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

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

$P(\omega_i|x)$

λ_{ij}

(functional units)

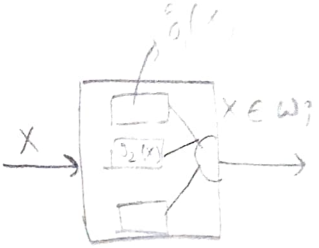
Taking into account the shape of the Gaussian function graph (Appendix A), the threshold for the minimum probability case will be

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

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.



Classifier
(black box)

$g_1(x), g_2(x) \ldots g_n(x)$
Discriminant function

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

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

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, $g_i(x) \equiv f(P(\omega_i|x))$, where $f(\cdot)$ is a monotonically increasing function. $g_i(x)$ is known as a *discriminant function*. The decision test (2.13) is now stated as

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

The decision surfaces, separating contiguous regions, are described by

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

$g_i(x) \equiv f(P(\omega_i|x))$

$f(\cdot)$ → monotonically inc. funⁿ

$g_i(x) > g_j(x)$
$\forall j \neq i$
$x \in \omega_i$

For minimum risk classifier :- $R(\alpha_i/x)$
$g_i(x) = -R(\alpha_i/x)$
Min error rate classification
$g_i(x) = P(\omega_i/x)$

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

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

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,

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

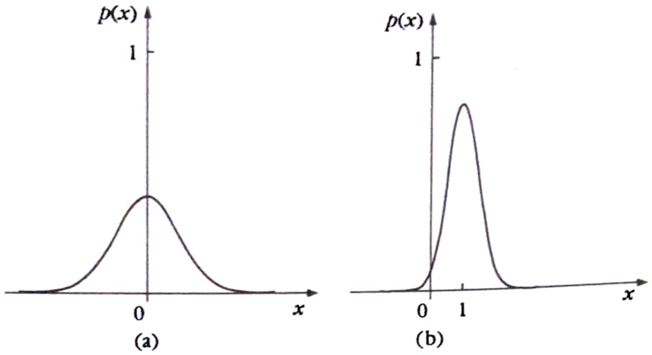


FIGURE 2.2

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

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

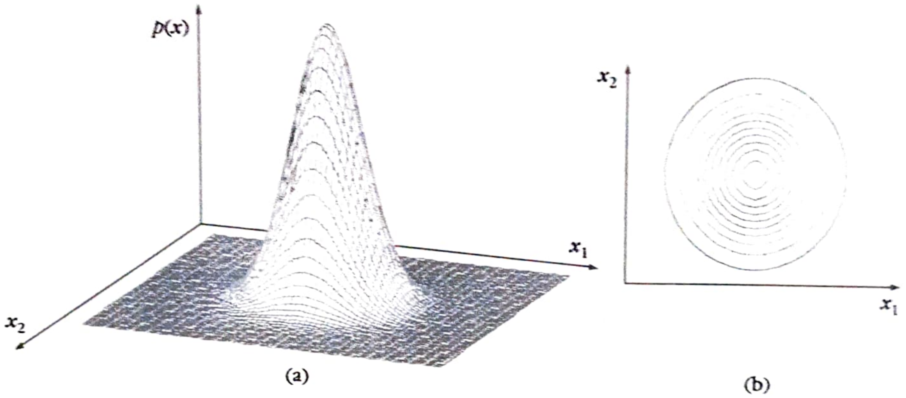


FIGURE 2.3

(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.

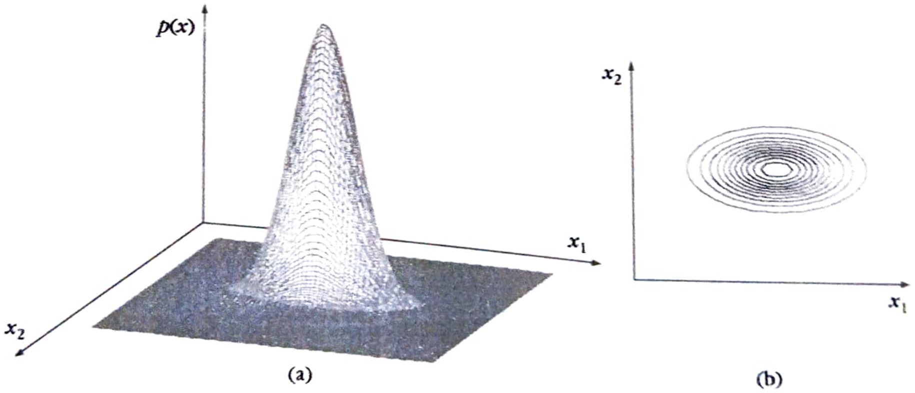


FIGURE 2.4

(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.

