Chapter 15

Discrimination and Classification

15.1. DISCRIMINATION VS. CLASSIFICATION

This chapter deals with the problem of discerning membership among some number of groups, on the basis of a K-dimensional vector $\boldsymbol{x}$ of attributes that is observed for each member of each group. It is assumed that the number of groups G is known in advance; that this collection of groups constitutes a MECE partition of the sample space; that each data vector belongs to one and only one group; and that a set of training data is available, in which the group membership of each of the data vectors $\boldsymbol{x}_i$, $i = 1, \ldots, n$, is known with certainty. The related problem, in which we know neither the group memberships of the data vectors nor the number of groups overall, is treated in Chapter 16.

The term *discrimination* refers to the process of estimating functions of the training data $\boldsymbol{x}_i$ that best describe the features separating the known group memberships of each $\boldsymbol{x}_i$. In cases where this can be achieved well with three or fewer functions, it may be possible to express the discrimination graphically. The statistical basis of discrimination is the notion that each of the G groups corresponds to a different multivariate PDF for the data, $f_g(\boldsymbol{x})$, $g = 1, \ldots, G$. It is not necessary to assume multinormality for these distributions, but informative connections can be made with the material presented in Chapter 12 when that assumption is supported by the data.

Classification refers to use of the discrimination rule(s) to assign data that were not part of the original training sample to one of the G groups, or to the estimation of probabilities $p_g\ (\boldsymbol{x})$, $g = 1, \ldots, G$, that the vector $\boldsymbol{x}$ belongs to group g. If the groupings of $\boldsymbol{x}$ pertain to a time after $\boldsymbol{x}$ itself has been observed, then classification is a natural tool to use for forecasting discrete events. That is, a forecast can be made by classifying the current observation $\boldsymbol{x}$ as belonging to the group that is forecast to occur, or by computing the probabilities $p_g\ (\boldsymbol{x})$ for the occurrence of each of the G events.

Most of this chapter describes well-established, mainly linear, methods for discrimination and classification. More exotic and flexible, but also more computationally demanding, approaches are presented in the final sections.

15.2. SEPARATING TWO POPULATIONS

15.2.1. Equal Covariance Structure: Fisher's Linear Discriminant

The simplest form of discriminant analysis involves distinguishing between $G = 2$ groups on the basis of K-dimensional vectors of observations $\boldsymbol{x}$. A training sample must exist, consisting of n_1 observations of $\boldsymbol{x}$ known to have come from Group 1, and n_2 observations of $\boldsymbol{x}$ known to have come from Group 2. That is, the basic data are the two matrices $[X_1]$, dimensioned $(n_1 \times K)$, and $[X_2]$, dimensioned $(n_2 \times K)$. The goal

Statistical Methods in the Atmospheric Sciences. https://doi.org/10.1016/B978-0-12-815823-4.00015-8
© 2019 Elsevier Inc. All rights reserved.

is to find a linear function of the K elements of the observation vector, that is, the linear combination $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}$, called the *discriminant function*, that will best allow a future K-dimensional vector of observations to be classified as belonging to either Group 1 or Group 2.

Assuming that the two populations corresponding to the groups have the same covariance structure, the approach to this problem taken by the statistician R.A. Fisher was to find the vector $\boldsymbol{a}$ as that direction in the K-dimensional space of the data that maximizes the separation in standard deviation units of the two vector means, when the data are projected onto $\boldsymbol{a}$. This criterion is equivalent to choosing $\boldsymbol{a}$ to maximize

$$\frac{\left(\boldsymbol{a}^T\overline{\boldsymbol{x}}_1 - \boldsymbol{a}^T\overline{\boldsymbol{x}}_2\right)^2}{\boldsymbol{a}^T\left[S_{\text{pool}}\right]\boldsymbol{a}}. \tag{15.1}$$

