CHAPTER 9

PATTERN CLASSIFICATION AND DIAGNOSTIC DECISION

An important final purpose of biomedical signal analysis is to classify a given signal into one of a few known categories and to arrive at a diagnostic decision regarding the condition of the patient. A physician or medical specialist may achieve this goal via visual or auditory analysis of the signal presented: Comparative analysis of the given signal with others of known diagnoses or established protocols and sets of rules assist in such a decision-making process. The basic knowledge, clinical experience, expertise, and intuition of the physician play significant roles in this process. Some measurements may also be made from the given signal to assist in its analysis, such as the QRS width from an ECG signal plot.

When signal analysis is performed via the application of computer algorithms, the typical result is the extraction of a number of numerical or quantitative features. When the numerical features relate directly to measures of the signal, such as the QRS width and RR interval of an ECG signal, the clinical specialist may be able to use the features in his or her diagnostic logic. Even indirect measures, such as the frequency content of PCG signals and murmurs, may find such direct use. However, when parameters such as AR-model coefficients and spectral statistics are derived, a human analyst is not likely to be able to comprehend and analyze the features. Furthermore, as the number of the computed features increases, the associated diagnostic logic may become too complicated and unwieldy for human analysis. Computer methods would then be desirable to realize the classification and decision process.

At the outset, it should be borne in mind that a biomedical signal forms but one piece of information in arriving at a diagnosis: The classification of a given signal into one of many categories may assist in the diagnostic procedure, but will almost never be the only factor. Regardless, pattern classification based upon signal analysis is an important aspect of biomedical signal analysis, and forms the theme of this chapter. Remaining within the realm of CAD as introduced in Figure 1.53 and Section 1.5, it would be preferable to design methods so as to assist a medical specialist in arriving at a diagnosis rather than to provide a decision.

9.1 Problem Statement

A number of measures and features have been derived from a biomedical signal. Explore methods to classify the signal into one of a few specified categories. Investigate the relevance of the features and the classification methods in arriving at a diagnostic decision about the patient.

Observe that the features may have been derived manually or by computer methods. Note the distinction between classifying the given signal and arriving at a diagnosis regarding the patient: The connection between the two tasks or steps may not always be direct. In other words, a pattern classification method may facilitate the labeling of a given signal as being a member of a particular class; arriving at a diagnosis of the condition of the patient will most likely require the analysis of several items of clinical and other information. Although it is common to work with a prespecified number of pattern classes, many problems do exist where the number of classes is not known *a priori.*

The following sections present a few illustrative case studies. A number of methods for pattern classification, decision making, and evaluation of the results of classification are reviewed and illustrated.

9.2 Illustration of the Problem with Case Studies

9.2.1 Diagnosis of bundle branch block

Bundle-branch block affects the propagation of the excitation pulse through the conduction system of the heart to the ventricles. A block in the left bundle branch results in delayed activation of the left ventricle as compared to the right; a block in the right bundle branch has the opposite effect. Essentially, contraction of the two ventricles becomes asynchronous. The resulting ECG typically displays a wider-than-normal QRS complex ($100 - 120\ ms$ or more), which could have a jagged or slurred shape as well [33]; see Figure 1.28.

The orientation of the cardiac electromotive forces is affected by bundle-branch block. The initial forces in left bundle-branch block are directed more markedly to the left-posterior, whereas the terminal forces are directed to the superior-left and posterior parts of the left ventricle [33]. Left bundle-branch block results in the loss of Q waves in leads I, V5, and V6. The following logic assists in the diagnosis of incomplete left bundle-branch block [472]:

IF (QRS duration $\geq 105\ ms$ and $\leq 120\ ms$) AND
(QRS amplitude is negative in leads V1 and V2) AND
(Q or S duration $\geq 80\ ms$ in leads V1 and V2) AND
(no Q wave is present in any two of leads I, V5, and V6) AND
(R duration $> 60\ ms$ in any two of leads I, aVL, V5, and V6) THEN
the patient has incomplete left bundle-branch block.

Incomplete right bundle-branch block is indicated by the following conditions [472]:

IF (QRS duration $\geq 91\ ms$ and $\leq 120\ ms$) AND
(S duration $\geq 40\ ms$ in any two of leads I, aVL, V4, V5, and V6) AND
in lead V1 or V2 EITHER
[(R duration $> 30\ ms$) AND (R amplitude $> 100\ \mu V$) AND
(no S wave is present)] OR
[(R$'$ duration $> 30\ ms$) AND (R$'$ amplitude $> 100\ \mu V$) AND
(no S$'$ wave is present)] THEN
the patient has incomplete right bundle-branch block.

[*Note:* The first positive deflection of a QRS complex is referred to as the R wave and the second positive deflection (if present) is referred to as the R$'$ wave. Similarly, S and S$'$ indicate the first and second (if present) negative deflections, respectively, after the R wave.]

Note that the logic or decision rules given above may be used either by a human analyst or in a computer algorithm after the durations and amplitudes of the various waves mentioned have been measured or computed. Cardiologists with extensive training and experience may arrive at such decisions via visual analysis of an ECG record without resorting to actual measurements.

9.2.2 Normal or ectopic ECG beat?

PVCs caused by ectopic foci could be precursors of more serious arrhythmia, and hence the detection of such beats is important in cardiac monitoring. As illustrated in Sections 5.4.2 and 5.7 as well as in Figures 5.1 and 5.11, PVCs possess shorter preceding RR intervals than normal beats and display bizarre waveshapes that are markedly different from those of the normal QRS complexes of the same subject. Therefore, a simple rule to detect PVCs or ectopic beats could be as follows:

IF (the RR interval of the beat is less than the normal at the current heart rate) AND
(the QRS waveshape is markedly different from the normal QRS of the patient)
THEN *the beat is a PVC.*

As in the preceding case study of bundle-branch block, the logic above may be easily applied for visual analysis of an ECG signal by a physician or a trained observer. Computer implementation of the first part of the rule relating in an objective or quantitative manner to the RR interval is simple. However, implementation of the second condition on waveshape, being qualitative and subjective, is neither direct nor easy. Regardless, we have seen in Chapter 5 how we may characterize waveshape. Figures 5.1 and 5.11 illustrate the application of waveshape analysis to quantify the differences between the shapes of normal QRS complexes and ectopic beats. Figure 5.2 suggests how a 2D feature space may be divided by a simple linear decision boundary to categorize beats as normal or ectopic. We shall study the details of such methods later in this chapter.

9.2.3 Is there an alpha rhythm?

The alpha rhythm appears in an EEG record as an almost-sinusoidal wave (see Figure 1.39); a trained EEG technologist or physician can readily recognize the pattern at a glance from an EEG record plotted at the standard scale. The number of cycles of the wave may be counted over one or two seconds of the plot if an estimate of the dominant frequency of the rhythm is desired.

In computer analysis of EEG signals, the ACF and PSD may be used to detect the presence of the alpha rhythm. We saw in Chapter 4 how these two functions demonstrate peaks at the basic period or dominant frequency of the rhythm, respectively (see Figure 4.9). A peak-detection algorithm may be applied to the ACF, and the presence of a significant peak in the range $75 - 125\ ms$ for the delay or lag may be used as an indication of the existence of the alpha rhythm. If the PSD is available, the fractional power of the signal in the band $8 - 12\ Hz$ (see Equation 6.40) may be computed: A high value of the fraction indicates the presence of the alpha rhythm. Note that the logic described above includes the qualifiers “significant” and “high”; experimentation with a number of signals that have been categorized by experts should assist in assigning a numerical value to represent the significance of the features described.

9.2.4 Is a murmur present?

Detection of the presence of a heart murmur is a fairly simple task for a trained physician or cardiologist: In performing auscultation of a patient with a stethoscope, the cardiologist needs to determine the existence of noise-like, high-frequency sounds between the low-frequency S1 and S2. It is necessary to exercise adequate care to reject high-frequency noise from other sources such as breathing, wheezing, and scraping of the stethoscope against the skin or hair. The cardiologist also has to distinguish between innocent physiological murmurs and those due to cardiovascular defects and diseases. Further discrimination between different types of murmurs requires more careful analysis: Figure 5.6 illustrates a decision tree to classify systolic murmurs based upon envelope analysis.

We have seen in Chapters 6 and 7 how we may derive frequency-domain parameters that relate to the presence of murmurs in the PCG signal. Once we have derived such numerical features for a number of signals of known categories of diseases (diagnoses), it becomes possible to design and train classifiers to categorize new signals into one of a few prespecified classes.

The preceding case studies suggest that the classification of patterns in a signal may, in some cases, be based upon thresholds applied to quantitative measurements obtained from the signal; in some other cases, it may be based upon objective measures derived from the signal that attempt to quantify certain notions regarding the characteristics of signals belonging to various categories. Classification may also be based upon the differences between certain measures derived from the signal on hand and those of established examples with known categorization. The succeeding sections of this chapter describe procedures for classification of signals based upon the approaches suggested above.

9.3 Pattern Classification

Pattern recognition or classification may be defined as categorization of input data into identifiable classes via the extraction of significant features or attributes of the data from a background of irrelevant detail [267, 268, 417, 473–476]. In biomedical signal analysis, after quantitative features have been extracted from the given signals, each signal may be represented by a feature vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$, which is also known as a measurement vector or a pattern vector. When the values x_i are real numbers, $\mathbf{x}$ is a point in an n-dimensional Euclidean space: Vectors of similar objects may be expected to form clusters as illustrated in Figure 9.1.

For efficient pattern classification, measurements that could lead to disjoint sets or clusters of feature vectors are desired. This point underlines the importance of appropriate design of the preprocessing and feature extraction procedures. Features or characterizing attributes that are common to all patterns belonging to a particular class are known as *intraset* or *intraclass features*. Discriminant features that represent differences between pattern classes are called *interset* or *interclass features*.

The pattern classification problem is that of generating optimal decision boundaries or decision procedures to separate the data into pattern classes based on the feature vectors. Figure 9.1 illustrates a simple linear decision function or boundary to separate 2D feature vectors into two classes.

9.4 Supervised Pattern Classification

Problem: *You are provided with a number of feature vectors with classes assigned to them. Propose techniques to characterize the boundaries that separate the classes.*

Solution: A given set of feature vectors of known categorization is often referred to as a *training set*. The availability of a training set facilitates the development of mathematical functions that can characterize the separation between the classes. The

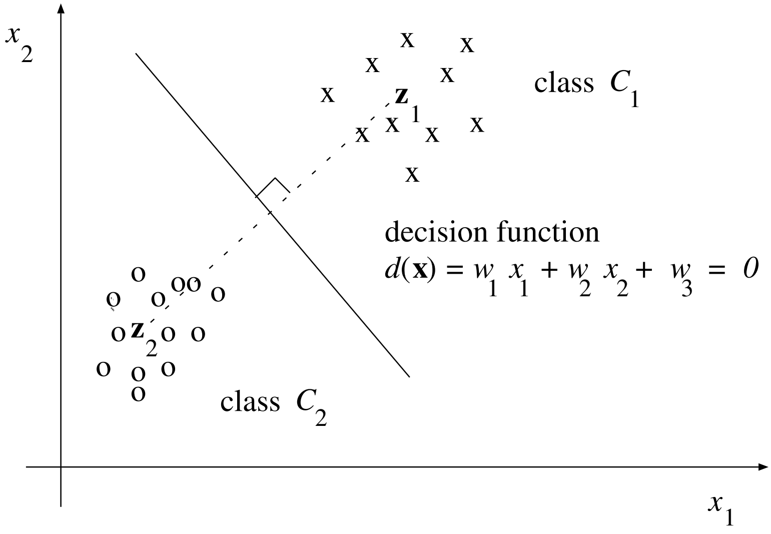

Figure 9.1 Two-dimensional feature vectors of two classes C_1 and C_2. The prototypes of the two classes are indicated by the vectors $\mathbf{z}_1$ and $\mathbf{z}_2$. The linear decision function shown $d(\mathbf{x})$ (solid line) is the perpendicular bisector of the straight line joining the two class prototypes (dashed line).

functions may then be applied to new feature vectors of unknown classes to classify or recognize them. This approach is known as *supervised pattern classification*. A set of feature vectors of known categorization that is used to evaluate a classifier designed in this manner is referred to as a *test set*. The following sections describe a few methods that can assist in the development of discriminant and decision functions.

9.4.1 Discriminant and decision functions

A general linear discriminant or decision function is of the form

$$d(\mathbf{x}) = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + w_{n+1} = \mathbf{w}^T \mathbf{x}, \tag{9.1}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_n, 1)^T$ is the feature vector augmented by an additional entry equal to unity, and $\mathbf{w} = (w_1, w_2, \ldots, w_n, w_{n+1})^T$ is a correspondingly augmented weight vector. A two-class pattern classification problem may be stated as

$$d(\mathbf{x}) = \mathbf{w}^T \mathbf{x} \begin{cases} > 0 & \text{if } \mathbf{x} \in C_1, \\ \leq 0 & \text{if } \mathbf{x} \in C_2, \end{cases} \tag{9.2}$$

where C_1 and C_2 represent the two classes. The discriminant function may be interpreted as the boundary separating the classes C_1 and C_2, as illustrated in Figure 9.1.

In the general case of an M-class pattern classification problem, we will need M weight vectors and M decision functions to perform the following decisions:

$$d_i(\mathbf{x}) = \mathbf{w}_i^T \, \mathbf{x} \begin{cases} > 0 & \text{if } \mathbf{x} \in C_i, \\ \leq 0 & \text{otherwise,} \end{cases} \tag{9.3}$$

for $i = 1, 2, \ldots, M$, where $\mathbf{w}_i = (w_{i1}, w_{i2}, \ldots, w_{in}, w_{i,n+1})^T$ is the weight vector for the class C_i.

Three cases arise in solving this problem [473]:

Case 1: Each class is separable from the rest by a single decision surface:

$$\text{if } \; d_i(\mathbf{x}) > 0, \;\; \text{then } \; \mathbf{x} \in C_i. \tag{9.4}$$

Case 2: Each class is separable from every other individual class by a distinct decision surface, that is, the classes are pairwise separable. There are $M(M-1)/2$ decision surfaces given by $d_{ij}(\mathbf{x}) = \mathbf{w}_{ij}^T \, \mathbf{x}$, such that

$$\text{if } \; d_{ij}(\mathbf{x}) > 0 \; \forall \; j \neq i, \;\; \text{then } \; \mathbf{x} \in C_i. \tag{9.5}$$

[*Note:* $d_{ij}(\mathbf{x}) = -d_{ji}(\mathbf{x})$.]

Case 3: There exist M decision functions $d_k(\mathbf{x}) = \mathbf{w}_k^T \, \mathbf{x}, \; k = 1, 2, \ldots, M$, with the property that

$$\text{if } \; d_i(\mathbf{x}) > d_j(\mathbf{x}) \; \forall \; j \neq i, \;\; \text{then } \; \mathbf{x} \in C_i. \tag{9.6}$$

This is a special instance of Case 2. We may define

$$d_{ij}(\mathbf{x}) = d_i(\mathbf{x}) - d_j(\mathbf{x}) = (\mathbf{w}_i - \mathbf{w}_j)^T \, \mathbf{x} = \mathbf{w}_{ij}^T \, \mathbf{x}. \tag{9.7}$$

If the classes are separable under Case 3, they are separable under Case 2; the converse, in general, is not true.

When patterns need to be separated into multiple categories, it may be possible to convert the problem into a series of binary decision problems by considering the separation of each class against the set of remaining classes.

Patterns that may be separated by linear decision functions as above are said to be *linearly separable.* In other situations, an infinite variety of complex decision boundaries may be formulated by using generalized decision functions based upon nonlinear functions of the feature vectors as

$$d(\mathbf{x}) = w_1 f_1(\mathbf{x}) + w_2 f_2(\mathbf{x}) + \cdots + w_K f_K(\mathbf{x}) + w_{K+1} \tag{9.8}$$

$$= \sum_{i=1}^{K+1} w_i \; f_i(\mathbf{x}). \tag{9.9}$$

Here, $\{f_i(\mathbf{x})\}, \; i = 1, 2, \ldots, K$, are real, single-valued functions of $\mathbf{x}$; $f_{K+1}(\mathbf{x}) = 1$.

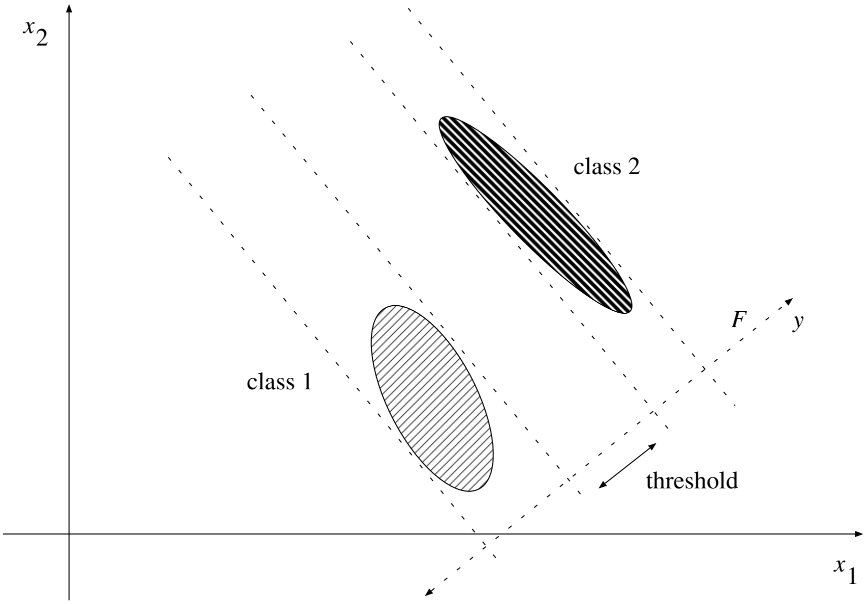

Figure 9.2 Illustration of a classifier based upon FLDA.

Whereas the functions $f_i(\mathbf{x})$ may be nonlinear in the n-dimensional space of $\mathbf{x}$, the decision function may be formulated as a linear function by defining a transformed feature vector $\mathbf{x}^\dagger = [f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_K(\mathbf{x}), 1]^T$. Then, $d(\mathbf{x}) = \mathbf{w}^T\mathbf{x}^\dagger$, with $\mathbf{w} = [w_1, w_2, \ldots, w_K, w_{K+1}]^T$. Once evaluated, $\{f_i(\mathbf{x})\}$ is just a set of numerical values, and $\mathbf{x}^\dagger$ is simply a K-dimensional vector augmented by an entry equal to unity.

See Section 9.11.1 for an illustration of application of linear discriminant analysis to an ECG signal with PVCs.

9.4.2 Fisher linear discriminant analysis

Fisher linear discriminant analysis (FLDA) is a technique that projects multidimensional data on to a 1D space or line which allows for improved discrimination [267]; dimensionality reduction is achieved in the process. In a pattern classification problem with two classes, the projection in FLDA is designed to maximize the distance between the means of the feature vectors for the two classes and minimize their variance within each class.

Figure 9.2 illustrates the concept of a linear classifier based on FLDA. Consider two classes, C_1 and C_2, in which each sample is represented by the features $\mathbf{x} = [x_1, x_2]^T$. If only the feature x_1 is used for classification, it is equivalent to the situation when all of the data points are projected on to the x_1 axis. Such a projection causes overlapping regions containing samples from both classes; thus, x_1

provides poor separation between the classes C_1 and C_2. The feature x_2 also shows overlap between the two classes and provides poor separation. However, by inspection of Figure 9.2, it is evident that the classes C_1 and C_2 can be clearly separated by a straight line oriented at about 130° to the abscissa. Hence, a classifier can be designed using projections of the feature vectors or samples provided on to the line labeled as F in Figure 9.2. FLDA provides the weights that define the projection as explained above [267, 282].

