$. Sometimes, these are also referred to as *a posteriori probabilities*. In words, each of them represents the probability that the unknown pattern belongs to the respective class ω_i, given that the corresponding feature vector takes the value $\boldsymbol{x}$. Who could then argue that these conditional probabilities are not sensible choices to quantify the term *most probable*? Indeed, the classifiers to be considered in this chapter compute either the maximum of these M values or, equivalently, the maximum of an appropriately defined function of them. The unknown pattern is then assigned to the class corresponding to this maximum.

The first task we are faced with is the computation of the conditional probabilities. The Bayes rule will once more prove its usefulness! A major effort in this chapter will be devoted to techniques for estimating probability density functions (pdf), based on the available experimental evidence, that is, the feature vectors corresponding to the patterns of the training set.

2.2 BAYES DECISION THEORY

We will initially focus on the two-class case. Let ω_1, ω_2 be the two classes in which our patterns belong. In the sequel, we assume that the *a priori probabilities*

$P(\omega_1), P(\omega_2)$ are known. This is a very reasonable assumption, because even if they are not known, they can easily be estimated from the available training feature vectors. Indeed, if N is the total number of available training patterns, and N_1, N_2 of them belong to ω_1 and ω_2, respectively, then $P(\omega_1) \approx N_1/N$ and $P(\omega_2) \approx N_2/N$.

The other statistical quantities assumed to be known are the class-conditional probability density functions $p(\boldsymbol{x}|\omega_i), i = 1, 2$, describing the distribution of the feature vectors in each of the classes. If these are not known, they can also be estimated from the available training data, as we will discuss later on in this chapter. The pdf $p(\boldsymbol{x}|\omega_i)$ is sometimes referred to as the *likelihood function of ω_i with respect to $\boldsymbol{x}$*. Here we should stress the fact that an implicit assumption has been made. That is, the feature vectors can take any value in the l-dimensional feature space. In the case that feature vectors can take only discrete values, density functions $p(\boldsymbol{x}|\omega_i)$ become probabilities and will be denoted by $P(\boldsymbol{x}|\omega_i)$.

We now have all the ingredients to compute our conditional probabilities, as stated in the introduction. To this end, let us recall from our probability course basics the *Bayes rule* (Appendix A)

$$P(\omega_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_i)P(\omega_i)}{p(\boldsymbol{x})} \tag{2.1}$$

where $p(\boldsymbol{x})$ is the pdf of $\boldsymbol{x}$ and for which we have (Appendix A)

$$p(\boldsymbol{x}) = \sum_{i=1}^{2} p(\boldsymbol{x}|\omega_i)P(\omega_i) \tag{2.2}$$

The *Bayes classification rule* can now be stated as

$$\begin{aligned} &\text{If} \quad P(\omega_1|\boldsymbol{x}) > P(\omega_2|\boldsymbol{x}), \quad \boldsymbol{x} \text{ is classified to } \omega_1 \\ &\text{If} \quad P(\omega_1|\boldsymbol{x}) < P(\omega_2|\boldsymbol{x}), \quad \boldsymbol{x} \text{ is classified to } \omega_2 \end{aligned} \tag{2.3}$$

The case of equality is detrimental and the pattern can be assigned to either of the two classes. Using (2.1), the decision can equivalently be based on the inequalities

$$p(\boldsymbol{x}|\omega_1)P(\omega_1) \gtrless p(\boldsymbol{x}|\omega_2)P(\omega_2) \tag{2.4}$$

$p(\boldsymbol{x})$ is not taken into account, because it is the same for all classes and it does not affect the decision. Furthermore, if the *a priori* probabilities are equal, that is, $P(\omega_1) = P(\omega_2) = 1/2$, Eq. (2.4) becomes

$$p(\boldsymbol{x}|\omega_1) \gtrless p(\boldsymbol{x}|\omega_2) \tag{2.5}$$

Thus, the search for the maximum now rests on the values of the conditional pdfs evaluated at $\boldsymbol{x}$. Figure 2.1 presents an example of two equiprobable classes and shows the variations of $p(x|\omega_i), i = 1, 2$, as functions of x for the simple case of a single feature ($l = 1$). The dotted line at x_0 is a threshold partitioning the feature space into two regions, R_1 and R_2. According to the Bayes decision rule, for all values of x in R_1 the classifier decides ω_1 and for all values in R_2 it decides ω_2. However, it is obvious from the figure that decision errors are unavoidable. Indeed, there is

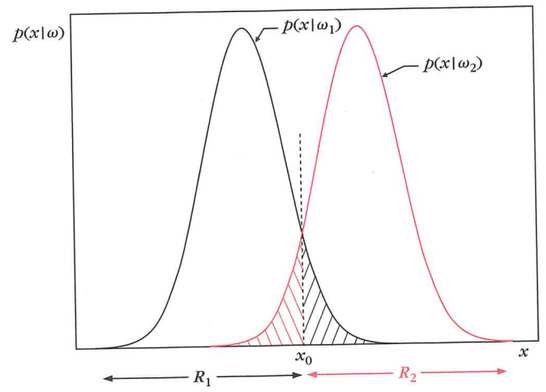

FIGURE 2.1
Example of the two regions R_1 and R_2 formed by the Bayesian classifier for the case of two equiprobable classes.

a finite probability for an x to lie in the R_2 region and at the same time to belong in class ω_1. Then our decision is in error. The same is true for points originating from class ω_2. It does not take much thought to see that the total probability, P_e, of committing a decision error for the case of two equiprobable classes, is given by

$$P_e = \frac{1}{2}\int_{-\infty}^{x_0} p(x|\omega_2)\,dx + \frac{1}{2}\int_{x_0}^{+\infty} p(x|\omega_1)\,dx \tag{2.6}$$

which is equal to the total shaded area under the curves in Figure 2.1. We have now touched on a very important issue. Our starting point to arrive at the Bayes classification rule was rather empirical, via our interpretation of the term *most probable*. We will now see that this classification test, though simple in its formulation, has a sounder mathematical interpretation.

Minimizing the Classification Error Probability

We will show that *the Bayesian classifier is optimal with respect to minimizing the classification error probability*. Indeed, the reader can easily verify, as an exercise, that moving the threshold away from x_0, in Figure 2.1, always increases the corresponding shaded area under the curves. Let us now proceed with a more formal proof.

Proof. Let R_1 be the region of the feature space in which we decide in favor of ω_1 and R_2 be the corresponding region for ω_2. Then an error is made if $\boldsymbol{x} \in R_1$, although it belongs to ω_2 or if $\boldsymbol{x} \in R_2$, although it belongs to ω_1. That is,

$$P_e = P(\boldsymbol{x} \in R_2, \omega_1) + P(\boldsymbol{x} \in R_1, \omega_2) \tag{2.7}$$

where $P(\cdot,\cdot)$ is the joint probability of two events. Recalling, once more, our probability basics (Appendix A), this becomes

$$\begin{aligned} P_e &= P(\boldsymbol{x} \in R_2|\omega_1)P(\omega_1) + P(\boldsymbol{x} \in R_1|\omega_2)P(\omega_2) \\ &= P(\omega_1)\int_{R_2} p(\boldsymbol{x}|\omega_1)\,d\boldsymbol{x} + P(\omega_2)\int_{R_1} p(\boldsymbol{x}|\omega_2)\,d\boldsymbol{x} \end{aligned} \tag{2.8}$$

or using the Bayes rule

$$P_e = \int_{R_2} P(\omega_1|\boldsymbol{x})p(\boldsymbol{x})\,d\boldsymbol{x} + \int_{R_1} P(\omega_2|\boldsymbol{x})p(\boldsymbol{x})\,d\boldsymbol{x} \tag{2.9}$$

It is now easy to see that the error is minimized if *the partitioning regions* R_1 and R_2 *of the feature space are chosen so that*

$$\begin{aligned} R_1&: P(\omega_1|\boldsymbol{x}) > P(\omega_2|\boldsymbol{x}) \\ R_2&: P(\omega_2|\boldsymbol{x}) > P(\omega_1|\boldsymbol{x}) \end{aligned} \tag{2.10}$$