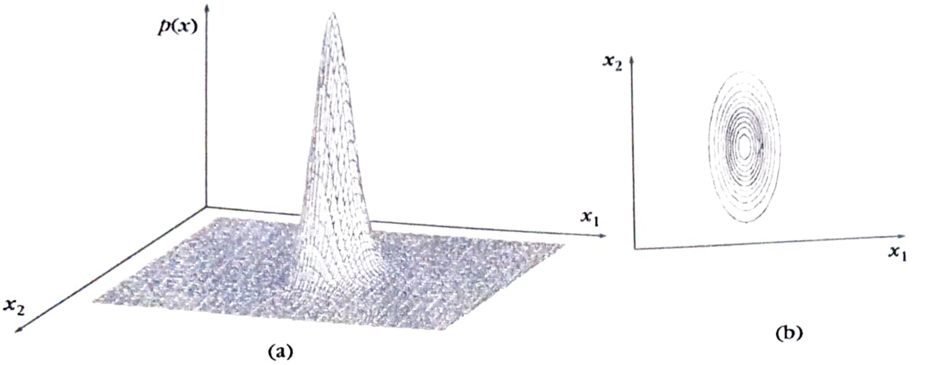


FIGURE 2.5

(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.

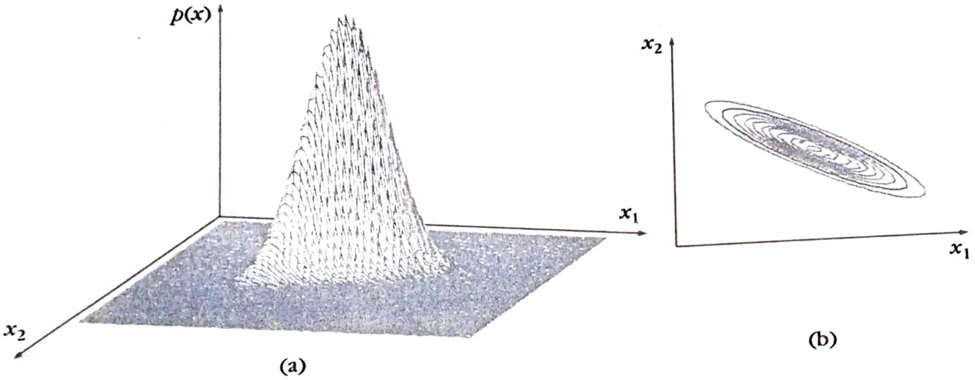


FIGURE 2.6

(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

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

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,

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

or

$$\frac{x_1^2}{\sigma_1^2} + \frac{x_2^2}{\sigma_2^2} = C \qquad (2.32)$$

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 $\ln(\cdot)$:

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

or

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

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

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

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

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

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.

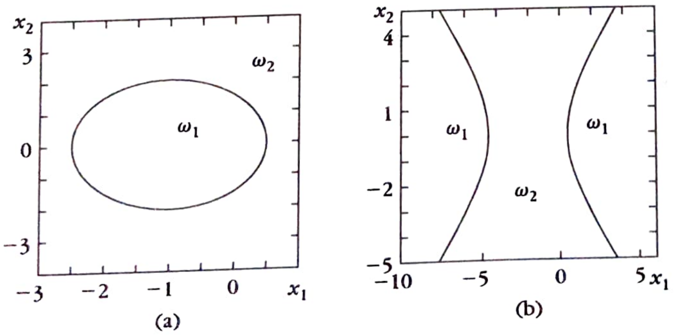


FIGURE 2.7

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.