Let a given set of feature vectors

$$X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} = \{X_1, X_2\} \tag{9.10}$$

be partitioned into N_1 training samples in subset X_1, corresponding to class C_1, and N_2 training samples in subset X_2, corresponding to class C_2, with $N_1 + N_2 = N$. The projection of $\mathbf{x}_i$ on to the FLDA discriminant line is

$$y_i = \mathbf{w}^T \mathbf{x}_i \,. \tag{9.11}$$

The projected value y_i belongs to the subset Y_1 or Y_2. The weight vector $\mathbf{w}$ defines the direction of the axis of the projected values y_i. The mean of all of the samples belonging to each class before projection is

$$\mathbf{m}_k = \frac{1}{N_k} \sum_{\mathbf{x}_i \in X_k} \mathbf{x}_i \,, \tag{9.12}$$

where $k = 1, 2$, and $i = 1, 2, \ldots, N$. The mean of the projected values in each class is

$$\begin{aligned} \tilde{m}_k &= \frac{1}{N_k} \sum_{y_i \in Y_k} y_i \\ &= \frac{1}{N_k} \sum_{\mathbf{x}_i \in X_k} \mathbf{w}^T \mathbf{x}_i, \\ &= \mathbf{w}^T \mathbf{m}_k, \quad k = 1, 2. \end{aligned} \tag{9.13}$$

The difference between the means of the classes after projection is

$$|\tilde{m}_1 - \tilde{m}_2| = |\mathbf{w}^T \mathbf{m}_1 - \mathbf{w}^T \mathbf{m}_2| \,. \tag{9.14}$$

The variance or scatter of the projected samples y_i in each class is

$$\begin{aligned} \tilde{s}_k^2 &= \sum_{y_i \in Y_k} (y_i - \tilde{m}_k)^2 \\ &= \sum_{\mathbf{x}_i \in X_k} (\mathbf{w}^T \mathbf{x}_i - \mathbf{w}^T \mathbf{m}_k)^2 \\ &= \sum_{\mathbf{x}_i \in X_k} \mathbf{w}^T (\mathbf{x}_i - \mathbf{m}_k)(\mathbf{x}_i - \mathbf{m}_k)^T \mathbf{w}, \quad k = 1, 2 \end{aligned}$$

$$= \mathbf{w}^T \mathbf{S}_k \mathbf{w}\,, \tag{9.15}$$

where

$$\mathbf{S}_k = \sum_{\mathbf{x}_i \in X_k} (\mathbf{x}_i - \mathbf{m}_k)(\mathbf{x}_i - \mathbf{m}_k)^T, \;\; k = 1, 2. \tag{9.16}$$

$\mathbf{S}_k$ is known as the within-class or intraclass scatter matrix for class k.

The total within-class scatter of the two classes is

$$\tilde{s}_1^2 + \tilde{s}_2^2 = \mathbf{w}^T \left\{\mathbf{S}_1 + \mathbf{S}_2\right\} \mathbf{w} = \mathbf{w}^T \mathbf{S}_W \mathbf{w}, \tag{9.17}$$

with $\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2$.

The FLDA criterion function is defined as

$$J(\mathbf{w}) = \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}\,. \tag{9.18}$$

Optimal separation between the classes is achieved when $J(\mathbf{w})$ is at its maximum. The denominator term of the function $J(\mathbf{w})$ is given by Equation 9.17. The numerator gives the separation between the means of the projected samples for the two classes, and is obtained as follows:

$$\begin{aligned}(\tilde{m}_1 - \tilde{m}_2)^2 &= (\mathbf{w}^T \mathbf{m}_1 - \mathbf{w}^T \mathbf{m}_2)^2 \\ &= \mathbf{w}^T (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \mathbf{w} \\ &= \mathbf{w}^T \mathbf{S}_B \mathbf{w}\,. \end{aligned} \tag{9.19}$$

Here, $\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T$ is the between-class or interclass scatter matrix. The FLDA criterion is given by

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}\,. \tag{9.20}$$

The weight vector $\mathbf{w}_o$ that maximizes $J(\mathbf{w})$ can be derived by solving a generalized eigenvalue problem. When a new sample is to be classified, its feature vector $\mathbf{x}$ is projected using Equation 9.11 and two-category classification is performed as

$$\mathbf{x} \in \begin{cases} C_1 & \text{if } y = \mathbf{w}_o^T \mathbf{x} < T_1, \\ C_2 & \text{otherwise}, \end{cases} \tag{9.21}$$

where T_1 is a threshold.

Illustration of application: Rangayyan and Wu [264] applied a number of statistical measures for screening of VAG signals (see Section 5.12.1 for a description of the data set used). The parameters used included FF computed for the full duration of each swing cycle and for the first and second halves (labeled as FF_1 and FF_2), corresponding to extension and flexion, as well as the entropy (H), skewness (S), and kurtosis (K). The features could not perform screening with high accuracy on their own: the six features $FF, FF_1, FF_2, S, K,$ and H individually provided A_z values of $0.72, 0.73, 0.68, 0.70, 0.61,$ and 0.60, respectively. Figure 9.3 shows the scatter plot of the feature vector $[FF_1, FF_2]$ for the set of 89 VAG signals. The straight line shown in the figure represents the decision function obtained via FLDA. The equation for the decision boundary is

$$0.0058\,FF_1 - 0.0010\,FF_2 - 0.0192 = 0. \tag{9.22}$$

Due to substantial overlap of the samples in the two categories, FLDA could correctly classify only $19/38$ abnormal signals and $40/51$ normal signals, yielding an average classification accuracy of 66%. The use of FLDA with all of the six features listed above and the LOO method for cross-validation did not provide any better results, with $A_z = 0.72$. The results obtained and the illustration in Figure 9.3 indicate that the samples are not linearly separable using the measures derived. Mu et al. [274] obtained much better classification performance with $A_z = 0.95$ using the same features as above but with a genetic algorithm for feature selection and the strict two-surface proximal classifier. See Section 9.8.1 for the results of application of a neural network to the same data set.

See Figure 5.2 and Section 9.11 for additional illustrations of application of linear decision functions to the classification of ECG signals.

9.4.3 Distance functions

Consider M pattern classes represented by their prototype patterns $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_M$. The prototype of a class is typically computed as the average of all of the feature vectors belonging to the class. Figure 9.1 illustrates schematically the prototypes $\mathbf{z}_1$ and $\mathbf{z}_2$ of the two classes shown.

The Euclidean distance between an arbitrary pattern vector $\mathbf{x}$ and the i^{th} prototype is given as

$$D_i = \|\mathbf{x} - \mathbf{z}_i\| = \sqrt{(\mathbf{x} - \mathbf{z}_i)^T\ (\mathbf{x} - \mathbf{z}_i)}. \tag{9.23}$$

A simple rule to classify the pattern vector $\mathbf{x}$ would be to choose that class for which the vector has the smallest distance:

$$\text{if } D_i < D_j\ \forall\ j \neq i, \text{ then } \mathbf{x} \in C_i. \tag{9.24}$$

A simple relationship may be established between discriminant functions and distance functions as follows [473]:

$$\begin{aligned} D_i^2 &= \|\mathbf{x} - \mathbf{z}_i\|^2 = (\mathbf{x} - \mathbf{z}_i)^T(\mathbf{x} - \mathbf{z}_i) \qquad (9.25) \\ &= \mathbf{x}^T\mathbf{x} - 2\mathbf{x}^T\mathbf{z}_i + \mathbf{z}_i^T\mathbf{z}_i = \mathbf{x}^T\mathbf{x} - 2(\mathbf{x}^T\mathbf{z}_i - \frac{1}{2}\mathbf{z}_i^T\mathbf{z}_i). \end{aligned}$$

Choosing the minimum of D_i^2 is equivalent to choosing the minimum of D_i (as all $D_i > 0$). Furthermore, from the equation given above, it follows that choosing

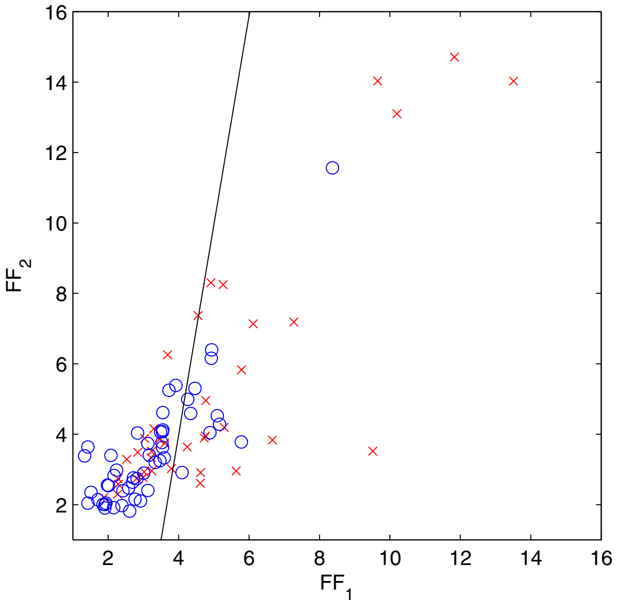

Figure 9.3 Illustration of classification of a data set of 89 VAG signals using FLDA. The circles represent normal samples of the feature vector $[FF_1, FF_2]^T$ and the crosses represent abnormal samples. The straight line is the FLDA decision function, given by $0.0058\, FF_1 - 0.0010\, FF_2 - 0.0192 = 0$. Figure courtesy of Tingting Mu, University of Liverpool, Liverpool, UK.

the minimum of D_i^2 is equivalent to choosing the maximum of $(\mathbf{x}^T \mathbf{z}_i - \frac{1}{2}\mathbf{z}_i^T \mathbf{z}_i)$. Therefore, we may define the decision function

$$d_i(\mathbf{x}) = (\mathbf{x}^T \mathbf{z}_i - \frac{1}{2}\mathbf{z}_i^T \mathbf{z}_i),\ i = 1, 2, \ldots, M. \tag{9.26}$$

A decision rule may then be stated as follows:

$$\text{if } d_i(\mathbf{x}) > d_j(\mathbf{x})\ \forall\ j \neq i,\ \text{then } \mathbf{x} \in C_i. \tag{9.27}$$

This is a linear discriminant function, which becomes obvious from the following representation: If $z_{ij},\ j = 1, 2, \ldots, n$, are the components of $\mathbf{z}_i$, let $w_{ij} = z_{ij}$, $j = 1, 2, \ldots, n$; $w_{i,n+1} = -\frac{1}{2}\mathbf{z}_i^T \mathbf{z}_i$; and $\mathbf{x} = [x_1, x_2, \ldots, x_n, 1]^T$. Then, $d_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x}$, $i = 1, 2, \ldots, M$, where $\mathbf{w}_i = [w_{i1}, w_{i2}, \ldots, w_{i,n+1}]^T$. Therefore, distance functions may be formulated as linear discriminant or decision functions.

9.4.4 The nearest neighbor rule

Suppose that we are provided with a set of N sample patterns $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N\}$ of known classification: Each pattern belongs to one of M classes $\{C_1, C_2, \ldots, C_M\}$.

We are then given a new feature vector $\mathbf{x}$ whose class needs to be determined. Let us compute a distance measure $D(\mathbf{s}_i, \mathbf{x})$ between the vector $\mathbf{x}$ and each sample pattern. Then, the nearest-neighbor rule states that the vector $\mathbf{x}$ is to be assigned to the class of the sample that is the closest to $\mathbf{x}$:

$$\mathbf{x} \in C_i \text{ if } D(\mathbf{s}_i, \mathbf{x}) = \min\{D(\mathbf{s}_l, \mathbf{x})\}, \ l = 1, 2, \ldots, N. \tag{9.28}$$

A major disadvantage of the above method is that the classification decision is made based upon a single sample vector of known classification. The nearest neighbor may happen to be an outlier that is not representative of its class. It would be more reliable to base the classification upon several samples: We may consider a certain number k of the nearest neighbors of the sample to be classified, and then seek a majority opinion. This leads to the so-called *k-nearest-neighbor* or *k-NN rule*: Determine the k nearest neighbors of $\mathbf{x}$, and use the majority of equal classifications in this group as the classification of $\mathbf{x}$.

9.5 Unsupervised Pattern Classification

Problem: *We are given a set of feature vectors with no categorization or classes attached to them. No prior training information is available. How may we group the vectors into multiple categories?*

Solution: The design of distance functions and decision boundaries requires a training set of feature vectors of known classes. The functions so designed may then be applied to a new set of feature vectors or samples to perform pattern classification. Such a procedure is known as *supervised* pattern classification due to the initial training step. In some situations a training step may not be possible, and we may be required to classify a given set of feature vectors into either a prespecified or unknown number of categories. Such a problem is labeled as *unsupervised* pattern classification and may be solved by cluster-seeking methods.

9.5.1 Cluster seeking methods

Given a set of feature vectors, we may examine them for the formation of inherent groups or clusters. This is a simple task in the case of 2D vectors, where we may plot them, visually identify groups, and label each group with a pattern class; see Figures 9.1 and 9.2. Allowance may have to be made to assign the same class to multiple disjoint groups. Such an approach may be used even when the number of classes is not known at the outset. When the vectors have a dimension higher than three, visual analysis will not be feasible. It then becomes necessary to define criteria to group the given vectors on the basis of similarity, dissimilarity, or distance measures. A few examples of such measures are as follows [473]:

- Euclidean distance

$$D_E^2 = \|\mathbf{x} - \mathbf{z}\|^2 = (\mathbf{x} - \mathbf{z})^T (\mathbf{x} - \mathbf{z}) = \sum_{i=1}^{n} (x_i - z_i)^2. \tag{9.29}$$

Here, $\mathbf{x}$ and $\mathbf{z}$ are two feature vectors; the latter could be a class prototype, if available. A small value of D_E indicates greater similarity between the two vectors than a large value of D_E.

- Mahalanobis distance

$$D_M^2 = (\mathbf{x} - \mathbf{m})^T \mathbf{C}^{-1} (\mathbf{x} - \mathbf{m}), \tag{9.30}$$

where $\mathbf{x}$ is a feature vector being compared to a pattern class for which $\mathbf{m}$ is the class mean vector and $\mathbf{C}$ is the covariance matrix. Inclusion of the (inverse of the) covariance of the distribution of the class in the distance measure allows consideration of the scatter of the constituent samples of the class: A large scatter of the population diminishes the distance measure as compared to a tighter cluster. A small value of D_M indicates a higher potential membership of the vector $\mathbf{x}$ in the class than a large value of D_M.

- Normalized dot product (cosine of the angle between the vectors $\mathbf{x}$ and $\mathbf{z}$)

$$D_d = \frac{\mathbf{x}^T \mathbf{z}}{\|\mathbf{x}\| \|\mathbf{z}\|}. \tag{9.31}$$

A large dot product value indicates a greater degree of similarity between the two vectors than a small value.

The covariance matrix is defined as

$$\mathbf{C} = E[(\mathbf{y} - \mathbf{m})(\mathbf{y} - \mathbf{m})^T], \tag{9.32}$$

where the expectation operation is performed over all feature vectors $\mathbf{y}$ that belong to the class. The covariance matrix provides the covariance of all possible pairs of the features in the feature vector over all samples belonging to the given class. The elements along the main diagonal of the covariance matrix provide the variance of the individual features that make up the feature vector. The covariance matrix represents the scatter of the features that belong to the given class. The mean and covariance need to be updated as more samples are added to a given class in a clustering procedure.

When the Mahalanobis distance needs to be calculated between a sample vector and a number of classes represented by their mean and covariance matrices, a pooled covariance matrix may be used if the numbers of members in the various classes are unequal and low [474]. For example, if the covariance matrices of two classes are $\mathbf{C}_1$ and $\mathbf{C}_2$, and the numbers of members in the two classes are N_1 and N_2, the pooled covariance matrix is given by

$$\mathbf{C} = \frac{(N_1 - 1)\mathbf{C}_1 + (N_2 - 1)\mathbf{C}_2}{N_1 + N_2 - 2}. \tag{9.33}$$

Various performance indices may be designed to measure the success of a clustering procedure [473]. A measure of the tightness of a cluster is the sum of the squared errors performance index:

$$J = \sum_{j=1}^{N_c} \sum_{\mathbf{x} \in S_j} \|\mathbf{x} - \mathbf{m}_j\|^2, \tag{9.34}$$

where N_c is the number of cluster domains, S_j is the set of samples in the j^{th} cluster,

$$\mathbf{m}_j = \frac{1}{N_j} \sum_{\mathbf{x} \in S_j} \mathbf{x} \tag{9.35}$$

is the sample mean vector of S_j, and N_j is the number of samples in S_j.

A few other examples of performance indices are:

- Average of the squared distances between the samples in a cluster domain.
- Intracluster variance.
- Average of the squared distances between the samples in different cluster domains.
- Intercluster distances.
- Scatter matrices.
- Covariance matrices.

A simple cluster-seeking algorithm [473]: Suppose we have N sample patterns $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$.

1. Let the first cluster center $\mathbf{z}_1$ be equal to any one of the samples, for example, $\mathbf{z}_1 = \mathbf{x}_1$.
2. Choose a nonnegative threshold θ.
3. Compute the distance D_{21} between $\mathbf{x}_2$ and $\mathbf{z}_1$. If $D_{21} < \theta$, assign $\mathbf{x}_2$ to the domain (class) of cluster center $\mathbf{z}_1$; otherwise, start a new cluster with its center as $\mathbf{z}_2 = \mathbf{x}_2$. For the subsequent steps, let us assume that a new cluster with center $\mathbf{z}_2$ has been established.
4. Compute the distances D_{31} and D_{32} from the next sample $\mathbf{x}_3$ to $\mathbf{z}_1$ and $\mathbf{z}_2$, respectively. If D_{31} and D_{32} are both greater than θ, start a new cluster with its center as $\mathbf{z}_3 = \mathbf{x}_3$; otherwise, assign $\mathbf{x}_3$ to the domain of the closer cluster.
5. Continue to apply steps 3 and 4 by computing and checking the distance from *every* new (unclassified) pattern vector to *every* established cluster center and applying the cluster-assignment or cluster-creation rule.
6. Stop when every given pattern vector has been assigned to a cluster.

Note that the procedure does not require knowledge of the number of classes *a priori*. Note also that the procedure does not assign a real-world class to each cluster: it merely groups the given vectors into disjoint clusters. A subsequent step is required to label each cluster with a class related to the actual problem. Multiple clusters may relate to the same real-world class and may have to be merged.

A major disadvantage of the simple cluster-seeking algorithm is that the results depend upon

- the first cluster center chosen for each domain or class,
- the order in which the sample patterns are considered,
- the value of the threshold θ, and
- the geometrical properties (distributions) of the data (or the feature-vector space).

The maximin-distance clustering algorithm [473]: This method is similar to the previous "simple" algorithm, but first identifies the cluster regions that are the farthest apart, as follows.

1. Let $\mathbf{x}_1$ be the first cluster center $\mathbf{z}_1$.
2. Determine the farthest sample from $\mathbf{x}_1$, and call it cluster center $\mathbf{z}_2$.
3. Compute the distance from each remaining sample to $\mathbf{z}_1$ and to $\mathbf{z}_2$. For every pair of these computations, save the minimum distance, and select the maximum of the minimum distances. If this "maximin" distance is an appreciable fraction of the distance between the cluster centers $\mathbf{z}_1$ and $\mathbf{z}_2$, label the corresponding sample as a new cluster center $\mathbf{z}_3$; otherwise stop forming new clusters and go to Step 5.
4. If a new cluster center was formed in Step 3, repeat Step 3 using a "typical" or the average distance between the established cluster centers for comparison.
5. Assign each remaining sample to the domain of its nearest cluster center.

The term "maximin" refers to the combined use of maximum and minimum distances between the given vectors and the centers of the clusters already formed.

The K-means algorithm [473]: The preceding "simple" and "maximin" algorithms are intuitive procedures. The K-means algorithm is based on iterative minimization of a performance index that is defined as the sum of the squared distances from all points in a cluster domain to the cluster center, as follows.

1. Choose K initial cluster centers $\mathbf{z}_1(1), \mathbf{z}_2(1), \ldots, \mathbf{z}_K(1)$. The index in parentheses represents the iteration number.
2. At the k^{th} iterative step, distribute the samples $\{\mathbf{x}\}$ among the K cluster domains, using the relation

$$\mathbf{x} \in S_j(k) \ \text{ if } \ \|\mathbf{x} - \mathbf{z}_j(k)\| < \|\mathbf{x} - \mathbf{z}_i(k)\| \ \forall\, i = 1, 2, \ldots, K, \ i \neq j, \tag{9.36}$$

where $S_j(k)$ denotes the set of samples whose cluster center is $\mathbf{z}_j(k)$.

3. From the results of Step 2, compute the new cluster centers $\mathbf{z}_j(k+1),\ j = 1, 2, \ldots, K$, such that the sum of the squared distances from all points in $S_j(k)$ to the new cluster center is minimized. In other words, the new cluster center $\mathbf{z}_j(k+1)$ is computed so that the performance index

$$J_j = \sum_{\mathbf{x} \in S_j(k)} \|\mathbf{x} - \mathbf{z}_j(k+1)\|^2, \quad j = 1, 2, \ldots, K, \tag{9.37}$$

is minimized. The $\mathbf{z}_j(k+1)$ that minimizes this performance index is simply the sample mean of $S_j(k)$. Therefore, the new cluster center is given by

$$\mathbf{z}_j(k+1) = \frac{1}{N_j} \sum_{\mathbf{x} \in S_j(k)} \mathbf{x}, \quad j = 1, 2, \ldots, K, \tag{9.38}$$

where N_j is the number of samples in $S_j(k)$. The name "K-means" is derived from the manner in which cluster centers are sequentially updated.