Indeed, since the union of the regions R_1, R_2 covers all the space, from the definition of a probability density function we have that

$$\int_{R_1} P(\omega_1|\boldsymbol{x})p(\boldsymbol{x})\,d\boldsymbol{x} + \int_{R_2} P(\omega_1|\boldsymbol{x})p(\boldsymbol{x})\,d\boldsymbol{x} = P(\omega_1) \tag{2.11}$$

Combining Eqs. (2.9) and (2.11), we get

$$P_e = P(\omega_1) - \int_{R_1} (P(\omega_1|\boldsymbol{x}) - P(\omega_2|\boldsymbol{x}))\,p(\boldsymbol{x})\,d\boldsymbol{x} \tag{2.12}$$

This suggests that the probability of error is minimized if R_1 is the region of space in which $P(\omega_1|\boldsymbol{x}) > P(\omega_2|\boldsymbol{x})$. Then, R_2 becomes the region where the reverse is true. □

So far, we have dealt with the simple case of two classes. Generalizations to the multiclass case are straightforward. In a classification task with M classes, $\omega_1, \omega_2, \ldots, \omega_M$, an unknown pattern, represented by the feature vector $\boldsymbol{x}$, is assigned to class ω_i if

$$P(\omega_i|\boldsymbol{x}) > P(\omega_j|\boldsymbol{x}) \quad \forall j \neq i \tag{2.13}$$

It turns out that such a choice also minimizes the classification error probability (Problem 2.1).

Minimizing the Average Risk

The classification error probability is not always the best criterion to be adopted for minimization. This is because it assigns the same importance to all errors. However, there are cases in which some wrong decisions may have more serious implications than others. For example, it is much more serious for a doctor to make a wrong decision and a malignant tumor to be diagnosed as a benign one, than the other way round. If a benign tumor is diagnosed as a malignant one, the wrong decision will be cleared out during subsequent clinical examinations. However, the results

from the wrong decision concerning a malignant tumor may be fatal. Thus, in such cases it is more appropriate to assign a penalty term to weigh each error. For our example, let us denote by ω_1 the class of malignant tumors and as ω_2 the class of the benign ones. Let, also, R_1, R_2 be the regions in the feature space where we decide in favor of ω_1 and ω_2, respectively. The error probability P_e is given by Eq. (2.8). Instead of selecting R_1 and R_2 so that P_e is minimized, we will now try to minimize a modified version of it, that is,

$$r = \lambda_{12}P(\omega_1)\int_{R_2} p(\boldsymbol{x}|\omega_1)d\boldsymbol{x} + \lambda_{21}P(\omega_2)\int_{R_1} p(\boldsymbol{x}|\omega_2)d\boldsymbol{x} \tag{2.14}$$

where each of the two terms that contributes to the overall error probability is weighted according to its significance. For our case, the reasonable choice would be to have $\lambda_{12} > \lambda_{21}$. Thus errors due to the assignment of patterns originating from class ω_1 to class ω_2 will have a larger effect on the cost function than the errors associated with the second term in the summation.

Let us now consider an M-class problem and let $R_j, j = 1, 2, \ldots, M$, be the regions of the feature space assigned to classes ω_j, respectively. Assume now that a feature vector $\boldsymbol{x}$ that belongs to class ω_k lies in $R_i, i \neq k$. Then this vector is misclassified in ω_i and an error is committed. A penalty term λ_{ki}, known as *loss*, is associated with this wrong decision. The matrix L, which has at its (k, i) location the corresponding penalty term, is known as the *loss matrix*.[1] Observe that in contrast to the philosophy behind Eq. (2.14), we have now allowed weights across the diagonal of the loss matrix (λ_{kk}), which correspond to correct decisions. In practice, these are usually set equal to zero, although we have considered them here for the sake of generality. The *risk or loss* associated with ω_k is defined as

$$r_k = \sum_{i=1}^{M} \lambda_{ki} \int_{R_i} p(\boldsymbol{x}|\omega_k)\,d\boldsymbol{x} \tag{2.15}$$

Observe that the integral is the overall probability of a feature vector from class ω_k being classified in ω_i. This probability is weighted by λ_{ki}. Our goal now is to choose the partitioning regions R_j so that the *average risk*

$$\begin{aligned} r &= \sum_{k=1}^{M} r_k P(\omega_k) \\ &= \sum_{i=1}^{M} \int_{R_i} \left(\sum_{k=1}^{M} \lambda_{ki}\, p(\boldsymbol{x}|\omega_k)P(\omega_k) \right) d\boldsymbol{x} \end{aligned} \tag{2.16}$$

is minimized. This is achieved if each of the integrals is minimized, which is equivalent to selecting partitioning regions so that

$$\boldsymbol{x} \in R_i \quad \text{if} \quad l_i \equiv \sum_{k=1}^{M} \lambda_{ki}\, p(\boldsymbol{x}|\omega_k)P(\omega_k) < l_j \equiv \sum_{k=1}^{M} \lambda_{kj}\, p(\boldsymbol{x}|\omega_k)P(\omega_k) \quad \forall j \neq i \tag{2.17}$$

[1] The terminology comes from the general decision theory.

It is obvious that if $\lambda_{ki} = 1 - \delta_{ki}$, where δ_{ki} is *Kronecker's delta* (0 if $k \neq i$ and 1 if $k = i$), then minimizing the average risk becomes equivalent to minimizing the classification error probability.

The two-class case. For this specific case we obtain

$$\begin{aligned} l_1 &= \lambda_{11}\, p(\boldsymbol{x}|\omega_1)P(\omega_1) + \lambda_{21}\, p(\boldsymbol{x}|\omega_2)P(\omega_2) \\ l_2 &= \lambda_{12}\, p(\boldsymbol{x}|\omega_1)P(\omega_1) + \lambda_{22}\, p(\boldsymbol{x}|\omega_2)P(\omega_2) \end{aligned} \tag{2.18}$$

We assign $\boldsymbol{x}$ to ω_1 if $l_1 < l_2$, that is,

$$(\lambda_{21} - \lambda_{22})p(\boldsymbol{x}|\omega_2)P(\omega_2) < (\lambda_{12} - \lambda_{11})p(\boldsymbol{x}|\omega_1)P(\omega_1) \tag{2.19}$$

It is natural to assume that $\lambda_{ij} > \lambda_{ii}$ (correct decisions are penalized much less than wrong ones). Adopting this assumption, the decision rule (2.17) for the two-class case now becomes

$$\boldsymbol{x} \in \omega_1 (\omega_2) \quad \text{if} \quad l_{12} \equiv \frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)} > (<) \frac{P(\omega_2)}{P(\omega_1)} \frac{\lambda_{21} - \lambda_{22}}{\lambda_{12} - \lambda_{11}} \tag{2.20}$$

The ratio l_{12} is known as the *likelihood ratio* and the preceding test as the *likelihood ratio test*. Let us now investigate Eq. (2.20) a little further and consider the case of Figure 2.1. Assume that the loss matrix is of the form

$$L = \begin{bmatrix} 0 & \lambda_{12} \\ \lambda_{21} & 0 \end{bmatrix}$$

If misclassification of patterns that come from ω_2 is considered to have serious consequences, then we must choose $\lambda_{21} > \lambda_{12}$. Thus, patterns are assigned to class ω_2 if

$$p(\boldsymbol{x}|\omega_2) > p(\boldsymbol{x}|\omega_1)\frac{\lambda_{12}}{\lambda_{21}}$$

where $P(\omega_1) = P(\omega_2) = 1/2$ has been assumed. That is, $p(\boldsymbol{x}|\omega_1)$ is multiplied by a factor less than 1 and the effect of this is to move the threshold in Figure 2.1 to the left of x_0. In other words, region R_2 is increased while R_1 is decreased. The opposite would be true if $\lambda_{21} < \lambda_{12}$.

An alternative cost that sometimes is used for two class problems is the Neyman-Pearson criterion. The error for one of the classes is now constrained to be fixed and equal to a chosen value (Problem 2.6). Such a decision rule has been used, for example, in radar detection problems. The task there is to detect a target in the presence of noise. One type of error is the so-called *false alarm*—that is, to mistake the noise for a signal (target) present. Of course, the other type of error is to miss the signal and to decide in favor of the noise (*missed detection*). In many cases the error probability of false alarm is set equal to a predetermined threshold.

