PATTERN RECOGNITION ENGINEERING

MORTON NADLER
Chief Scientist, Image Processing Technologies

ERIC P. SMITH
Professor of Statistics
Virginia Polytechnic Institute and State University

A Wiley-Interscience Publication
JOHN WILEY & SONS INC.
New York / Chichester / Brisbane / Toronto / Singapore

CHAPTER 7

BAYES' OPTIMAL DECISIONS

In certain classes of problems we can use statistical techniques to obtain optimal functions for classifying objects. Figure 7.1 illustrates a typical commercial system. It displays a BioSonics, Inc., optical pattern recognition system. The system consists of a microscope, a video camera and monitor, a real-time video frame grabber, a digitizing tablet, and a hard-disk microcomputer with menu-driver software. The software allows full-frame processing of video images, interactive or automated measurement and extraction of features from the video image, and the analysis of these features using a variety of statistical pattern recognition techniques.

The system can be used to obtain a classifier in the following way. The basic objective is to obtain a function that will allow classification of an object into one of g classes with high accuracy. The starting point is a training set, a set of objects whose true classes are known. On this set of objects a number of features are extracted. As we have seen, these features may correspond to such simple measurements as lengths of objects or to more complex measurements. According to our basic model, it is the optical system that yields the measurements. The microscope and video camera provide an optical image in the form of an analog video signal, which is then converted to digital form. The image, now stored in the computer, can be processed and features extracted; the digital image can be converted back to analog and displayed, or it can be manipulated by digital processes and relevant features displayed. Some relevant features in this case may be shape descriptors or a luminance profile along a single line.

For example, Fig. 7.2 shows a video image that looks like a fingerprint. If a line is drawn from the center outward, one notices changes in luminance. The screen can be thought of as an array of pixels (say 512×512). Associated with each pixel will be a shade of gray ranging from 0 to 255. Thus, the displayed

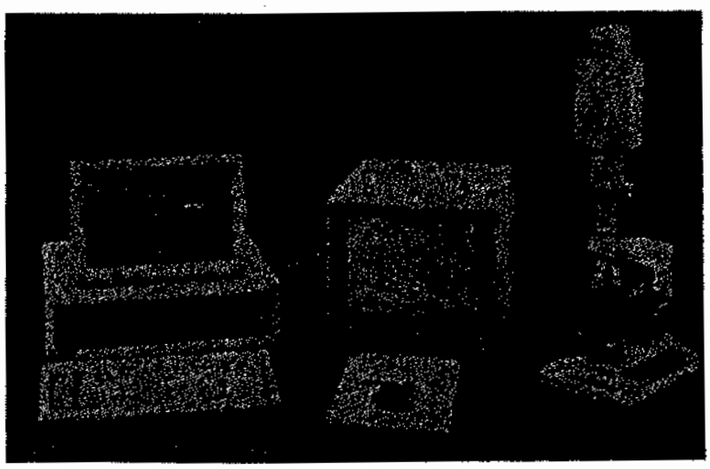


FIG. 7.1. BioSonics optical pattern recognition system (With permission, BioSonics, Inc., Seattle, Washington).

image is a 512 × 512 array of 8-bit numbers, with each number representing a shade of gray, or luminance, at a particular point on the screen. Since the luminance changes along a line, one may use as a set of features the "luminance profile." If we use the pixel values along the line as features, there would be 512 of them, corresponding to the 512 measurements. The number of features may be reduced by reducing the profile. Below the video screen image in Fig. 7.2 is a smoothed profile (the points have been connected and a moving average applied). Using the Fast Fourier Transform (FFT, see Chapter 4), we can represent this smoothed profile by a Fourier series of cosine functions of regularly increasing frequency, with various amplitudes and phase shifts. The profile is thus represented by an $N \times 2$ dimensional array $[a_i, b_i]$, $i = 0, 1, 2, \ldots, N-1$, where a_i is the amplitude, b_i the phase shift of the ith cosine function, and N the number of sampling units along the line. The index i corresponds to the harmonics, or spatial frequencies; $i = 0$ is the mean value, $i = 1$ represents the fundamental frequency, and so on. The quantities $[a_i, b_i]$ often provide an adequate summary of the information in the image. In the general case many of these harmonics will be small, so that a small subset of these features will be sufficient to analyze the patterns and separate the objects.

While the image displayed in Fig. 7.2 looks like a fingerprint, it is in fact a fish scale from a Chinook salmon. One of the uses of this pattern recognition system is to classify salmon and other fish species in terms of such characteristics as age, genetic stock, and so on. To develop the classification functions, we would collect fish from the various groups of interest and for each fish measure the features obtained from the FFT. Thus the design set consists of a number of vectors of harmonics for each of the groups. These vectors are then processed using a statis-

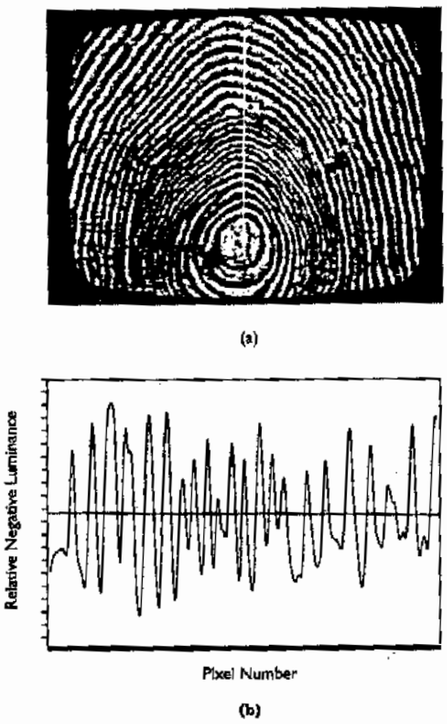


FIG. 7.2. Typical BioSonics images. (a) A fish scale. (b) Luminance profile of the section shown on the scale. (Reproduced with permission, BioSonics, Inc., Seattle, Washington).

tical procedure to obtain a classification function. There are many approaches for obtaining this rule, many of which involve application of Bayes' rule. There are three main methods commonly used: linear discriminant analysis, quadratic discriminant analysis, and nonparametric discriminant analysis. Differences among the methods are related to the assumptions that the user is willing to make about the information in the design set.

As we have seen, in statistical pattern recognition the starting point is the idea that measurements are observed with variation. In many situations the training set actually used is only one of many possible design sets. Animal measurements vary from animal to animal, EKG readings vary from subject to subject among patients from a heart disease category, and a letter may look different depending on the typewriter that is used, even when the type style is the same. In statistical pattern recognition we use information about the random components and the deterministic components of the design set to obtain optimal discriminants. In particular, this information is in the form of the measurement vectors in the design set. By

analogy to the training of a person or an animal to perform certain tasks, we often call the design set a "training" set (see Chapter 11).

7.1. MULTIVARIATE NORMAL DISTRIBUTION

For continuous measurements the most commonly used statistical distribution is the normal distribution (Chapter 6). In statistical pattern recognition, a common assumption is that the features in the training set follow a multivariate normal distribution. This assumption is important (especially for small training sets) because the distribution will affect the form of the classifier. It is therefore important to know what this assumption means in terms of the training set and how this relates to the classifier. Let $\mathbf{x}$ denote the $d \times 1$ feature vector and denote its transpose by $\mathbf{x}^T = (x_1, \ldots, x_d)$. We commonly view the feature vector $\mathbf{x}^T$ as being a sample from a d-variate normal distribution $N_d(\boldsymbol{\mu}, \Sigma)$. The density function is

$$f(x_1, x_2, \ldots, x_d) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\{-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\} \quad (7.1)$$

Here the parameter $\boldsymbol{\mu}$ is the vector of the feature means,[1] Σ is the covariance matrix, and $|\Sigma|$ is its determinant. This matrix has elements σ_{ij} which define the covariance between feature i and feature j. On the diagonal, σ_{ii} is just the variance of feature i.[2] It is assumed that Σ is nonsingular, so the inverse exists. This means that there are no linear redundancies in the data (i.e., no linear combination of the features is equal to a constant). When there is only one feature the model reduces to the univariate normal model

$$f(x) = \frac{\exp\{\frac{1}{2}(x - \mu)\sigma^{-2}(x - \mu)\}}{\sqrt{2\pi\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-(x - \mu)^2/2\sigma^2\} \quad (7.2)$$

The general shape of this function is important as it relates to the distribution of the measurements in feature space. The multivariate normal distribution has the well-known "bell" shape. This is illustrated in Fig. 7.3 for the case of two features.

The shape of the bell is primarily determined by the exponent term in the equation. The exponent $(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$ is called a "quadratic form." A quadratic form $Q(\mathbf{x})$ in the d features $x_1, x_2, \ldots, x_d$ is $Q(\mathbf{x}) = \mathbf{x}^T A \mathbf{x}$, where $\mathbf{x}^T = (x_1, x_2, \ldots, x_d)$ and A is a $d \times d$ symmetric matrix. If we write out this expression, the

[1] In the previous chapter we used $\bar{\mathbf{x}}$ to represent an estimate of this parameter from the data.

[2] We have here another case of differing notational conventions. In the case of a single feature we usually write variance as σ^2; it is the double subscript on σ_{ij} that stands in for the quadratic exponent.

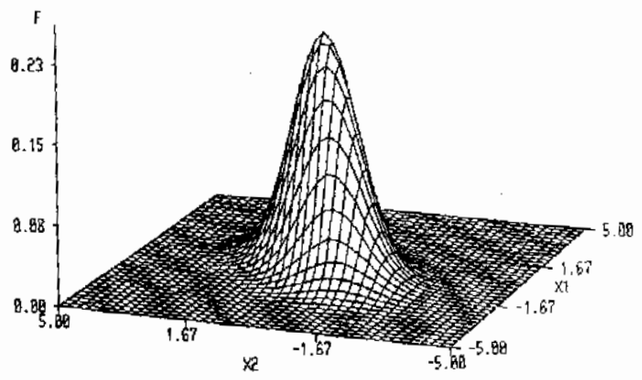


FIG. 7.3. The bivariate normal distribution.

quadratic form becomes $Q(\mathbf{x}) = \Sigma_{i=1}^{d} \Sigma_{j=1}^{d} a_{ij} x_i x_j$ or, for the exponent term, $Q(\mathbf{x}) = \Sigma_{i=1}^{d} \Sigma_{j=1}^{d} a_{ij}(x_i - \mu_i)(x_j - \mu_j)$, where a_{ij} is the ijth element of Σ^{-1}.

For two variables we get the bivariate normal $N_2(\boldsymbol{\mu}, \Sigma)$. A convenient way to write this density is as

$$f(x_1, x_2) = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1 - \rho_{12})}} \exp\left\{-\frac{1}{2(1 - \rho_{12}^2)} \cdot \left(\frac{(x_1 - \mu_1)^2}{\sigma_{11}} - 2\rho_{12}\frac{(x_1 - \mu_1)(x_2 - \mu_2)}{\sqrt{\sigma_{11}\sigma_{22}}} + \frac{(x_2 - \mu_2)^2}{\sigma_{22}}\right)\right\} \tag{7.3}$$

where ρ is the correlation between features 1 and 2. When there are several groups, each group can be represented by its individual distribution. If the groups are labeled $c_1, c_2, \ldots, c_g$ then the distributions can be represented by $f_i(\mathbf{x})$ or $p(\mathbf{x}|c_i)$. The latter is read "the probability distribution of $\mathbf{x}$ given class i" and is referred to as a *conditional distribution*—because it is conditional on the class. A common assumption in statistical pattern recognition is that the distributions are multivariate normal with a common covariance matrix but with means that depend on the class, that is, $p(\mathbf{x}|c_i)$ is $N_d(\boldsymbol{\mu}, \Sigma)$. We expect graphs of data x_i versus x_j to reflect the underlying distribution. If $\mathbf{x} \sim N_d(\boldsymbol{\mu}, \Sigma)$ (read "if the distribution of $\mathbf{x}$ is . . .)" one expects the graph of x_i versus x_j to be elliptical in shape and concentrated about the mean. The shape is of course determined by σ_{ii}, σ_{jj}, and σ_{ij}.