Here the two mean vectors are calculated separately for each group, as would be expected, according to

$$\overline{\boldsymbol{x}}_g = \frac{1}{n_g}\left[X_g\right]^T\mathbf{1} = \begin{bmatrix} \dfrac{1}{n_g}\displaystyle\sum_{i=1}^{n_g} x_{i,1} \\ \dfrac{1}{n_g}\displaystyle\sum_{i=1}^{n_g} x_{i,2} \\ \vdots \\ \dfrac{1}{n_g}\displaystyle\sum_{i=1}^{n_g} x_{i,K} \end{bmatrix}, \; g = 1,2; \tag{15.2}$$

where $\mathbf{1}$ is a $(n \times 1)$ vector containing only 1's, and n_g is the number of training-data vectors $\boldsymbol{x}$ in the gth group. The estimated common covariance matrix for the two groups, $[S_{\text{pool}}]$, is calculated using Equation 12.42b. If $n_1 = n_2$, the result is that each element of $[S_{\text{pool}}]$ is the simple average of the corresponding elements of $[S_1]$ and $[S_2]$. Note that multivariate normality has not been assumed for either of the groups. Rather, regardless of their distributions and whether or not those distributions are of the same form, all that has been assumed is that their underlying population covariance matrices $[\Sigma_1]$ and $[\Sigma_2]$ are equal.

Finding the direction $\boldsymbol{a}$ maximizing Equation 15.1 reduces the discrimination problem from one of sifting through and comparing relationships among the K elements of the data vectors, to looking at a single number. That is, the data vector $\boldsymbol{x}$ is transformed to a new scalar variable, $\delta_1 = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}$, known as *Fisher's linear discriminant function*. The groups of K-dimensional multivariate data are essentially reduced to groups of univariate data with different means (but equal variances), distributed along the $\boldsymbol{a}$ axis. The discriminant vector locating this direction of maximum separation is given by

$$\boldsymbol{a} = \left[S_{\text{pool}}\right]^{-1}\left(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2\right), \tag{15.3}$$

so that Fisher's linear discriminant function is

$$\delta_1 = \boldsymbol{a}^T\boldsymbol{x} = \left(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2\right)^T\left[S_{\text{pool}}\right]^{-1}\boldsymbol{x}. \tag{15.4}$$

As indicated in Equation 15.1, this transformation to Fisher's linear discriminant function maximizes the scaled distance between the two sample means in the training sample, which is

$$\boldsymbol{a}^T\left(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2\right) = \left(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2\right)^T\left[S_{\text{pool}}\right]^{-1}\left(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2\right) = D^2. \tag{15.5}$$

That is, this maximum distance between the projections of the two sample means is exactly the Mahalanobis distance between them, according to $[S_{\text{pool}}]$.

A decision to classify a future observation $\boldsymbol{x}$ as belonging to either Group 1 or Group 2 can now be made according to the value of the scalar $\delta_1 = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}$. This dot product is a one-dimensional (i.e., scalar) projection of the vector $\boldsymbol{x}$ onto the direction of maximum separation, $\boldsymbol{a}$. The discriminant function δ_1 is essentially a new variable, analogous to the new variable u in PCA and the new variables v and w in CCA, produced as a linear combination of the elements of a data vector $\boldsymbol{x}$. The simplest way to classify an observation $\boldsymbol{x}$ is to assign it to Group 1 if the projection $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}$ is closer to the projection of the Group 1 mean onto the direction $\boldsymbol{a}$, and assign it to Group 2 if $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}$ is closer to the projection of the mean of Group 2. Along the $\boldsymbol{a}$ axis, the midpoint between the means of the two groups is given by the projection of the average of these two mean vectors onto the vector $\boldsymbol{a}$,

$$\hat{m} = \frac{1}{2}\left(\boldsymbol{a}^T\overline{\boldsymbol{x}}_1 + \boldsymbol{a}^T\overline{\boldsymbol{x}}_2\right) = \frac{1}{2}\boldsymbol{a}^T(\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2) = \frac{1}{2}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T\left[S_{pool}\right]^{-1}(\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2). \tag{15.6}$$

Given an observation $\boldsymbol{x}_0$ whose group membership is unknown, this simple midpoint criterion classifies it according to the rule

$$\text{Assign } \boldsymbol{x}_0 \text{ to Group 1 if } \ \boldsymbol{a}^T\boldsymbol{x}_0 \geq \hat{m}, \tag{15.7a}$$

or

$$\text{Assign } \boldsymbol{x}_0 \text{ to Group 2 if } \ \boldsymbol{a}^T\boldsymbol{x}_0 < \hat{m}. \tag{15.7b}$$

This classification rule divides the K-dimensional space of $\boldsymbol{x}$ into two regions, according to the (hyper-) plane perpendicular to $\boldsymbol{a}$ at the midpoint given by Equation 15.6. In two dimensions, the plane is divided into two regions according to the line perpendicular to $\boldsymbol{a}$ at this point. The volume in three dimensions is divided into two regions according to the plane perpendicular to $\boldsymbol{a}$ at this point, and so on for higher dimensions.