Example 2.1

In a two-class problem with a single feature x the pdfs are Gaussians with variance $\sigma^2 = 1/2$ for both classes and mean values 0 and 1, respectively, that is,

$$p(x|\omega_1) = \frac{1}{\sqrt{\pi}} \exp(-x^2)$$

$$p(x|\omega_2) = \frac{1}{\sqrt{\pi}} \exp(-(x-1)^2)$$

If $P(\omega_1) = P(\omega_2) = 1/2$, compute the threshold value x_0 (a) for minimum error probability and (b) for minimum risk if the loss matrix is

$$L = \begin{bmatrix} 0 & 0.5 \\ 1.0 & 0 \end{bmatrix}$$

Taking into account the shape of the Gaussian function graph (Appendix A), the threshold for the minimum probability case will be

$$x_0 : \exp(-x^2) = \exp(-(x-1)^2)$$

Taking the logarithm of both sides, we end up with $x_0 = 1/2$. In the minimum risk case we get

$$x_0 : \exp(-x^2) = 2\exp(-(x-1)^2)$$

or $x_0 = (1 - \ln 2)/2 < 1/2$; that is, the threshold moves to the left of 1/2. If the two classes are not equiprobable, then it is easily verified that if $P(\omega_1) > (<) P(\omega_2)$ the threshold moves to the right (left). That is, we expand the region in which we decide in favor of the most probable class, since it is better to make fewer errors for the most probable class.

2.3 DISCRIMINANT FUNCTIONS AND DECISION SURFACES

It is by now clear that minimizing either the risk or the error probability or the Neyman-Pearson criterion is equivalent to partitioning the feature space into M regions, for a task with M classes. If regions R_i, R_j happen to be contiguous, then they are separated by a *decision surface* in the multidimensional feature space. For the minimum error probability case, this is described by the equation

$$P(\omega_i|\boldsymbol{x}) - P(\omega_j|\boldsymbol{x}) = 0 \tag{2.21}$$

From the one side of the surface this difference is positive, and from the other it is negative. Sometimes, instead of working directly with probabilities (or risk functions), it may be more convenient, from a mathematical point of view, to work with an equivalent function of them, for example, $g_i(\boldsymbol{x}) \equiv f(P(\omega_i|\boldsymbol{x}))$, where $f(\cdot)$ is a monotonically increasing function. $g_i(\boldsymbol{x})$ is known as a *discriminant function*. The decision test (2.13) is now stated as

$$\text{classify } \boldsymbol{x} \text{ in } \omega_i \quad \text{if} \quad g_i(\boldsymbol{x}) > g_j(\boldsymbol{x}) \quad \forall j \neq i \tag{2.22}$$

The decision surfaces, separating contiguous regions, are described by

$$g_{ij}(\boldsymbol{x}) \equiv g_i(\boldsymbol{x}) - g_j(\boldsymbol{x}) = 0, \quad i,j = 1,2,\ldots,M, \quad i \neq j \tag{2.23}$$

So far, we have approached the classification problem via Bayesian probabilistic arguments and the goal was to minimize the classification error probability or the risk. However, as we will soon see, not all problems are well suited to such approaches. For example, in many cases the involved pdfs are complicated and their estimation is not an easy task. In such cases, it may be preferable to compute decision surfaces *directly by means of alternative costs*, and this will be our focus in Chapters 3 and 4. Such approaches give rise to discriminant functions and decision surfaces, which are entities with no (necessary) relation to Bayesian classification, and they are, in general, suboptimal with respect to Bayesian classifiers.

In the following we will focus on a particular family of decision surfaces associated with the Bayesian classification for the specific case of Gaussian density functions.

2.4 BAYESIAN CLASSIFICATION FOR NORMAL DISTRIBUTIONS

2.4.1 The Gaussian Probability Density Function

One of the most commonly encountered probability density functions in practice is the Gaussian or normal probability density function. The major reasons for its popularity are its computational tractability and the fact that it models adequately a large number of cases. One of the most celebrated theorems in statistics is the *central limit theorem*. The theorem states that if a random variable is the outcome of a summation of a number of *independent* random variables, its pdf approaches the Gaussian function as the number of summands tends to infinity (see Appendix A). In practice, it is most common to assume that the sum of random variables is distributed according to a Gaussian pdf, for a sufficiently large number of summing terms.

The one-dimensional or the univariate Gaussian, as it is sometimes called, is defined by

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{2.24}$$

The parameters μ and σ^2 turn out to have a specific meaning. The mean value of the random variable x is equal to μ, that is,

$$\mu = E[x] \equiv \int_{-\infty}^{+\infty} xp(x)dx \tag{2.25}$$

where $E[\cdot]$ denotes the mean (or expected) value of a random variable. The parameter σ^2 is equal to the variance of x, that is,

$$\sigma^2 = E[(x-\mu)^2] \equiv \int_{-\infty}^{+\infty} (x-\mu)^2 p(x)dx \tag{2.26}$$

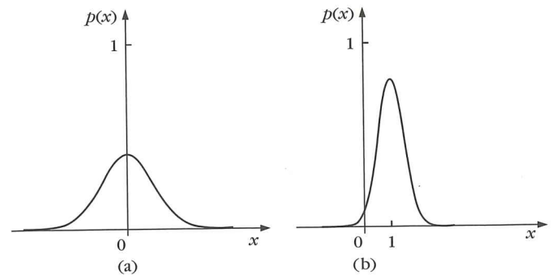

FIGURE 2.2

Graphs for the one-dimensional Gaussian pdf. (a) Mean value $\mu = 0$, $\sigma^2 = 1$, (b) $\mu = 1$ and $\sigma^2 = 0.2$. The larger the variance the broader the graph is. The graphs are symmetric, and they are centered at the respective mean value.

Figure 2.2a shows the graph of the Gaussian function for $\mu = 0$ and $\sigma^2 = 1$, and Figure 2.2b the case for $\mu = 1$ and $\sigma^2 = 0.2$. The larger the variance the broader the graph, which is symmetric, and it is always centered at μ (see Appendix A, for some more properties).

The multivariate generalization of a Gaussian pdf in the l-dimensional space is given by

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{l/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{2.27}$$

where $\boldsymbol{\mu} = E[\boldsymbol{x}]$ is the mean value and Σ is the $l \times l$ *covariance matrix* (Appendix A) defined as

$$\Sigma = E[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T] \tag{2.28}$$

where $|\Sigma|$ denotes the determinant of Σ. It is readily seen that for $l = 1$ the multivariate Gaussian coincides with the univariate one. Sometimes, the symbol $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ is used to denote a Gaussian pdf with mean value $\boldsymbol{\mu}$ and covariance Σ.

To get a better feeling on what the multivariate Gaussian looks like, let us focus on some cases in the two-dimensional space, where nature allows us the luxury of visualization. For this case we have

$$\Sigma = E\left[\begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}\begin{bmatrix} x_1-\mu_1, & x_2-\mu_2 \end{bmatrix}\right] \tag{2.29}$$

$$= \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \tag{2.30}$$

where $E[x_i] = \mu_i$, $i = 1, 2$, and by definition $\sigma_{12} = E[(x_1-\mu_1)(x_2-\mu_2)]$, which is known as the covariance between the random variables x_1 and x_2 and it is a measure

of their mutual statistical correlation. If the variables are statistically independent, their covariance is zero (Appendix A). Obviously, the diagonal elements of Σ are the variances of the respective elements of the random vector.

Figures 2.3–2.6 show the graphs for four instances of a two-dimensional Gaussian probability density function. Figure 2.3a corresponds to a Gaussian with a diagonal covariance matrix

$$\Sigma = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$$

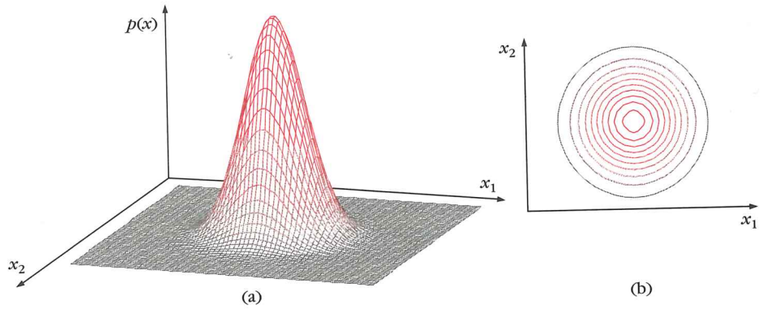

FIGURE 2.3
(a) The graph of a two-dimensional Gaussian pdf and (b) the corresponding isovalue curves for a diagonal Σ with $\sigma_1^2 = \sigma_2^2$. The graph has a spherical symmetry showing no preference in any direction.