Contours of constant density for the d-dimensional normal distribution are ellipsoids defined by $\boldsymbol{x}$ such that $(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c^2$. These ellipsoids are centered at $\boldsymbol{\mu}$ and have axes defined by $\pm c\sqrt{\lambda_j}\boldsymbol{e}_j$, where the λ's are the ordered eigenvalues of Σ and the $\boldsymbol{e}_i$'s are the associated eigenvectors.

For the bivariate normal, the contours describe an ellipse. Note that the ellipse has an angle determined by $\boldsymbol{e}_1$ and is centered at (μ_1, μ_2). The length of the major

axis is $c\sqrt{\lambda_1}$, and the length of the minor axis is $c\sqrt{\lambda_2}$. The direction of the major axis is given by e_1, and the direction of the minor axis is given by e_2.

Some contours are graphed in Figs. 7.4–7.7 for the bivariate normal distribution. These contours are not the boundaries of compact clusters; they do not mean that all members of the classes indicated are inside the contour. Rather, they represent contours of constant probability density. Choosing other probability densities, we would have concentric ellipses inside or outside the ones shown.

In Fig. 7.4 the variances are equal and the correlation zero, so that if the axes have the same scale the contours are circles. The difference in the two circles that are displayed is one of location, with the means for the two distributions being the only difference. Figure 7.5 illustrates the stretching effect caused by a change in the variance. Now the variance in the x_2 direction is four times the variance in the x_1 direction. The effect of increasing the variance of a feature is to stretch the ellipse in that direction. A nonzero correlation will also stretch the ellipse, but at the same time it causes the ellipse to rotate in the direction defined by the magnitude and sign of the correlation. Figure 7.6 illustrates contours for two bivariate normal distributions with the same variance–covariance matrix, but with a nonzero correlation. Here the correlation is positive, and the effect is to rotate and stretch the ellipse. The angle of rotation and magnitude of the stretch depends on the covariance matrix.

Note also that in Fig. 7.6 the contours overlap. This shows that while the class means are well separated, some of the data have similar values. From a pattern recognition viewpoint, this suggests that objects falling in the overlap region will be difficult to classify.

Figure 7.7 illustrates the case where the variance–covariance matrices and the feature mean vectors differ. The differences between Fig. 7.6 and Fig. 7.7 are

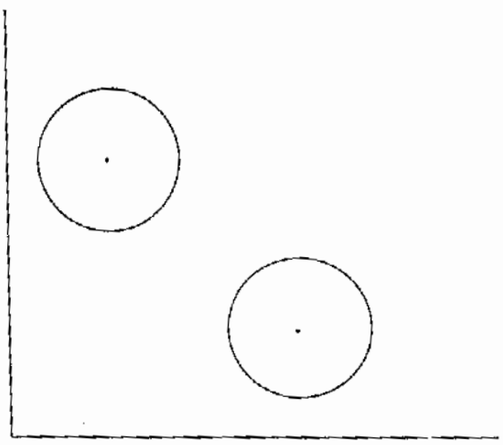


FIG. 7.4. Equiprobable contour plots of two bivariate normal distributions with equal variances and zero covariances and zero correlation.

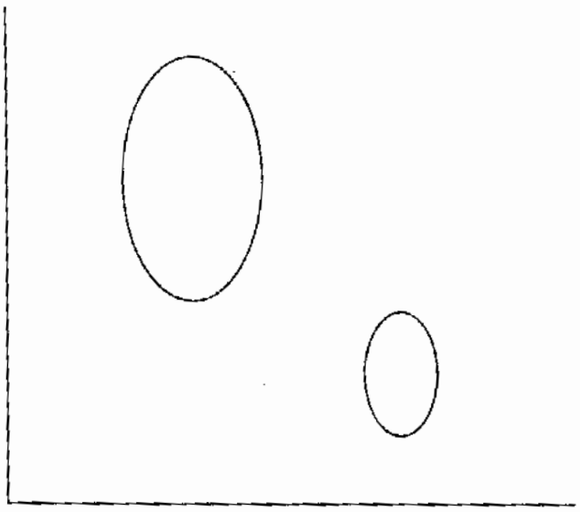


FIG. 7.5. Equiprobable contour plots of two bivariate normal distributions with unequal variances but zero covariances and zero correlation.

important. As will be seen below, groups that have contours similar to those of Fig. 7.6 lead to a classification rule that is linear and relatively simple to interpret. In the case of Fig. 7.7, a linear rule will generally not be optimal and one must resort to a more complex quadratic rule.

There are several distributions related to the multivariate normal that are useful. The first is the *marginal distribution*. By a marginal distribution, we mean the

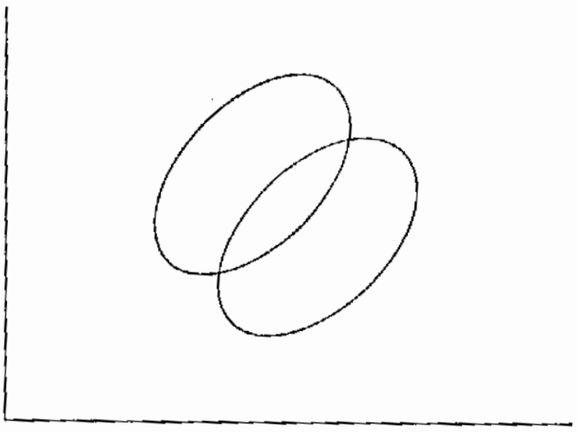


FIG. 7.6. Equiprobable contour plots of two bivariate normal distributions with equal covariance matrices but non-zero correlation.

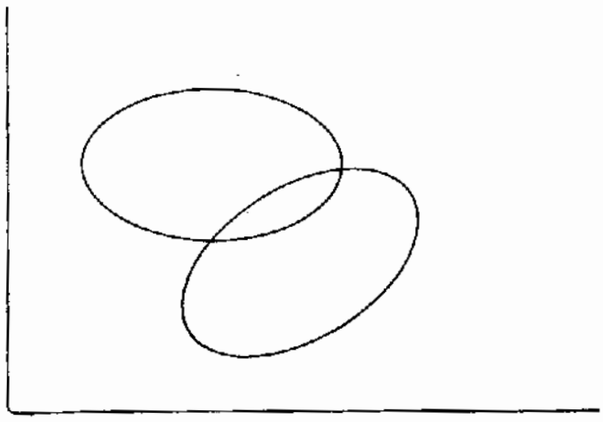


FIG. 7.7. Distributions leading to quadratic discriminant functions. Equiprobable contour plots of two bivariate normal distributions with unequal covariances.

distribution of any subset of the feature vector. Geometrically, the marginal distribution can be viewed as a projection of the distribution onto an axis defined by a feature or a plane defined by a pair of features, etc. If there are two features, and they are described by a bivariate normal distribution, then the marginal distribution of the first feature is what one would see if you stood on the axis at the mean and looked straight ahead at the bivariate distribution.

If x is multivariate normal, then each marginal distribution is normal.

This result has a number of uses. First it suggests that if the features are multivariate normal, then the individual features follow a univariate normal distribution. Secondly, each pair of features follows a bivariate normal. However, it is not necessarily true that if each univariate distribution is normal, then the joint distribution is multivariate normal. For example, if a bivariate normal distribution has two quarters cut out of it and is then rescaled, the marginal distributions may still be normal but the joint distribution is no longer a bivariate normal. Figure 7.8 shows a bivariate normal that has been altered in this manner. Note that the view shows only one of the quarters that is removed; the other is hidden. The marginal distribution for x_2 is seen to be normal. While this is an obviously contrived case, it illustrates the pitfalls inherent in the statistical assumptions that might be made.

Another important concept is a *conditional distribution*. A conditional distribution for a particular feature (or group of features), say x_i, is the distribution of the feature assuming that the other features (or a subset of the other features) are known. We discussed conditional distributions in the last chapter, where we asked the following question: What is the probability of a discrete event A, given another discrete event B? For continuous distributions we can ask for the distribution of a, given b, as the limit as the range for b passes to zero in the limit.

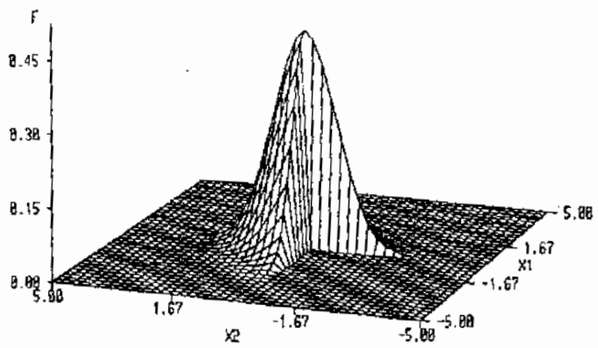


FIG. 7.8. Graph of a density that is not bivariate normal although both univariate marginal distributions are normal.

If $\mathbf{x}$ *is distributed as* $N_d(\boldsymbol{\mu}, \Sigma)$, *then every conditional distribution is normal.*

Conditional distributions are most commonly used in regression analysis, where we are interested in the dependence of one feature on a set of features.

A third type of distribution is called a *component distribution*. This is the distribution of any linear function of the features.

If $\mathbf{x}$ *is distributed as* $N_d(\boldsymbol{\mu}, \Sigma)$, *then component distributions are also normal.*

This can be used to obtain an important result: If $\mathbf{x}$ is distributed as $N_d(\boldsymbol{\mu}, \Sigma)$ and $\mathbf{a}$ is a vector of constants, then $Y = \mathbf{a}^T\mathbf{x}$ is distributed as $N(\mathbf{a}^T\boldsymbol{\mu}, \mathbf{a}^T\Sigma\mathbf{a})$. For example, $Y = 2x_1$ is normally distributed with mean $2\mu_1$ and variance $4\sigma_{11}$. The Karhunen–Loève expansion is a special case of a component chosen so that $\mathbf{a}^T\Sigma\mathbf{a}$ is a diagonal matrix.

7.2. PARAMETER ESTIMATION

The parameters of the multivariate normal distribution are the vector of feature means $\boldsymbol{\mu}$ and the variance–covariance matrix Σ. The estimate of $\boldsymbol{\mu}$ from a set of feature vectors is the vector of sample means $\bar{\mathbf{x}}^T = (\bar{x}_1, \ldots, \bar{x}_d)$, where d is the number of features, $\bar{x}_i = (1/N)\sum_{k=1}^{N} x_{ik}$, and N is the number of objects. The estimate of Σ is the sample variance–covariance matrix S. The elements of S are the sample variances $s_{ii} = [1/(N-1)]\sum_{k=1}^{N}(x_{ik} - \bar{x}_i)^2$ and the sample covariances $s_{ij} = [1/(N-1)]\sum_{k=1}^{N}(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)$. We use the factor $1/(N-1)$ to obtain an unbiased estimate, as we explained in the last chapter. Also, the estimated correlation between feature i and feature j is given by $r = s_{ij}/\sqrt{s_{ii}s_{jj}}$.

In the general case, $s_{ij} \neq 0$, if we assume multivariate normality, we need to find the d means of the x_i and the d^2 s_{ij}. But since $s_{ij} = s_{ji}$, there are only $d(d+3)/2$ parameters.

Example 7.1 One of the first data sets used for classification purposes was one with three species of iris: *setosa*, *versicolor*, and *virginica*. There were four features measured on each species, with 50 objects per group. The four features describe the shape and size of the flower and are sepal and petal length and width. The actual observations are given in Table 7.1. The mean vectors and covariance matrices are given in Table 7.2.

Note that the means for *setosa* are quite different from the means for the other two species (especially for the petal variables). Also, the covariance matrices appear to differ for all three of the species. In Fig. 7.9, the objects are plotted for the petal features. Note that the first species (*setosa*) appears to be well separated from the other two species. Also notice that there is overlap for *versicolor* (labeled □) and *virginica* (labeled △). Objects in the overlapping region are the ones that will be difficult to classify unless they are separated on another feature. The third feature of interest in the display is the shape of the data. The plots of the *versicolor* and *virginica* data are similar in shape and mostly differ in location. The *setosa* data have a different orientation and are compressed more. This reflects the differences in the variance–covariance matrices. The three-dimensional plot (Fig. 7.9(b)) gives a similar view.