Example 15.1. Linear Discrimination in $K = 2$ Dimensions

Table 15.1 presents average July temperature and precipitation for cities in three regions of the United States. The data vectors include $K = 2$ elements each: one temperature element and one precipitation element. Consider the problem of distinguishing between membership in Group 1 vs. Group 2. This problem might arise if the locations in Table 15.1 represented the core portions of their respective climatic regions, and on the basis of these data we wanted to classify stations not listed in this table as belonging to one or the other of these two groups.

The mean vectors for the $n_1 = 10$ and $n_2 = 9$ data vectors in Groups 1 and 2 are

$$\overline{\boldsymbol{x}}_1 = \begin{bmatrix} 80.6°\text{F} \\ 5.67\,\text{in.} \end{bmatrix} \ \text{and} \ \overline{\boldsymbol{x}}_2 = \begin{bmatrix} 78.7°\text{F} \\ 3.57\,\text{in.} \end{bmatrix}, \tag{15.8a}$$

and the two sample covariance matrices are

$$[S_1] = \begin{bmatrix} 1.47 & 0.65 \\ 0.65 & 1.45 \end{bmatrix} \ \text{and} \ [S_2] = \begin{bmatrix} 2.01 & 0.06 \\ 0.06 & 0.17 \end{bmatrix}. \tag{15.8b}$$

Since $n_1 \neq n_2$ the pooled estimate for the common variance–covariance matrix is obtained by the weighted average specified by Equation 12.42b. The vector $\boldsymbol{a}$ pointing in the direction of maximum separation of the two sample mean vectors is then computed using Equation 15.3 as

TABLE 15.1 Average July Temperature (°F) and Precipitation (inches) for Locations in Three Regions of the United States, for the Period 1951–1980

Group 1: Southeast United States (o)			Group 2: Central United States (×)			Group 3: Northeast United States (+)		
Station	Temp.	Ppt.	Station	Temp.	Ppt.	Station	Temp.	Ppt.
Athens, GA	79.2	5.18	Concordia, KS	79.0	3.37	Albany, NY	71.4	3.00
Atlanta, GA	78.6	4.73	Des Moines, IA	76.3	3.22	Binghamton, NY	68.9	3.48
Augusta, GA	80.6	4.4	Dodge City, KS	80.0	3.08	Boston, MA	73.5	2.68
Gainesville, FL	80.8	6.99	Kansas City, MO	78.5	4.35	Bridgeport, CT	74.0	3.46
Huntsville, AL	79.3	5.05	Lincoln, NE	77.6	3.2	Burlington, VT	69.6	3.43
Jacksonville, FL	81.3	6.54	Springfield, MO	78.8	3.58	Hartford, CT	73.4	3.09
Macon, GA	81.4	4.46	St. Louis, MO	78.9	3.63	Portland, ME	68.1	2.83
Montgomery, AL	81.7	4.78	Topeka, KS	78.6	4.04	Providence, RI	72.5	3.01
Pensacola, FL	82.3	7.18	Wichita, KS	81.4	3.62	Worcester, MA	69.9	3.58
Savannah, GA	81.2	7.37						
Averages:	80.6	5.67		78.7	3.57		71.3	3.17

$$\begin{aligned} \boldsymbol{a} &= \begin{bmatrix} 1.73 & 0.37 \\ 0.37 & 0.84 \end{bmatrix}^{-1} \left(\begin{bmatrix} 80.6 \\ 5.67 \end{bmatrix} - \begin{bmatrix} 78.7 \\ 3.57 \end{bmatrix} \right) \\ &= \begin{bmatrix} 0.640 & -.283 \\ -.283 & 1.309 \end{bmatrix} \begin{bmatrix} 1.9 \\ 2.10 \end{bmatrix} = \begin{bmatrix} 0.62 \\ 2.21 \end{bmatrix}. \end{aligned} \tag{15.9}$$

Figure 15.1 illustrates the geometry of this problem. Here the data for the warmer and wetter southeastern stations of Group 1 are plotted as circles, and the central U.S. stations of Group 2 are plotted as ×'s. The vector means for the two groups are indicated by the heavy symbols. The direction $\boldsymbol{a}$ is not, and in general will not be, parallel to the line segment connecting the two group means. The projections of these two means onto $\boldsymbol{a}$ are indicated by the lighter dashed lines. The midpoint between these two projections locates the dividing point between the two groups in the one-dimensional discriminant space defined by $\boldsymbol{a}$. The heavy dashed line perpendicular to the discriminant function δ_1 at this point divides the (temperature, precipitation) plane into two regions. Future points of unknown group membership falling above and to the right of this heavy dashed line would be classified as belonging to Group 1, and points falling below and to the left would be classified as belonging to Group 2.