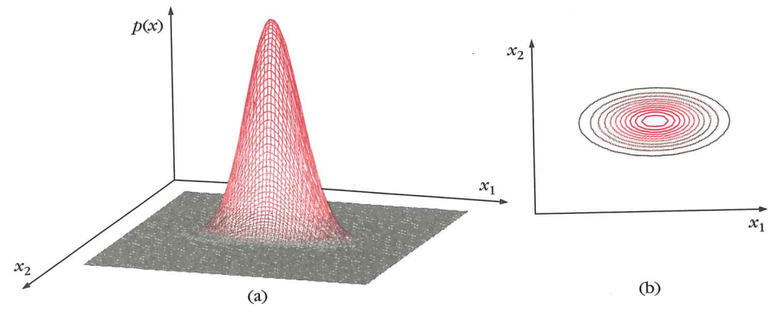

FIGURE 2.4
(a) The graph of a two-dimensional Gaussian pdf and (b) the corresponding isovalue curves for a diagonal Σ with $\sigma_1^2 >> \sigma_2^2$. The graph is elongated along the x_1 direction.

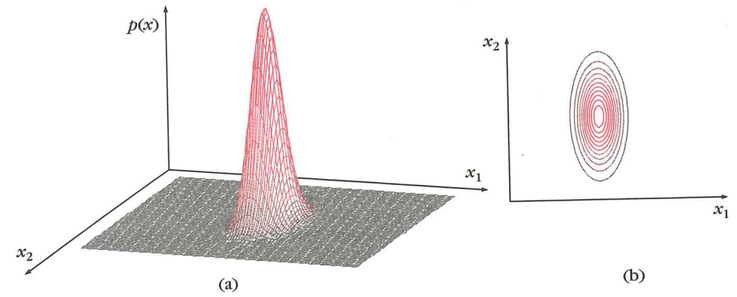

FIGURE 2.5
(a) The graph of a two-dimensional Gaussian pdf and (b) the corresponding isovalue curves for a diagonal Σ with $\sigma_1^2 << \sigma_2^2$. The graph is elongated along the x_2 direction.

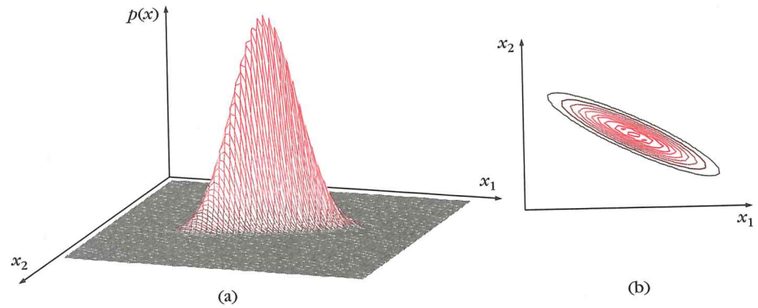

FIGURE 2.6
(a) The graph of a two-dimensional Gaussian pdf and (b) the corresponding isovalue curves for a case of a nondiagonal Σ. Playing with the values of the elements of Σ one can achieve different shapes and orientations.

that is, both features, x_1, x_2 have variance equal to 3 and their covariance is zero. The graph of the Gaussian is symmetric. For this case the isovalue curves (i.e., curves of equal probability density values) are circles (hyperspheres in the general l-dimensional space) and are shown in Figure 2.3b. The case shown in Figure 2.4a corresponds to the covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

with $\sigma_1^2 = 15 >> \sigma_2^2 = 3$. The graph of the Gaussian is now elongated along the x_1-axis, which is the direction of the larger variance. The isovalue curves, shown

in Figure 2.4b, are ellipses. Figures 2.5a and 2.5b correspond to the case with $\sigma_1^2 = 3 << \sigma_2^2 = 15$. Figures 2.6a and 2.6b correspond to the more general case where

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

and $\sigma_1^2 = 15$, $\sigma_2^2 = 3$, $\sigma_{12} = 6$. Playing with σ_1^2, σ_2^2 and σ_{12} one can achieve different shapes and different orientations.

The isovalue curves are ellipses of different orientations and with different ratios of major to minor axis lengths. Let us consider, as an example, the case of a zero mean random vector with a diagonal covariance matrix. To compute the isovalue curves is equivalent to computing the curves of constant values for the exponent, that is,

$$\boldsymbol{x}^T \Sigma^{-1} \boldsymbol{x} = [x_1, x_2] \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = C \tag{2.31}$$

or

$$\frac{x_1^2}{\sigma_1^2} + \frac{x_2^2}{\sigma_2^2} = C \tag{2.32}$$

for some constant C. This is the equation of an ellipse whose axes are determined by the the variances of the involved features. As we will soon see, the principal axes of the ellipses are controlled by the eigenvectors/eigenvalues of the covariance matrix. As we know from linear algebra (and it is easily checked), the eigenvalues of a diagonal matrix, which was the case for our example, are equal to the respective elements across its diagonal.

2.4.2 The Bayesian Classifier for Normally Distributed Classes

Our goal in this section is to study the optimal Bayesian classifier when the involved pdfs, $p(\boldsymbol{x}|\omega_i)$, $i = 1, 2, \ldots, M$ (likelihood functions of ω_i with respect to $\boldsymbol{x}$), describing the data distribution in each one of the classes, are multivariate normal distributions, that is, $\mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i)$, $i = 1, 2, \ldots, M$. Because of the exponential form of the involved densities, it is preferable to work with the following discriminant functions, which involve the (monotonic) logarithmic function $\ln(\cdot)$:

$$g_i(\boldsymbol{x}) = \ln(p(\boldsymbol{x}|\omega_i)P(\omega_i)) = \ln p(\boldsymbol{x}|\omega_i) + \ln P(\omega_i) \tag{2.33}$$

or

$$g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i) + \ln P(\omega_i) + c_i \tag{2.34}$$

where c_i is a constant equal to $-(l/2)\ln 2\pi - (1/2)\ln|\Sigma_i|$. Expanding, we obtain

$$g_i(\boldsymbol{x}) = -\frac{1}{2}\boldsymbol{x}^T \Sigma_i^{-1} \boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^T \Sigma_i^{-1} \boldsymbol{\mu}_i - \frac{1}{2}\boldsymbol{\mu}_i^T \Sigma_i^{-1} \boldsymbol{\mu}_i + \frac{1}{2}\boldsymbol{\mu}_i^T \Sigma_i^{-1} \boldsymbol{x} + \ln P(\omega_i) + c_i \tag{2.35}$$

In general, this is a nonlinear quadratic form. Take, for example, the case of $l = 2$ and assume that

$$\Sigma_i = \begin{bmatrix} \sigma_i^2 & 0 \\ 0 & \sigma_i^2 \end{bmatrix}$$

Then (2.35) becomes

$$g_i(\boldsymbol{x}) = -\frac{1}{2\sigma_i^2}(x_1^2 + x_2^2) + \frac{1}{\sigma_i^2}(\mu_{i1}x_1 + \mu_{i2}x_2) - \frac{1}{2\sigma_i^2}(\mu_{i1}^2 + \mu_{i2}^2) + \ln P(\omega_i) + c_i \tag{2.36}$$

and obviously the associated decision curves $g_i(\boldsymbol{x}) - g_j(\boldsymbol{x}) = 0$ are *quadrics* (i.e., ellipsoids, parabolas, hyperbolas, pairs of lines). That is, in such cases, the Bayesian classifier is a *quadratic classifier*, in the sense that the partition of the feature space is performed via quadric decision surfaces. For $l > 2$ the decision surfaces are *hyperquadrics*. Figure 2.7a shows the decision curve corresponding to $P(\omega_1) = P(\omega_2)$, $\boldsymbol{\mu}_1 = [0, 0]^T$ and $\boldsymbol{\mu}_2 = [4, 0]^T$. The covariance matrices for the two classes are

$$\Sigma_1 = \begin{bmatrix} 0.3 & 0.0 \\ 0.0 & 0.35 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 1.2 & 0.0 \\ 0.0 & 1.85 \end{bmatrix}$$

For the case of Figure 2.7b the classes are also equiprobable with $\boldsymbol{\mu}_1 = [0, 0]^T$, $\boldsymbol{\mu}_2 = [3.2, 0]^T$ and covariance matrices

$$\Sigma_1 = \begin{bmatrix} 0.1 & 0.0 \\ 0.0 & 0.75 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 0.75 & 0.0 \\ 0.0 & 0.1 \end{bmatrix}$$

Figure 2.8 shows the two pdfs for the case of Figure 2.7a. The red color is used for class ω_1 and indicates the points where $p(\boldsymbol{x}|\omega_1) > p(\boldsymbol{x}|\omega_2)$. The gray color is similarly used for class ω_2. It is readily observed that the decision curve is an ellipse, as shown in Figure 2.7a. The setup corresponding to Figure 2.7b is shown in Figure 2.9. In this case, the decision curve is a hyperbola.