In statistical pattern recognition, the multivariate normal distribution is commonly assumed as the distribution describing the features. One reason for its use is that it is well studied and the properties related to marginal, conditional, and component distributions are simple. While it is possible to use the methods described below on features that follow other distributions, most computer packages provide programs only for the normal case. Hence it may be difficult to implement such methods when data are non-normal.

In practice, non-normality is not all that uncommon. For example, if we want to develop a classifier for deciding if an individual should get a bank loan, there are certainly going to be questions leading to binary features. For example:

"Have you ever had a previous loan?"

"Have you ever defaulted on a loan?"

have yes/no, or binary responses. Such discrete-valued features are subject to the *binomial distribution* or, under certain conditions, to the *Poisson distribution*.

How important then is the assumption made in practice of multivariate normality? Multivariate normality is used in the theory described below in two ways. First, in the Bayes framework of pattern recognition, it is used as a measure of likelihood that an object came from a particular class. However, it is possible to develop a framework not based on the multivariate normal assumption but which leads to the same rules for classification when the data do in fact satisfy that distribution. An example of this would be Fisher's linear discriminant function, which we will consider later in this chapter. In any case, as we shall see below, the Bayes framework is perfectly general and does not require normality [Hand81].

TABLE 7.1. Iris Data[a]

Iris setosa				*Iris versicolor*				*Iris virginica*			
Sepal		Petal		Sepal		Petal		Sepal		Petal	
l	*w*	*l*	*w*	*l*	*w*	*l*	*w*	*l*	*w*	*l*	*w*
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.2	2.9	5.6	1.8
5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3.0	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3.0	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0	6.5	3.2	5.1	2.0
4.8	3.4	1.6	0.2	5.9	3.0	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3.0	1.4	0.1	6.0	2.2	4.0	1.0	6.8	3.0	5.5	2.1
4.3	3.0	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5.0	2.0
5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5	6.5	3.0	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6.0	2.2	5.0	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4.0	1.3	5.6	2.8	4.9	2.0
4.6	3.6	1.0	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2.0
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5.0	3.0	1.6	0.2	6.6	3.0	4.4	1.4	7.2	3.2	6.0	1.8
5.0	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3.0	5.0	1.7	6.1	3.0	4.9	1.8
5.2	3.4	1.4	0.2	6.0	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1.0	7.2	3.0	5.8	1.6
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1.0	7.9	3.8	6.4	2.0
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
5.5	4.2	1.4	0.2	6.0	2.7	5.1	1.6	6.3	2.8	5.1	1.5
4.9	3.1	1.5	0.2	5.4	3.0	4.5	1.5	6.1	2.6	5.6	1.4
5.0	3.2	1.2	0.2	6.0	3.4	4.5	1.6	7.7	3.0	6.1	2.3
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3.0	1.3	0.2	5.6	3.0	4.1	1.3	6.0	3.0	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4.0	1.3	6.9	3.1	5.4	2.1
5.0	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3.0	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4.0	1.2	5.8	2.7	5.1	1.9

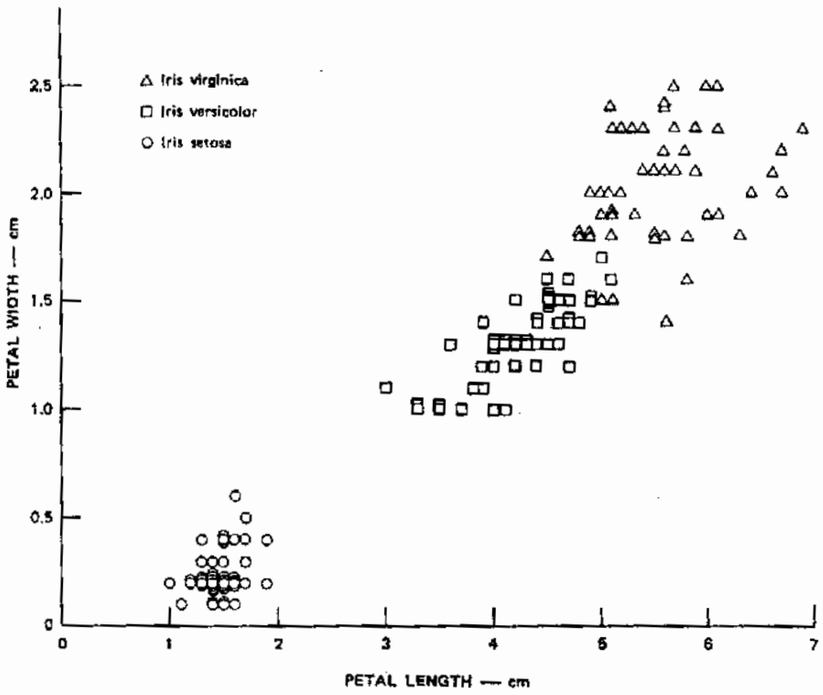
TABLE 7.1. (*Continued*)

Iris setosa				*Iris versicolor*				*Iris virginica*			
Sepal		Petal		Sepal		Petal		Sepal		Petal	
l	*w*	*l*	*w*	*l*	*w*	*l*	*w*	*l*	*w*	*l*	*w*
5.0	3.5	1.6	0.6	5.0	2.3	3.3	1.0	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3.0	1.4	0.3	5.7	3.0	4.2	1.2	6.7	3.0	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5.0	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3.0	5.2	2.0
5.3	3.7	1.5	0.2	5.1	2.5	3.0	1.1	6.2	3.4	5.4	2.3
5.0	3.3	1.4	0.2	5.7	2.8	4.1	1.5	5.9	3.0	5.1	1.8

[a]Reprinted with permission from R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, **7**, 179–188, 1936.

TABLE 7.2. Summary Statistics for Iris Data

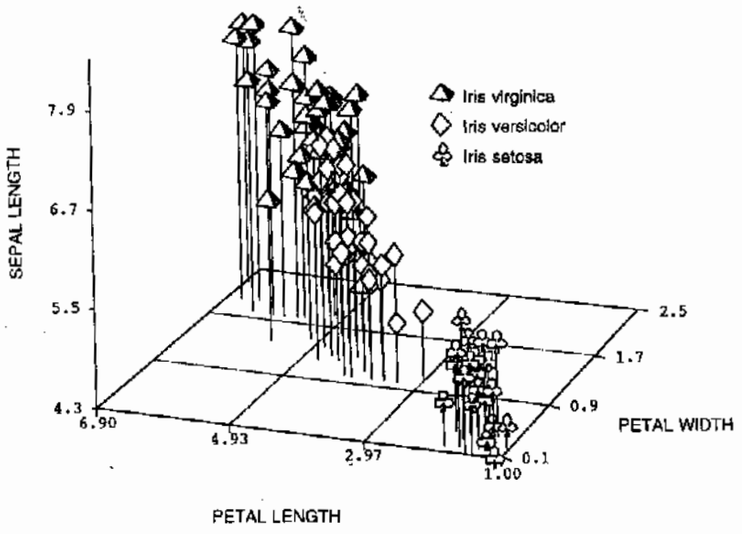
	Variable			
Means	Sepal Length	Sepal Width	Petal Length	Petal Width
Iris setosa	5.00	3.43	1.46	0.25
Iris versicolor	5.94	2.77	4.26	1.33
Iris virginica	6.59	2.97	5.55	2.03
	Variable			
Covariance Matrices	Sepal Length	Sepal Width	Petal Length	Petal Width
Iris setosa	0.124	0.099	0.016	0.010
	0.099	0.143	0.011	0.009
	0.016	0.011	0.030	0.006
	0.010	0.009	0.006	0.011
Iris versicolor	0.266	0.085	0.183	0.056
	0.085	0.098	0.082	0.041
	0.183	0.082	0.220	0.073
	0.056	0.041	0.073	0.039
Iris virginica	0.405	0.094	0.303	0.049
	0.094	0.104	0.071	0.047
	0.303	0.071	0.304	0.049
	0.049	0.047	0.049	0.075



(a)

FIG. 7.9. Data plots for the three iris species: setosa, versicolor, virginica. (a) Plot of petal width vs. petal length. (Reproduced with permission from [Duda73; copyright 1973 by John Wiley and Sons.) (b) Three-dimensional plot for three variables.

A second use of normality is in inference about parameters and differences between classes. Also the classification functions result in relatively simple linear or quadratic discriminant functions. When the number of features is large or the number of objects in the training set is small, the assumption of normality is important. In any case, it is important to check training sets prior to development of a classifier. Often, simple checks such as scatter plots of the data are adequate for noticing strong differences in variance and covariance and are also useful for locating odd data values. Other checks, such as drawing univariate histograms or stem-and-leaf displays (Fig. 7.9(b)) or boxplots, are useful for checking univariate normality (Walpo85, Koopm87]). If all the features are reasonably normal, then the approach based on the multivariate normality assumption is reasonable. Problems such as skewness and other evidence of non-normality can often be detected in these displays and adjusted via transformation of features.



(b)

FIG. 7.9. (*Continued*)

7.3. BAYES' OPTIMAL DECISIONS

The objective of this section is to describe an approach to statistical pattern recognition that is based on information from the assumed distribution of the features. The most commonly used approach involves the multivariate normal distribution, although methods that do not require a distributional assumption exist and will be discussed later. The approach that is commonly used is an optimization method based on error rates and Bayes' rule.

In statistical pattern recognition, we recognize that features may be measured with error and that some of the features are useful for identification of the class while others are not. Our goals are then to obtain useful sets of features and to use these features in such a way that the identification is as accurate as possible—that is, that the number of classification errors (or probability of misclassification) are as small as possible. If there is an object that is to be classified on the basis of a feature vector $\mathbf{x}$, into one of g possible classes $(c_1, c_2, \ldots, c_g)$, then the probability that the object is classified into class i when $\mathbf{x}$ is observed can be described by $P(c_i|\mathbf{x})$. This probability is obtained by Bayes' theorem, which we shall now derive. We start from the "theorem on compound probabilities" [Felle57]. This states that if there are random variables x and y, then

$$P(x\&y) = P(y|x)P(x) \tag{7.4}$$

which is the same relation we have already seen in Eq. (6.2), or, equivalently,

$$P(y|x) = P(x\&y)/P(x) \tag{7.5}$$

This can be diagrammed as shown in Fig. 7.10. In our situation, x is the set of features and y represents the class variable c_i. But also

$$P(x|y) = P(x\&y)/P(y) \tag{7.5'}$$

From these two relations we obtain

$$P(x\&y) = P(x|y) \cdot P(y) = P(y|x) \cdot P(x) \tag{7.4'}$$

Finally,

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \tag{7.5''}$$

Substituting for x and y in Eq. (7.5''), we obtain the probability that the class is i when feature x is observed, $P(c_i|x)$. To evaluate this probability, we need to know the probability of observing x when the class is c_i, $P(x|c_i)$. We get

$$P(c_i|x) = \frac{P(x|c_i)P(c_i)}{P(x)} \tag{7.6}$$

This is Bayes' theorem, which gives the probability of a class i being present when a feature x is observed, provided we know the probability of the feature being observed when the class is present, the probability of that class being present, and the probability of that feature. The probability of a class being present is called the "prior probability" or, in statisticians' jargon, the "prior."

Since, in fact, the particular feature may be observed in a number of classes, the denominator on the right-hand side is given by a sum. Also, in the general

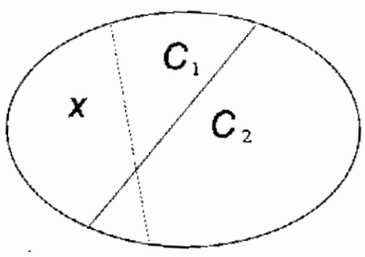


FIG. 7.10. Conditional and joint probabilities.

case we have a feature vector $\mathbf{x}$. So, finally, we get

$$P(c_i|\mathbf{x}) = \frac{p(\mathbf{x}|c_i)P(c_i)}{\sum_{j=1}^{g} p(\mathbf{x}|c_j)P(c_j)} \tag{7.7}$$

where $P(c_i)$ or P_i, the prior probability, is the unconditional probability that the class is i independently of the information provided by $\mathbf{x}$ and $p(\mathbf{x}|c_i)$ is the density of $\mathbf{x}$ when the class is c_i. $P(c_i|\mathbf{x})$ is referred to as the "posterior probability for class i."