4. If $\mathbf{z}_j(k+1) = \mathbf{z}_j(k)$ for $j = 1, 2, \ldots, K$, the algorithm has converged: Terminate the procedure. Otherwise, go to Step 2.

The behavior of the K-means algorithm is influenced by:

- the number of cluster centers specified,
- the choice of the initial cluster centers,
- the order in which the sample patterns are considered, and
- the geometrical properties (distributions) of the data (or the feature-vector space).

See Section 9.11.3 for an illustration of application of the K-means method to an ECG signal with PVCs.

9.6 Probabilistic Models and Statistical Decision

Problem: *Pattern classification methods such as discriminant functions are dependent upon the set of training samples provided. Their success, when applied to new cases, will depend upon the accuracy of representation of the various pattern classes by the training samples. How can we design pattern classification techniques that are independent of specific training samples and optimal in a broad sense?*

Solution: Probability functions and probabilistic models may be developed to represent the occurrence and statistical attributes of classes of patterns. Such functions may be based upon large collections of data, historical records, or mathematical models of pattern generation. In the absence of information as above, a training step with samples of known categorization is required to estimate the required model parameters. It is common practice to assume a Gaussian PDF to represent the distribution of the features for each class, and to estimate the required mean and variance parameters from the training sets. When PDFs are available to characterize pattern classes and their features (see Sections 5.12.2 and 5.12.3), optimal decision functions may be designed based upon statistical functions and decision theory. The following sections describe a few methods that fall into this category.

9.6.1 Likelihood functions and statistical decision

Let $P(C_i)$ be the probability of occurrence of class $C_i, \ i = 1, 2, \ldots, M$; this is known as the *a priori*, *prior*, or unconditional probability. The *a posteriori* or *posterior* probability that an observed sample pattern $\mathbf{x}$ comes from C_i is expressed as $P(C_i|\mathbf{x})$. If a classifier decides that $\mathbf{x}$ comes from C_j when it actually came from C_i, then the classifier is said to incur a *loss* L_{ij}, with $L_{ii} = 0$ or a fixed operational cost and $L_{ij} > L_{ii} \ \forall \ j \neq i$.

Since $\mathbf{x}$ may belong to any of M classes under consideration, the expected loss, known as the *conditional average risk* or *loss*, in assigning $\mathbf{x}$ to C_j is [473]

$$R_j(\mathbf{x}) = \sum_{i=1}^{M} L_{ij} \ P(C_i|\mathbf{x}). \tag{9.39}$$

A classifier could compute $R_j(\mathbf{x}), \ j = 1, 2, \ldots, M$, for each sample $\mathbf{x}$ and then assign $\mathbf{x}$ to the class with the smallest conditional loss. Such a classifier will minimize the total expected loss over all decisions, and is called the *Bayes classifier*. From a statistical point of view, the Bayes classifier represents an optimal classifier.

According to Bayes formula, we have [417, 473]

$$P(C_i|\mathbf{x}) = \frac{P(C_i) \ p(\mathbf{x}|C_i)}{p(\mathbf{x})}, \tag{9.40}$$

where $p(\mathbf{x}|C_i)$ is called the *likelihood function* of class C_i or the *state-conditional PDF* of $\mathbf{x}$, and $p(\mathbf{x})$ is the PDF of $\mathbf{x}$ regardless of class membership (unconditional). [*Note:* $P(y)$ is used to represent the probability of occurrence of an event y; $p(y)$ is used to represent the PDF of a random variable y. Probabilities and PDFs involving a multidimensional feature vector are multivariate functions with dimension equal to that of the feature vector.] Bayes formula shows how observing the sample $\mathbf{x}$ changes the *a priori* probability $P(C_i)$ to the *a posteriori* probability $P(C_i|\mathbf{x})$. In other words, Bayes formula provides a mechanism to update the *a priori* probability $P(C_i)$ to the *a posteriori* probability $P(C_i|\mathbf{x})$ due to the observation of the sample $\mathbf{x}$. Then, we can express the expected loss as [473]

$$R_j(\mathbf{x}) = \frac{1}{p(\mathbf{x})} \sum_{i=1}^{M} L_{ij} \ p(\mathbf{x}|C_i) \ P(C_i). \tag{9.41}$$

As $\frac{1}{p(\mathbf{x})}$ is common for all j, we could modify $R_j(\mathbf{x})$ to

$$r_j(\mathbf{x}) = \sum_{i=1}^{M} L_{ij}\; p(\mathbf{x}|C_i)\; P(C_i). \tag{9.42}$$

In a two-class case with $M = 2$, we obtain the following expressions [473]:

$$r_1(\mathbf{x}) = L_{11}\; p(\mathbf{x}|C_1)\; P(C_1) + L_{21}\; p(\mathbf{x}|C_2)\; P(C_2). \tag{9.43}$$

$$r_2(\mathbf{x}) = L_{12}\; p(\mathbf{x}|C_1)\; P(C_1) + L_{22}\; p(\mathbf{x}|C_2)\; P(C_2). \tag{9.44}$$

$$\mathbf{x} \in C_1 \;\text{ if }\; r_1(\mathbf{x}) < r_2(\mathbf{x}), \tag{9.45}$$

that is,

$$\begin{aligned} \mathbf{x} \in C_1 \;\text{ if }\quad & L_{11}\; p(\mathbf{x}|C_1)\; P(C_1) + L_{21}\; p(\mathbf{x}|C_2)\; P(C_2) \\ < \quad & L_{12}\; p(\mathbf{x}|C_1)\; P(C_1) + L_{22}\; p(\mathbf{x}|C_2)\; P(C_2), \end{aligned} \tag{9.46}$$

or equivalently,

$$\mathbf{x} \in C_1 \;\text{ if }\; (L_{21} - L_{22})\; p(\mathbf{x}|C_2)\; P(C_2) < (L_{12} - L_{11})\; p(\mathbf{x}|C_1)\; P(C_1). \tag{9.47}$$

This expression may be rewritten as [473]

$$\mathbf{x} \in C_1 \;\text{ if }\; \frac{p(\mathbf{x}|C_1)}{p(\mathbf{x}|C_2)} > \frac{P(C_2)}{P(C_1)} \frac{(L_{21} - L_{22})}{(L_{12} - L_{11})}. \tag{9.48}$$

The LHS of the inequality above, which is a ratio of two likelihood functions, is often referred to as the *likelihood ratio*:

$$l_{12}(\mathbf{x}) = \frac{p(\mathbf{x}|C_1)}{p(\mathbf{x}|C_2)}. \tag{9.49}$$

Then, Bayes decision rule for $M = 2$ is [473]:

1. Assign $\mathbf{x}$ to class C_1 if $l_{12}(\mathbf{x}) > \theta_{12}$, where θ_{12} is a threshold given by $\theta_{12} = \frac{P(C_2)}{P(C_1)} \frac{(L_{21}-L_{22})}{(L_{12}-L_{11})}$.

2. Assign $\mathbf{x}$ to class C_2 if $l_{12}(\mathbf{x}) < \theta_{12}$.

3. Make an arbitrary or heuristic decision if $l_{12}(\mathbf{x}) = \theta_{12}$.

The rule may be generalized to the M-class case as [473]:

$$\mathbf{x} \in C_i \;\text{ if }\; \sum_{k=1}^{M} L_{ki}\; p(\mathbf{x}|C_k)\; P(C_k) < \sum_{q=1}^{M} L_{qj}\; p(\mathbf{x}|C_q)\; P(C_q), \tag{9.50}$$

$j = 1, 2, \ldots, M,\; j \neq i.$

In most pattern classification problems, the loss is nil for correct decisions. The loss could be assumed to be equal to a certain quantity for all erroneous decisions. Then, $L_{ij} = 1 - \delta_{ij}$, where

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise} \end{cases} \tag{9.51}$$

and

$$\begin{aligned} r_j(\mathbf{x}) &= \sum_{i=1}^{M} (1 - \delta_{ij})\, p(\mathbf{x}|C_i)\, P(C_i) \\ &= p(\mathbf{x}) - p(\mathbf{x}|C_j)\, P(C_j), \end{aligned} \tag{9.52}$$

since

$$\sum_{i=1}^{M} p(\mathbf{x}|C_i)\, P(C_i) = p(\mathbf{x}). \tag{9.53}$$

The Bayes classifier will assign a pattern $\mathbf{x}$ to class C_i if

$$p(\mathbf{x}) - p(\mathbf{x}|C_i)P(C_i) < p(\mathbf{x}) - p(\mathbf{x}|C_j)P(C_j),\ j = 1, 2, \ldots, M,\ j \neq i, \tag{9.54}$$

that is,

$$\mathbf{x} \in C_i \ \text{ if } \ p(\mathbf{x}|C_i)P(C_i) > p(\mathbf{x}|C_j)P(C_j),\ j = 1, 2, \ldots, M,\ j \neq i. \tag{9.55}$$

This is nothing more than using the decision functions

$$d_i(\mathbf{x}) = p(\mathbf{x}|C_i)\, P(C_i),\ i = 1, 2, \ldots, M, \tag{9.56}$$

where a pattern $\mathbf{x}$ is assigned to class C_i if $d_i(\mathbf{x}) > d_j(\mathbf{x})\ \forall\ j \neq i$ for that pattern. Using Bayes formula, we get

$$d_i(\mathbf{x}) = P(C_i|\mathbf{x})\, p(\mathbf{x}),\ i = 1, 2, \ldots, M. \tag{9.57}$$

Since $p(\mathbf{x})$ does not depend upon the class index i, this can be reduced to

$$d_i(\mathbf{x}) = P(C_i|\mathbf{x}),\ i = 1, 2, \ldots, M. \tag{9.58}$$

The different decision functions given above provide alternative yet equivalent approaches, depending upon whether $p(\mathbf{x}|C_i)$ or $P(C_i|\mathbf{x})$ is used (or available). Estimation of $p(\mathbf{x}|C_i)$ would require a training set for each class C_i. It is common to assume a Gaussian distribution and estimate its mean and variance using the training set.

9.6.2 Bayes classifier for normal patterns

The univariate normal or Gaussian PDF for a single random variable x is given by

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2\right], \tag{9.59}$$

which is completely specified by two parameters: the mean

$$m = E[x] = \int_{-\infty}^{\infty} x\, p(x)\, dx, \tag{9.60}$$

and the variance

$$\sigma^2 = E[(x-m)^2] = \int_{-\infty}^{\infty} (x-m)^2\, p(x)\, dx. \tag{9.61}$$

In the case of M pattern classes and pattern vectors $\mathbf{x}$ of dimension n governed by multivariate normal PDFs, we have

$$p(\mathbf{x}|C_i) = \frac{1}{(2\pi)^{n/2}|\mathbf{C}_i|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T\mathbf{C}_i^{-1}(\mathbf{x}-\mathbf{m}_i)\right], \tag{9.62}$$

$i = 1, 2, \ldots, M$, where each PDF is completely specified by its mean vector $\mathbf{m}_i$ and its $n \times n$ covariance matrix $\mathbf{C}_i$, with

$$\mathbf{m}_i = E_i[\mathbf{x}], \tag{9.63}$$

and

$$\mathbf{C}_i = E_i[(\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^T]. \tag{9.64}$$

Here, $E_i[\,]$ denotes the expectation operator over the patterns belonging to class C_i.

Normal distributions occur frequently in nature and have the advantage of analytical tractability. A multivariate normal PDF reduces to a product of univariate normal PDFs when the elements of $\mathbf{x}$ are mutually independent (then the covariance matrix is a diagonal matrix).

Earlier, we formulated the decision functions

$$d_i(\mathbf{x}) = p(\mathbf{x}|C_i)\; P(C_i),\; i = 1, 2, \ldots, M. \tag{9.65}$$

Given the exponential in the normal PDF, it is convenient to use

$$d_i(\mathbf{x}) = \ln\,[p(\mathbf{x}|C_i)\; P(C_i)] = \ln p(\mathbf{x}|C_i) + \ln P(C_i), \tag{9.66}$$

which is equivalent in terms of classification performance as the natural logarithm ln is a monotonically increasing function. Then [473],

$$d_i(\mathbf{x}) = \ln P(C_i) - \frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{C}_i| - \frac{1}{2}[(\mathbf{x}-\mathbf{m}_i)^T\mathbf{C}_i^{-1}(\mathbf{x}-\mathbf{m}_i)], \tag{9.67}$$

$i = 1, 2, \ldots, M$. The second term does not depend upon i; therefore, we can simplify $d_i(\mathbf{x})$ to

$$d_i(\mathbf{x}) = \ln P(C_i) - \frac{1}{2}\ln|\mathbf{C}_i| - \frac{1}{2}[(\mathbf{x}-\mathbf{m}_i)^T \mathbf{C}_i^{-1}(\mathbf{x}-\mathbf{m}_i)],\ i = 1, 2, \ldots, M. \tag{9.68}$$

The decision functions above are hyperquadrics; hence, the best that a Bayes classifier for normal patterns can do is to place a general second-order (quadratic) decision surface between each pair of pattern classes. In the case of true normal distributions of patterns, the decision functions as above are optimal on an average basis: They minimize the expected loss with the simplified loss function $L_{ij} = 1 - \delta_{ij}$ [473].

If all the covariance matrices are equal, that is, $\mathbf{C}_i = \mathbf{C},\ i = 1, 2, \ldots, M$, we get

$$d_i(\mathbf{x}) = \ln P(C_i) + \mathbf{x}^T \mathbf{C}^{-1}\mathbf{m}_i - \frac{1}{2}\mathbf{m}_i^T \mathbf{C}^{-1}\mathbf{m}_i,\ i = 1, 2, \ldots, M, \tag{9.69}$$

after omitting terms independent of i. The Bayesian classifier is now represented by a set of linear decision functions.

Before one may apply the decision functions as above, it would be appropriate to verify the Gaussian nature of the PDFs of the variables on hand by conducting statistical tests [6, 268]. Furthermore, it would be necessary to derive or estimate the mean vector and covariance matrix for each class; sample statistics computed from a training set may serve this purpose.

See Section 9.11.2 for an illustration of application of the Bayesian method to an ECG signal with PVCs.

9.7 Logistic Regression Analysis

Logistic classification is a statistical technique based on a logistic regression model that estimates the probability of occurrence of an event [477–479]. The technique is designed for problems where patterns are to be classified into one of two classes. When the response variable is binary, theoretical and empirical considerations indicate that the response function is often curvilinear. The typical response function is shaped as a forward or backward tilted "S" and is known as a sigmoidal function. The function has asymptotes at 0 and 1.

In logistic pattern classification, an event is defined as the membership of a pattern vector in one of the two classes. The method computes a variable that depends upon the given parameters and is constrained to the range $[0, 1]$ so that it may be interpreted as a probability. The probability of the pattern vector belonging to the second class is simply the difference between unity and the estimated value.

For the case of a single feature or parameter, the logistic regression model is given as

$$P(\text{event}) = \frac{\exp(b_0 + b_1 x)}{1 + \exp(b_0 + b_1 x)}, \tag{9.70}$$

or equivalently,

$$P(\text{event}) = \frac{1}{1 + \exp[-(b_0 + b_1 x)]}, \tag{9.71}$$

where b_0 and b_1 are coefficients estimated from the data, and x is the independent (feature) variable. The relationship between the independent variable and the estimated probability is nonlinear; it follows an S-shaped curve that resembles the integral of a Gaussian function. In the case of an n-dimensional feature vector $\mathbf{x}$, the model can be written as

$$P(\text{event}) = \frac{1}{1 + \exp(-z)}, \tag{9.72}$$

where z is the linear combination

$$z = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n = \langle \mathbf{b}, \mathbf{x} \rangle, \tag{9.73}$$

that is, z is the dot product of the augmented feature vector $\mathbf{x}$ with a coefficient or weight vector $\mathbf{b}$.

In linear regression, the coefficients of the model are estimated using the method of least squares; the selected regression coefficients are those that result in the smallest sum of squared distances between the observed and the predicted values of the dependent variable. In logistic regression, the parameters of the model are estimated using the maximum likelihood method [268, 477]; the coefficients that make the observed results "most likely" are selected. Since the logistic regression model is nonlinear, an iterative algorithm is necessary for estimation of the coefficients [478,479]. A training set is required to design a classifier based upon logistic regression.

9.8 Neural Networks

In many practical problems, we may have no knowledge of the prior probabilities of patterns belonging to one class or another. No general classification rules may exist for the patterns on hand. Clinical knowledge may not yield symbolic knowledge bases that could be used to classify patterns that demonstrate exceptional behavior. In such situations, conventional pattern classification methods as described in the preceding sections may not be well suited for classification of pattern vectors. Artificial neural networks (ANNs), with the properties of experience-based learning and fault tolerance, should be effective in solving such classification problems [267, 475, 476, 480–483].

Figure 9.4 illustrates a two-layer perceptron with one hidden layer and one output layer for pattern classification. The network learns the similarities among patterns directly from their instances in the training set that is provided initially. Classification rules are inferred from the training data without prior knowledge of the pattern class distributions in the data. Training of an ANN classifier is typically achieved by the *back-propagation* algorithm [267, 475, 476, 480–483]. The actual output of the ANN

y_k is calculated as

$$y_k = f\left(\sum_{j=1}^{J} w_{jk}^{\#} \, x_j^{\#} - \theta_k^{\#} \right), \quad k = 1, 2, \ldots, K, \tag{9.74}$$

where

$$x_j^{\#} = f\left(\sum_{i=1}^{I} w_{ij} \, x_i - \theta_j \right), \quad j = 1, 2, \ldots, J, \tag{9.75}$$

and

$$f(\beta) = \frac{1}{1 + \exp(-\beta)}. \tag{9.76}$$

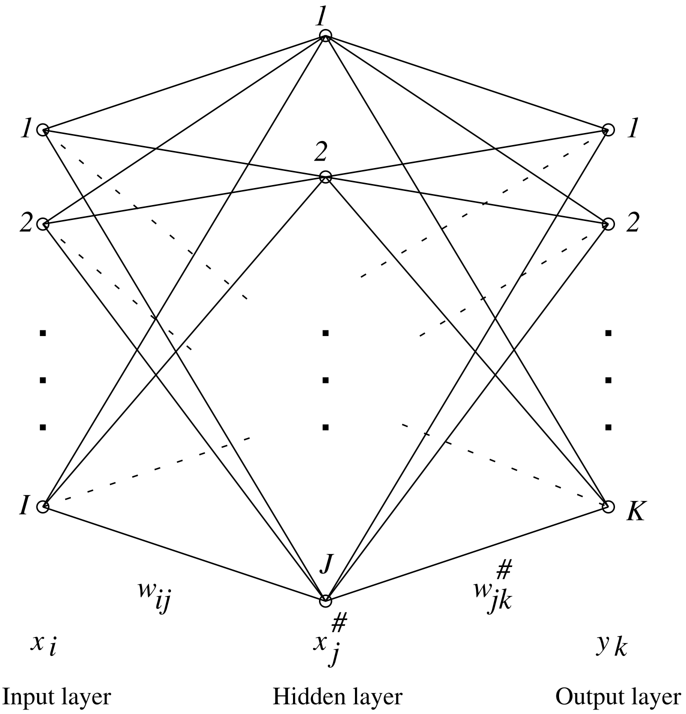

Figure 9.4 A two-layer perceptron.

In the equations given above, θ_j and $\theta_k^{\#}$ are node offsets; w_{ij} and $w_{jk}^{\#}$ are node weights; x_i are the elements of the pattern vectors (input parameters); and I, J, and K are the numbers of nodes in the input, hidden, and output layers, respectively. The weights and offsets are updated by

$$w_{jk}^{\#}(n+1) = w_{jk}^{\#}(n) + \eta[y_k(1-y_k)(d_k-y_k)]x_j^{\#} + \alpha[w_{jk}^{\#}(n) - w_{jk}^{\#}(n-1)], \tag{9.77}$$

$$\theta_k^{\#}(n+1) = \theta_k^{\#}(n) + \eta[y_k(1-y_k)(d_k-y_k)](-1) + \alpha[\theta_k^{\#}(n) - \theta_k^{\#}(n-1)], \quad (9.78)$$

$$\begin{aligned} w_{ij}(n+1) &= w_{ij}(n) \quad (9.79) \\ &+ \eta \left[x_j^{\#}(1-x_j^{\#}) \sum_{k=1}^{K} \{y_k(1-y_k)(d_k-y_k)w_{jk}^{\#}\} \right] x_i \\ &+ \alpha[w_{ij}(n) - w_{ij}(n-1)], \end{aligned}$$

and

$$\begin{aligned} \theta_j(n+1) &= \theta_j(n) \quad (9.80) \\ &+ \eta \left[x_j^{\#}(1-x_j^{\#}) \sum_{k=1}^{K} \{y_k(1-y_k)(d_k-y_k)w_{jk}^{\#}\} \right] (-1) \\ &+ \alpha[\theta_j(n) - \theta_j(n-1)], \end{aligned}$$

where d_k are the desired outputs, α is a momentum term, η is a gain term, and n refers to the iteration number. Equations 9.77 and 9.78 represent the back-propagation steps, with $y_k(1-y_k)x_j^{\#}$ being the sensitivity of y_k to $w_{jk}^{\#}$, that is, $\frac{\partial y_k}{\partial w_{jk}^{\#}}$.