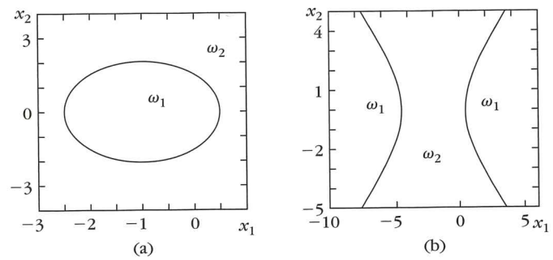

FIGURE 2.7

Examples of quadric decision curves. Playing with the covariance matrices of the Gaussian functions, different decision curves result, that is, ellipsoids, parabolas, hyperbolas, pairs of lines.

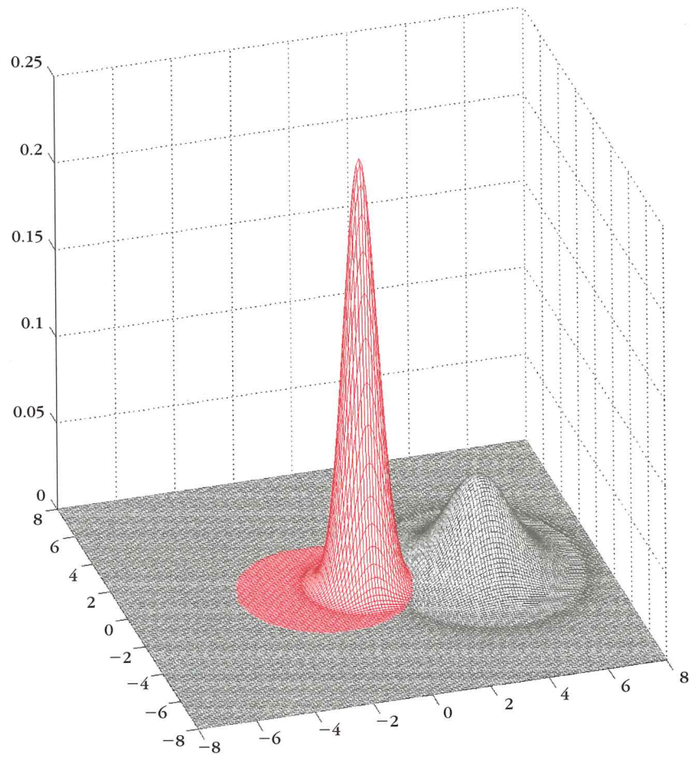

FIGURE 2.8

An example of the pdfs of two equiprobable classes in the two-dimensional space. The feature vectors in both classes are normally distributed with different covariance matrices. In this case, the decision curve is an ellipse and it is shown in Figure 2.7a. The coloring indicates the areas where the value of the respective pdf is larger.

Decision Hyperplanes

The only quadratic contribution in (2.35) comes from the term $\boldsymbol{x}^T\Sigma_i^{-1}\boldsymbol{x}$. If we now assume that the covariance matrix is the same in all classes, that is, $\Sigma_i = \Sigma$, the quadratic term will be the same in all discriminant functions. Hence, it does not enter into the comparisons for computing the maximum, and it cancels out in the decision surface equations. The same is true for the constants c_i. Thus, they can be omitted and we may redefine $g_i(\boldsymbol{x})$ as

$$g_i(\boldsymbol{x}) = \boldsymbol{w}_i^T\boldsymbol{x} + w_{i0} \tag{2.37}$$

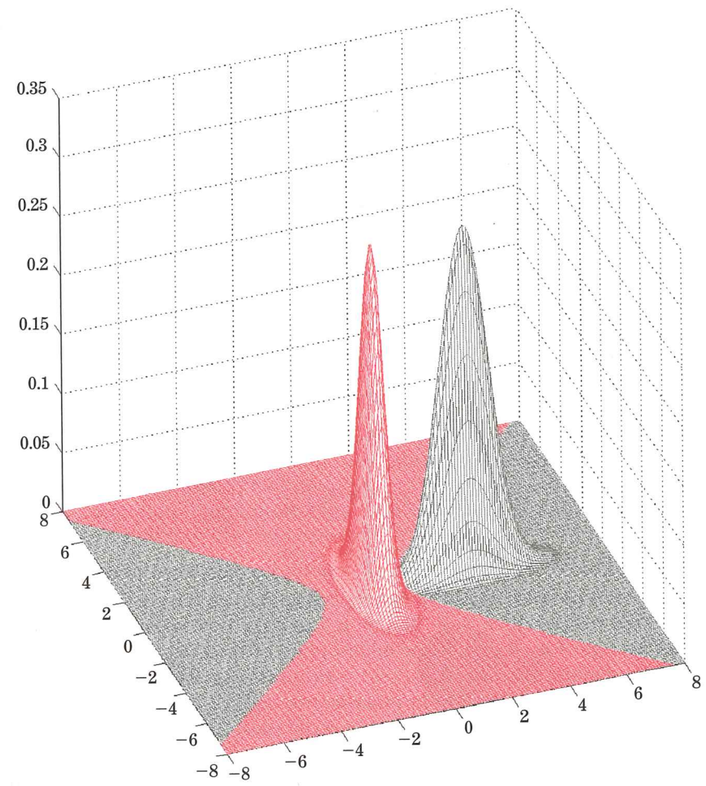

FIGURE 2.9

An example of the pdfs of two equiprobable classes in the two-dimensional space. The feature vectors in both classes are normally distributed with different covariance matrices. In this case, the decision curve is a hyperbola and it is shown in Figure 2.7b.

where

$$\boldsymbol{w}_i = \Sigma^{-1}\boldsymbol{\mu}_i \tag{2.38}$$

and

$$w_{i0} = \ln P(\omega_i) - \frac{1}{2}\boldsymbol{\mu}_i^T\Sigma^{-1}\boldsymbol{\mu}_i \tag{2.39}$$

Hence $g_i(\boldsymbol{x})$ is a *linear function* of $\boldsymbol{x}$ and the respective decision surfaces are *hyperplanes*. Let us investigate this a bit more.

- *Diagonal covariance matrix with equal elements*: Assume that the individual features, constituting the feature vector, are *mutually uncorrelated and of the same variance* ($E[(x_i - \mu_i)(x_j - \mu_j)] = \sigma^2\delta_{ij}$). Then, as discussed in Appendix A, $\Sigma = \sigma^2 I$, where I is the l-dimensional identity matrix, and (2.37) becomes

$$g_i(\boldsymbol{x}) = \frac{1}{\sigma^2}\boldsymbol{\mu}_i^T\boldsymbol{x} + w_{i0} \tag{2.40}$$

Thus, the corresponding decision hyperplanes can now be written as (verify it)

$$g_{ij}(\boldsymbol{x}) \equiv g_i(\boldsymbol{x}) - g_j(\boldsymbol{x}) = \boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{x}_0) = 0 \tag{2.41}$$

where

$$\boldsymbol{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j \tag{2.42}$$

and

$$\boldsymbol{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \sigma^2 \ln\left(\frac{P(\omega_i)}{P(\omega_j)}\right)\frac{\boldsymbol{\mu}_i - \boldsymbol{\mu}_j}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2} \tag{2.43}$$

where $\|\boldsymbol{x}\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_l^2}$ denotes the Euclidean norm of $\boldsymbol{x}$. Thus, the decision surface is a *hyperplane* passing through the point $\boldsymbol{x}_0$. Obviously, if $P(\omega_i) = P(\omega_j)$, then $\boldsymbol{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j)$, and the hyperplane passes through the average of $\boldsymbol{\mu}_i, \boldsymbol{\mu}_j$, that is, the middle point of the segment joining the mean values. On the other hand, if $P(\omega_j) > P(\omega_i)$ ($P(\omega_i) > P(\omega_j)$) the hyperplane is located closer to $\boldsymbol{\mu}_i(\boldsymbol{\mu}_j)$. In other words, the area of the region where we decide in favor of the more probable of the two classes is increased.