Bayes' theorem is extremely important, because the probability of group membership is written in terms of distributions that can be evaluated on the basis of data. It can be used to derive decision rules that are optimum for fairly broad classes of optimum criteria. *Even when the necessary data cannot be obtained, Bayes' rule will define the limiting performance that any suboptimal decision rule can attain.*

In pattern recognition we are interested in obtaining a rule based on a set of features that is useful for classification of objects into one of the g possible classes. Such a rule is called an "allocation" or "decision" rule. In statistical pattern recognition, the probability distribution of the features and Bayes' formula are used to obtain a decision rule.

As a simple example, consider Fig. 7.11, where the distributions of a single feature are represented. Note that the populations overlap. What this means is that some of the data from population c_1 will fall close to those of population c_2, and vice versa. One approach to developing a rule is based on the overlap between the distributions. Suppose, for example, that the rule we decide to use selects population c_1 if $x \leq A$ and population c_2 if $x > A$. Then sometimes the decision will be incorrect. Let $P(c_1|c_2)$ be the probability that population c_1 is selected when in fact the true population is c_2. This represents the probability of an incorrect decision, the probability of misclassification for group c_2. There is a similar probability for group c_1, $P(c_2|c_1)$.

These errors can also be viewed in a decision theory context. Assume the true class is c_1. The Type I error is the error of assigning the object to the wrong class $P(c_2|c_1)$. We see that in Fig. 7.11 this is the area under the curve for c_1 to the right of the cutoff point. The second type of error, the Type II error, is to "accept" the object as belonging to group c_1 when in fact it belongs to group c_2. This is just $P(c_1|c_2)$, the area under the curve for c_2 to the left of the cutoff point. The total probability of misclassification is the sum of these areas.

One criterion for obtaining an optimal rule is to minimize the probability of misclassification, a rule sometimes called "Bayes' optimal rule." It is important to know that the optimality of a Bayesian decision rule is independent of the form of the distribution. However, to apply the rule it is necessary to know the distribution. Since $P(c_1|c_2)$ and $P(c_2|c_1)$ are areas, they may be expressed as integrals.

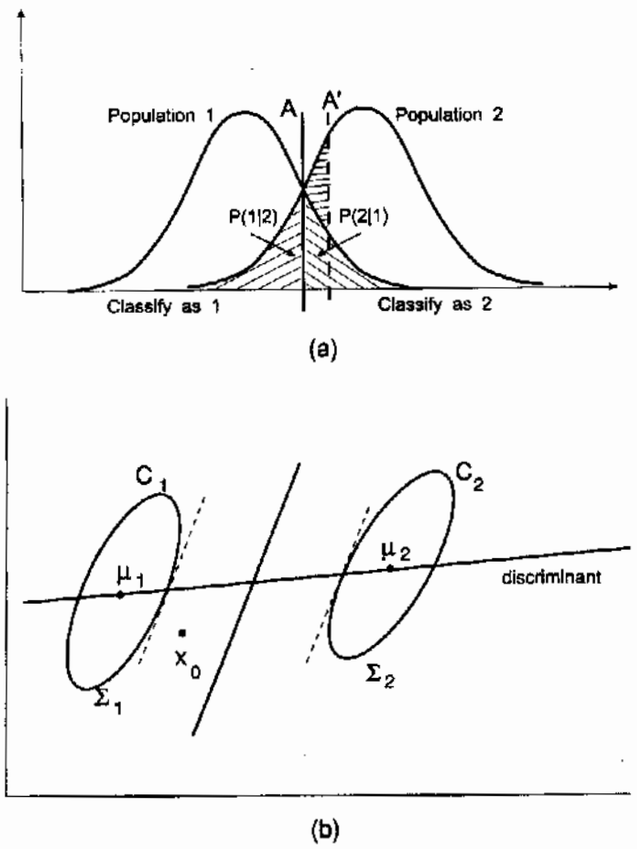


FIG. 7.11. Plot of probability densities for two classes with equal variances and discriminant boundary. (a) Derivation of the minimum-error Bayes' decision surface. (b) The minimum-error decision surface.

Thus

$$P(c_1|c_2) = \int_{R_1} p(\mathbf{x}|c_2)\, d\mathbf{x} \tag{7.8}$$

is the area associated with the density $p(\mathbf{x}|c_2)$ in the region R_1; $P(c_2|c_1)$ is defined analogously. Then the total probability of misclassification may be expressed as

$$T = P(c_2) \int_{R_1} p(\mathbf{x}|c_2)\, d\mathbf{x} + P(c_1) \int_{R_2} p(\mathbf{x}|c_1)\, d(\mathbf{x}) \tag{7.9}$$

Seber [1984] shows that this integral is minimized when $R_1 = \{\mathbf{x}|P(c_1)p(\mathbf{x}|c_1) - P(c_2)p(\mathbf{x}|c_2) > 0\}$. The optimal Bayes rule is then to assign the object with feature vector $\mathbf{x}_0$ to class c_1 if

$$\frac{p(\mathbf{x}|c_1)}{p(\mathbf{x}|c_2)} > \frac{P(c_2)}{P(c_1)} \tag{7.10}$$

and to assign $\mathbf{x}_0$ to class c_2 otherwise.

We can see this graphically in Fig. 7.11(a). The right side of (7.10) shows that the optimal decision boundary is at A, the crossover point of the two distributions. If we shift the decision boundary away from this point, say to A', the error areas can only increase.

The conditional density $p(\mathbf{x}_0|c_i)$ is known as the "likelihood" of $\mathbf{x}_0$ with respect to c_i. The expression on the left-hand side of Eq. (7.10) is therefore known as the "likelihood ratio."

Minimum error is the simplest optimum criterion that can be used with Bayes' rule. It is possible to assign weights of various kinds to the various types of errors and to obtain optimal decision rules for more complex situations. For example, we can assign costs, that is, penalties assessed for each type of error. In Bayesian terminology these are generally called "risks."

It is reasonable to assume that errors cost more than correct decisions: that is, if r_{ij} is the risk connected with deciding c_j when the true class is c_i and r_{ii} is the risk connected with deciding c_i when in fact the class is c_i, then $r_{ij} > r_{ii}$ and $r_{ji} > r_{jj}$. Let R_i be the region in feature space where we decide $\mathbf{x} \in c_i$ and let R_j be the region where we decide c_j. We want to choose R_i and R_j in such a way as to minimize the total risk [Fukun90]. Then the decision is to decide c_i if

$$(r_{ij} - r_{ii})p(\mathbf{x}|c_i)P(c_i) > (r_{ji} - r_{jj})p(\mathbf{x}|c_j)P(c_j) \qquad \text{for } \mathbf{x} \in \mathrm{R}_i \tag{7.11}$$

where, by virtue of our assumption about relative costs, $r_{ij} > r_{ii}$ always, so the left and right expressions in the inequality are always positive. A symmetric rule applies to the decision for c_j; in the case of equality (i.e., of equal costs), it is immaterial which decision is made. See also the discussion of the general case at the end of the chapter.

To return to (7.10), it is often used in the form $\log p(\mathbf{x}|c_1) - \log p(\mathbf{x}|c_2) > \log p(c_2) - \log p(c_1)$, because the computations are easier. The natural logarithm, denoted ln, is often used here, because of the exp function that enters into the definitions.

We shall now consider the case of normal populations, first under the assumption that the covariance matrices are equal and then under the assumption that they are not.

7.4. NORMAL POPULATIONS WITH EQUAL VARIANCE-COVARIANCE MATRICES

In this section we examine the case where the populations are assumed to be normal with equal variance-covariance matrices. Then

$$p(\mathbf{x}|c_1)/p(\mathbf{x}|c_2) = \exp\{-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\}$$

$$= \exp\{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} + \tfrac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\} \tag{7.12}$$

If we take the logarithm of the ratio we find a linear rule, as the terms that are quadratic in x sum to zero:

$$\ln p(\mathbf{x}|c_1) - \ln p(\mathbf{x}|c_2) = (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} - \tfrac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2) \tag{7.13}$$

To see that this is a linear rule in x, note that if we put $\boldsymbol{l} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, then

$$\ln p(\mathbf{x}|c_1) - \ln p(\mathbf{x}|c_2) = \boldsymbol{l}^{\mathrm{T}}(\mathbf{x} - \tfrac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)) \tag{7.14}$$

This leads to the allocation rule: Assign the object with feature vector $\mathbf{x}_0$ to group c_1 if

$$\boldsymbol{l}^{\mathrm{T}}(\mathbf{x}_0 - \tfrac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)) > \ln(p_2/p_1) \tag{7.15}$$

This rule may be expressed in terms of the individual features as

$$l_1x_{01} + l_2x_{02} + \cdots + l_dx_{0d} - c_0 > 0 \tag{7.16}$$

where $c_0 = \frac{1}{2}\boldsymbol{l}^{\mathrm{T}}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2) + \ln(p_2/p_1)$. In using the rule with the training set, we use estimates of the mean and covariance matrix. The rule then becomes: Assign the object with feature $\mathbf{x}_0$ to class c_1 if $(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^{\mathrm{T}}\boldsymbol{S}_p^{-1}\mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^{\mathrm{T}}\boldsymbol{S}_p^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2) > \ln(p_1/p_2)$ and to class c_2 otherwise. Here $\boldsymbol{S}_p$ is the pooled covariance matrix:

$$\boldsymbol{S}_p = \frac{1}{n_1+n_2-2}[(n_1-1)\boldsymbol{S}_1 + (n_2-1)\boldsymbol{S}_2] \tag{7.17}$$

Now consider the classification problem from the viewpoint of assigning the object to the class that yields the maximum likelihood, given the data or the max-

imum posterior probability. We find from Eq. (7.7) the posterior probability as

$$P(c_1|\mathbf{x}_0) = \frac{p(\mathbf{x}_0|c_1)P(c_1)}{p(\mathbf{x}_0|c_1)P(c_1) + p(\mathbf{x}_0|c_2)P(c_2)} \tag{7.18}$$

Symmetrically,

$$P(c_2|\mathbf{x}_0) = \frac{p(\mathbf{x}_0|c_2)P(c_2)}{p(\mathbf{x}_0|c_1)P(c_1) + p(\mathbf{x}_0|c_2)P(c_2)} \tag{7.18'}$$

Since the denominators are the same for both the posterior probabilities, they cancel out in the decision process and need not be computed.

We classify the object as c_1 if $P(c_1|\mathbf{x}_0) > P(c_2|\mathbf{x}_0)$, that is, in the group with the higher posterior probability. By considering the numerators, it is easily seen that this rule is equivalent to using $p(\mathbf{x}_0|c_1)P(c_1) > p(\mathbf{x}_0|c_2)P(c_2)$, which is the previous rule.

For normal populations with equal covariance matrices, the posterior probability for class c_i can be computed in the general multiclass problem (of g classes) as[3]

$$P(c_i|\mathbf{x}_0) = \frac{P(c_i)e^{-1/2D_i^2}}{\sum_j P(c_j)e^{-1/2D_j^2}} \tag{7.19}$$

where $D_i^2 = (\mathbf{x}_0 - \boldsymbol{\mu}_i)^T \Sigma^- (\mathbf{x}_0 - \boldsymbol{\mu}_i)$; D_i^2 measures how far $\mathbf{x}_0$ is from the feature mean vector for class i, and the sum in the denominator is over all the classes. It is sometimes useful to look at posterior probabilities for cases that do not clearly define one of the groups. To estimate these probabilities, one substitutes the parameter estimates for $\boldsymbol{\mu}_i$ and Σ. For example, if we found that $p(c_1|\mathbf{x}_0) = 0.35$, $p(c_2|\mathbf{x}_0) = 0.32$, and $p(c_3|\mathbf{x}_0) = 0.33$, we would normally assign the object to c_1. But it is clear that there is high probability of an incorrect assignment. In such cases additional information may be useful to improve the classification reliability. In this connection, see again Fisher's z-transformation, in section 6.5.4.