The training algorithm is repeated until the errors between the desired outputs and the actual outputs for the training data are smaller than a predetermined threshold value. Shen et al. [483] present a LOO approach to determine the most suitable values for the parameters J, η, and α.

9.8.1 ANNs with radial basis functions

An ANN with RBF [484] or RBF network (RBFN) with a feed-forward hidden layer (see Figure 9.5) applies a nonlinear transformation from the input space to a high-dimensional hidden space, and then produces responses through a linear output transformation. Cover [485] showed that it is possible to transform a given a set of samples that is not linearly separable into one that is linearly separable by applying a nonlinear transformation to project it into a higher-dimensional space. This approach forms the motivation for the use of nonlinear kernel-based methods in pattern classification and machine learning applications.

Consider a set of N labeled input feature vectors, $\mathbf{x}_n$, $n = 1, 2, \ldots, N$, characterizing the N signals in a study, each of which is an $M \times 1$ vector. Let y_n be the desired output or classification response for the n^{th} signal, represented by its feature vector $\mathbf{x}_n$. With reference to the RBFN shown in Figure 9.5, we have the output of the network as

$$\widetilde{y_n} = \sum_{j=1}^{J} w_j \, \phi(\mathbf{x}_n, \mathbf{c}_j) + w_0, \quad (9.81)$$

where $\widetilde{y_n}$ indicates an estimate of y_n, the RBF ϕ is defined as

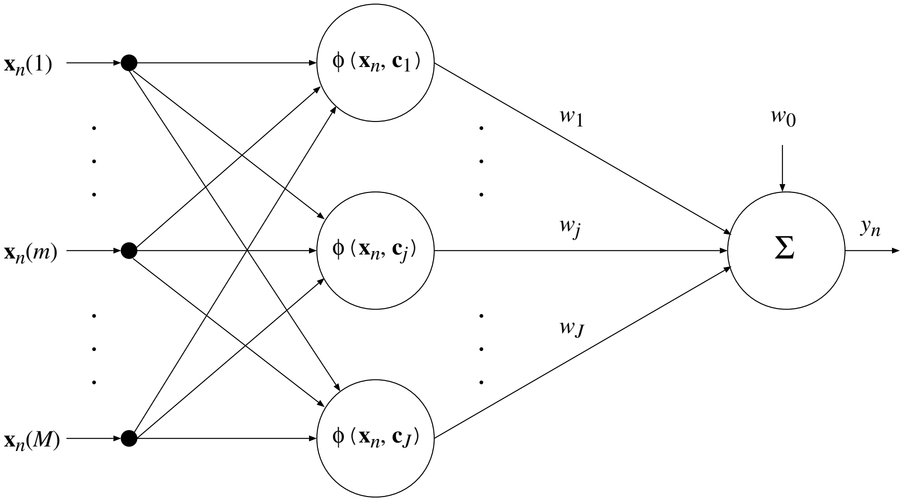
$\mathbf{x}_n(1)$
$\mathbf{x}_n(m)$
$\mathbf{x}_n(M)$
$\phi\,(\mathbf{x}_n, \mathbf{c}_1)$
$\phi\,(\mathbf{x}_n, \mathbf{c}_j)$
$\phi\,(\mathbf{x}_n, \mathbf{c}_J)$
w_1
w_j
w_J
w_0
Σ
y_n

Figure 9.5 Schematic representation of an RBF network. The inputs to the network, $\mathbf{x}_n(1), \mathbf{x}_n(2), \ldots, \mathbf{x}_n(M)$, are the M components of the feature vector $\mathbf{x}_n$ of a signal to be classified. The functions shown as ϕ are the RBFs. The hidden layer has J neurons.

$$\phi(\mathbf{x}_n, \mathbf{c}_j) = \exp\left(- \log_e(2) \frac{\|\mathbf{x}_n - \mathbf{c}_j\|^2}{\sigma^2} \right), \tag{9.82}$$

w_j is the weight and $\mathbf{c}_j$ is the center vector for the j^{th} neuron in the hidden layer, J is the number of neurons in the hidden layer, w_0 is the bias, and σ is the spread parameter that determines the width of the area in the input space to which each hidden neuron responds.

The major challenge in the design of an RBFN is the selection of the centers. The selection of the centers in a random fashion commonly leads to a relatively large network with high computation complexity. Rangayyan and Wu [264] applied the orthogonal least-squares (OLS) method [486], a systematic method for center selection which can reduce the size of the RBFN, for screening of VAG signals.

According to Equation 9.81, the mapping performed by the RBFN can be viewed as a regression model, expressed in matrix form as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & \phi(\mathbf{x}_1, \mathbf{c}_1) & \cdots & \phi(\mathbf{x}_1, \mathbf{c}_J) \\ 1 & \phi(\mathbf{x}_2, \mathbf{c}_1) & \cdots & \phi(\mathbf{x}_2, \mathbf{c}_J) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \phi(\mathbf{x}_N, \mathbf{c}_1) & \cdots & \phi(\mathbf{x}_N, \mathbf{c}_J) \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_J \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}, \tag{9.83}$$

which is equivalent to

$$\mathbf{y} = \Phi\mathbf{w} + \mathbf{e}, \tag{9.84}$$

where Φ is the $N\times(J+1)$ regression matrix with the RBFs; $\mathbf{y}$ represents the vectorial form of the corresponding values y_n for $n = 1, 2, \ldots, N$; $\mathbf{w} = [w_0, w_1, \cdots, w_J]^T$; and $\mathbf{e}$ is the approximation error.

The centers of the RBFN are chosen from the set of input feature vectors (a total of N candidates). The task of the OLS method is to perform a systematic selection of fewer than N centers so that the network size can be reduced with minimal degradation of performance during the learning procedure. From Equation 9.83, we can see that there is a one-to-one correspondence between the centers of the RBFN and the coefficients in the regression matrix Φ. At each step of the OLS regression, a new center can be selected in such a manner that the incremental variance of the desired output is maximized. Suppose that there are $Q < N$ centers selected. The OLS solution yielding the weights is given by [486]

$$\widetilde{\mathbf{w}} = \left(\Phi^T\Phi\right)^{-1}\Phi^T\mathbf{y} = \Phi^+\mathbf{y}, \tag{9.85}$$

where Φ^+ represents the pseudoinverse of the regression matrix Φ. The output of the RBFN is then expressed as

$$\widetilde{\mathbf{y}} = \Phi\widetilde{\mathbf{w}} = [\phi_1, \phi_2, \ldots, \phi_Q]\widetilde{\mathbf{w}}, \tag{9.86}$$

where $\widetilde{\mathbf{y}}$ denotes the portion of $\mathbf{y}$ that is within the vectorial space spanned by the columns ϕ_q of the regression matrix Φ.

By using Gram–Schmidt orthogonalization [484], the regression matrix can be decomposed as

$$\Phi = \mathbf{BA} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_Q]\begin{bmatrix} 1 & a_{11} & a_{12} & \cdots & a_{1Q} \\ 0 & 1 & a_{23} & \cdots & a_{2Q} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}, \tag{9.87}$$

where $\mathbf{A}$ is a $Q \times Q$ upper-triangular matrix with 1s on the main diagonal, and $\mathbf{B}$ is an $N \times Q$ matrix with mutually orthogonal columns $\mathbf{b}_q$ such that

$$\mathbf{B}^T\mathbf{B} = \mathbf{H} = diag[h_1, h_2, \cdots, h_Q], \tag{9.88}$$

where the $Q \times Q$ matrix $\mathbf{H}$ is a diagonal matrix with elements h_k given by

$$h_k = \mathbf{b}_k^T\mathbf{b}_k = \sum_{q=1}^{N} b_{kq}^2. \tag{9.89}$$

By substituting Equation 9.87 into Equation 9.84, we obtain

$$\mathbf{y} = \mathbf{B}\,\mathbf{A}\,\mathbf{w} + \mathbf{e} = \mathbf{B}\,\mathbf{g} + \mathbf{e}, \tag{9.90}$$

where $\mathbf{g} = \mathbf{A}\mathbf{w}$. In Equation 9.90, the desired output vector $\mathbf{y}$ is expressed as a linear combination of the mutually orthogonal columns of the matrix $\mathbf{B}$. The OLS solution for the coordinate vector $\mathbf{g}$ is given by

$$\widetilde{\mathbf{g}} = (\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T \mathbf{y} = \mathbf{B}^{+}\mathbf{z} = \mathbf{H}^{-1} \mathbf{B}^T \mathbf{y}. \tag{9.91}$$

The k^{th} component of the vector $\widetilde{\mathbf{g}}$ is given by

$$g_k = \frac{\mathbf{b}_k^T \mathbf{y}}{\mathbf{b}_k^T \mathbf{b}_k}. \tag{9.92}$$

Because Gram–Schmidt orthogonalization ensures the orthogonality between the approximation error $\mathbf{e}$ and $\mathbf{Bg}$ in Equation 9.90, we have

$$\mathbf{y}^T \mathbf{y} = \mathbf{g}^T \mathbf{B}^T \mathbf{B} \mathbf{g} + \mathbf{e}^T \mathbf{e} = \mathbf{g} \mathbf{H} \mathbf{g} + \mathbf{e}^T \mathbf{e} = \sum_{k=1}^{Q} h_k \, g_k^2 + \mathbf{e}^T \mathbf{e}. \tag{9.93}$$

Because there is a one-to-one correspondence between the elements of the regression vector $\mathbf{g}$ and the RBF centers $\mathbf{c}_j$, each term in the summation above reflects the contribution of each of the RBF centers. We can, therefore, define an error reduction ratio (ϵ) with respect to the j^{th} RBF center as [486]

$$\epsilon_j = \frac{h_j \, g_j^2}{\mathbf{y}^T \mathbf{y}}. \tag{9.94}$$

The error reduction ratio offers an effective criterion for the selection of RBF centers in a regression model. At each step of the regression, an RBF center is selected so as to maximize the error reduction ratio toward a tolerance value.

The input layer of the RBFN used by Rangayyan and Wu [264] contained $M = 5$ nodes to accept the set of five features (FF_1, FF_2, S, K, H) extracted from each VAG signal (see Sections 5.12 and 9.4.2). The spread parameter σ was varied over the range $[1, 6]$, and the number of hidden nodes J was varied over the range $[1, 30]$. The resulting output values were used to derive ROC curves and the associated A_z values. The highest A_z value obtained was 0.82 using the RBFN classifier with $\sigma = 6$ and $J = 23$ hidden nodes using $N = 89$ VAG signals. In another work, Rangayyan and Wu [265] used an RBFN with the turns count computed for the first and second halves of each VAG signal, and obtained better classification performance with $A_z = 0.92$.

Regardless of the good results they provide and their popularity, RBFNs pose problems related to the derivation of optimal network parameters and generalization from a training set to a test set.

9.9 Measures of Diagnostic Accuracy and Cost

Pattern recognition or classification decisions that are made in the context of medical diagnosis have implications that go beyond statistical measures of accuracy and validity. We need to provide a clinical or diagnostic interpretation of statistical or rule-based decisions made with signal pattern vectors.

Consider the simple situation of *screening*, which represents the use of a test to detect the presence or absence of a specific disease in a certain study population. The decision to be made is binary: A given subject has or does not have the disease in question. Let us represent by A the event that a subject has the particular pathology, and by N the event that the subject does not have the disease. Let the prior probabilities $P(A)$ and $P(N)$ represent the fractions of subjects with the disease and the subjects without the disease, respectively, in the test population. Let T^+ represent a positive screening test result (indicative of the presence of the disease) and T^- a negative result. The following possibilities arise [487]:

- A *true positive* (TP) is the situation when the test is positive for a subject with the disease (also known as a *hit*). The true-positive fraction (TPF) or *sensitivity* S^+ is given as $P(T^+|A)$ or

  $$S^+ = \frac{\text{number of TP decisions}}{\text{number of subjects with the disease}} . \tag{9.95}$$

  The sensitivity of a test represents its capability to detect the presence of the disease of concern.

- A *true negative* (TN) represents the case when the test is negative for a subject who does not have the disease. The true-negative fraction (TNF) or *specificity* S^- is given as $P(T^-|N)$ or

  $$S^- = \frac{\text{number of TN decisions}}{\text{number of subjects without the disease}} . \tag{9.96}$$

  The specificity of a test indicates its accuracy in identifying the absence of the disease of concern.

- A *false negative* (FN) is said to occur when the test is negative for a subject who has the disease of concern; that is, the test has missed the case. The probability of this error, known as the false-negative fraction (FNF), is $P(T^-|A)$.

- A *false positive* (FP) is defined as the case where the result of the test is positive when the individual being tested does not have the disease. The probability of this type of error or false alarm, known as the false-positive fraction (FPF), is $P(T^+|N)$.

Table 9.1 summarizes the classification possibilities. Note that

- $FNF + TPF = 1$,
- $FPF + TNF = 1$,
- $S^- = 1 - FPF = TNF$, and
- $S^+ = 1 - FNF = TPF$.

A summary measure of accuracy may be defined as [487]

$$\text{accuracy} = S^{+}\ P(A) + S^{-}\ P(N), \tag{9.97}$$

where $P(A)$ is the fraction of the study population that actually has the disease (that is, the prevalence of the disease) and $P(N)$ is the fraction of the study population that is actually free of the disease.

If the prior probabilities are not available, the accuracy of classification can be estimated as

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{9.98}$$

In the same way, we have

$$TPF = \frac{TP}{TP + FN} \tag{9.99}$$

and

$$TNF = \frac{TN}{TN + FP}. \tag{9.100}$$

Actual Group	Predicted Group	
	Without the Disease	With the Disease
Without the Disease	$S^{-} = TNF$	FPF
With the Disease	FNF	$S^{+} = TPF$

Table 9.1 Schematic representation of a classification matrix. S^{-} denotes the specificity (true-negative fraction or TNF), FPF denotes the false-positive fraction, FNF denotes the false-negative fraction, and S^{+} denotes the sensitivity (true-positive fraction or TPF).

The efficiency of a test may also be indicated by its predictive values. The *positive predictive value* PPV of a test, defined as

$$PPV = 100\ \frac{TP}{TP + FP}\ , \tag{9.101}$$

represents the percentage of the cases labeled as positive by the test that are actually positive. The *negative predictive value* NPV, defined as

$$NPV = 100\ \frac{TN}{TN + FN}\ , \tag{9.102}$$

represents the percentage of cases labeled as negative by the test that are actually negative.

When a new test or method of diagnosis is being developed and tested, it is necessary to use another previously established method as a reference to confirm the presence or absence of the disease. Such a reference method is often called the *gold standard*, and its results are referred to as the *ground truth*. When computer-based methods need to be tested, it is common practice to use the diagnosis or classification provided by an expert in the field as the gold standard. Results of biopsy, other established laboratory or investigative procedures, or long-term clinical follow-up in the case of normal subjects may also serve this purpose. The term "actual group" in Table 9.1 indicates the result of the gold standard, and the term "predicted group" refers to the result of the test conducted.

Health-care professionals (and the general public) would be interested in knowing the probability that a subject with a positive test result actually has the disease: This is given by the conditional probability $P(A|T^+)$. The question could be answered by using Bayes theorem [268] as

$$P(A|T^+) = \frac{P(A)\ P(T^+|A)}{P(A)P(T^+|A) + P(N)P(T^+|N)}\,. \qquad (9.103)$$

Note that $P(T^+|A) = S^+$ and $P(T^+|N) = 1 - S^-$. In order to determine the posterior probability as above, the sensitivity and specificity of the test and the prior probabilities of negative cases and positive cases (the rate of prevalence of the disease) should be known.

A cost matrix may be defined, as in Table 9.2, to reflect the overall cost effectiveness of a test or method of diagnosis. The cost of conducting the test and arriving at a TN decision is indicated by C_N; this is the cost of subjecting a normal person to the test for the purposes of screening for a disease. The cost of the test when a TP is found is shown as C_A; this might include the costs of further tests, treatment, and follow-up, which are secondary to the test itself but part of the screening and health-care program. The value C^+ indicates the cost of an FP result; this represents the cost of erroneously subjecting an individual without the disease to further tests or therapy. Whereas it may be easy to identify the costs of clinical tests or treatment procedures, it is difficult to quantify the traumatic and psychological effects of an FP result and the consequent procedures on a normal subject. The cost C^- is the cost of an FN result: The presence of the disease in a patient is not diagnosed, the condition worsens with time, the patient faces more complications of the disease, and the health-care system or the patient has to bear the costs of further tests and delayed therapy.

A loss factor due to misclassification may be defined as

$$L = FPF \times C^+ \ + \ FNF \times C^-. \qquad (9.104)$$

The total cost of the screening program may be computed as

$$C_S = TPF \times C_A + TNF \times C_N + FPF \times C^+ \ + \ FNF \times C^-. \qquad (9.105)$$

Metz [487] provides more details on the computation of the costs of diagnostic tests.

Actual Group	Predicted Group	
	Without the Disease	With the Disease
Without the Disease	C_N	C^+
With the Disease	C^-	C_A

Table 9.2 Schematic representation of the cost matrix of a diagnostic method.

9.9.1 Receiver operating characteristics

Measures of overall correct classification of patterns as percentages provide limited indications of the accuracy of a diagnostic method. The provision of separate percentage correct classification for each category, such as sensitivity and specificity, can facilitate improved analysis. However, these measures do not indicate the dependence of the results upon the decision threshold. Furthermore, the effect of the rate of incidence or prevalence of the particular disease is not considered.

From another perspective, it is desirable to have a screening or diagnostic test that is both highly sensitive and highly specific. In reality, however, such a test is usually not achievable. Most tests are based on clinical measurements that can assume limited ranges of a variable (or a few variables) with an inherent trade-off between sensitivity and specificity. The relationship between sensitivity and specificity is illustrated by the ROC curve, which facilitates improved analysis of the classification accuracy of a diagnostic method [338, 487–489].

Consider the situation illustrated in Figure 9.6. For a given diagnostic test with the decision variable z, we have predetermined state-conditional PDFs of the decision variable z for actually negative cases indicated as $p(z|N)$ and for actually positive indicated as $p(z|A)$. As indicated in Figure 9.6, the two PDFs will almost always overlap, given that no method can be perfect. The user or operator needs to determine a decision threshold (indicated by the vertical line) so as to strike a compromise between sensitivity and specificity. Lowering the decision threshold will increase TPF at the cost of increased FPF. (*Note:* TNF and FNF may be derived easily from FPF and TPF, respectively.)

An ROC curve is a graph that plots (FPF, TPF) points obtained for a range of decision threshold or cut points of the decision method (see Figure 9.7). The cut point could correspond to the threshold of the probability of prediction. By varying the decision threshold, we get different decision fractions, within the range $[0, 1]$. An ROC curve describes the inherent detection (diagnostic or discriminant) characteristics of a test or method: A receiver (user) may choose to operate at any point along the curve. The ROC curve is independent of the prevalence of the disease or disorder being investigated as it is based upon normalized decision fractions. As all cases may be simply labeled as negative or all may be labeled as positive, an ROC curve has to pass through the points $(0, 0)$ and $(1, 1)$.

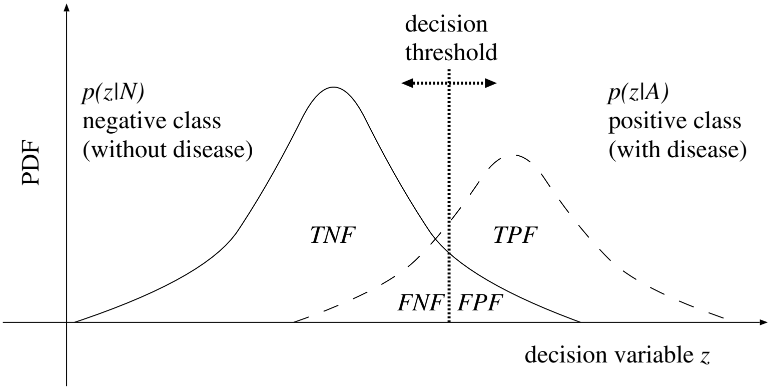

Figure 9.6 State-conditional PDFs of a diagnostic decision variable z for negative and positive cases. The vertical line represents the decision threshold. Based on a similar figure in T.M. Cabral and R.M. Rangayyan, *Fractal Analysis of Breast Masses in Mammograms*, Morgan & Claypool, 2012.

