Chapter 3

HYPOTHESIS TESTING

The purpose of pattern recognition is to determine to which category or class a given sample belongs. Through an observation or measurement process, we obtain a set of numbers which make up the observation vector. The observation vector serves as the input to a decision rule by which we assign the sample to one of the given classes. Let us assume that the observation vector is a random vector whose conditional density function depends on its class. If the conditional density function for each class is known, then the pattern recognition problem becomes a problem in statistical hypothesis testing.

3.1 Hypothesis Tests for Two Classes

In this section, we discuss two-class problems, which arise because each sample belongs to one of two classes, ω_1 or ω_2. The conditional density functions and the *a priori* probabilities are assumed to be known.

The Bayes Decision Rule for Minimum Error

Bayes test: Let X be an observation vector, and let it be our purpose to determine whether X belongs to ω_1 or ω_2. A decision rule based simply on probabilities may be written as follows:

$$q_1(X) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} q_2(X) , \tag{3.1}$$

where $q_i(X)$ is *a posteriori probability* of ω_i given X. Equation (3.1) indicates that, if the probability of ω_1 given X is larger than the probability of ω_2, X is classified to ω_1, and vice versa. The *a posteriori* probability $q_i(X)$ may be calculated from the *a priori* probability P_i and the conditional density function $p_i(X)$, using *Bayes theorem*, as

$$q_i(X) = \frac{P_i p_i(X)}{p(X)} \tag{3.2}$$

where $p(X)$ is the mixture density function. Since $p(X)$ is positive and common to both sides of the inequality, the decision rule of (3.1) can be expressed as

$$P_1 p_1(X) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} P_2 p_2(X) \tag{3.3}$$

or

$$\ell(X) = \frac{p_1(X)}{p_2(X)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{P_2}{P_1} . \tag{3.4}$$

The term $\ell(X)$ is called the *likelihood ratio* and is the basic quantity in hypothesis testing. We call P_2/P_1 the *threshold value* of the likelihood ratio for the decision. Sometimes it is more convenient to write the *minus-log likelihood ratio* rather than writing the likelihood ratio itself. In that case, the decision rule of (3.4) becomes

$$h(X) = -\ln \ell(X) = -\ln p_1(X) + \ln p_2(X) \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} \ln \frac{P_1}{P_2} . \tag{3.5}$$

The direction of the inequality is reversed because we have used the negative logarithm. The term $h(X)$ is called *the discriminant function*. Throughout this book, we assume $P_1 = P_2$, and set the threshold $\ln P_1/P_2 = 0$ for simplicity, unless otherwise stated.

Equation (3.1), (3.4), or (3.5) is called the *Bayes test for minimum error*.

Bayes error: In general, the decision rule of (3.5), or any other decision rule, does not lead to perfect classification. In order to evaluate the performance of a decision rule, we must calculate the *probability of error*, that is, the probability that a sample is assigned to the wrong class.

The *conditional error* given X, $r(X)$, due to the decision rule of (3.1) is either $q_1(X)$ or $q_2(X)$ whichever smaller. That is,

$$r(X) = \min[q_1(X), q_2(X)] \ . \tag{3.6}$$

The total error, which is called the *Bayes error*, is computed by $E\{r(\mathbf{X})\}$.

$$\begin{aligned}
\varepsilon = E\{r(\mathbf{X})\} &= \int r(X)p(X)dX \\
&= \int \min[P_1 p_1(X), P_2 p_2(X)]dX \\
&= P_1 \int_{L_2} p_1(X)dX + P_2 \int_{L_1} p_2(X)dX \\
&= P_1\varepsilon_1 + P_2\varepsilon_2 \ ,
\end{aligned} \tag{3.7}$$

where

$$\varepsilon_1 = \int_{L_2} p_1(X)\, dX \quad \text{and} \quad \varepsilon_2 = \int_{L_1} p_2(X)\, dX \ . \tag{3.8}$$

Equation (3.7) shows several ways to express the Bayes error, ε. The first line is the definition of ε. The second line is obtained by inserting (3.6) into the first line and applying the Bayes theorem of (3.2). The integral regions L_1 and L_2 of the third line are the regions where X is classified to ω_1 and ω_2 by this decision rule, and they are called the ω_1- and ω_2*–regions*. In L_1, $P_1p_1(X) > P_2p_2(X)$, and therefore $r(X) = P_2p_2(X)/p(X)$. Likewise, $r(X) = P_1p_1(X)/p(X)$ in L_2 because $P_1p_1(X) < P_2p_2(X)$ in L_2. In (3.8), we distinguish two types of errors: one results from misclassifying samples from ω_1 and the other results from misclassifying samples from ω_2. The total error is a weighted sum of these errors.