Example 7.2. Consider again the iris data and assume that we want to classify an iris as either a *setosa* or a *versicolor* variety. The pooled covariance matrix is given by Table 7.3. The linear discriminant function (assuming equal priors) is

$$(\mathbf{x}_1 - \mathbf{x}_2)^T S_p^{-1} = 3.03l_s + 18.01w_s - 21.76l_p - 30.84w_p \tag{7.20}$$

where the l_s, w_s, l_p, w_p are the sepal and petal length and width, respectively. The value of c_0 is 13.95. If the priors are taken to be equal, the classification rule is to

[3]Here is an example where the symbol Σ has suddenly changed back to its other significance. To warn the reader, we have included the index of summation, j, which runs over all the classes from c_1 to c_g.

TABLE 7.3. Pooled Iris Covariance Matrix for *setosa* and *versicolor*

	Sepal Length	Sepal Width	Petal Length	Petal Width
$S_p =$	0.195	0.092	0.099	0.033
	0.092	0.121	0.047	0.025
	0.099	0.047	0.125	0.039
	0.033	0.025	0.039	0.025

assign the plant with feature vector $\mathbf{x}_0$ to iris species *setosa* if

$$3.03l_s + 18.01w_s - 21.76l_p - 30.84w_p + 13.95 > 0 \qquad (7.21)$$

A two-dimensional representation of this is given in Fig. 7.11(b). As before the ellipses represent the equiprobable contours of the two distributions with equal variance–covariance matrices and different means. The optimal discriminant is any line parallel to the line joining the means of the two distributions, while the optimal decision boundary is a line through the midpoint of the line joining the means, and parallel to the tangents to the equiprobable contours where the line joining the midpoints intersects them.

7.5. FISHER'S METHOD

In d-space the line joining the means (i.e., the linear discriminant function) is not so easy to visualize, but in the case of two groups there always is one. Fisher [1936] designed a method in which points in d-space are projected onto that line (or a line parallel to it). He viewed separation and classification together and used an approach based on the distance between groups. One requirement is that the rule obtained be a linear rule. Hence we are interested in linear combinations $Y = l^T\mathbf{x}$. From our previous results, the mean of the combination is $\mu_Y = l^T\mu_x$ while the variance of Y is $V(Y) = l^T \Sigma l$. Using the linear combination changes a multivariate problem into a univariate problem because Y has dimension one. Fisher's idea was to pick l so as to maximize

$$\frac{(l^T\mu_1 - l^T \mu_2)^2}{l^T \Sigma l} = \frac{\text{squared distance between the } Y \text{ means}}{\text{variance of } Y} \qquad (7.22)$$

The maximum is obtained when $l = c\Sigma^{-1} (\mu_1 - \mu_2)$, where c is an arbitrary constant. The constant indicates that the solution is not unique but depends on a scaling factor. A commonly used scaling factor is to choose c so that the variance of each Y variable is unity. The maximum possible value is $D^2 = (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2)$, which is the Mahalanobis distance between the class means that we discussed earlier.

To use this we substitute the parameter estimates for the parameters. Hence

$$\hat{l} = cS_p^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2); \tag{7.23}$$

$\hat{l}$ is sometimes referred to as "Fisher's linear discriminant function" (Fig. 7.12).

It is sometimes useful to report the sample distances between the class means as an indication of the degree of separation between pairs of classes. These values are computed as

$$\hat{D}^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathrm{T}} S_p^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = \hat{l}^{\mathrm{T}}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \tag{7.24}$$

The quantity $T^2 = [N_1N_2/(N_1 + N_2)]D^2$ where N_i is the size of the sample from class i, is called "Hotelling's T^2 statistic" and is useful for testing differences between the feature mean vectors for the two groups. To test the hypothesis H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ versus the alternative H_1: $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, we convert T^2 to Snedecor's F [Burin70][4] by

$$F = \frac{N_1 + N_2 - d - 1}{d(N_1 + N_2 - 2)} \times \frac{N_1N_2}{N_1N_2}D^2 \tag{7.25}$$

We compare F with the critical F value using a table of the F-distribution [Burin70], entering the table with d and $N_1 + N_2 - d - 1$ degrees of freedom. For the two iris species, $D^2 = 103.23$. The F statistic is 625.45 with 4 and 95 degrees of freedom. We enter the table for F at the confidence level 0.01 [Burin70] with 4 and 100 degrees of freedom, since that is the nearest to our 95; we find that the critical F value $F_{4,100,0.01}$ equals 13.57, so the result is very much greater than would be given by pure chance and indicates good separation between the species.

The midpoint between the two groups on the two derived variables is $\frac{1}{2}(\mu_{Y1} + \mu_{Y2})$, where $\mu_{Yi} = \hat{l}^{\mathrm{T}}\boldsymbol{\mu}_i$. To allocate an object with feature vector $\mathbf{x}_0$ the first step is to compute $y_0 = \hat{l}^{\mathrm{T}}\mathbf{x}_0$. The allocation rule is to assign the object with feature vector $\mathbf{x}_0$ to group c_1 if $y_0 > \frac{1}{2}(\mu_{Y1} + \mu_{Y2})$, else to c_2. If this rule is converted back to the original scale, it becomes: assign the object with feature vector $\mathbf{x}_0$ to group c_1 if

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}\Sigma^{-1}\mathbf{x}_0 > \tfrac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \tag{7.26}$$

In terms of data the decision rule is to assign to population c_1 if

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathrm{T}} S_p^{-1}\bar{\mathbf{x}}_0 > \tfrac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathrm{T}} S_p^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \tag{7.27}$$

else assign to c_2. You can see that this is just Bayes' rule with equal priors. *When Bayes' linear rule holds, Fisher's method gives identical results.* When the Bayes

[4]The F distribution was so named after Ronald Fisher by Snedecor.

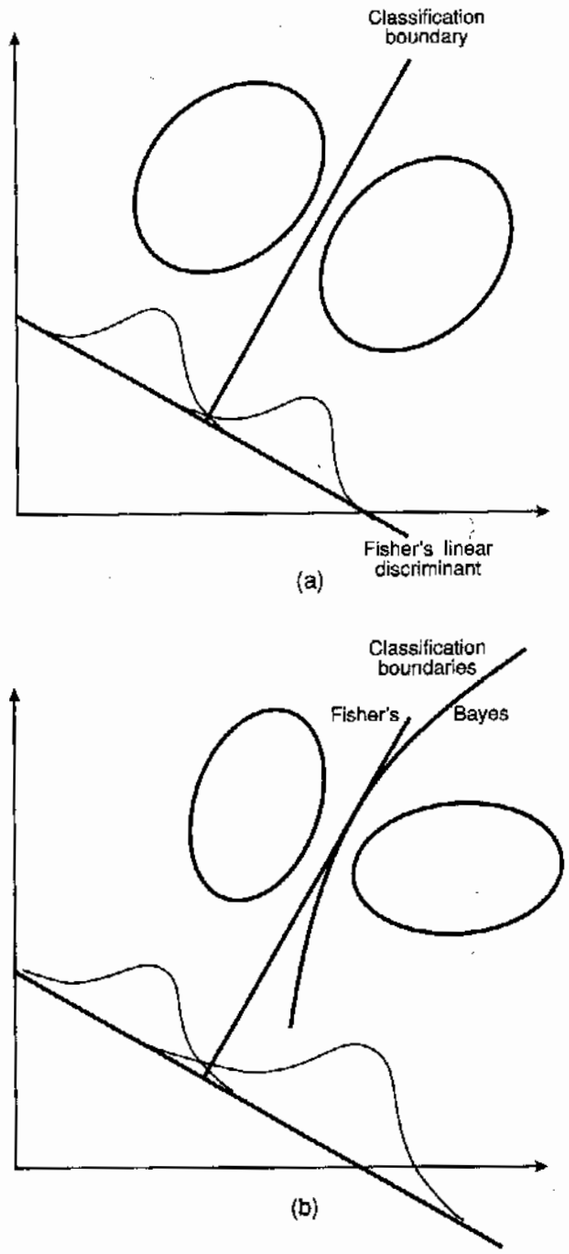


FIG. 7.12. Fisher's linear discriminant. (a) When the Bayes optimal discriminant is linear. (b) When the Bayes optimal discriminant is quadratic.

linear rule does not hold, Fisher's linear discriminant will give only suboptimal results (Fig. 7.12(b)).

7.5.1. Geometry of Fisher's Rule

Fisher's method is interesting from a geometrical viewpoint. We can gain some intuitive understanding of the formulas from Figs. 7.4–7.7. In Fig. 7.4 a simple case is displayed, because the variances are equal. In this case $\Sigma^{-1} = 1/\sigma^2 I$, where I is the identity matrix and we classify according to the rule defined by Eq. (7.26). If this is expanded, it becomes

$$\frac{\mu_{11} - \mu_{21}}{\sigma^2} x_{01} + \frac{\mu_{12} - \mu_{22}}{\sigma^2} x_{02} > \frac{1}{2}\left(\frac{\mu_{11}^2 - \mu_{21}^2}{\sigma^2} + \frac{\mu_{12}^2 - \mu_{22}^2}{\sigma^2}\right) \quad (7.26')$$

or

$$(\mu_{11} - \mu_{21})x_{01} + (\mu_{12} - \mu_{22})x_{02} > (\mu_{11} - \mu_{21})\bar{\mu}_1 + (m_{12} - \mu_{22})\bar{\mu}_2 \quad (7.26'')$$

where $\bar{\mu} = \frac{1}{2}(\mu_1 + \mu_2)$. Note that $\bar{\mu}$ is the midpoint of the line segment joining the means of the two classes, while the slope of the line segment is $(\mu_{12} - \mu_{22})/(\mu_{11} - \mu_{21})$. We see that in the simple case of Fig. 7.4 the discriminant function is the equation of any line parallel to the line connecting the means. If $y_0 = l^T \mathbf{x}_0$, then y_0 can be thought of as a projection onto that line. The quantity y_0 is sometimes called a "discriminant score." By using y_0 for classification we have simplified the classification problem, as we have reduced it to a univariate problem. The object can be classified by comparison of y_0 with $\mu_y = l^T(\mu_1 - \mu_2)/2$. In this way Fisher's method can be viewed as a distance method.

When the variance–covariance matrices are equal but the covariance is not zero, the geometry is a little more complicated. Because the contours are not circular, but rather elliptical, a first step is to transform the ellipses to circles by an appropriate shrinking and rotation. The matrix that will produce the appropriate transformation is $\Sigma^{-1/2}$. This is somewhat similar to what is done in a two-sample Student's t-test. One wants to compare the difference between two means, but because the absolute difference may change with changes of scale, one divides by the appropriate standard deviation to remove scale effects. Once in the scaled space, the ellipses become circles and the problem is like the one in Fig. 7.4.

The slope of the line that separates the means in the scaled space is defined by $(\mu_1 - \mu_2)^T \Sigma^{-1/2}$. In the unscaled space, this corresponds to a line with slope defined by $(\mu_1 - \mu_2)^T \Sigma^{-1/2} \Sigma^{-1/2}$, or $(\mu_1 - \mu_2)^T \Sigma^{-1}$. One possible line is illustrated in Fig. 7.12(a). The contours in this figure have equal variance covariance matrices and nonzero correlation. The line that is drawn there is a line with slope defined by $(\mu_1 - \mu_2)^T \Sigma^{-1}$. Note that the scaling on this new line is arbitrary. From a classification point of view, it is only the location of a point relative to the others that is important, not the numerical value of that point. A common scaling to use is in terms of standard deviation units. Thus l may be adjusted so that Var(y)

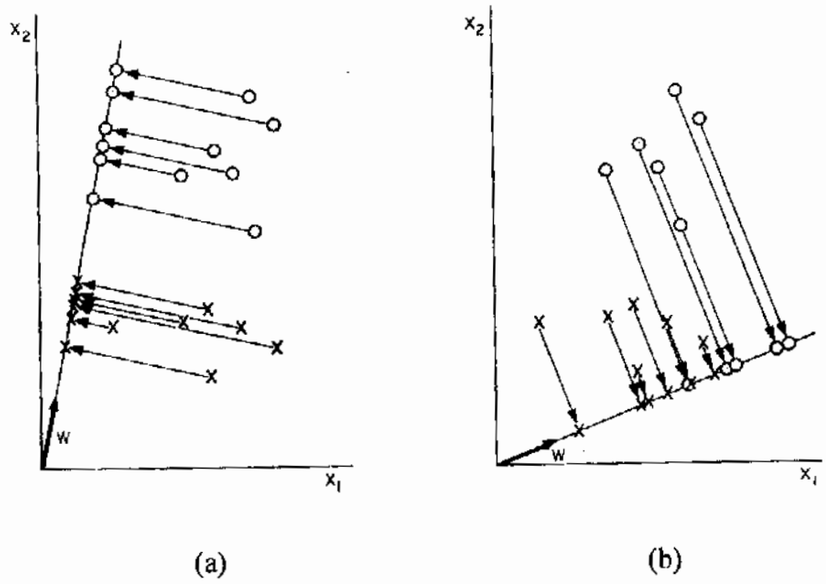


FIG. 7.13. Scatter plot of data, discriminant line using two features, and projection of data onto discriminant line. (a) Fisher's linear discriminant. (b) Nonoptimal discriminant. (Reproduced with permission from [Duda73]; copyright 1973 by John Wiley & Sons.)