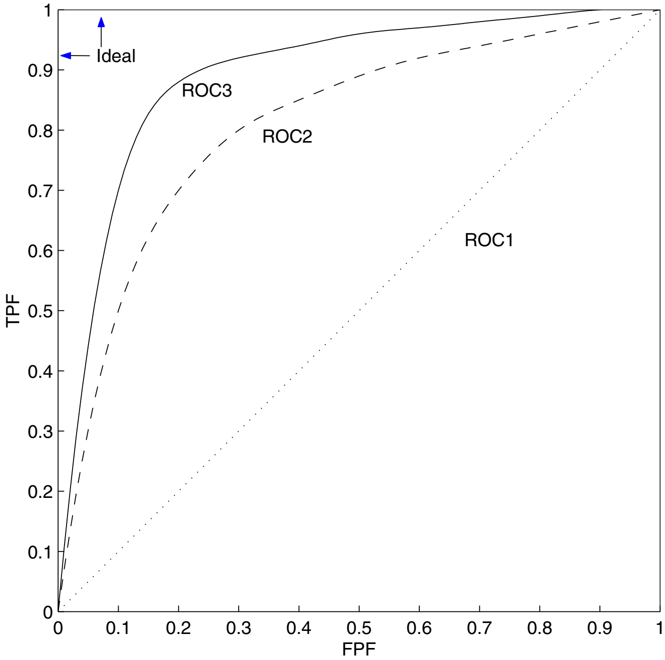

Figure 9.7 Examples of receiver operating characteristic curves.

In a diagnostic situation where a human operator or specialist is required to provide the diagnostic decision, ROC analysis is usually conducted by requiring the specialist to rank each case as one of five possibilities [487]:

1. definitely or almost definitely negative,
2. probably negative,
3. possibly positive,
4. probably positive,
5. definitely or almost definitely positive.

Item 3 above may be replaced by "indeterminate" if appropriate. Various values of TPF and FPF are then calculated by varying the decision threshold from level 5 to level 1 according to the decision items listed above. The resulting (FPF, TPF) points are then plotted to form an ROC curve. The maximum likelihood estimation method [490] is commonly used to fit a binormal curve to data as above.

A summary measure of effectiveness of a test is given by the area under the ROC curve, traditionally labeled as A_z. It is clear from Figure 9.7 that A_z is limited to the range $[0, 1]$. A test that gives a larger area under the ROC curve indicates a better method than one with a smaller area: in Figure 9.7, the method corresponding to ROC3 is better than the method corresponding to ROC2; both are better than the method represented by ROC1 with $A_z = 0.5$. An ideal method will have an ROC curve that follows the vertical line from $(0, 0)$ to $(0, 1)$ and then the horizontal line from $(0, 1)$ to $(1, 1)$, with $A_z = 1$: The method has $TPF = 1$ with $FPF = 0$, which is ideal. (*Note:* This would require the PDFs represented in Figure 9.6 not to overlap.)

9.9.2 McNemar's test of symmetry

Problem: *Suppose we have two methods to perform a certain diagnostic test. How may we compare the classification performance of one against that of the other?*

Solution: Measures of overall classification accuracies such as a percentage of correct classification or the area under the ROC curve provide simple measures to compare two or more diagnostic methods. If more details are required as to how the classifications of groups of cases vary from one method to another, McNemar's test of symmetry [491, 492] would be an appropriate tool.

McNemar's test is based on the construction of contingency tables that compare the results of two classification methods. The rows of a contingency table represent the outcomes of one of the methods used as the reference, possibly a gold standard (labeled as Method A in Table 9.3); the columns represent the outcomes of the other method, which is usually a new method (Method B) to be evaluated against the gold standard. The entries in the table are counts that correspond to particular diagnostic categories, which in Table 9.3 are labeled as normal, indeterminate, and abnormal. A separate contingency table should be prepared for each true category of the patterns;

for example, normal and abnormal. (The class "indeterminate" may not be applicable as a true category.) The true category of each case may have to be determined by a third method (for example, biopsy or surgery).

	Method B			
Method A	Normal	Indeterminate	Abnormal	Total
Normal	a (1)	b (2)	c (3)	$R1$
Indeterminate	d (4)	e (5)	f (6)	$R2$
Abnormal	g (7)	h (8)	i (9)	$R3$
Total	$C1$	$C2$	$C3$	N

Table 9.3 Schematic representation of a contingency table for McNemar's test of asymmetry [448, 493].

In Table 9.3, the variables a, b, c, d, e, f, g, h, and i denote the counts in each cell, and the numbers in parentheses denote the cell number. The variables $C1$, $C2$, and $C3$ denote the total numbers of counts in the corresponding columns; $R1$, $R2$, and $R3$ denote the total numbers of counts in the corresponding rows. The total number of cases in the true category represented by the table is $N = C1 + C2 + C3 = R1 + R2 + R3$.

Each cell in a contingency table represents a paired outcome. For example, in evaluating the diagnostic efficiency of Method B versus Method A, cell number 3 will contain the number of samples that were classified as normal by Method A but as abnormal by Method B. The row totals ($R1$, $R2$, and $R3$) and the column totals ($C1$, $C2$, and $C3$) may be used to determine the sensitivity and specificity of the methods.

High values along the main diagonal (a, e, i) of a contingency table (see Table 9.3) indicate no change or difference in diagnostic performance with Method B as compared to Method A. In a contingency table for truly abnormal cases, a high value in the upper-right portion (cell number 3) will indicate an improvement in diagnosis (higher sensitivity) with Method B as compared to Method A. In evaluating a contingency table for truly normal cases, Method B will have a higher specificity than Method A if a large value is found in cell 7. McNemar's method may be used to perform detailed statistical analysis of improvement in performance based upon contingency tables if large numbers of cases are available in each category [491,492].

Illustration of application: Krishnan et al. [493] proposed two methods for auditory display and sonification of processed VAG signals. Table 9.4 shows the contingency table for 18 abnormal VAG signals comparing their classification by listening to direct playback of the signal data or after a sonification technique based on the IMF and envelope of a given VAG signal. The table shows substantial improvement in sensitivity with the sonification method $(15/18)$ as compared to direct playback (4/18). Ten abnormal signals that were called normal with direct playback were correctly classified as abnormal with the sonification method.

	Sonification			
Direct Playback	Normal	Indeterminate	Abnormal	Total
Normal	2	0	10	12
Indeterminate	1	0	1	2
Abnormal	0	0	4	4
Total	3	0	15	18

Table 9.4 Contingency table for a method of sonification of VAG signals versus direct playback with 18 abnormal signals [448, 493].

Table 9.5 shows the contingency table for 19 normal VAG signals comparing their classification by listening to direct playback of the signal data or after sonification. The results indicate poorer specificity (8/19) of the results of sonification as compared to direct playback (13/19). The results indicate that the higher sensitivity of the sonification method was gained at the expense of a decrease in specificity. See Krishnan et al. [493] for further details.

	Sonification			
Direct Playback	Normal	Indeterminate	Abnormal	Total
Normal	8	2	3	13
Indeterminate	0	0	1	1
Abnormal	0	0	5	5
Total	8	2	9	19

Table 9.5 Contingency table for a method of sonification of VAG signals versus direct playback with 19 normal signals [448, 493].

9.10 Reliability of Features, Classifiers, and Decisions

In most practical applications of biomedical signal analysis, the researcher is presented with the problem of designing a system for pattern classification and decision making using a small number of training samples (signals), with no knowledge of the distributions of the features or parameters computed from the signals. The size of the training set, relative to the number of features used in the pattern classification system, determines the accuracy and reliability of the decisions made [494,495]. One should not increase the number of features to be used without a simultaneous increase in the number of training samples, as the two quantities together affect the bias and variance of the classifier. On the other hand, when the training set has a fixed number of samples, the addition of more features beyond a certain limit will lead to poorer performance of the classifier: this is known as the "curse of dimensionality." It is desirable to be able to analyze the bias and variance of a classification rule while isolating the effects of the functional form of the distributions of the features used.

Raudys and Jain [495] give a rule-of-thumb table for the number of training samples required in relation to the number of features used in order to remain within certain limits of classification errors for five pattern classification methods. When the available features are ordered in terms of their individual classification performance, the optimal number of features to be used with a certain classification method and training set may be determined by obtaining unbiased estimates of the classification accuracy with the number of features increased one at a time in order. A point is reached when the performance deteriorates, which will indicate the optimal number of features to be used. This method, however, cannot take into account the joint performance of various combinations of features: Exhaustive combinations of all features may have to be evaluated to take this aspect into consideration. Software packages such as the Statistical Package for the Social Sciences (SPSS) [478,479] provide programs to facilitate feature evaluation and selection as well as the estimation of classification accuracies.

Durand et al. [331] reported on the design and evaluation of several pattern classification systems for the assessment of bioprosthetic heart valves based upon 18 features computed from PCG spectra (see Section 6.5). Based upon the rule of thumb that the number of training samples should be five or more times the number of features used, and with the number of training samples limited to data from 20 normal and 28 degenerated valves, exhaustive combinations of the 18 features taken $2, 3, 4, 5,$ and 6 at a time were used to design and evaluate pattern classification systems. The Bayes method was seen to provide the best performance (98% correct classification) with six features; as many as 511 combinations of the 18 features taken six at a time provided correct classification between 90% and 98%. The nearest-neighbor algorithm with the Mahalanobis distance provided 94% correct classification with only three features and did not perform any better with more features.

9.10.1 Separability of features

Normalized distance between PDFs: Consider a feature x that has the means m_1 and m_2 and SD values σ_1 and σ_2 for the two classes C_1 and C_2. Assume that the PDFs $p(x|C_1)$ and $p(x|C_2)$ overlap; then, the area of overlap is related to the error of classification. If the SD values σ_1 and σ_2 remain constant, the overlap between the

PDFs decreases as $|m_1 - m_2|$ increases. If the means remain constant, the overlap increases as σ_1 and σ_2 increase (the dispersion of the features increases). These observations are captured by the normalized distance between the means, defined as [496]

$$d_n = \frac{|m_1 - m_2|}{\sigma_1 + \sigma_2}\,. \tag{9.106}$$

The measure d_n provides an indicator of the statistical separability of the PDFs. A limitation of d_n is that $d_n = 0$ if $m_1 = m_2$ regardless of σ_1 and σ_2.

See Section 5.12.3 for a discussion on the KLD.

Divergence: Let us rewrite the likelihood ratio in Equation 9.49 as

$$l_{ij}(\mathbf{x}) = \frac{p(\mathbf{x}|C_i)}{p(\mathbf{x}|C_j)}. \tag{9.107}$$

Applying the logarithm, we get

$$l'_{ij}(\mathbf{x}) = \ln[l_{ij}(\mathbf{x})] = \ln[p(\mathbf{x}|C_i)] - \ln[p(\mathbf{x}|C_j)]. \tag{9.108}$$

The divergence D_{ij} between the PDFs $p(\mathbf{x}|C_i)$ and $p(\mathbf{x}|C_j)$ is defined as [496]

$$D_{ij} = E[l'_{ij}(\mathbf{x})|C_i] + E[l'_{ji}(\mathbf{x})|C_j], \tag{9.109}$$

where

$$E[l'_{ij}(\mathbf{x})|C_i] = \int_{\mathbf{x}} l'_{ij}(\mathbf{x})\; p(\mathbf{x}|C_i)\; d\mathbf{x}. \tag{9.110}$$

Divergence has the following properties [496]:

- $D_{ij} > 0$;
- $D_{ii} = 0$;
- $D_{ij} = D_{ji}$; and
- if the individual features $x_1, x_2, \ldots, x_n$ are statistically independent, then $D_{ij}(x_1, x_2, \ldots, x_n) = \sum_{k=1}^{n} D_{ij}(x_k)$.

It follows that adding more features that are statistically independent of one another will increase divergence and statistical separability.

In the case of multivariate Gaussian PDFs, we have [496]

$$\begin{aligned} D_{ij} &= \frac{1}{2}\, Tr[(\mathbf{C}_i - \mathbf{C}_j)(\mathbf{C}_j^{-1} - \mathbf{C}_i^{-1})] \\ &+ \frac{1}{2}\, Tr[(\mathbf{C}_i^{-1} + \mathbf{C}_j^{-1})(\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T]. \end{aligned} \tag{9.111}$$

The second term in the equation given above is similar to the normalized distance d_n as defined in Equation 9.106 and is zero for PDFs with identical means; however, due to the first term, $D_{ij} \neq 0$ unless the covariance matrices are identical.

In the case of the existence of multiple classes $C_i, i = 1, 2, \ldots, m$, the pairwise divergence values may be averaged to obtain a single measure across all of the m classes as [496]

$$D_{\text{av}} = \sum_{i=1}^{m} \sum_{j=1}^{m} p(C_i)\, p(C_j)\, D_{ij}. \tag{9.112}$$

A limitation of both d_n and D_{ij} is that they increase without an upper bound as the separation between the means increases. On the other hand, the error of classification is limited to the range $[0, 100]\%$ or $[0, 1]$.

Jeffries–Matusita distance: The Jeffries–Matusita (JM) distance provides an improved measure of the separation between PDFs as compared to the normalized distance and divergence. The JM distance between the PDFs $p(\mathbf{x}|C_i)$ and $p(\mathbf{x}|C_j)$ is defined as [496]

$$J_{ij} = \left\{ \int_{\mathbf{x}} \left[\sqrt{p(\mathbf{x}|C_i)} - \sqrt{p(\mathbf{x}|C_j)} \right]^2 d\mathbf{x} \right\}^{1/2}. \tag{9.113}$$

In the case of multivariate Gaussian PDFs, we have [496]

$$J_{ij} = \sqrt{2[1 - \exp(-\alpha)]}, \tag{9.114}$$

where

$$\begin{aligned} \alpha &= \frac{1}{8} (\mathbf{m}_i - \mathbf{m}_j)^T \left(\frac{\mathbf{C}_i + \mathbf{C}_j}{2} \right)^{-1} (\mathbf{m}_i - \mathbf{m}_j) \\ &+ \frac{1}{2} \ln \left[\frac{|(\mathbf{C}_i + \mathbf{C}_j)|/2}{(|\mathbf{C}_i|\, |\mathbf{C}_j|)^{1/2}} \right], \end{aligned} \tag{9.115}$$

where $|\mathbf{C}_i|$ is the determinant of $\mathbf{C}_i$.

An advantage of the JM distance is that it is limited to the range $[0, \sqrt{2}]$. $J_{ij} = 0$ when the means of the PDFs are equal and the covariance matrices are zero matrices. Pairwise JM distances may be averaged over multiple classes similar to the averaging of divergence as in Equation 9.112. The JM distance determines the upper and lower bounds on the error of classification [496].

It should be observed that divergence and JM distance are defined for a given feature vector $\mathbf{x}$. The measures would have to be computed for all combinations of features in order to select the best feature set for a particular pattern classification problem.

See Section 5.12.3 for a description of the KLD between PDFs.

The t-test and the p-value: The t-test is used to assess whether the means of a feature, x, for two groups or classes are statistically different from each other [10, 282, 497, 498]. The test is performed by assessing the difference between the means of the two groups of the features relative to their spread or variability. If it is assumed that the variance, σ^2, is the same for the two groups, the t-statistic, t_s, is computed as

$$t_s = \frac{\mu_1 - \mu_2}{\sigma_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{9.116}$$

where μ_1 and n_1 are the mean and size of the first group, μ_2 and n_2 are the mean and size of the second group, and σ_p is the pooled SD, defined as

$$\sigma_p = \sqrt{\frac{(n_1 - 1)\sigma^2 + (n_2 - 1)\sigma^2}{n_1 + n_2 - 2}} . \tag{9.117}$$

If the variance is not the same for the two groups, t_s is computed as

$$t_s = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} , \tag{9.118}$$

where σ_1^2 is the variance of the first group and σ_2^2 is the variance of the second group. The number of degrees of freedom is equal to $n_1 + n_2 - 2$.

The significance level is usually set at 0.05; this means that, five times out of 100, the difference between the means is concluded to be statistically significant when it is not actually so. This is the probability of being incorrect if the null hypothesis (which assumes that $\mu_1 = \mu_2$) is rejected. Given the degrees of freedom and significance level, the T-critical value, T_c, can be determined using the t-distribution look-up table. When t_s, computed with either Equation 9.118 or Equation 9.116, is larger than T_c, the null hypothesis is rejected.

The t-distribution curve shown in Figure 9.8 illustrates the significance level and p-value in terms of the area under the t-distribution curve. The significance level is the sum of the area under the curve to the right of $+T_c$ and the area to the left of $-T_c$ (all of the areas labeled as A and B in Figure 9.8). The p-value is the sum of the area under the curve to the right of $+t_s$ and the area to the left of $-t_s$ (all of the areas labeled as B in Figure 9.8).

The p-value is a statistical measure of the probability that the results observed in a study could have occurred by chance. A small p-value is a rejection of the null hypothesis in favor of the alternative hypothesis (which assumes that $\mu_1 \neq \mu_2$) because it indicates how unlikely it is that a test statistic as extreme as or more extreme than that given by the data will be observed from the population if the null hypothesis were true. For example, a p-value of 0.01 indicates that there is a one out of a hundred chance that the result occurred by chance. A test resulting in a p-value less than 0.05 is considered to indicate that the difference between the means is statistically significant. A p-value less than 0.01 is taken to indicate that the difference between the means is statistically highly significant.

9.10.2 Feature selection

In a practical application of pattern classification, several parameters may be required to discriminate between multiple classes. Because most features lead to limited discrimination between classes due to the overlap in the ranges of their values for the various classes, it is common to use several features. However, various costs are associated with the measurement, derivation, and computation of each feature. It would be advantageous to be able to assess the contribution of each feature to the

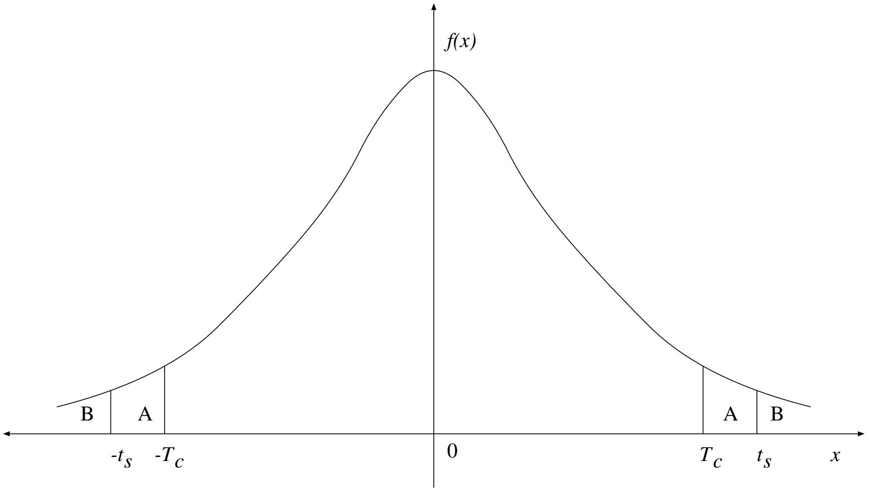

Figure 9.8 The t-distribution curve. Reproduced with permission from T.M. Cabral and R.M. Rangayyan, *Fractal Analysis of Breast Masses in Mammograms*, Morgan & Claypool, 2012. ©Morgan & Claypool.

task of discriminating between the classes of interest, and to select the feature set that provides the best separation between classes and the lowest classification error. The notion of statistical separability of features between classes is useful in addressing these concerns [496]. Various parameters related to the separation of features between classes may be used to select features, such as A_z [338,487,488] and the p-value [498].

Popular methods for feature selection [8,267,293,499] include sequential forward selection, sequential backward selection, and stepwise regression. In sequential forward selection, starting with an empty set, features are sequentially included in the set until the inclusion of additional features does not improve the classification performance. To begin, the best single feature in terms of separation of the classes of interest, such as the p-value, is determined and included in the set of selected features. The selected feature is then used in combination with each remaining feature, in turn, to create pairs. The classification performance of each pair is evaluated and the best pair is selected for the next iteration. The process of including more and more features and evaluation of performance is continued until a prespecified number of features is selected or the classification performance does not improve much with the inclusion of more features.