Figure 3-1 shows an example of this decision rule for a simple one-dimensional case. The decision boundary is set at $x = t$ where $P_1p_1(x) = P_2p_2(x)$, and $x < t$ and $x > t$ are designated to L_1 and L_2 respectively. The resulting errors are $P_1\varepsilon_1 = B + C$, $P_2\varepsilon_2 = A$, and $\varepsilon = A + B + C$, where A, B, and C indicate the areas, for example, $B = \int_t^{t'} P_1p_1(x)\,dx$.

This decision rule gives the smallest probability of error. This may be demonstrated easily from the one-dimensional example of Fig. 3-1. Suppose that the boundary is moved from t to t', setting up the new ω_1- and ω_2–regions as L_1' and L_2'. Then, the resulting errors are $P_1\varepsilon_1' = C$, $P_2\varepsilon_2' = A + B + D$, and $\varepsilon' = A + B + C + D$, which is larger than ε by D. The same is true when the

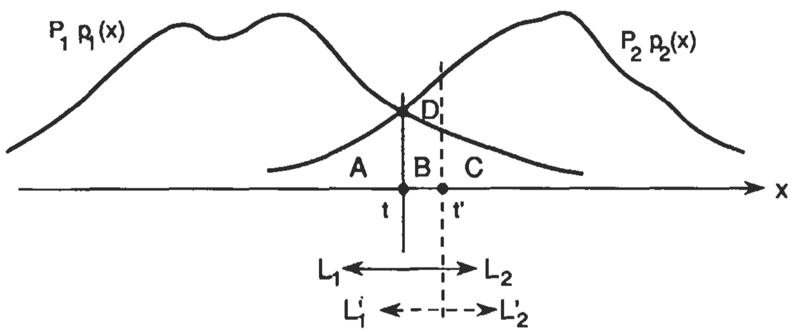

Fig. 3-1 Bayes decision rule for minimum error.

boundary is shifted to the left. This argument can be extended to a general n-dimensional case.

The computation of the Bayes error is a very complex problem except in some special cases. This is due to the fact that ε is obtained by integrating high-dimensional density functions in complex regions as seen in (3.8). Therefore, it is sometimes more convenient to integrate the density function of $\mathbf{h} = h(\mathbf{X})$ of (3.5), which is one-dimensional:

$$\varepsilon_1 = \int_{\ln (P_1/P_2)}^{+\infty} p_h(h \mid \omega_1) dh \,, \tag{3.9}$$

$$\varepsilon_2 = \int_{-\infty}^{\ln (P_1/P_2)} p_h(h \mid \omega_2) dh \,, \tag{3.10}$$

where $p_h(h \mid \omega_i)$ is the conditional density of $\mathbf{h}$ for ω_i. However, in general, the density function of $\mathbf{h}$ is not available, and very difficult to compute.

Example 1: When the $p_i(X)$'s are normal with expected vectors M_i and covariance matrices Σ_i, the decision rule of (3.5) becomes

$$\begin{aligned} h(X) &= -\ln \ell(X) \\ &= \frac{1}{2}(X-M_1)^T \Sigma_1^{-1} (X-M_1) - \frac{1}{2}(X-M_2)^T \Sigma_2^{-1} (X-M_2) + \frac{1}{2} \ln \frac{|\Sigma_1|}{|\Sigma_2|} \\ &\underset{\omega_2}{\overset{\omega_1}{\gtrless}} \ln \frac{P_1}{P_2} \,. \end{aligned} \tag{3.11}$$

Equation (3.11) shows that the decision boundary is given by a quadratic form in X. When $\Sigma_1 = \Sigma_2 = \Sigma$, the boundary becomes a linear function of X as