$= \text{Var}(l^T\mathbf{x}) = l^T l = 1$. When this scaling is used, y is sometimes referred to as a "canonical variate" [Kleck80].

We have drawn the discriminant line in Fig. 7.13 with the data points projected onto this line. The slope is determined from Fisher's linear discriminant function. The other line in Fig. 7.12(a) defines the classification boundary. This line is perpendicular to the discriminant function and passes through the point associated with $\frac{1}{2}(\mu_1 - \mu)^T \Sigma^{-1} (\mu_1 + \mu_2)$, thus separating the space into the regions Γ_1 and Γ_2. The effect of the prior probability is to shift the boundary line up or down the discriminant line. If, for example, $P(c_1)$ is larger than $P(c_2)$ we shift the boundary towards the second class, thus improving classification for the group with the higher prior. Thus for the case of two groups, Bayes' method represents a slight modification of Fisher's method. We emphasize that *in Fisher's method, the normality assumption is not used*, although it works best (i.e., is equivalent to Bayes' rule) when the assumption is valid. Indeed, when the optimal decision surface is far from linear it may give very bad results [Hand81].

In the general case it is necessary to find the principal axis of the data (see the discussion on the K-L transformation, Chapter 4) on which to project the data (Fig. 7.13(a)). When this is not done, the result can be quite disappointing (Fig. 7.13(b)).

7.6. NORMAL POPULATIONS WITH UNEQUAL VARIANCE-COVARIANCE MATRICES

When the populations are normal but the variance–covariance matrices are not equal, the allocation problem is more difficult. In the case of equal variance–covariance matrices, the logarithm of the ratio of the distribution functions was simplified by the cancellation of a quadratic term. However, in the unequal variance–covariance case, that term cannot be removed and the result is called a "quadratic rule." When $\Sigma_1 \neq \Sigma_2$,

$$\begin{aligned}\log(f_1/f_2) &= \tfrac{1}{2}\log(|\Sigma_2|/|\Sigma_1|) - \tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^{\mathrm{T}}\Sigma_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) \\ &\quad + \tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^{\mathrm{T}}\Sigma_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) \\ &= C_0 - \tfrac{1}{2}[\mathbf{x}^{\mathrm{T}}(\Sigma_1^{-1}-\Sigma_2^{-1})\mathbf{x} - 2\mathbf{x}^{\mathrm{T}}(\Sigma_1^{-1}\boldsymbol{\mu}_1 - \Sigma_2^{-1}\boldsymbol{\mu}_2)] \qquad (7.28)\end{aligned}$$

where $C_0 = \frac{1}{2}\ln(|\Sigma_1|/|\Sigma_2|) - \frac{1}{2}(\boldsymbol{\mu}_1^{\mathrm{T}}\Sigma_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^{\mathrm{T}}\Sigma_2^{-1}\boldsymbol{\mu}_2)$. This is a *quadratic* rule because the middle term of Eq. (7.28) is a quadratic form in the features.

The estimates of the covariance matrices and the feature mean vectors from the data are substituted for the parameters. The posterior probability is computed as

$$p(c_i|\mathbf{x}_0) = \frac{P(c_i)e^{-1/2D_i^2}}{\sum_j P(c_j)e^{-1/2D_j^2}} \qquad (7.29)$$

where $D_j^2 = (\mathbf{x}_0 - \bar{\mathbf{x}}_j)^{\mathrm{T}} S_j^{-1} (\mathbf{x}_0 - \bar{\mathbf{x}}_j) + \ln|S_j|$.

Example 7.3. Our next example deals with ground-water tests.[5] The training set is part of a larger set collected under the National Uranium Resource Evaluation Program, which had the goal of estimating the nation's uranium resources. Water samples were taken from five different groundwater formations, and 40 measurements were taken. Twelve of these measurements are given for 123 sites in the data set. Additional details are given in Beauchamp et al. [1980]. This training set is a good example because it illustrates some of the difficulties that can be encountered when trying to build a classifier on the basis of measurements.

To illustrate, let us consider classification for two groups, ATRD and TPO, using two features: uranium (U) and arsenic (AS). In Fig. 7.14, the 58 observations for the two groups are plotted using A for observations from ATRD and T for observations from the TPO group. In making field measurements, levels of a chemical are sometimes below detection. In the training set, these are given by negative values. Here, they have been set to the lower limit of detection.

[5]Instructors interested in setting up exercises using these data may obtain the complete table on diskette from Professor Eric Smith at the Department of Statistics, VPI&SU, Blacksburg, VA 24061-0439.

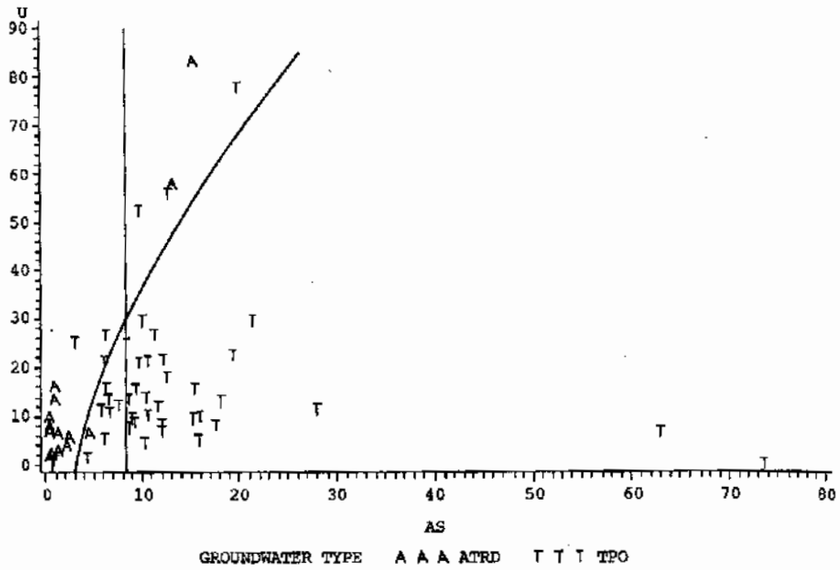


FIG. 7.14. Scatter plot of arsenic versus uranium for sites in ATRD (labeled A) and TOP (labeled T), and the boundaries for linear and quadratic classification.

The boundaries of the linear and quadratic rule are also drawn in Fig. 7.14. The straight line is the boundary for the linear rule, of course, and the curved is for the quadratic rule. Notice that these observations have a quite different shape than do the iris observations. In particular, there are two sites with extremely large arsenic values from the TPO group and several with large uranium values from both groups. Also, the variation in arsenic is considerably larger from the TPO group. The covariance matrix for ATRD is

	U	AS
U	539	103
AS	103	21

while TPO has a covariance matrix

	U	AS
U	228	−14
AS	−14	180

The variance is thus roughly nine times as large for the TPO group for arsenic. This can have a great effect on the linear classifier. As the linear rule uses a pooled covariance matrix, the low variance for ATRD for arsenic does not correctly model

the relationship with arsenic (note that s_p^2 is 140). The quadratic rule uses the information in both matrices separately and hence should better reflect the importance of arsenic. The linear classification rule assigns an object to group ATRD if $0.0755AS + 0.00062U < 0.618$.

You can see that the quadratic rule results in better classification. Sites to the left of the boundary are classified as ATRD, whereas sites to the right are classified as TPO. There is a total of 13 observations misclassified using the linear rule and only 8 with the quadratic rule. Notice also that classification could be improved simply by moving the linear boundary to the left. Unfortunately, if one tries to do this by using estimates of prior probabilities (i.e., p_1 and p_2), difficulties can occur. For example, if the prior probabilities are estimated in proportion to sample sizes (i.e., 15/58 and 43/58) the effect will be too much and all observations would be classified as belonging to TPO. Sometimes problems such as differences in variances can be diminished by using a feature transformation. With these features, classification can be improved by using a logarithmic transformation (Fig. 7.15). The classification function now would assign to group ATRD if $3.051 \ln(AS) - 0.48 \ln(U) < 6.09$. Both the linear and quadratic rules result in the same number of misclassified sites. Also, the variances of the ln(AS) values for ATRD and for TPO are much more nearly equal, and the assumption of equal covariance matrices is much more tenable for the transformed set. Additional information on transformations may be found in Koopmans [1987].

In a small problem like this, the use of quadratic discriminants does not seem to pose any combinatorial problems. However, by a preliminary transformation of

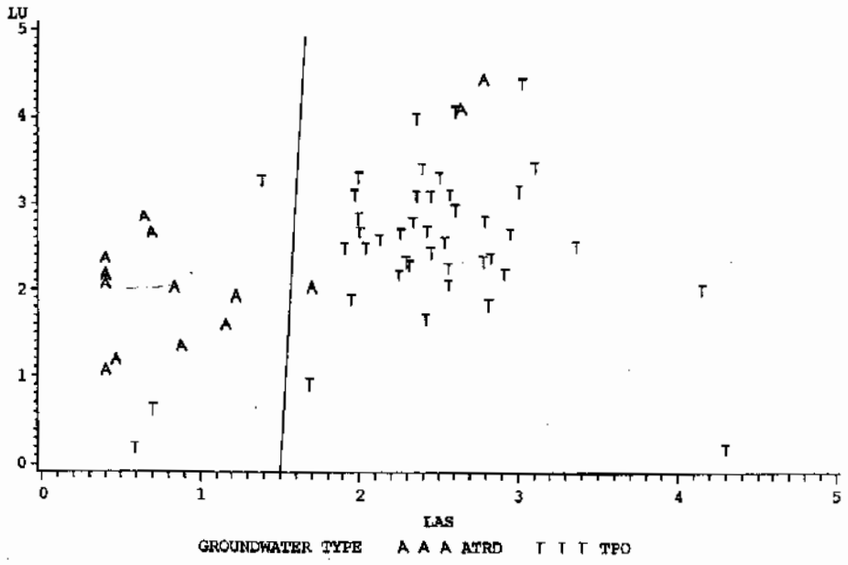


FIG. 7.15. Transformed scatter plot of arsenic and uranium and the classification boundary.

variables the quadratic discriminant function can be made to look like a linear one [Nilss65]. This is the concept of the ϕ-machine, discussed briefly in Section 6.5.2.

If we compare Figs. 7.14 and 7.15, we see that the linear discriminant on the log-transformed data is much more successful than the quadratic discriminant on the original data. But the use of the variance–covariance matrix and the normality assumption is risky in this case. Inspection of Fig. 7.14 shows that a different linear discriminant will result in only four misclassifications, whereas a quadratic discriminant in the log-measurement space might have resulted in only two, not to mention the possibilities presented by piecewise-linear solutions. The data are just not normal. Furthermore, as we shall see in Chapter 11, on "learning," just because we can draw a perfect classifier on a graph, this does not mean that we shall have achieved perfect recognition on any new data set. We shall return to these concepts in the next chapters.