Sequential backward selection starts with the full set of features and eliminates features until the removal of more features deteriorates the classification performance. To begin with, the feature whose elimination leads to the largest separation of classes is identified and removed. The process is continued by removing one feature at a time until a prespecified number of features remain or any further removal of features causes deterioration in the classification performance.

Stepwise regression [293, 500] may be seen as a variant of forward selection. At each stage, a new feature is added and the obtained model is checked for any possible elimination of the selected features without substantially increasing an error measure. The process is continued until the error is minimized or the improvement achieved is below a limit. The method starts with an initial model and then evaluates the discriminating power of incrementally larger and smaller models at each iteration [293, 500]. At each iteration, the p-values of an F-statistic are computed to evaluate the models with and without a potential feature. The feature is included if sufficient evidence is found to support its inclusion; otherwise, it is removed.

Advanced methods for feature selection include genetic algorithms and genetic programming [274, 501]. Instead of selecting a subset of a given set of features, one could also take the approach of reducing the dimension of the given feature vector via PCA (see Section 8.8.1). PCA helps in removing redundancy in the given features. It should be noted that the result of PCA does not include any of the given features directly but provides transformed values derived from the given feature vector. The transform itself is derived by using statistical measures of a population of feature vectors.

In general, it is expected that selecting a reduced set of features as above will lead to higher classification accuracy than the case with the full set of features. A smaller set of features will also facilitate the design and implementation of more efficient classifiers at lower cost than a larger set. See Banik et al. [293, 502], Nandi et al. [501], and Mu et al. [274, 503] for additional discussions on feature selection and examples of application.

9.10.3 The training and test steps

In the situation when the number of available sample vectors with known classification is limited, questions arise as to how many of the samples may be used to design or train a classifier, with the understanding that the classifier so designed needs to be tested using an independent set of samples of known classification as well. When a sufficiently large number of samples are available, they may be randomly split into two approximately equal sets, one for use as the training set and the other to be used as the test set. The random splitting procedure may be repeated a number of times to generate several classifiers. Finally, one of the classifiers so designed may be selected based upon its performance in both the training and test steps.

When the available number of labeled samples is small, the bootstrap method [504, 505] may be used. In this procedure, a training set is created by drawing at random a large number of samples from the available pool of labeled samples with replacement; that is, a given sample may be drawn and included a number of times in the training set. Similarly, a test set is created by random drawing of samples with replacement. The resampling procedure may be repeated a number of times to obtain multiple estimates of various measures of classification performance.

In the procedure known as k-fold cross-validation, the original labeled data set is partitioned randomly into k subsets of equal size with no overlap. For example, in 10-fold cross-validation with $k = 10$, the given set of samples is randomly divided into 10 subsets with no overlap. Out of the k subsets, one subset is set aside to test the classifier being designed, and the remaining $(k - 1)$ subsets are used together as the training data. The training and testing process, also known as cross-validation, is repeated k times, with each of the k subsets used only once as the test data. The k results from the k-fold cross-validation process may be averaged to obtain an overall measure of performance of the features used and classification method chosen. In k-fold cross-validation, all data samples are used for both training and testing, and each sample is used only once for testing; this cannot be guaranteed in the random resampling procedure described in the preceding paragraphs.

The leave-one-out method: The LOO method [268] is suitable for the estimation of the classification accuracy of a pattern classification technique, particularly when the number of available samples is small. In this method, one of the available samples is excluded, the classifier is designed with the remaining samples, and then the classifier is applied to the excluded sample. The validity of the classification so performed is noted. This procedure is repeated with each available sample: if N training samples are available, N classifiers are designed and tested. The training and test sets for any one classifier so designed and tested are independent. However, whereas the training set for each classifier has $N - 1$ samples, the test set has only one sample. This procedure is equivalent to k-fold cross-validation with $k = N$. In the final analysis, every sample will have served $(N - 1)$ times as a training sample but only once as a test sample. An average classification accuracy is then computed using all of the test results.

Let us consider a simple case in which the covariances of the sample sets of two classes are equal. Assume that two sample sets, $S_1 = \{\mathbf{x}_1^{(1)}, \mathbf{x}_2^{(1)}, \ldots, \mathbf{x}_{N_1}^{(1)}\}$ from class C_1 and $S_2 = \{\mathbf{x}_1^{(2)}, \mathbf{x}_2^{(2)}, \ldots, \mathbf{x}_{N_2}^{(2)}\}$ from class C_2, are given. Here, N_1 and N_2 are the numbers of samples in the sets S_1 and S_2, respectively. Assume also that the prior probabilities of the two classes are equal to each other. Then, according to the Bayes classifier and assuming $\mathbf{x}$ to be governed by a multivariate Gaussian PDF, a sample $\mathbf{x}$ is assigned to class C_1 if

$$(\mathbf{x} - \mathbf{m}_1)^T(\mathbf{x} - \mathbf{m}_1) - (\mathbf{x} - \mathbf{m}_2)^T(\mathbf{x} - \mathbf{m}_2) > \theta, \tag{9.119}$$

where θ is a threshold, and the sample mean $\tilde{\mathbf{m}}_i$ is given by

$$\tilde{\mathbf{m}}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbf{x}_j^{(i)}. \tag{9.120}$$

In the LOO method, one sample $\mathbf{x}_k^{(i)}$ is excluded from the training set and then used as the test sample. The mean estimate for class C_i without $\mathbf{x}_k^{(i)}$, labeled as $\tilde{\mathbf{m}}_{ik}$, may be computed as

$$\tilde{\mathbf{m}}_{ik} = \frac{1}{N_i - 1} \left[\sum_{j=1}^{N_i} \mathbf{x}_j^{(i)} - \mathbf{x}_k^{(i)} \right], \tag{9.121}$$

which leads to

$$\mathbf{x}_k^{(i)} - \tilde{\mathbf{m}}_{ik} = \frac{N_i}{N_i - 1}(\mathbf{x}_k^{(i)} - \tilde{\mathbf{m}}_i). \tag{9.122}$$

Then, testing a sample $\mathbf{x}_k^{(1)}$ from C_1 can be carried out as

$$\begin{aligned} & (\mathbf{x}_k^{(1)} - \tilde{\mathbf{m}}_{1k})^T(\mathbf{x}_k^{(1)} - \tilde{\mathbf{m}}_{1k}) - (\mathbf{x}_k^{(1)} - \tilde{\mathbf{m}}_2)^T(\mathbf{x}_k^{(1)} - \tilde{\mathbf{m}}_2) \qquad (9.123) \\ = & \left(\frac{N_1}{N_1 - 1}\right)^2 (\mathbf{x}_k^{(1)} - \tilde{\mathbf{m}}_1)^T(\mathbf{x}_k^{(1)} - \tilde{\mathbf{m}}_1) - (\mathbf{x}_k^{(1)} - \tilde{\mathbf{m}}_2)^T(\mathbf{x}_k^{(1)} - \tilde{\mathbf{m}}_2) > \theta. \end{aligned}$$

Note that when $\mathbf{x}_k^{(1)}$ is tested, only $\tilde{\mathbf{m}}_1$ is changed and $\tilde{\mathbf{m}}_2$ is not changed. Likewise, when a sample $\mathbf{x}_k^{(2)}$ from C_2 is tested, the decision rule is

$$\begin{aligned} & (\mathbf{x}_k^{(2)} - \tilde{\mathbf{m}}_1)^T(\mathbf{x}_k^{(2)} - \tilde{\mathbf{m}}_1) - (\mathbf{x}_k^{(2)} - \tilde{\mathbf{m}}_{2k})^T(\mathbf{x}_k^{(2)} - \tilde{\mathbf{m}}_{2k}) \qquad (9.124) \\ = & (\mathbf{x}_k^{(2)} - \tilde{\mathbf{m}}_1)^T(\mathbf{x}_k^{(2)} - \tilde{\mathbf{m}}_1) - \left(\frac{N_2}{N_2 - 1}\right)^2 (\mathbf{x}_k^{(2)} - \tilde{\mathbf{m}}_2)^T(\mathbf{x}_k^{(2)} - \tilde{\mathbf{m}}_2) < \theta. \end{aligned}$$

The LOO method provides the least biased (practically unbiased) estimate of the classification accuracy of a given classification method for a given training set; it is useful when the number of samples available with known classification is small. The LOO approach may be applied on a sample (signal or feature vector) basis or on a patient basis if multiple signals are included in the data set for each patient.

9.11 Application: Normal versus Ectopic ECG Beats

We have seen the distinctions between normal and ectopic (PVC) beats in the ECG in several different contexts (see Sections 1.2.5, 5.4.2, 5.7, and 9.2.2, as well as Figures 5.1 and 5.11). We shall now see how we can put together several of the topics we have studied so far for the purpose of detecting PVCs in an ECG signal.

9.11.1 Classification with a linear discriminant function

Training step: Figure 9.9 shows the ECG signal of a patient with several ectopic beats, including episodes of bigeminy (alternating normal beats and PVCs). The beats in the portion of the signal in Figure 9.9 were manually labeled as normals ('$\circ$' marks) or PVCs ('$\times$' marks) and were used to train a pattern classification system. The training set includes 121 normal beats and 39 PVCs.

The following procedure was applied to the signal to detect each beat, compute features, and develop a pattern classification rule:

1. The signal was filtered with a Butterworth lowpass filter of order 8 and cutoff frequency $70\ Hz$ to remove noise (see Section 3.6.1); the sampling rate is $200\ Hz$.

2. The Pan–Tompkins algorithm was applied to detect each beat (see Section 4.3.2).

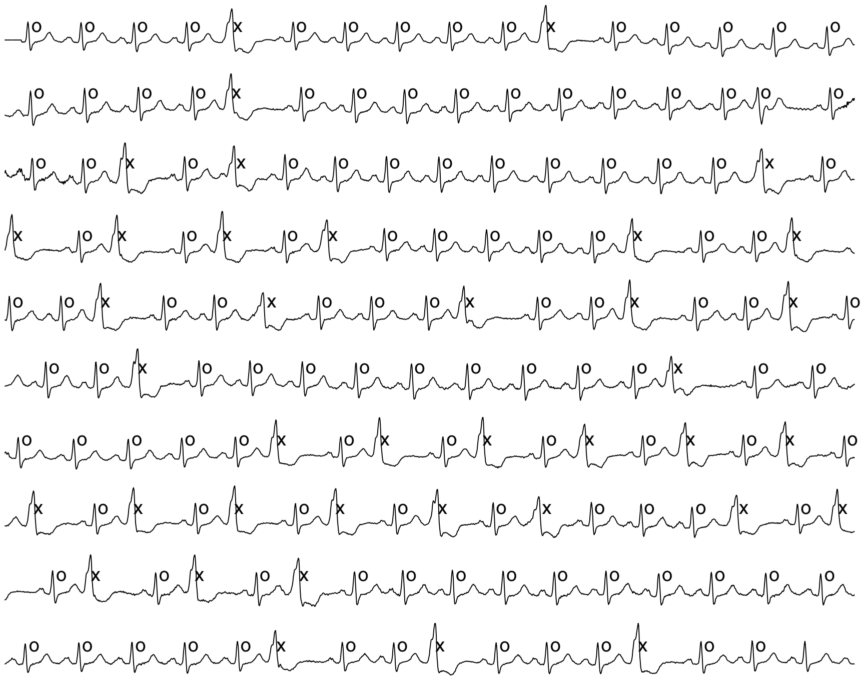

Figure 9.9 The ECG signal of a patient (male, 65 years) with PVCs (training set). Each strip is of duration 10 s; the signal continues from top to bottom. The second half of the seventh strip and the first half of the eighth strip illustrate an episode of bigeminy. Each beat was manually labeled as normal ('$\circ$') or PVC ('$\times$'). The last beat was not processed.

3. The QRS – T portion of each beat was segmented by selecting the interval from the sample 160 ms before the peak of the Pan–Tompkins output to the sample 240 ms after the peak (see Figure 5.11).

4. The RR interval and form factor FF were computed for each beat (see Sections 5.6.4 and 5.7 and Figure 5.11). Figure 9.10 illustrates the feature vector plot for the training set.

5. The prototype (mean) feature vectors were computed for the normal and PVC groups in the training set. The prototype vectors are $[RR, FF] = [0.66, 1.58]$ and $[RR, FF] = [0.45, 2.74]$ for the normal and PVC classes, respectively.

6. The equations of the straight line joining the two prototype vectors and its normal bisector were determined; the latter is a linear decision function (see Section 9.4.1 and Figure 9.1). Figure 9.10 illustrates the two lines.

7. The equation of the linear decision function was obtained as $RR - 5.56FF + 11.44 = 0$. The decision rule was derived as

$$\text{if } \; RR - 5.56FF + 11.44 \begin{cases} > 0 & \text{normal beat,} \\ \leq 0 & \text{PVC.} \end{cases} \tag{9.125}$$

All of the beats in the training set were correctly classified by the decision rule in Equation 9.125.

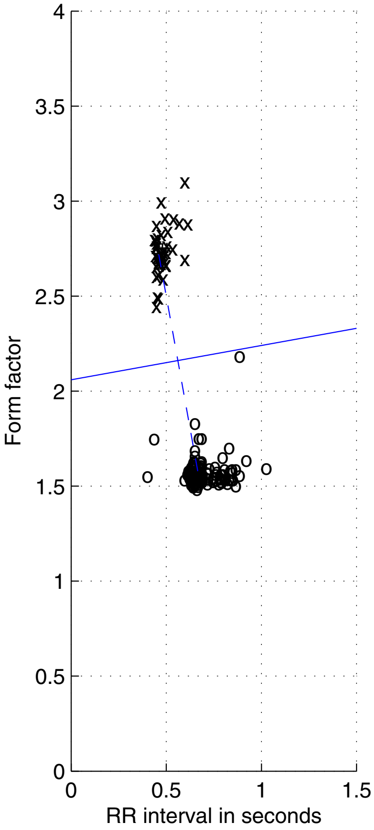

Figure 9.10 The $[RR, FF]$ feature-vector space corresponding to the ECG in Figure 9.9 (training set). Normal: '○', PVC: '×'. The straight line joining the two prototype vectors (dashed) and its normal bisector (solid) are also shown; the latter is a linear decision function.

Observe from Figure 9.10 that a simple threshold on FF alone can effectively separate the PVCs from the normals in the training set. A viable classification rule to detect PVCs may also be stated in a manner similar to that in Section 9.2.2. The example given here is intended to serve as a simple illustration of the design of a 2D linear decision function.

Test step: Figure 9.11 illustrates an ECG segment immediately following that in Figure 9.9. The same procedure as described above was applied to detect the beats in the signal in Figure 9.11 and to compute their features, which were used as the test set. The decision rule in Equation 9.125 was applied to the feature vectors and the beats in the signal were automatically classified as normal or PVC. Figure 9.12 illustrates the feature-vector space of the beats in the test set, along with the decision boundary given by Equation 9.125. Figure 9.11 shows the automatically applied labels for each beat: All of the 37 PVCs were correctly classified, and only one of the 120 normal beats was misclassified as a PVC (that is, there was one FP).

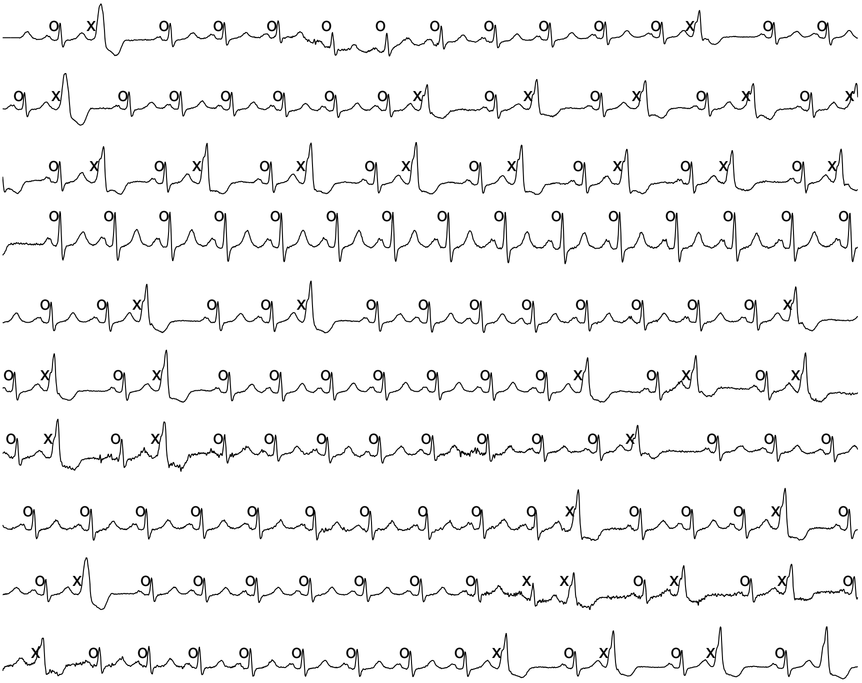
o x o o o o o o o o o o x o o
o x o o o o o o x o x o x o x o x
o x o x o x o x o x o x o x o x
o o o o o o o o o o o o o o o
o o x o o x o o o o o o o o x
o x o x o o o o o o o x o x o x
o x o x o o o o o o o o x o o o
o o o o o o o o o o x o o o x o
o x o o o o o o o x x o x o x o
x o o o o o o o o x o x o x o

Figure 9.11 The ECG signal of a patient with PVCs (test set); this portion immediately follows that in Figure 9.9. Each strip is of duration 10 s; the signal continues from top to bottom. Each beat was automatically labeled as normal ('o') or PVC ('×') by the decision rule stated in Equation 9.125. The 10^{th} beat in the 9^{th} strip with $[RR, FF] = [0.66, 2.42]$ was misclassified. The last beat was not processed.

It should be observed that a PVC has, by definition, an RR interval that is less than that for a normal beat (at the same heart rate). However, the heart rate of a subject will vary over time, and the reference RR interval to determine the prematurity of PVCs needs to be updated periodically. A decision rule as in Equation 9.125 cannot be applied on a continuing basis even to the same subject. Note that the proposed method can be extended for the identification of sinus beats (originating from

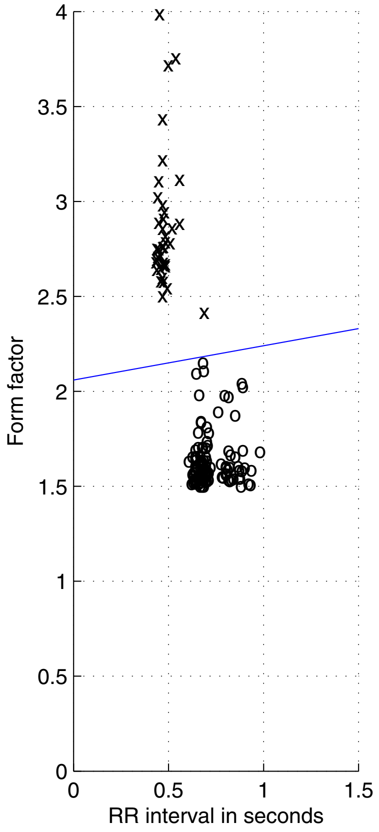

Figure 9.12 $[RR, FF]$ feature-vector space corresponding to the ECG in Figure 9.11 (test set). Normal: '$\circ$', PVC: '$\times$'. The straight line is the linear decision function given in Equation 9.125. The '$\times$' mark closest to the decision boundary with $[RR, FF] = [0.66, 2.42]$ corresponds to an FP classification.

the SA node) that meet the prematurity condition due to sinus arrhythmia but are, nevertheless, normal in waveshape.

The FF values will depend upon the waveshape of each ECG beat, which will vary from one ECG lead to another. Therefore, the same decision rule based upon waveshape cannot be applied to all ECG leads of even the same subject. Furthermore, a given subject may have PVCs originating from various ectopic foci resulting in widely different waveshapes even in the same ECG lead. A shape factor to be used for pattern classification must be capable of maintaining different values for (a) PVCs of various waveshapes as one group and (b) normal beats as another group.

The preceding illustration is intended to serve as a simple example of the design of a pattern classification system; in practice, more complex decision rules based upon more than two features are required. Furthermore, it should be observed that a pattern classification procedure as described above provides beat-by-beat labeling; the overall diagnosis of the patient's condition requires many other items of clinical information and the expertise of a cardiologist.

9.11.2 Application of the Bayes classifier

In another classification experiment with the same ECG signal as in the preceding section, the waveform for each beat was segmented using the Pan–Tompkins algorithm. Each signal was normalized by subtracting its mean and dividing by the maximum value of the result. The area under each segmented and normalized QRS–T wave, referred to as $QRSTA$, was computed as follows (see Section 5.4.3). The baseline value of each beat was obtained as the average of its starting and ending sample values. The baseline was subtracted and the signal was rectified, that is, its absolute value was obtained. $QRSTA$ was computed as the sum of all of the rectified values multiplied by the sampling interval. The feature vector $[QRSTA, FF]^T$ was computed for each beat.