The geometry is illustrated in Figure 2.10 for the two-dimensional case and for two cases, that is, $P(\omega_j) = P(\omega_i)$ (black line) and $P(\omega_j) > P(\omega_i)$ (red line). We observe that for both cases the decision hyperplane (straight line) is orthogonal to $\boldsymbol{\mu}_i - \boldsymbol{\mu}_j$. Indeed, for any point $\boldsymbol{x}$ lying on the decision hyperplane, the vector $\boldsymbol{x} - \boldsymbol{x}_0$ also lies on the hyperplane and

$$g_{ij}(\boldsymbol{x}) = 0 \Rightarrow \boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{x}_0) = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T(\boldsymbol{x} - \boldsymbol{x}_0) = 0$$

That is, $\boldsymbol{\mu}_i - \boldsymbol{\mu}_j$ is orthogonal to the decision hyperplane. Furthermore, if σ^2 is small with respect to $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|$, the location of the hyperplane is rather insensitive to the values of $P(\omega_i), P(\omega_j)$. This is expected, because small variance indicates that the random vectors are clustered within a small radius around their mean values. Thus a small shift of the decision hyperplane has a small effect on the result.

Figure 2.11 illustrates this. For each class, the circles around the means indicate regions where samples have a high probability, say 98%,

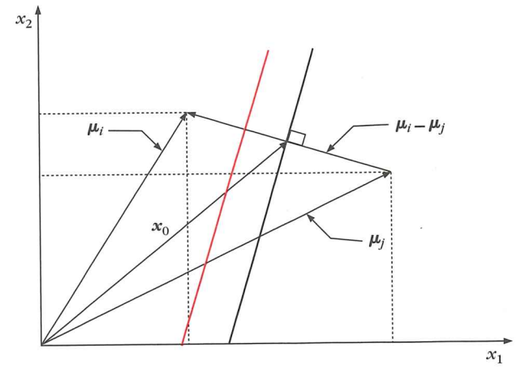

FIGURE 2.10

Decision lines for normally distributed vectors with $\Sigma = \sigma^2 I$. The black line corresponds to the case of $P(\omega_j) = P(\omega_i)$ and it passes through the middle point of the line segment joining the mean values of the two classes. The red line corresponds to the case of $P(\omega_j) > P(\omega_i)$ and it is closer to $\boldsymbol{\mu}_i$, leaving more "room" to the more probable of the two classes. If we had assumed $P(\omega_j) < P(\omega_i)$, the decision line would have moved closer to $\boldsymbol{\mu}_j$.

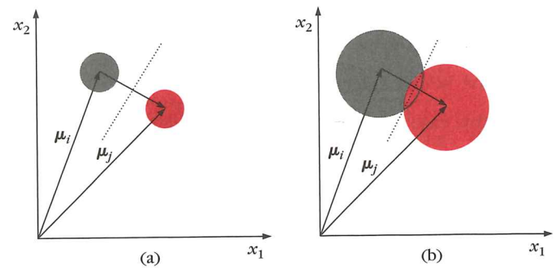

FIGURE 2.11

Decision line (a) for compact and (b) for noncompact classes. When classes are compact around their mean values, the location of the hyperplane is rather insensitive to the values of $P(\omega_1)$ and $P(\omega_2)$. This is not the case for noncompact classes, where a small movement of the hyperplane to the right or to the left may be more critical.

of being found. The case of Figure 2.11a corresponds to small variance, and that of Figure 2.11b to large variance. No doubt the location of the decision hyperplane in Figure 2.11b is much more critical than that in Figure 2.11a.

- *Nondiagonal covariance matrix*: Following algebraic arguments similar to those used before, we end up with hyperplanes described by

$$g_{ij}(\boldsymbol{x}) = \boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{x}_0) = 0 \tag{2.44}$$

where

$$\boldsymbol{w} = \Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \tag{2.45}$$

and

$$\boldsymbol{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \ln\left(\frac{P(\omega_i)}{P(\omega_j)}\right)\frac{\boldsymbol{\mu}_i - \boldsymbol{\mu}_j}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2_{\Sigma^{-1}}} \tag{2.46}$$

where $\|\boldsymbol{x}\|_{\Sigma^{-1}} \equiv (\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x})^{1/2}$ denotes the so-called Σ^{-1} norm of $\boldsymbol{x}$. The comments made before for the case of the diagonal covariance matrix are still valid, with one exception. *The decision hyperplane is no longer orthogonal to the vector $\boldsymbol{\mu}_i - \boldsymbol{\mu}_j$ but to its linear transformation $\Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$.*

Figure 2.12 shows two Gaussian pdfs with equal covariance matrices, describing the data distribution of two equiprobable classes. In both classes, the data are distributed around their mean values in *exactly* the same way and the optimal decision curve is a straight line.

Minimum Distance Classifiers

We will now view the task from a slightly different angle. Assuming equiprobable classes with the same covariance matrix, $g_i(\boldsymbol{x})$ in (2.34) is simplified to

$$g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^T\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i) \tag{2.47}$$

where constants have been neglected.

- $\Sigma = \sigma^2 I$: In this case maximum $g_i(\boldsymbol{x})$ implies minimum

$$\textit{Euclidean distance:}\quad d_\epsilon = \|\boldsymbol{x} - \boldsymbol{\mu}_i\| \tag{2.48}$$

Thus, feature vectors are assigned to classes according to their Euclidean distance from the respective mean points. Can you verify that this result ties in with the geometry of the hyperplanes discussed before?

Figure 2.13a shows curves of equal distance $d_\epsilon = c$ from the mean points of each class. They are obviously circles of radius c (hyperspheres in the general case).

- Nondiagonal Σ: For this case maximizing $g_i(\boldsymbol{x})$ is equivalent to minimizing the Σ^{-1} norm, known as the

$$\textit{Mahalanobis distance:}\quad d_m = \left((\boldsymbol{x} - \boldsymbol{\mu}_i)^T\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i)\right)^{1/2} \tag{2.49}$$

In this case, the constant distance $d_m = c$ curves are ellipses (hyperellipses). Indeed, the covariance matrix is symmetric and, as discussed in Appendix B, it can always be diagonalized by a unitary transform

$$\Sigma = \Phi\Lambda\Phi^T \tag{2.50}$$

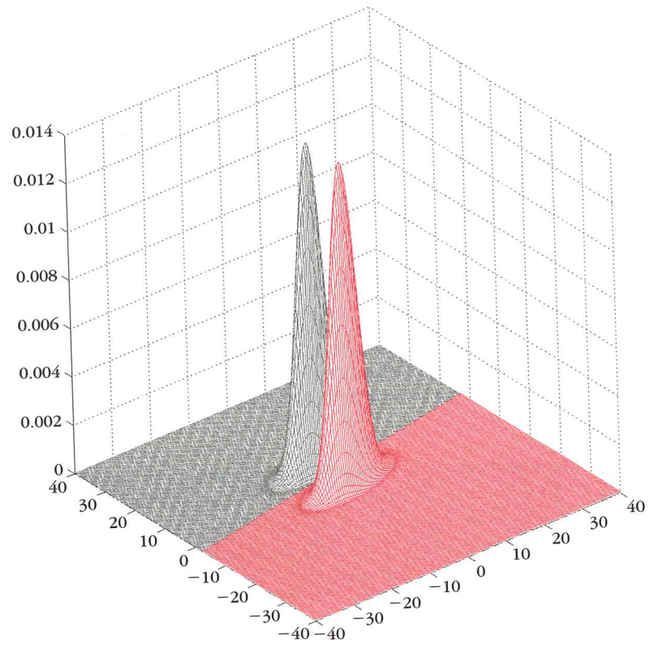

FIGURE 2.12
An example of two Gaussian pdfs with the same covariance matrix in the two-dimensional space. Each one of them is associated with one of two equiprobable classes. In this case, the decision curve is a straight line.