7.6.1. The Case of *g* Groups

Now we turn to the more general situation where there are g classes. Assume a training set of features to be available; we wish to obtain a rule that minimizes the total misclassification. Let $c_1, c_2, \ldots, c_g$ denote the g classes with $P(c_i)$, the prior probability of class i occurring in a sample. Given an object with feature vector $\mathbf{x}_0$ from group i, it can be correctly classified or incorrectly classified to one of the $g - 1$ other groups. Furthermore, let N_i denote the number of objects from class i in the training set and let $N = \sum_{i=1}^{g} N_i$ be the total training set. Let $P(c_k|c_i)$ be the probability of misclassifying an object into class k, given that the object is from class i, $k \neq i$. Note that $P(c_k|c_i) = \int_{\Gamma_k} p(x|c_i)\, dx$, where Γ_k is the region where objects are classified as being from c_k.

Our objective is then to find a rule that results in minimizing the total misclassification probability (TMP). Now

$$P(\text{misclassify object from } c_i) = \sum_{\substack{k=1 \\ k \neq i}}^{g} P(c_k|c_i) \tag{7.30}$$

where we have omitted the symbol $\leq$ for simplicity. The unconditional probability is $P(c_i)\ \Sigma_{k \neq i}\ P(c_k|c_i)$. We would like to minimize this unconditionally over all groups. Hence we minimize

$$TMP = \sum_{i=1}^{g} P(c_i) \sum_{k=i}^{g} P(c_k|c_i) \tag{7.31}$$

It can be shown that this function is minimized by assigning $\mathbf{x}_0$ to the group c_k for which $p_k f_k(\mathbf{x}_0)$ is maximum. Notice that for two groups we compare $P(c_2)p(\mathbf{x}_0|c_2)$ and $P(c_1)p(\mathbf{x}_0|c_1)$. The rule is to assign $\mathbf{x}_0$ to class c_1 if

$$\frac{p(\mathbf{x}_0|c_1)}{p(\mathbf{x}_0|c_2)} \geq \frac{P(c_2)}{P(c_1)} \tag{7.32}$$

[i.e., if $P(c_2)p(\mathbf{x}_0|c_2) \leq P(c_1)p(x|c_0)$] and to assign $\mathbf{x}_0$ to group c_2 otherwise. Thus this rule reduces to the previous rule for two groups. When the priors are equal, the rule simply assigns the object to the class that it is most likely to have come from—that is, the group for which $f_k(\mathbf{x}_0)$ is largest.

Now, assuming normality we obtain

$$\ln P(c_k)p(\mathbf{x}_0|c_k) = \ln P(c_k) - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(\mathbf{x}_0 - \boldsymbol{\mu}_k)^{\mathrm{T}}\Sigma_k^{-1}(\mathbf{x}_0 - \boldsymbol{\mu}_k) \tag{7.33}$$

First we consider the case when the variance–covariance matrices are equal. Then $\ln P(c_k)p(\mathbf{x}_0|c_k)$ is, eliminating unessential terms,

$$\ln P(c_k) - \tfrac{1}{2}(\mathbf{x}_0 - \boldsymbol{\mu}_k)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}_0 - \boldsymbol{\mu}_k) \tag{7.34}$$

This is just the standardized distance from $\mathbf{x}_0$ to $\boldsymbol{\mu}_k$ plus an adjustment for the prior. Now, less a constant

$$\ln P(c_k)p(\mathbf{x}_0|c_k) = \ln P(c_k) - \tfrac{1}{2}[\mathbf{x}_0^{\mathrm{T}}\Sigma^{-1}\mathbf{x}_0 - \boldsymbol{\mu}_k^{\mathrm{T}}\Sigma^{-1}\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_k + \boldsymbol{\mu}_k^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_k] \tag{7.35}$$

Since $\mathbf{x}_0^{\mathrm{T}}\Sigma^{-1}\mathbf{x}_0$ is the same for all the groups, it is not needed for computations. Hence, a linear function is obtained that can be used for ordering the distances without doing all the computations:

$$d_k = \ln P(c_k) + \boldsymbol{\mu}_k^{\mathrm{T}}\Sigma^{-1}\mathbf{x}_0 - \tfrac{1}{2}\boldsymbol{\mu}_k^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_k \tag{7.36}$$

Using the data, we estimate the d_k as

$$\hat{d}_k = \ln P(c_k) + \bar{\mathbf{x}}_k^{\mathrm{T}}S_p^{-1}\mathbf{x}_0 - \tfrac{1}{2}\bar{\mathbf{x}}_k^{\mathrm{T}}S_p^{-1}\bar{\mathbf{x}}_k \tag{7.37}$$

where S_p is the pooled covariance matrix:

$$S_p = \frac{\sum_{i=1}^{g}(N_i - 1)S_i}{\sum_{i=1}^{g}(N_i - 1)} \tag{7.38}$$

Note that a large value of d_k means that $\mathbf{x}_0$ is close to $\bar{\mathbf{x}}_k$.

Our rule is thus to assign $\mathbf{x}_0$ to class k if $\hat{d}_k$ is the largest among $\hat{d}_1, \hat{d}_2, \ldots, \hat{d}_g$. This is a linear estimate, $\bar{\mathbf{x}}^{\mathrm{T}}S_p^{-1}$ is the vector of coefficients for the linear equation, and $\ln p_k - \frac{1}{2}\bar{\mathbf{x}}_k^{\mathrm{T}}S_p^{-1}\bar{\mathbf{x}}_k$ is the constant or intercept term.

Example 7.4. The Iris data: Consider classification for all three of the iris species. The pooled covariance matrix is

Sepal Length	Sepal Width	Petal Length	Petal Width
0.265	0.093	0.167	0.039
0.093	0.115	0.055	0.033
0.167	0.055	0.185	0.043
0.039	0.033	0.043	0.042

The distances to each of the three groups can be estimated from the linear functions with coefficients (assuming equal priors):

	Iris setosa	*Iris versicolor*	*Iris virginica*
Constant	−85.11	−71.68	−103.17
Sepal Length	23.43	15.60	12.32
Sepal Width	23.65	7.13	3.75
Petal Length	−16.30	5.31	12.88
Petal Width	−17.56	6.32	20.97

For the first iris $\mathbf{x}_0^T = (5.1, 3.5, 1.4, 0.2)$. Using Eq. (7.37), $\hat{d}_1 = 90.82$, $\hat{d}_2 = 41.52$, and $\hat{d}_3 = 6.98$. Since $\hat{d}_1$ is the largest, the first iris is classified as *setosa*. Three objects are misclassified. One of the *virginica* is classified as a *versicolor*, and two of the *versicolor* are classified as *virginica*. The misclassification was suggested by the plot of the features in Fig. 7.9. The posterior probabilities (Eq. (7.29)) for these objects are:

			Posterior probabilities for:		
Object	True Class	Decided Class	*Iris setosa*	*Iris versicolor*	*Iris virginica*
71	*versicolor*	*virginica*	0.0	0.252	0.748
84	*versicolor*	*virginica*	0.0	0.143	0.857
134	*virginica*	*versicolor*	0.0	0.730	0.270

Thus, object 71 is assigned to *Iris virginica* because the posterior probability (0.748) for that class is the largest posterior probability. The high posterior probabilities for misclassified observations indicate that the classes overlap, and additional features need to be measured if classification is to be improved. For this type of problem, the classification rate is quite good (3 misclassifications out of 150 possible).

7.6.2. Unequal Covariance Matrices

When the covariance matrices are not equal, Eq. (7.36) does not reduce to a linear function for obtaining distances. In this case we need to use

$$\ln p_k \hat{f}_k(\mathbf{x}_0) = \ln p_k - \frac{d}{2}\ln(2\pi) - \tfrac{1}{2}\ln|S_k| - \tfrac{1}{2}(\mathbf{x}_0 - \bar{\mathbf{x}}_k)^T S_k^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_k) \quad (7.39)$$

or

$$\hat{d}_k^* = \ln p_k - \tfrac{1}{2}\ln|S_k| - \tfrac{1}{2}(\mathbf{x}_0 - \bar{\mathbf{x}}_k)^T S_k^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_k) \quad (7.40)$$

to classify objects. Again, the object with feature vector $\mathbf{x}_0$ is assigned to the class to which it is closest. Note that $\hat{d}_k^*$ is always negative; hence the group that is closest to $\mathbf{x}_0$ will have $\hat{d}_k^*$ closest to zero.

7.7. BAYES' RULE WITH EMPIRICAL P.D.F. AND UNKNOWN DISTRIBUTIONS

The methods developed in the previous sections all involved the assumption that the distribution of the data is known. Indeed, we assumed the distributions to be multivariate normal. In practice this assumption may not always be valid. For example, rather than having a unimodal, symmetric distribution, it may be bimodal or even multimodal or may be skewed. Also, some of the data may be discrete while other data are continuous. We saw at the end of the last chapter how great could be the divergence between an assumed bivariate normal and the actual distribution of the observed data points. In this section we describe two alternative approaches to the estimation of the required p.d.f., namely, the kernel method and the nearest-neighbor method. These techniques are nonparametric; that is, we make no assumptions about the distributions belonging to a family of functions, described by distinct parameter values.

These two main types of nonparametric methods, the kernel method and nearest-neighbor method, while appearing to be quite different, are related [Hand81]. The general approach to classification involves estimation of $p(\mathbf{x}_0)$, the value of the distribution density at $\mathbf{x}_0$. This, then, is the important quantity to estimate. In the previous sections we assumed f to be known, so that we only needed to estimate parameters to obtain an estimate of the density. With the nonparametric methods we directly estimate $p(\mathbf{x}_0)$.

Suppose for simplicity that there are only two classes and a single variable. Then a simple approach is to estimate p by a histogram. From the histograms for both classes we can estimate $p(\mathbf{x}_0|c_1)$ and $p(\mathbf{x}_0|c_2)$ and make an assignment. However, when the number of features is large, it is difficult to obtain a meaningful histogram. Consider so simple a case as a set of two-valued (binary) variables. Then the number of bins in which the histogram is to be evaluated will be 2^d,

where d is the number of features. For $d = 20$ (a modest number of features in a real pattern recognition problem), there would be over one million bins, most of which would necessarily be empty! We see easily that other methods are required. This points up one of the advantages if we can use the variance–covariance matrix Σ. We assume only first-order interactions; there are only d^2 values to estimate, in this case only 400.

The kernel and nearest-neighbor methods are refinements to the histogram method, differing in the way the density is estimated. With the kernel method, one fixes a volume (or in one dimension a bin width) and counts the number of points inside the volume to compute the density estimate. The order-k nearest-neighbor method estimates the volume that contains a point $\mathbf{x}_0$ and its k nearest neighbors, and from this computes a density estimate.

Because these methods are based on the actual observations, with no assumptions made as to the underlying distributions, they are purely empirical. Since we need know nothing at all about the underlying distributions and associated parameters in these methods, they are called "nonparametric."

7.7.1. Kernel Methods

The task of the kernel method is to estimate the empirical p.d.f. It attempts to estimate the density at $\mathbf{x}_0$—that is, $p(\mathbf{x}_0|c_i)$ as an average of simpler functions, evaluated at $\mathbf{x}_0$. These simpler functions are called "kernels" and have the property that they are themselves density functions (i.e., they integrate to 1 and are non-negative). The task of the kernel function is to smooth the histogram and to normalize it. It is generally a symmetric function as well. Essentially we divide up the pattern space into elementary volumes of a given shape, pick a smoothing function in those volumes, and count the number of observations that fall into each elementary volume. Essentially, what we are doing is computing a sliding weighted average over the empirical histogram.

We can then estimate $p(c_i|\mathbf{x}_0)$ by $\hat{p}(\mathbf{x}_0|c_i)$, the density of class i about the point $\mathbf{x}_0$ using the kernel estimate.

To illustrate the basic ideas let us consider the two-group case when there is only one feature used to classify; we will use the first seven observations for petal length for two iris species. The first step in the approximation is to choose the kernel. One simple one is the quadratic kernel:

$$k(x) = \begin{cases} \frac{3}{4}(1 - x^2), & -1 \le x \le 1 \\ 0, & \text{otherwise} \end{cases} \tag{7.41}$$

where x is the measurement, normalized to the range $-1 < x < 1$, and the factor $\frac{3}{4}$ is needed for this function to integrate to 1. The kernel simply describes the shape of a function that smooths the data around each data point.