The mean and SD of each feature were computed using the samples in the training set with 123 normal beats and 39 PVCs. Each feature value was normalized by dividing by its SD. Using the training set only, the 1D conditional and posterior probability functions were estimated. Figures 9.13 and 9.14 show the estimated functions using equal prior probabilities for the normal and PVC classes. In a real ECG signal, the number of PVCs could be expected to be far lower than the number of normal beats. Figures 9.15 and 9.16 show the estimated functions using prior probabilities of 0.999 and 0.001 for the normal and PVC classes, respectively. The four figures listed above show how the features vary in their probabilities across the two classes and also how the assumed prior probabilities can affect the posterior probabilities derived.

The 2D scatter of the feature vectors $[QRSTA, FF]^T$ for the training set is shown in Figure 9.17. Assuming the prior probabilities for the two classes to be equal, the Bayesian classifier was derived (see Section 9.6.1). Ellipses are shown for each cluster (normal or PVC) to show the boundaries of the 2D Gaussian functions estimated at $\sigma, 2\sigma,$ and 3σ in Figure 9.17; the black contour shown is the Bayesian decision boundary, which misclassifies only one normal beat as a PVC.

Figure 9.18 demonstrates the application of the Bayesian decision function shown in Figure 9.17 to the test set of ECG signals with 183 normal beats and 53 PVCs. It is seen that the classifier correctly recognizes all PVCs but fails to identify a small number of normal beats. However, inspection of the two scatter plots indicates that the samples are linearly separable; a linear classifier without the need to assume prior probabilities could possibly lead to better results.

9.11.3 Classification using the K means method

The K-means clustering method (see Section 9.5.1) was also applied to the ECG signal described in the preceding sections. Figures 9.19 to 9.22 show the evolution

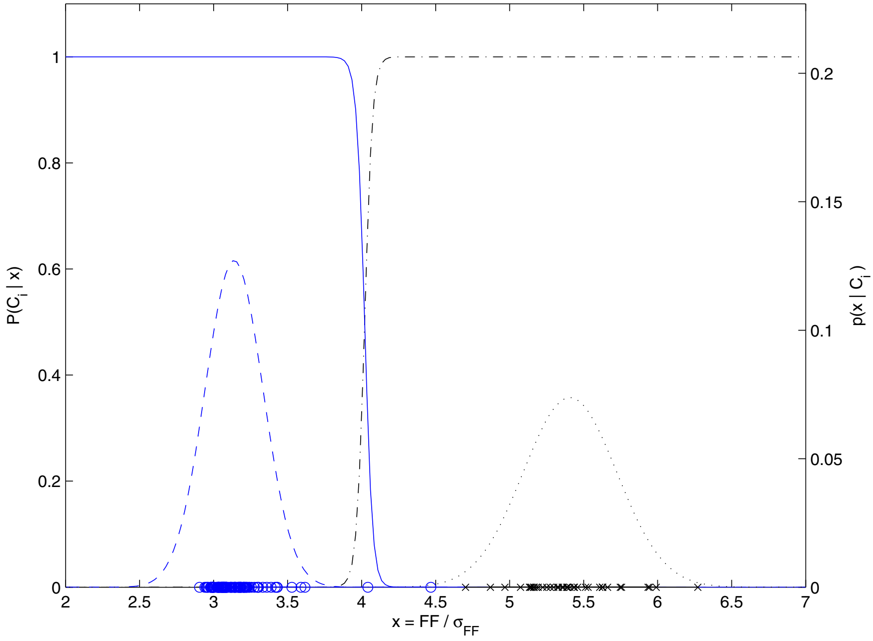

Figure 9.13 Conditional and posterior probability functions estimated for the feature FF for the training set of normal ECG signals ('○') and PVCs ('×'). The prior probabilities using the two classes were assumed to be equal. See also Figure 9.15. Figure courtesy of Fábio José Ayres, University of Calgary, Calgary, Alberta, Canada.

of class means and the separating boundary over four iterations. It is seen that, after the fourth and final iteration, all of the samples are correctly classified.

9.12 Application: Detection of Knee joint Cartilage Pathology

Moussavi et al. [94], Krishnan et al. [95], and Rangayyan et al. [96] proposed a series of adaptive segmentation, modeling, and pattern classification techniques for the detection of knee-joint cartilage pathology using VAG signals (see Sections 1.2.14, 5.12, 6.6, 8.2.3, 8.6, and 8.14). In consideration of the fact that VAG signals are nonstationary, each VAG signal was first divided into locally stationary segments using the RLS or the RLSL algorithm (see Sections 8.6.1 and 8.6.2). Each segment was considered as a separate signal and modeled by the forward–backward linear prediction or the Burg-lattice method (see Section 8.6.2). The model coefficients or poles were used as parameters for pattern classification.

A striking difference that may be appreciated visually and aurally between normal and abnormal VAG signals is that abnormal signals are much more variable in amplitude across a swing cycle than normal signals. However, this difference is

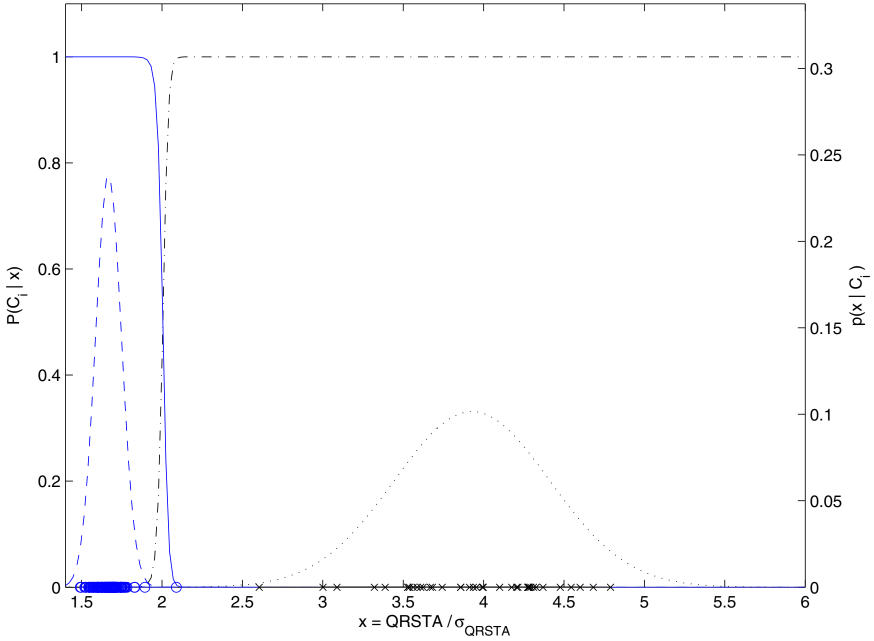

Figure 9.14 Conditional and posterior probability functions estimated for the feature $QRSTA$ using the training set of normal ECG signals ('○') and PVCs ('×'). The prior probabilities for the two classes were assumed to be equal. See also Figure 9.16. Figure courtesy of Fábio José Ayres, University of Calgary, Calgary, Alberta, Canada.

lost in the process of dividing the signals into segments and considering each segment as a separate signal. To overcome this problem, the means (time averages) of the segments of each subject's signal were computed, and then the variance of the means was computed across the various segments of the same signal. The variance of the means represents the above-mentioned difference, and it was used as one of the discriminant features. (The MS or RMS values of VAG segments may be more suitable for this purpose than their mean values.)

In addition to quantitative parameters derived from VAG signal analysis, clinical parameters (to be described in the following paragraphs) related to the subjects were also investigated for possible discriminant capabilities. At the outset, as shown in Figure 9.23, knee joints of the subjects in the study were categorized into two groups: normal and abnormal. The normal group was divided into two subgroups: normal-silent and normal-noisy. If no sound was heard during auscultation, a normal knee was considered to be normal-silent; otherwise, it was considered to be normal-noisy. All knees in the abnormal group used were examined by arthroscopy (see Section 8.2.3 and Figure 8.2) and divided into two groups: arthroscopically normal and arthroscopically abnormal.

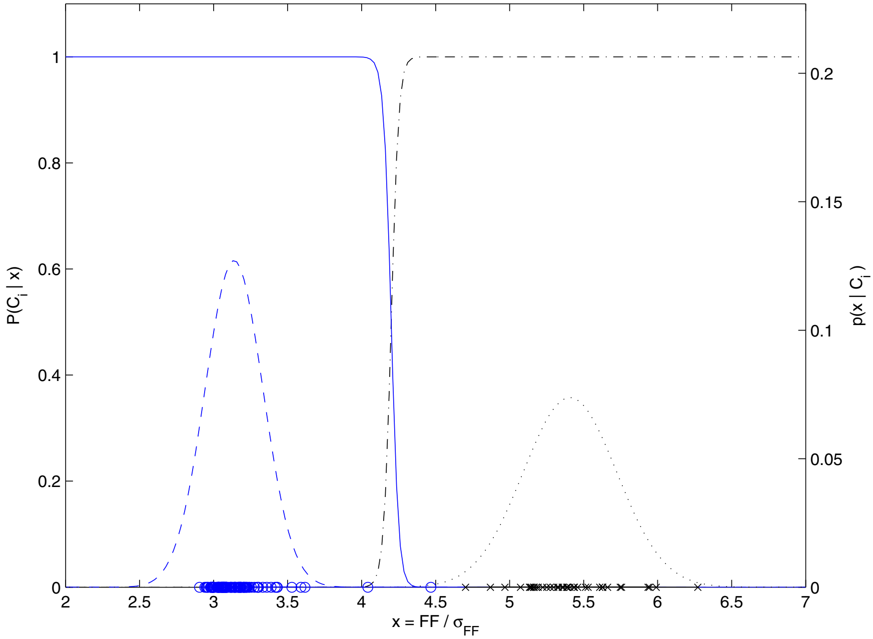

Figure 9.15 Conditional and posterior probability functions estimated for the feature FF using the training set of normal ECG signals ('∘') and PVCs ('×'). The prior probabilities of 0.999 and 0.001 were assumed for the normal and PVC classes, respectively. See also Figure 9.13. Figure courtesy of Fábio José Ayres, University of Calgary, Calgary, Alberta, Canada.

Labeling of VAG signal segments was achieved by comparing the auscultation and arthroscopy results of each patient with the corresponding segmented VAG and joint angle signals. Localization of the pathology was performed during arthroscopy and the joint angle ranges where the affected areas could come into contact with other joint surfaces were estimated. These results were then compared with the auscultation reports to determine whether the joint angle(s) at which pathology existed corresponded to the joint angle(s) at which sound was heard. For example, if it was found from the arthroscopy report of a patient that the abnormal parts of the patient's knee could cause contact in the range $30^\circ - 90^\circ$, VAG signal segments of the subject corresponding to the angle range of $30^\circ - 90^\circ$ were labeled as arthroscopically abnormal; the rest of the segments of the signal were labeled as arthroscopically normal.

Categorization into four groups as above was done based upon the presumptions that normal-noisy and arthroscopically abnormal signals might be distinguishable in their characteristics, and that normal-silent and arthroscopically normal knees would also be distinguishable. The possibilities of arthroscopically normal knees being associated with sounds, normal-noisy knees not having any associated pathology, and

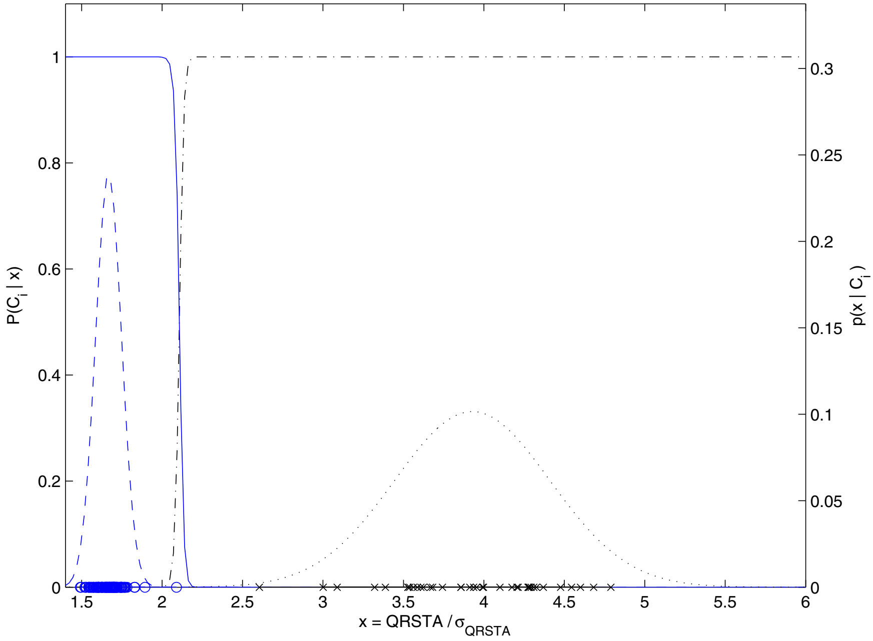

Figure 9.16 Conditional and posterior probability functions estimated for the feature $QRSTA$ using the training set of normal ECG signals ('○') and PVCs ('×'). The prior probabilities of 0.999 and 0.001 were assumed for the normal and PVC classes, respectively. See also Figure 9.14. Figure courtesy of Fábio José Ayres, University of Calgary, Calgary, Alberta, Canada.

normal-silent knees having undetermined pathologies were also admitted. Krishnan et al. [95] further subdivided the arthroscopically normal and arthroscopically abnormal categories into silent and noisy categories, thereby having a total of six categories; this is not shown in Figure 9.23.

Based on clinical reports and auscultation of knee joints, the following clinical parameters were chosen as features (in addition to AR-model parameters) for classification:

Sound: The sound heard by auscultation during flexion and extension movement of the knee joint was coded as

0– silent,

1– click,

2– pop,

3– grinding, or

4– a mixture of the above-mentioned sounds or other sounds.

Each segment of the VAG signals was labeled with one of the above codes.

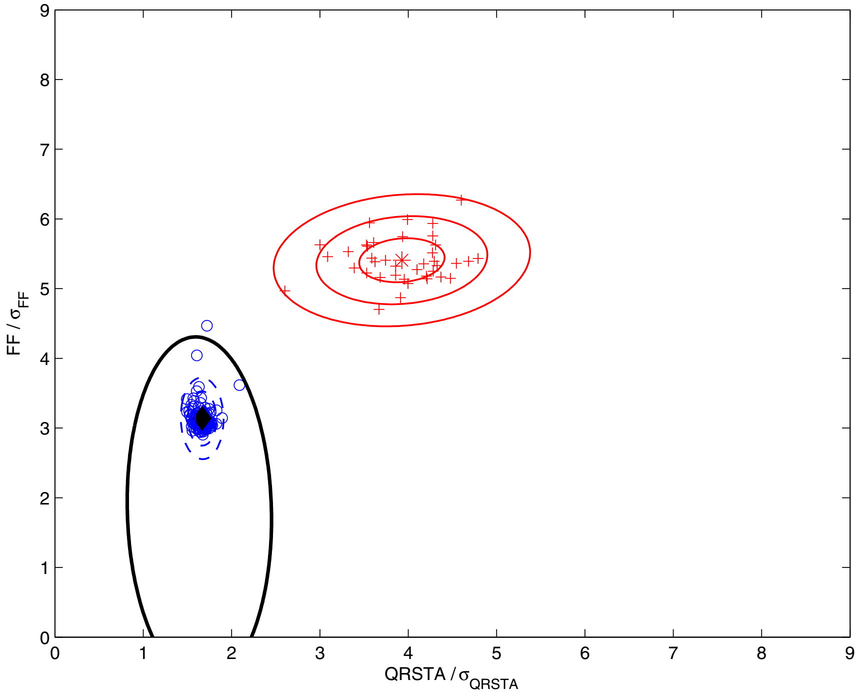

Figure 9.17 2D scatter plot of the feature vectors $[QRSTA, FF]^T$ for the training set of ECG signals. The circles represent normal beats and the plus marks represent PVCs; the diamond and star symbols represent the corresponding average or prototype vectors. The ellipses show the boundaries of the 2D Gaussian functions estimated at σ, 2σ, and 3σ for each class. The partial elliptical contour is the Bayesian decision boundary. See also Figure 9.18. Figure courtesy of Fábio José Ayres, University of Calgary, Calgary, Alberta, Canada.

Activity level: The activity level of each subject was coded as

1– exercising once per week or less,

2– exercising two or three times per week, or

3– exercising more than three times per week.

Age: The age of the subject in years.

Gender: The gender of the subject, which was coded as

0– female, or

1– male.

Among the parameters mentioned above, gender may not be a discriminant parameter; however, it is customary to record gender in clinical analysis. Note that

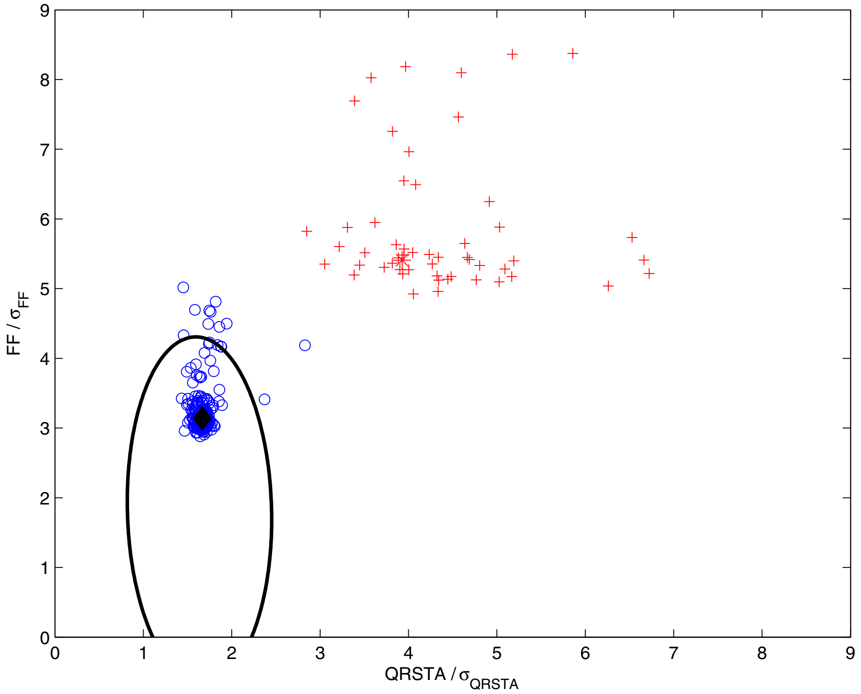

Figure 9.18 2D scatter plot of the feature vectors $[QRSTA, FF]^T$ for the test set of ECG signals. The circles represent normal beats and the plus marks represent PVCs; the diamond and star symbols represent the corresponding average or prototype vectors. The partial elliptical contour is the Bayesian decision boundary derived from the training set shown in Figure 9.17. Figure courtesy of Fábio José Ayres, University of Calgary, Calgary, Alberta, Canada.

among the four parameters listed above, only the first one can vary between the different segments of a given subject's VAG signal.

Moussavi et al. [94] compared the performance of various sets of features in the classification of VAG signals into two groups and four groups (see Figure 9.23) with random selections of cases. Using a set of 540 segments obtained from 20 normal subjects and 16 subjects with cartilage pathology, different numbers of segments were randomly selected for use in the training step of designing a discriminant function, and finally the selection which provided the best result was chosen for the final classification system. Two-group classification accuracies in the range $77 - 91\%$ and four-group classification accuracies in the range $65 - 88\%$ were obtained.