where $\Phi^T = \Phi^{-1}$ and Λ is the diagonal matrix whose elements are the eigenvalues of Σ. Φ has as its columns the corresponding (orthonormal) eigenvectors of Σ

$$\Phi = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_l] \tag{2.51}$$

Combining (2.49) and (2.50), we obtain

$$(\boldsymbol{x} - \boldsymbol{\mu}_i)^T\Phi\Lambda^{-1}\Phi^T(\boldsymbol{x} - \boldsymbol{\mu}_i) = c^2 \tag{2.52}$$

Define $\boldsymbol{x}' = \Phi^T\boldsymbol{x}$. The coordinates of $\boldsymbol{x}'$ are equal to $\boldsymbol{v}_k^T\boldsymbol{x}, k = 1, 2, \ldots, l$, that is, the projections of $\boldsymbol{x}$ onto the eigenvectors. In other words, they are the coordinates of $\boldsymbol{x}$ with respect to a new coordinate system whose axes are determined by $\boldsymbol{v}_k, k = 1, 2, \ldots, l$. Equation (2.52) can now be written as

$$\frac{(x_1' - \mu_{i1}')^2}{\lambda_1} + \cdots + \frac{(x_l' - \mu_{il}')^2}{\lambda_l} = c^2 \tag{2.53}$$

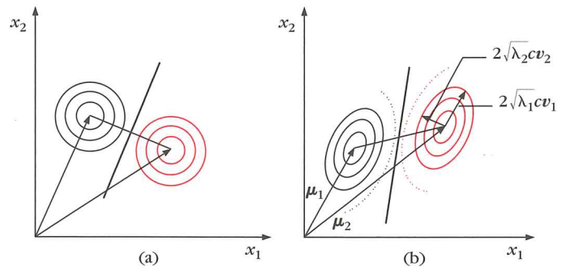

FIGURE 2.13

Curves of (a) equal Euclidean distance and (b) equal Mahalanobis distance from the mean points of each class. In the two-dimensional space, they are circles in the case of Euclidean distance and ellipses in the case of Mahalanobis distance. Observe that in the latter case the decision line is no longer orthogonal to the line segment joining the mean values. It turns according to the shape of the ellipses.

This is the equation of a hyperellipsoid in the new coordinate system. Figure 2.13b shows the $l = 2$ case. The center of mass of the ellipse is at $\boldsymbol{\mu}_i$, and the principal axes are aligned with the corresponding eigenvectors and have lengths $2\sqrt{\lambda_k}c$, respectively. Thus, *all points having the same Mahalanobis distance from a specific point are located on an ellipse*.

Example 2.2

In a two-class, two-dimensional classification task, the feature vectors are generated by two normal distributions sharing the same covariance matrix

$$\Sigma = \begin{bmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{bmatrix}$$

and the mean vectors are $\boldsymbol{\mu}_1 = [0, 0]^T, \boldsymbol{\mu}_2 = [3, 3]^T$, respectively.

(a) Classify the vector $[1.0, 2.2]^T$ according to the Bayesian classifier.

It suffices to compute the Mahalanobis distance of $[1.0, 2.2]^T$ from the two mean vectors. Thus,

$$d_m^2(\boldsymbol{\mu}_1, \boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu}_1)^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_1)$$

$$= [1.0, 2.2] \begin{bmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{bmatrix} \begin{bmatrix} 1.0 \\ 2.2 \end{bmatrix} = 2.952$$

Similarly,

$$d_m^2(\boldsymbol{\mu}_2, \boldsymbol{x}) = [-2.0, -0.8] \begin{bmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{bmatrix} \begin{bmatrix} -2.0 \\ -0.8 \end{bmatrix} = 3.672 \qquad (2.54)$$

Thus, the vector is assigned to the class with mean vector $[0, 0]^T$. *Notice that the given vector* $[1.0, 2.2]^T$ *is closer to* $[3, 3]^T$ *with respect to the Euclidean distance.*

(b) Compute the principal axes of the ellipse centered at $[0, 0]^T$ that corresponds to a constant Mahalanobis distance $d_m = \sqrt{2.952}$ from the center.

To this end, we first calculate the eigenvalues of Σ.

$$\det\left(\begin{bmatrix} 1.1 - \lambda & 0.3 \\ 0.3 & 1.9 - \lambda \end{bmatrix}\right) = \lambda^2 - 3\lambda + 2 = 0$$

or $\lambda_1 = 1$ and $\lambda_2 = 2$. To compute the eigenvectors we substitute these values into the equation

$$(\Sigma - \lambda I)\boldsymbol{v} = \boldsymbol{0}$$

and we obtain the unit norm eigenvectors

$$\boldsymbol{v}_1 = \begin{bmatrix} \frac{3}{\sqrt{10}} \\ -\frac{1}{\sqrt{10}} \end{bmatrix}, \quad \boldsymbol{v}_2 = \begin{bmatrix} \frac{1}{\sqrt{10}} \\ \frac{3}{\sqrt{10}} \end{bmatrix}$$

It can easily be seen that they are mutually orthogonal. The principal axes of the ellipse are parallel to $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ and have lengths 3.436 and 4.859, respectively.

Remarks

- In practice, it is quite common to assume that the data in each class are adequately described by a Gaussian distribution. As a consequence, the associated Bayesian classifier is either linear or quadratic in nature, depending on the adopted assumptions concerning the covariance matrices. That is, if they are all equal or different. In statistics, this approach to the classification task is known as *linear discriminant analysis* (LDA) or *quadratic discriminant analysis* (QDA), respectively. *Maximum likelihood* is usually the method mobilized for the estimation of the unknown parameters that define the mean values and the covariance matrices (see Section 2.5 and Problem 2.19).
- A major problem associated with LDA and even more with QDA is the large number of the unknown parameters that have to be estimated in the case of high-dimensional spaces. For example, there are l parameters in each of the mean vectors and approximately $l^2/2$ in each (symmetric) covariance matrix. Besides the high demand for computational resources, obtaining good estimates of a large number of parameters dictates a large number of training points, N. This is a major issue that also embraces the design of other types of classifiers, for most of the cases, and we will come to it in greater detail in Chapter 5. In an effort to reduce the number of parameters to be estimated, a number of approximate techniques have been suggested over the years, including [Kimu 87, Hoff 96, Frie 89, Liu 04]. Linear discrimination will be approached from a different perspective in Section 5.8.
- LDA and QDA exhibit good performance in a large set of diverse applications and are considered to be among the most popular classifiers. No doubt, it is hard to accept that in all these cases the Gaussian assumption provides a reasonable modeling for the data statistics. The secret of the success seems

to lie in the fact that linear or quadratic decision surfaces offer a reasonably good partition of the space, from the classification point of view. Moreover, as pointed out in [Hast 01], the estimates associated with Gaussian models have some good statistical properties (i.e., bias variance trade-off, Section 3.5.3) compared to other techniques.

2.5 ESTIMATION OF UNKNOWN PROBABILITY DENSITY FUNCTIONS

So far, we have assumed that the probability density functions are known. However, this is not the most common case. In many problems, the underlying pdf has to be estimated from the available data. There are various ways to approach the problem. Sometimes we may know the type of the pdf (e.g., Gaussian, Rayleigh), but we do not know certain parameters, such as the mean values or the variances. In contrast, in other cases we may not have information about the type of the pdf but we may know certain statistical parameters, such as the mean value and the variance. Depending on the available information, different approaches can be adopted. This will be our focus in the next subsections.

2.5.1 Maximum Likelihood Parameter Estimation

Let us consider an M-class problem with feature vectors distributed according to $p(\boldsymbol{x}|\omega_i), i = 1, 2, \ldots, M$. We assume that these likelihood functions are given in a *parametric* form and that the corresponding parameters form the vectors $\boldsymbol{\theta}_i$ which are unknown. To show the dependence on $\boldsymbol{\theta}_i$ we write $p(\boldsymbol{x}|\omega_i; \boldsymbol{\theta}_i)$. Our goal is to estimate the unknown parameters using a set of known feature vectors in each class. If we further assume that data from one class do not affect the parameter estimation of the others, we can formulate the problem independent of classes and simplify our notation. At the end, one has to solve one such problem for each class independently.

Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ be random samples drawn from pdf $p(\boldsymbol{x}; \boldsymbol{\theta})$. We form the joint pdf $p(X; \boldsymbol{\theta})$, where $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ is the set of the samples. Assuming *statistical independence* between the different samples, we have

$$p(X; \boldsymbol{\theta}) \equiv p(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N; \boldsymbol{\theta}) = \prod_{k=1}^{N} p(\boldsymbol{x}_k; \boldsymbol{\theta}) \tag{2.55}$$

This is a function of $\boldsymbol{\theta}$, and it is also known as the likelihood function of $\boldsymbol{\theta}$ with respect to X. *The maximum likelihood (ML) method estimates $\boldsymbol{\theta}$ so that the likelihood function takes its maximum value*, that is,

$$\hat{\boldsymbol{\theta}}_{ML} = \arg\max_{\boldsymbol{\theta}} \prod_{k=1}^{N} p(\boldsymbol{x}_k; \boldsymbol{\theta}) \tag{2.56}$$

A necessary condition that $\hat{\boldsymbol{\theta}}_{ML}$ must satisfy in order to be a maximum is the gradient of the likelihood function with respect to $\boldsymbol{\theta}$ to be zero, that is

$$\frac{\partial \prod_{k=1}^{N} p(\boldsymbol{x}_k; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0} \tag{2.57}$$

Because of the monotonicity of the logarithmic function, we define the *log-likelihood function* as

$$L(\boldsymbol{\theta}) \equiv \ln \prod_{k=1}^{N} p(\boldsymbol{x}_k; \boldsymbol{\theta}) \tag{2.58}$$

and (2.57) is equivalent to

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{k=1}^{N} \frac{\partial \ln p(\boldsymbol{x}_k; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{k=1}^{N} \frac{1}{p(\boldsymbol{x}_k; \boldsymbol{\theta})} \frac{\partial p(\boldsymbol{x}_k; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0} \tag{2.59}$$

Figure 2.14 illustrates the method for the single unknown parameter case. The ML estimate corresponds to the peak of the log-likelihood function.

Maximum likelihood estimation has some very desirable properties. If $\boldsymbol{\theta}_0$ is the true value of the unknown parameter in $p(\boldsymbol{x}; \boldsymbol{\theta})$, it can be shown that under generally valid conditions the following are true [Papo 91].

- The ML estimate is *asymptotically unbiased*, which by definition means that

$$\lim_{N \to \infty} E[\hat{\boldsymbol{\theta}}_{ML}] = \boldsymbol{\theta}_0 \tag{2.60}$$

Alternatively, we say that the estimate *converges in the mean* to the true value. The meaning of this is as follows. The estimate $\hat{\boldsymbol{\theta}}_{ML}$ is itself a random vector, because for different sample sets X different estimates will result. An estimate is called *unbiased* if its mean is the true value of the unknown parameter. In the ML case this is true only asymptotically ($N \to \infty$).

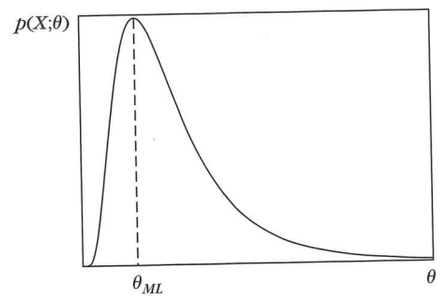

FIGURE 2.14
The maximum likelihood estimator θ_{ML} corresponds to the peak of $p(X; \theta)$.

- The ML estimate is *asymptotically consistent*, that is, it satisfies

$$\lim_{N\to\infty} \text{prob}\{\|\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0\| \leq \epsilon\} = 1 \tag{2.61}$$

where ϵ is arbitrarily small. Alternatively, we say that the estimate converges *in probability*. In other words, for large N it is highly probable that the resulting estimate will be arbitrarily close to the true value. A stronger condition for consistency is also true:

$$\lim_{N\to\infty} E[\|\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0\|^2] = 0 \tag{2.62}$$

In such cases we say that the estimate converges in the *mean square*. In words, for large N, the variance of the ML estimates tends to zero.

Consistency is very important for an estimator, because it may be unbiased, but the resulting estimates exhibit large variations around the mean. In such cases we have little confidence in the result obtained from a single set X.

- The ML estimate is asymptotically *efficient*; that is, it achieves the Cramer–Rao lower bound (Appendix A). This is the lowest value of variance, which *any* estimate can achieve.

- The pdf of the ML estimate as $N \to \infty$ approaches the Gaussian distribution with mean $\boldsymbol{\theta}_0$ [Cram 46]. This property is an offspring of (a) the central limit theorem (Appendix A) and (b) the fact that the ML estimate is related to the *sum* of random variables, that is, $\partial \ln(p(\boldsymbol{x}_k; \boldsymbol{\theta}))/\partial\boldsymbol{\theta}$ (Problem 2.16).

In summary, the ML estimator is unbiased, is normally distributed, and has the minimum possible variance. However, all these nice properties are valid *only* for large values of N.

Example 2.3

Assume that N data points, $x_1, x_2, \ldots, x_N$, have been generated by a one-dimensional Gaussian pdf of known mean, μ, but of unknown variance. Derive the ML estimate of the variance.

The log-likelihood function for this case is given by

$$L(\sigma^2) = \ln \prod_{k=1}^{N} p(x_k; \sigma^2) = \ln \prod_{k=1}^{N} \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2}} \exp\left(-\frac{(x_k - \mu)^2}{2\sigma^2}\right)$$

or

$$L(\sigma^2) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{k=1}^{N}(x_k - \mu)^2$$

Taking the derivative of the above with respect to σ^2 and equating to zero, we obtain

$$-\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{k=1}^{N}(x_k - \mu)^2 = 0$$

and finally the ML estimate of σ^2 results as the solution of the above,

$$\hat{\sigma}^2_{ML} = \frac{1}{N}\sum_{k=1}^{N}(x_k - \mu)^2 \tag{2.63}$$

Observe that, for finite N, $\hat{\sigma}^2_{ML}$ in Eq. (2.63) is a biased estimate of the variance. Indeed,

$$E[\hat{\sigma}^2_{ML}] = \frac{1}{N}\sum_{k=1}^{N} E[(x_k - \mu)^2] = \frac{N-1}{N}\sigma^2$$

where σ^2 is the true variance of the Gaussian pdf. However, for large values of N, we have

$$E[\hat{\sigma}^2_{ML}] = (1 - \frac{1}{N})\sigma^2 \approx \sigma^2$$

which is in line with the theoretical result of asymptotic consistency of the ML estimator.

Example 2.4

Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ be vectors stemmed from a normal distribution with known covariance matrix and unknown mean, that is,

$$p(\boldsymbol{x}_k; \boldsymbol{\mu}) = \frac{1}{(2\pi)^{l/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_k - \boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}_k - \boldsymbol{\mu})\right)$$

Obtain the ML estimate of the unknown mean vector.

For N available samples we have

$$L(\boldsymbol{\mu}) \equiv \ln \prod_{k=1}^{N} p(\boldsymbol{x}_k; \boldsymbol{\mu}) = -\frac{N}{2}\ln((2\pi)^l|\Sigma|) - \frac{1}{2}\sum_{k=1}^{N}(\boldsymbol{x}_k - \boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}_k - \boldsymbol{\mu}) \tag{2.64}$$

Taking the gradient with respect to $\boldsymbol{\mu}$, we obtain

$$\frac{\partial L(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} \equiv \begin{bmatrix} \frac{\partial L}{\partial \mu_1} \\ \frac{\partial L}{\partial \mu_2} \\ \vdots \\ \frac{\partial L}{\partial \mu_l} \end{bmatrix} = \sum_{k=1}^{N}\Sigma^{-1}(\boldsymbol{x}_k - \boldsymbol{\mu}) = 0 \tag{2.65}$$

or

$$\hat{\boldsymbol{\mu}}_{ML} = \frac{1}{N}\sum_{k=1}^{N}\boldsymbol{x}_k \tag{2.66}$$

That is, the ML estimate of the mean, for Gaussian densities, is the sample mean. However, this very "natural approximation" is not necessarily ML optimal for non-Gaussian density functions.