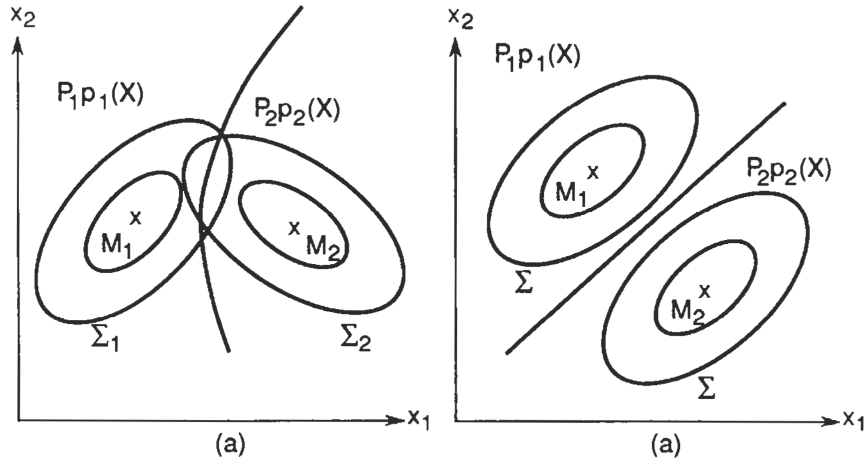

Fig. 3-2 Decision boundaries for normal distributions: (a) $\Sigma_1 \neq \Sigma_2$; (b) $\Sigma_1 = \Sigma_2$.

$$h(X) = (M_2-M_1)^T\Sigma^{-1}X + \frac{1}{2}(M_1^T\Sigma^{-1}M_1 - M_2^T\Sigma^{-1}M_2)$$

$$\underset{\omega_2}{\overset{\omega_1}{\gtrless}} \ln \frac{P_1}{P_2} . \tag{3.12}$$

Figure 3-2 shows two-dimensional examples for $\Sigma_1 \neq \Sigma_2$ and $\Sigma_1 = \Sigma_2$.

Example 2: Let us study a special case of (3.11) where

$$M_i = 0 \quad \text{and} \quad \Sigma_i = \begin{bmatrix} 1 & \rho_i & \dots & \rho_i^{n-1} \\ \rho_i & 1 & & \vdots \\ \vdots & & \ddots & \rho_i \\ \rho_i^{n-1} & \dots & \rho_i & 1 \end{bmatrix} . \tag{3.13}$$

This type of covariance matrix is often seen, for example, when *stationary random processes* are time-sampled to form random vectors. The explicit expressions for Σ_i^{-1} and $|\Sigma_i|$ are known for this covariance matrix as

$$\Sigma_i^{-1} = \frac{1}{1-\rho_i^2} \begin{bmatrix} 1 & -\rho_i & 0 & \dots & 0 \\ -\rho_i & 1+\rho_i^2 & -\rho_i & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & & 1+\rho_i^2 & -\rho_i \\ 0 & \dots & 0 & -\rho_i & 1 \end{bmatrix}, \tag{3.14}$$

$$|\Sigma_i| = (1-\rho_i^2)^{n-1} . \tag{3.15}$$

Therefore, the quadratic equation of (3.11) becomes

$$\left\{ \frac{1+\rho_1^2}{1-\rho_1^2} - \frac{1+\rho_2^2}{1-\rho_2^2} \right\} \sum_{i=1}^{n} x_i^2 - \left\{ \frac{\rho_1^2}{1-\rho_1^2} - \frac{\rho_2^2}{1-\rho_2^2} \right\} (x_1^2 + x_n^2)$$

$$- \left\{ \frac{2\rho_1}{1-\rho_1^2} - \frac{2\rho_2}{1-\rho_2^2} \right\} \sum_{i=1}^{n-1} x_i x_{i+1} + (n-1) \ln \frac{1-\rho_1^2}{1-\rho_2^2} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \ln \frac{P_1}{P_2} , \tag{3.16}$$

where the second term shows the edge effect of terminating the observation of random processes within a finite length, and this effect diminishes as n gets large. If we could ignore the second and fourth terms and make $\ln(P_1/P_2) = 0$ $(P_1 = P_2)$, the decision rule becomes $(\Sigma x_i x_{i+1})/(\Sigma x_i^2) \gtrless t$; that is, the decision is made by estimating the correlation coefficient and thresholding the estimate. Since $\rho_1 \neq \rho_2$ is the only difference between ω_1 and ω_2 in this case, this decision rule is reasonable.