By combining the steps of classification into two groups and four groups, a two-step method was proposed by Moussavi et al. [94]; a block diagram of this method is illustrated in Figure 9.24. The algorithm first uses training sets to design classifiers for two and four groups. The resulting discriminant functions are used as Classifier 1 (two groups) and Classifier 2 (four groups), respectively. An unknown signal, which

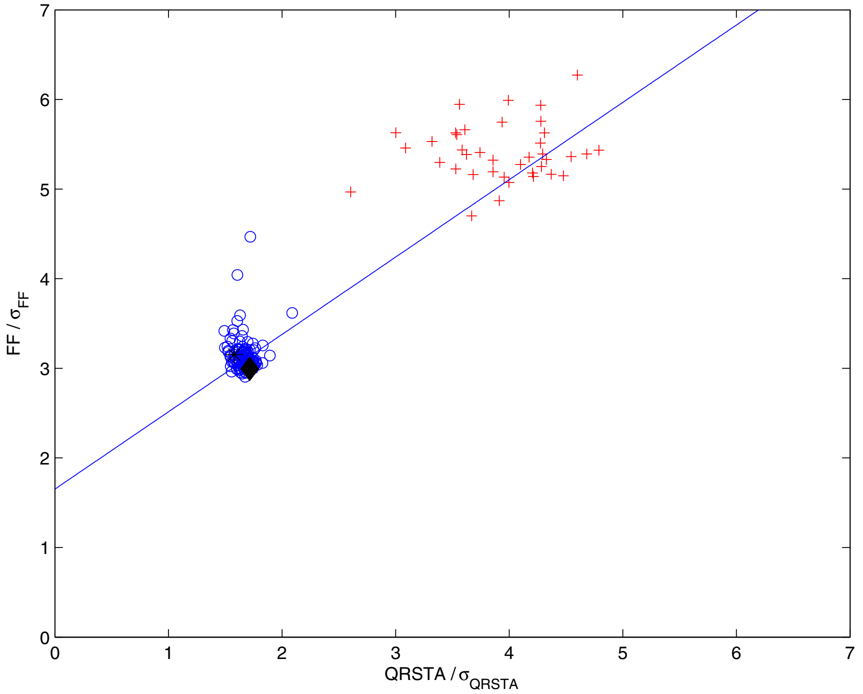

Figure 9.19 2D scatter plot of the feature vectors $[QRSTA, FF]^T$ for the training set of ECG signals with the results of the K-means method after one iteration. The circles represent normal beats and the plus marks represent PVCs; the diamond and star symbols represent the corresponding mean vectors. The straight line is the separating boundary. Figure courtesy of Fábio José Ayres, University of Calgary, Calgary, Alberta, Canada.

has been adaptively divided into segments, enters Classifier 1. If segments spanning more than 90% of the duration of the signal are classified as being normal, the signal (or subject) is considered to be normal. On the other hand, if more than 90% of the duration of the signal is classified as being abnormal, the signal (or subject) is considered to be abnormal. If more than 10% but less than 90% of the signal duration is classified as abnormal, the signal goes to Classifier 2, which classifies the signal into four groups (see Figure 9.23). In the second step, if more than 10% of the duration of the signal is classified as being arthroscopically abnormal, the signal is considered to be abnormal; otherwise it is considered to be normal. At this stage, information on the numbers of segments belonging to the four categories shown in Figure 9.23 is available, but the final decision is on the normality of the whole signal (subject or knee joint).

The two-step diagnosis method was trained with 262 segments obtained from 10 normal subjects and eight subjects with cartilage pathology, and it was tested with 278 segments obtained from a different set of 10 normal subjects and eight subjects

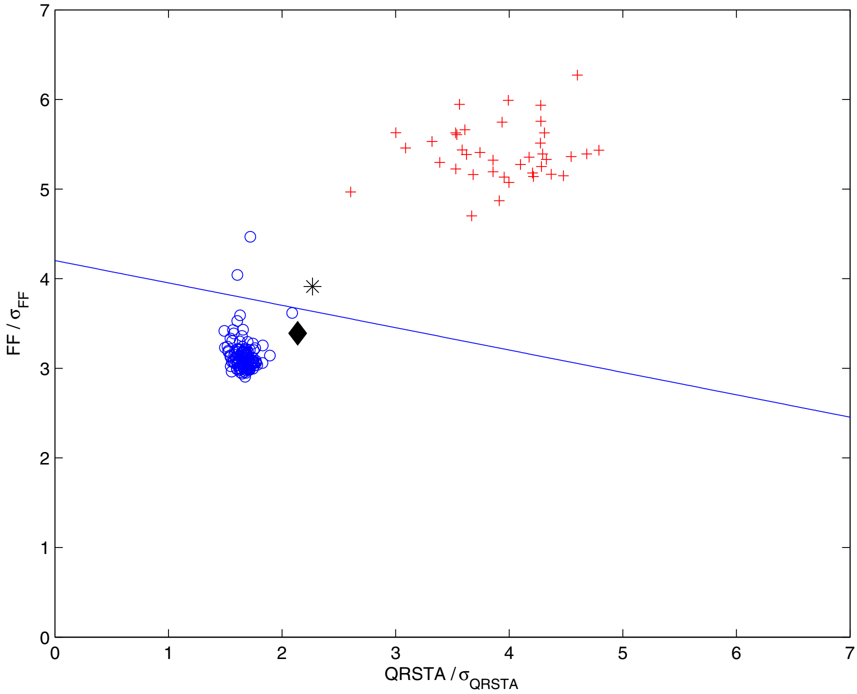

Figure 9.20 2D scatter plot of the feature vectors $[QRSTA, FF]^T$ for the training set of ECG signals with the results of the K-means method after two iterations. The circles represent normal beats and the plus marks represent PVCs; the diamond and star symbols represent the corresponding mean vectors. The straight line is the separating boundary. Figure courtesy of Fábio José Ayres, University of Calgary, Calgary, Alberta, Canada.

with cartilage pathology but without any restriction on the kind of abnormality. Except for one normal signal which was indicated as being abnormal over 12% of its duration, all of the signals were correctly classified. The results also showed that all of the abnormal signals including signals associated with chondromalacia grades I to IV (see Section 8.2.3 and Figure 8.2) were classified correctly. Based upon this result, it was indicated that the method has the ability to detect chondromalacia patella at its early stages as well as advanced stages. Krishnan et al. [95] and Rangayyan et al. [96] reported on further work along these directions.

9.13 Remarks

The subject of pattern classification is a vast area by itself. The topics presented in this chapter provide a brief introduction to the subject with several illustrations of application to biomedical signals.

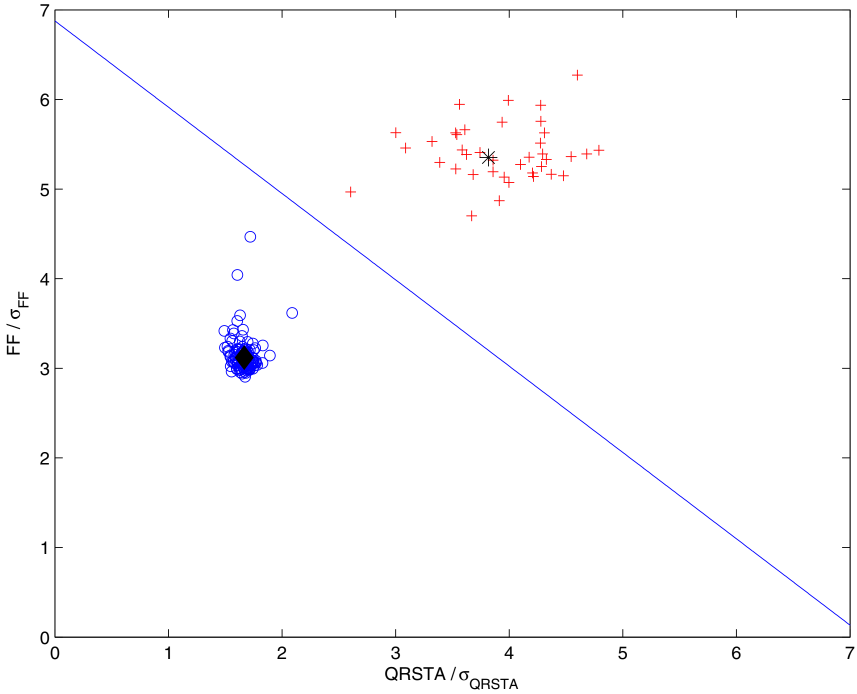

Figure 9.21 2D scatter plot of the feature vectors $[QRSTA, FF]^T$ for the training set of ECG signals with the results of the K-means method after three iterations. The circles represent normal beats and the plus marks represent PVCs; the diamond and star symbols represent the corresponding mean vectors. The straight line is the separating boundary. Figure courtesy of Fábio José Ayres, University of Calgary, Calgary, Alberta, Canada.

We have now seen how biomedical signals may be processed and analyzed to extract quantitative features that may be used to classify the signals as well as to design diagnostic decision functions. Practical development of such techniques is usually hampered by a number of limitations related to the extent of discriminant information present in the signals selected for analysis, as well as the limitations of the features designed and computed. Artifacts inherent in the signal or caused by the signal acquisition systems impose further limitations.

A pattern classification system that is designed with limited data and information about the chosen signals and features will provide results that should be interpreted with due care. Above all, it should be borne in mind that the final diagnostic decision requires far more information than that provided by biomedical signals and their analysis: this aspect is best left to the physician or health-care specialist in the spirit of computer-*aided* diagnosis.

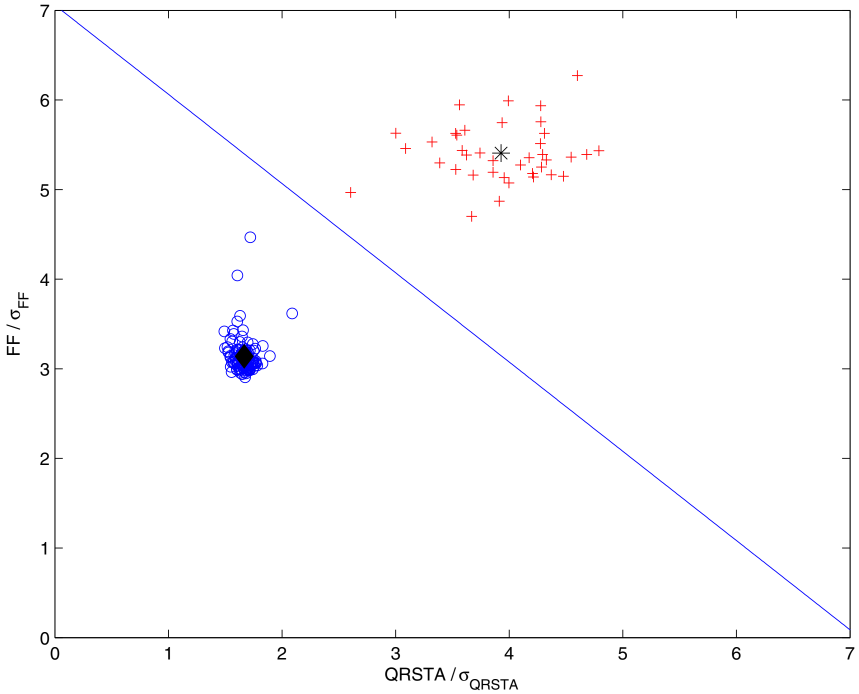

Figure 9.22 2D scatter plot of the feature vectors $[QRSTA, FF]^T$ for the training set of ECG signals with the results of the K-means method after the fourth and final iteration. The circles represent normal beats and the plus marks represent PVCs; the diamond and star symbols represent the corresponding mean vectors. The straight line is the separating boundary. Figure courtesy of Fábio José Ayres, University of Calgary, Calgary, Alberta, Canada.

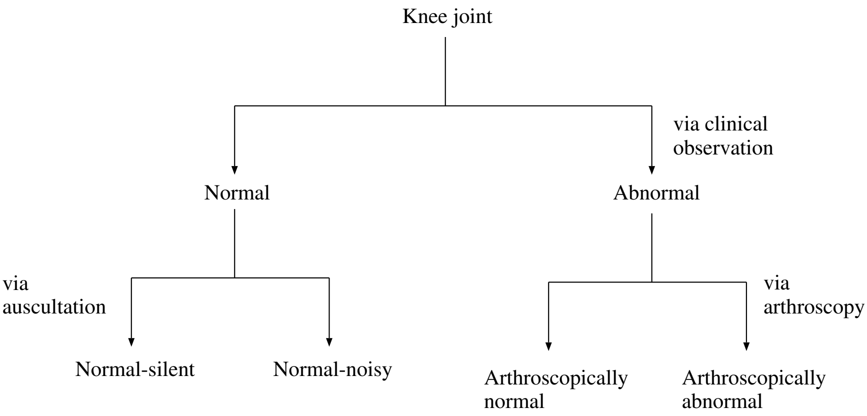

Figure 9.23 Categorization of knee joints based upon auscultation and arthroscopy.

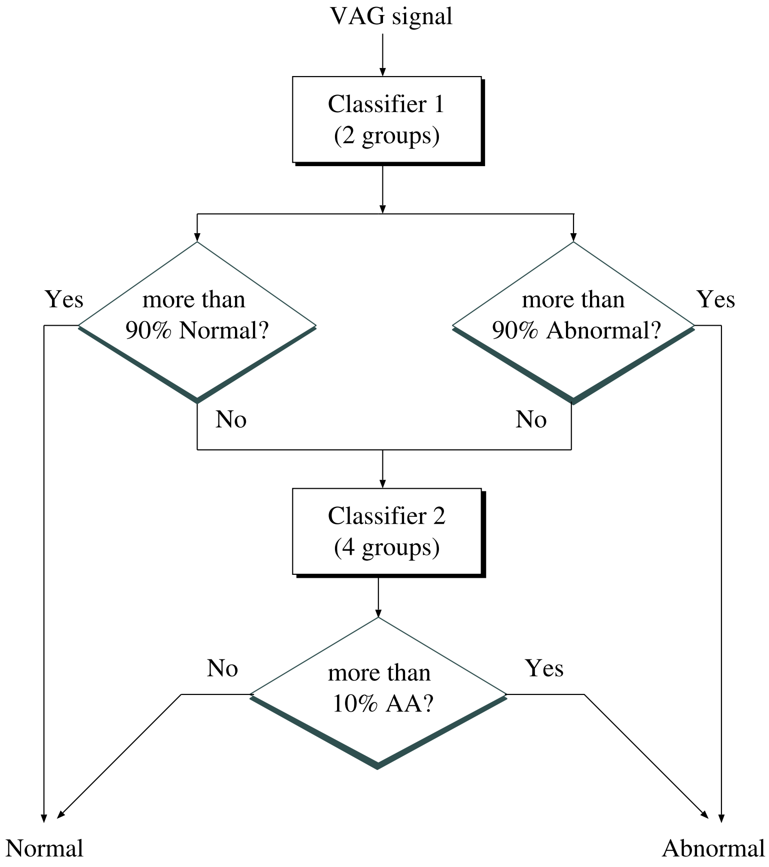

Figure 9.24 A two-step classification method for the diagnosis of cartilage pathology. AA, arthroscopically abnormal. See also Figure 9.23. Reproduced with permission from Z.M.K. Moussavi, R.M. Rangayyan, G.D. Bell, C.B. Frank, K.O. Ladly, and Y.T. Zhang, Screening of vibroarthrographic signals via adaptive segmentation and linear prediction modeling, *IEEE Transactions on Biomedical Engineering,* 43(1):15–23, 1996. ©IEEE.

9.14 Study Questions and Problems

1. The prototype vectors of two classes of signals are specified as Class 1: $[1, 0.5]$, and Class 2: $[3, 3]$. A new sample vector is given as $[2, 1]$. Plot the feature-vector space and describe your observations. Give the equations for two measures of similarity or dissimilarity, compute the measures for the sample vector, and classify the sample as Class 1 or Class 2 using each measure.

2. In a three-class pattern classification problem, the three decision boundaries are $d_1(\mathbf{x}) = -x_1 + x_2$, $d_2(\mathbf{x}) = x_1 + x_2 - 5$, and $d_3(\mathbf{x}) = -x_2 + 1$.

   Draw the decision boundaries on a sheet of graph paper.

   Classify the sample pattern vector $\mathbf{x} = [6, 5]$ using the decision functions.

3. Two pattern class prototype vectors are given to you as $\mathbf{z}_1 = [3, 4]$ and $\mathbf{z}_2 = [10, 2]$. Classify the sample pattern vector $\mathbf{x} = [4, 5]$ using (a) the normalized dot product, and (b) the Euclidean distance. Plot the feature-vector space and describe your observations.

4. A researcher makes two measurements per sample on a set of 10 normal and 10 abnormal samples. The set of feature vectors for the normal samples is
   $\{[2, 6], [22, 20], [10, 14], [10, 10], [24, 24], [8, 10], [8, 8], [6, 10], [8, 12], [6, 12]\}$.
   The set of feature vectors for the abnormal samples is
   $\{[4, 10], [24, 16], [16, 18], [18, 20], [14, 20], [20, 22], [18, 16], [20, 20], [18, 18], [20, 18]\}$.

   Plot the scatter diagram of the samples in both classes in the feature-vector space. Draw a linear decision function to classify the samples with the least error of misclassification. Write the decision function as a mathematical rule.

   How many (if any) samples are misclassified by your decision function? Mark the misclassified samples on the plot.

   Two new observation sample vectors are provided to you as $\mathbf{x}_1 = [12, 15]$ and $\mathbf{x}_2 = [14, 15]$. Classify the samples using your decision rule.

   Now, classify the samples $\mathbf{x}_1$ and $\mathbf{x}_2$ using the k-nearest-neighbor method, with $k = 7$. Measure distances graphically on your graph paper plot and mark the neighbors used in this decision process for each sample.

   Comment upon the results — whether the two methods resulted in the same classification result or not — and provide reasons.

5. A researcher makes measurements of RR intervals (in seconds) and FF for a number of ECG beats including (i) normal beats, (ii) PVCs, and (iii) normal beats with a compensatory pause (NBCP). The values (training set) are given in Table 9.6.

Normal Beats		PVCs		NBCPs	
RR	FF	RR	FF	RR	FF
0.700	1.5	0.600	5.5	0.800	1.2
0.720	1.0	0.580	6.1	0.805	1.1
0.710	1.2	0.560	6.4	0.810	1.6
0.705	1.3	0.570	5.9	0.815	1.3
0.725	1.4	0.610	6.3	0.790	1.4

Table 9.6 Training set of $[RR, FF]$ feature vectors.

   (a) Plot the $[RR, FF]$ feature-vector points for the three classes of beats. (b) Compute the prototype vectors for each class as the class means. Indicate the prototypes on the plot. (c) Derive linear discriminant functions (or decision functions) as the perpendicular bisectors of the straight lines joining the prototypes. State the decision rule(s) for each type of beat. (d) Three new beats are observed to have the parameters listed in Table 9.7. Classify each beat using the decision functions derived in part (c).

6. For the training data given in the preceding problem, compute the mean and covariance matrices of the feature vectors for each class, as well as the pooled covariance matrix. Design a classifier based upon the Mahalanobis distance using the pooled covariance matrix.

Beat No.	RR	FF
1	0.650	5.5
2	0.680	1.9
3	0.820	1.8

Table 9.7 Test set of $[RR, FF]$ feature vectors.

7. You have won a contract to design a software package for CAD of cardiovascular diseases using the heart sound signal (PCG) as the main source of information. The main task is to identify the presence of murmurs in systole and/or diastole. You may use other signals for reference.

   Propose a signal processing system to (i) acquire the required signals; (ii) preprocess them as required; (iii) extract at least two features for classification; and (iv) classify the PCG signals as: class 1, normal (no murmurs); class 2, systolic murmur; class 3, diastolic murmur; or class 4, systolic and diastolic murmur.

   Provide a block diagram of the complete procedure. Explain the reason behind the application of each step and state the expected results or benefits. Provide algorithmic details and/or mathematical definitions for at least two major steps in your procedure.

   Draw a schematic plot of the feature-vector space and indicate where samples from the four classes listed above would fall. Propose a framework of decision rules to classify an incoming signal as belonging to one of the four classes.

9.15 Laboratory Exercises and Projects

Note: Data files related to the exercises are available at the site

```
http://people.ucalgary.ca/~ranga/enel563
```

1. The data file ecgpvc.dat contains the ECG signal of a patient with PVCs (see Figures 9.9 and 9.11). Refer to the file ecgpvc.m for details. Use the first 40% of the signal as training data to develop a PVC detection system (see Section 9.11). Develop code to segment the QRS – T portion of each beat using the Pan–Tompkins method (see Section 4.3.2), and compute the RR interval, QRS width (see Figure 4.6), and FF for each beat (see Section 5.6.4). Design linear discriminant functions using (i) RR and QRS width and (ii) QRS width and FF as the features; see Figure 9.10. Analyze the results in terms of TPF and FPF.

   Code the decision function into your program as a classification rule. Test the pattern classifier program with the remaining 60% of the signal as the test signal. Compute the test-stage classification accuracy in terms of TPF and FPF.

2. Repeat the previous exercise replacing the linear discriminant function with the k-nearest-neighbor method, with $k = 1, 3, 5$, and 7. Evaluate the method with feature sets composed as (a) RR and QRS width, (b) QRS width and FF, and (c) RR, FF, and QRS width.

   Compare the performances of the three classifiers and provide reasons for any differences between them.