To illustrate the kernel approach we have assumed that there is one feature and two groups. Further suppose the training set has seven observations from each group: (31.0, 34.0, 35.0, 32.0, 40.5, 41.0, 39.5) and (41.0, 44.0, 45.0, 42.0,

41.5, 45.5, 43.0). In Fig. 7.16, at the top, we display the histograms for the two groups. Recall that the way a histogram is formed is to define a bin width (say $2h$), then pick an endpoint, form bins, and count the number of observations falling in each bin. An estimate of $p(\mathbf{x}_0)$ is then given by the proportion of $\mathbf{x}_0$ values in the bin. (Note that we have used here for simplicity the vector notation for $\mathbf{x}$, although, strictly speaking, with one feature x is a scalar.)

In forming histograms we must choose the bin width and the endpoint. The kernel estimator can be viewed as a method which removes the dependence on the endpoint by smoothing the histogram. This smoothing is accomplished by the choice of kernel function and bin width, and the density at any point is estimated as an average of functions. If these are smooth, then the resulting density estimate will also be reasonably smooth. In the single feature case

$$\hat{p}(x) = \frac{1}{Nh} \sum_{j=1}^{N} k\left(\frac{x - x_j}{h}\right) \tag{7.42}$$

where N is the number of observations and h is a smoothing parameter (also called the "bandwidth" or "window width"). There are many kernel functions that have been suggested, besides Eq. (7.42) [Duda73]. An especially simple one is

$$k_1(x) = \begin{cases} \frac{1}{2}, & \text{if } |x| \leq 1 \\ 0, & \text{otherwise} \end{cases} \tag{7.43}$$

The graph of k_1 is a rectangle of height $\frac{1}{2}$ and endpoints at ± 1. We can call Eq. (7.41) "k_2"; its graph is a parabola. We are interested in the data in the window $k(x - x_{ij})/h$. This kernel is symmetric about x_{ij}, the jth observation in group i, and has endpoints determined by h. In Fig. 7.16 a histogram is plotted for each group. Below the histogram is a plot of the 14 kernel functions, 7 for each group. Only 13 functions can be seen; two of them overlap because they are based on the same value, 41.0. Below this figure, the sum is given for each group, representing the estimates of the densities for each group. The estimate using the kernel is seen to be much smoother than the histogram estimate. The functions defining the bottom figure are given by

$$\begin{aligned} \hat{p}(x|c_i) &= \frac{1}{N_i h} \sum_{j=1}^{N_i} k_2\left(\frac{x - x_{ij}}{h}\right) \\ &= \frac{1}{N_i h} \sum_{j=1}^{N_i} \begin{cases} \frac{3}{4}\left[1 - \left(x - \frac{x_{ij}}{h}\right)^2\right], & |x - x_{ij}| \leq 1 \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{7.44}$$

where h, the scaling parameter, has been set to 1.

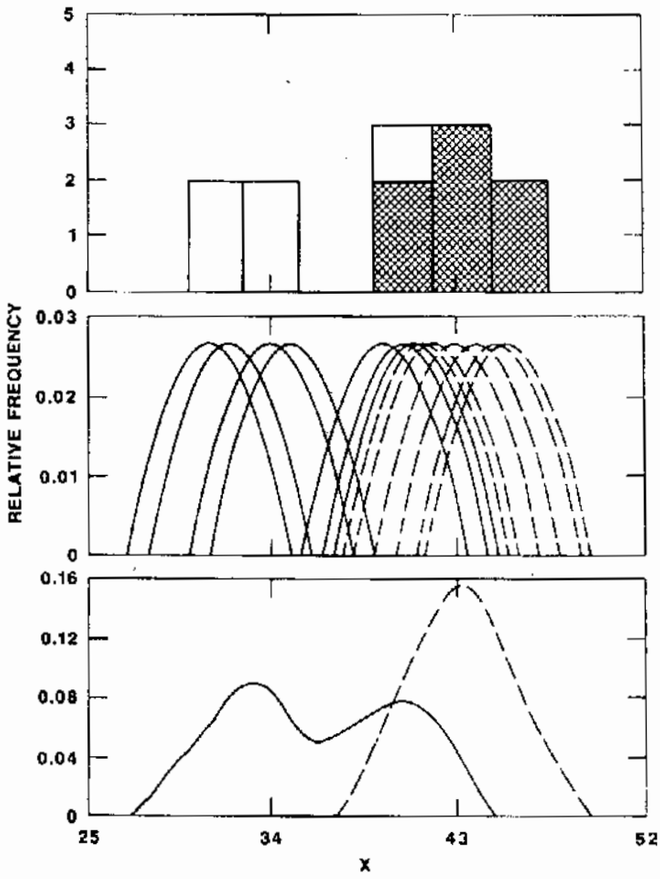


FIG. 7.16. The kernel method; the kernel functions and their sums.

In practice, there is more than one feature and usually more than two groups. While the basic approach remains the same, the kernels that are most often used are related to the multivariate normal distribution. In the general multivariate setting,

$$\hat{p}(\mathbf{x}|c_i) = \frac{1}{N}\sum_{j=1}^{N_i} k\left(\frac{1}{h}(\mathbf{x} - \mathbf{x}_{ij})\right) \tag{7.45}$$

where $1 \leq i \leq g$. We see that there is one kernel-smoothed histogram (i.e., p.d.f. estimate) for each of the K classes. The feature vectors $\mathbf{x}$ have d features each.

Another common choice for the kernel is the multivariate normal density with

a scale adjustment $k_1(\mathbf{x}) = (2\pi h)^{-d/2}|S|^{-1/2}\exp\{-(\mathbf{x}-\mathbf{x}_{ij})^{\mathrm{T}} S^{-1}(\mathbf{x}-\mathbf{x}_{ij})/2h^2\}$, where S is an estimate of the covariance matrix or

$$k_2(\mathbf{x}) = (2\pi h)^{-d/2}\left(\prod_{k=1}^{d} S_k\right)^{-1}\exp\left\{-\frac{1}{2}\sum_{k=1}^{d}\left(\frac{x_k - x_{ijk}}{s_k}\right)^2\right\} \tag{7.46}$$

which is the same as k_1 with S a diagonal matrix. The use of the multivariate normal density as the *kernel* is not related to any assumption of normality of the *entire* p.d.f. It assumes that *locally* the distribution may be approximated in this way. It is used for convenience because it is a non-negative function that decays exponentially from a center point.

To classify an object with feature vector $\mathbf{x}_0$, choose a kernel function and scaling parameter; compute $P(c_i)\hat{p}(\mathbf{x}_0|c_i)$, where $P(c_i)$ is the prior probability and $\hat{p}(\mathbf{x}_0|c_i)$ is the kernel estimate. Then assign the object to the population with the largest value of $p_i\hat{p}(\mathbf{x}_0|c_i)$.

The kernel method was seen above as a way to obtain the empirical p.d.f. Classification is then based on evaluation of the p.d.f. for $\mathbf{x}_0$ over all the groups.

Instead of choosing an elementary volume and counting the objects that it contains, nearest-neighbor techniques find the volume that contains the k nearest neighbors of the same class as the observation $\mathbf{x}_0$. Then $p(\mathbf{x}_0|c_i) = k/V_{\mathbf{x}0,j}$, where $\mathbf{x}_0$ is the feature vector of class i at which we wish to estimate the p.d.f. But, unlike the kernel methods, the nearest-neighbor methods can be used directly to establish a decision rule. We shall study these in the next chapter.

7.7.2. Bayes' Optimal Decisions: The General Case

Up to here we have been dealing with very neat problems. Our two-dimensional plots of elliptical or circular probability contours showing how the normal multivariate assumption leads to nice linear or quadratic discriminants, the nice graphs of p.d.f. showing how we can get minimum error decisions, and even the Fisher linear discriminant theory may have led you astray. The material in the earlier chapters should have warned you that maybe things here are just too neat. The fact is that from the viewpoint of the minimum error criterion, Bayes' optimal decisions are just that. *No other decision procedure can yield a better result. This is independent of whether or not the basic assumptions of Bayes' optimal decision theory are actually satisfied in a given concrete problem.*

What are the necessary conditions for a Bayes' optimal decision procedure to be implemented? There are only two:

1. We have available or can compute $p(c_i|\mathbf{x}_0)$, the probability distribution of class c_i, given the observation $\mathbf{x}_0$.
2. We also have available $P(c_i)$, the "prior" probability of the class c_i.

In actual practice, if we do implement a decision procedure based on the theory we will use estimates or even assume a form for these probability functions that

may be wildly divergent from the evidence, such as in the ground-water example above. The discussion of the kernel and nearest-neighbor methods should have shown that we are not hostages of the normality assumption. Nor need we restrict our attention to minimum error criteria, even in the form of a reject option, as discussed earlier.

A more general criterion is that of cost or risk or penalty, as it has been called by different authors. For each of the possible errors in a given problem we can assign a cost factor specific to each kind of error. In other words, for each class we can assign individual costs to the Type I and Type II errors. In a two-class problem, for example, we would have r_{12},[6] the cost of deciding class c_1 when in fact class c_2 is present and, symmetrically r_{21}.

. We will also need r_{11} and r_{22}, the costs of the corresponding correct decisions. Of course, it will surely always be true that $r_{12} \gg r_{22}$ and $r_{21} \gg r_{11}$. But we do not even require any of the costs of incorrect decisions to be equal in the general case. In the words of one author [Fukun90], "the misclassification of a cancer patient to normal may have a more damaging effect than the misclassification of a normal patient to cancer"—quite ignoring the fact that many treatments for cancer are exceedingly drastic and would be terribly damaging to a healthy patient, leaving the physician open to a high-cost malpractice suit. Given an empirical distribution in feature space, the cost R of decision when class c_1 is present will be $R(c_1) = (r_{21} + r_{11})P(y_1)$, and symmetrically for class c_2. How will we find a decision boundary when we have only empirical data on $P(c_i|\mathbf{x}_0)$?

Using either kernel or nearest-neighbor methods, we can estimate $p(c_i)dr$, where dr is an elementary volume in d-space. Now, integrating R over the entire space individually for each class, we find:

$$R_{c_1} = \int_{\Gamma} R(c_1)\, dr$$
$$R_{c_2} = \int_{\Gamma} R(c_2)\, dr \tag{7.47}$$

If we divide the space Γ into two half-spaces by a boundary, separating the region where we decide class c_1 from the region where we decide class c_2, r_1, and r_2, and integrate $R(c_1)$ and $R(c_2)$ in the corresponding half-spaces, we want to select the boundary in such a way as to minimize

$$R = \int_{\Gamma_1} R(c_1)\, dr + \int_{\Gamma_2} R(c_2)\, dr \tag{7.47'}$$

These considerations lead easily to the following decision rules:

1. Decide c_1 if $(r_{21} - r_{11})P(c_1)|\mathbf{x}_0 > (r_{12} - r_{22})P(c_2|\mathbf{x}_0)$, else decide c_2 (7.48)

[6]We use "r," for "risk instead of "c," for "cost," according to the usual convention. This also avoids confusion with the class variable c.

or

2. Decide c_1 if $\frac{p(\mathbf{x}_0|c_1)}{p(\mathbf{x}_0|c_2)} > \frac{(r_{12} - r_{22})P(c_2)}{(r_{21} - r_{11})P(c_1)}$, else decide c_2 (7.48′)

7.8. CONCLUSION

In this chapter we have examined Bayes' optimal decision theory from the most restrictive assumptions of the multivariate normal distribution to successively more and more general assumptions about the underlying statistics of the patterns to be classified. Some concrete examples were discussed, showing how the assumption of a simple parametric distribution affected the quality of the classification when the assumption was even quite far from the actual distribution.

Regardless of the state of our knowledge (i.e., the possibility of effectively constructing a Bayes optimal decision rule in a given problem), no decision rule can give a more favorable result for the prescribed optimality condition. In the next chapter we shall examine some nonparametric and non-Bayesian methods and will estimate the degree to which they can approach the Bayes optimal decision performance as a bound.