Example 3: When $\mathbf{x}_k$'s are mutually independent and exponentially distributed,

$$p_i(X) = \prod_{k=1}^{n} \frac{1}{\alpha_{ik}} \exp\left[-\frac{1}{\alpha_{ik}} x_k \right] u(x_k) \quad (i=1,2) , \tag{3.17}$$

where α_{ik} is the parameter of the exponential distribution for $\mathbf{x}_k$ and ω_i, and $u(\cdot)$ is the step function. Then, $h(X)$ of (3.5) becomes

$$h(X) = \sum_{k=1}^{n} \left[\frac{1}{\alpha_{1k}} - \frac{1}{\alpha_{2k}} \right] x_k + \sum_{k=1}^{n} \ln \frac{\alpha_{1k}}{\alpha_{2k}} . \tag{3.18}$$

The Bayes decision rule becomes a linear function of x_k's.

The Bayes Decision Rule for Minimum Cost

Often in practice, minimizing the probability of error is not the best criterion to design a decision rule because the misclassifications of ω_1- and ω_2-samples may have different consequences. For example, the misclassification of a cancer patient to normal may have a more damaging effect than the misclassification of a normal patient to cancer. Therefore, it is appropriate to assign a cost to each situation as

$$c_{ij} = \text{cost of deciding } X \in \omega_i \text{ when } X \in \omega_j . \tag{3.19}$$

Then, the *conditional cost of deciding* $X \in \omega_i$ given X, $r_i(X)$, is

$$r_i(X) = c_{i1} q_1(X) + c_{i2} q_2(X) . \tag{3.20}$$

The decision rule and the resulting *conditional cost* given X, $r(X)$, are

$$r_1(X) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} r_2(X) \tag{3.21}$$

and

$$r(X) = \min[r_1(X), r_2(X)] . \tag{3.22}$$

The total *cost* of this decision is

$$r = E\{r(\mathbf{X})\} = \int \min[r_1(X), r_2(X)] p(X)\, dX$$

$$= \int \min[c_{11} q_1(X) + c_{12} q_2(X), c_{21} q_1(X) + c_{22} q_2(X)] p(X)\, dX$$

$$= \int \min[c_{11} P_1 p_1(X) + c_{12} P_2 p_2(X), c_{21} P_1 p_1(X) + c_{22} P_2 p_2(X)]\, dX$$

$$= \int_{L_1} [c_{11} P_1 p_1(X) + c_{12} P_2 p_2(X)]\, dX$$

$$+ \int_{L_2} [c_{21} P_1 p_1(X) + c_{22} P_2 p_2(X)]\, dX , \tag{3.23}$$

where L_1 and L_2 are determined by the decision rule of (3.21).

The boundary which minimizes r of (3.23) can be found as follows. First, rewrite (3.23) as a function of L_1 alone. This is done by replacing $\int_{L_2} p_i(X)dX$ with $1 - \int_{L_1} p_i(X)dX$, since L_1 and L_2 do not overlap and cover the entire domain. Thus,

$$r = (c_{21}P_1 + c_{22}P_2) + \int_{L_1} [(c_{11}-c_{21})P_1p_1(X) + (c_{12}-c_{22})P_2p_2(X)]dX \ . \tag{3.24}$$

Now our problem becomes one of choosing L_1 such that r is minimized. Suppose, for a given value of X, that the integrand of (3.24) is negative. Then we can decrease r by assigning X to L_1. If the integrand is positive, we can decrease r by assigning X to L_2. Thus the *minimum cost decision rule* is to assign to L_1 those X's and only those X's, for which the integrand of (3.24) is negative. This decision rule can be stated by the following inequality:

$$(c_{12}-c_{22})P_2p_2(X) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} (c_{21}-c_{11})P_1p_1(X) \tag{3.25}$$

or

$$\frac{p_1(X)}{p_2(X)} \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} \frac{(c_{12}-c_{22})P_2}{(c_{21}-c_{11})P_1} \ . \tag{3.26}$$

This decision rule is called the *Bayes test for minimum cost*.

Comparing (3.26) with (3.4), we notice that the Bayes test for minimum cost is a likelihood ratio test with a different threshold from (3.4), and that the selection of the cost functions is equivalent to changing the *a priori* probabilities P_i. Equation (3.26) is equal to (3.4) for the special selection of the cost functions

$$c_{21} - c_{11} = c_{12} - c_{22} \ . \tag{3.27}$$

This is called a *symmetrical cost function*. For a symmetrical cost function, the cost becomes the probability of error, and the test of (3.26) minimizes the probability of error.

Different cost functions are used when a wrong decision for one class is more critical than one for the other class.