Since the average of the mean vectors for Groups 1 and 2 is $[79.65, 4.62]^{\mathrm{T}}$, the value of the dividing point is $\hat{m} = (0.62)(79.65) + (2.21)(4.62) = 59.59$. Of the 19 points in this training data, only that for Atlanta has been misclassified. For this station, $\delta_1 = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} = (0.62)(78.6) + (2.20)(4.73) = 59.18$. Since this value of δ_1 is slightly less than the midpoint value, Atlanta would be incorrectly classified as belonging to Group 2 (Equation 15.7). By contrast, the point for Augusta lies just to the Group 1 side of the heavy dashed line. For Augusta, $\delta_1 = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} = (0.62)(80.6) + (2.20)(4.40) = 59.70$, which is slightly greater than $\hat{m}$.

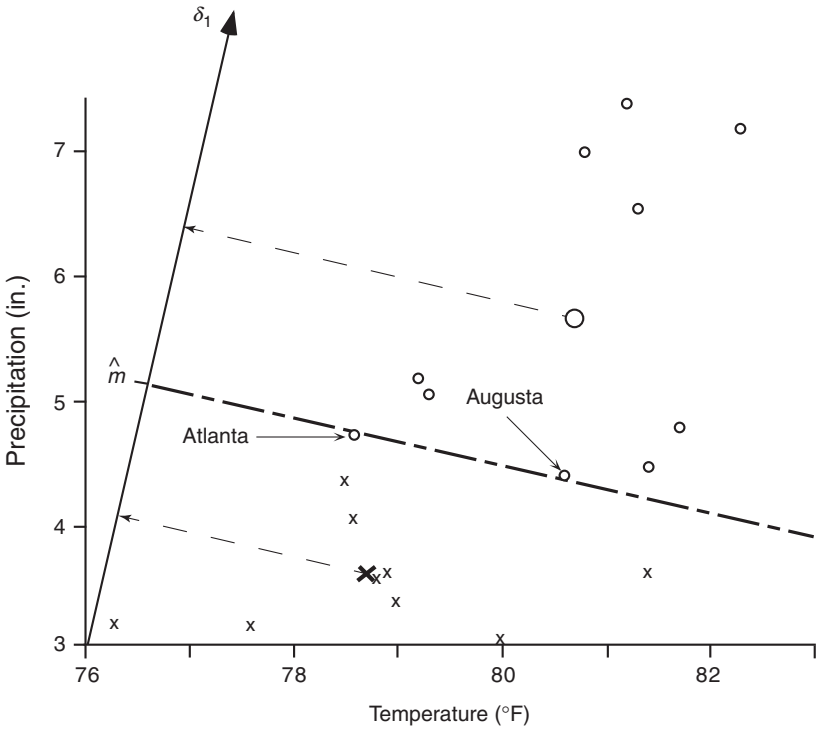

FIGURE 15.1 Illustration of the geometry of linear discriminant analysis applied to the southeastern (circles) and central (X's) U. S. data in Table 15.1. The (vector) means of the two groups of data are indicated by the heavy symbols, and their projections onto the discriminant function are indicated by the light dashed lines. The midpoint between these two projections, $\hat{m}$, defines the dividing line (heavier dashed line) used to assign future (temperature, precipitation) pairs to the groups. Of these training data, only the data point for Atlanta has been misclassified. Note that the discriminant function has been shifted to the right (i.e., does not pass through the origin, but is parallel to the vector $\boldsymbol{a}$ in Equation 15.9) in order to improve the clarity of the plot, but this does not affect the relative positions of the projections of the data points onto it.

Consider now the assignment to either Group 1 or Group 2 of two stations not listed in Table 15.1. For New Orleans, Louisiana, the average July temperature is 82.1°F, and the average July precipitation is 6.73 in. Applying Equation 15.7, we find $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} = (0.62)(82.1) + (2.20)(6.73) = 65.78 > 59.59$. Therefore New Orleans would be classified as belonging to Group 1. Similarly the average July temperature and precipitation for Columbus, Ohio, are 74.7°F and 3.37 in., respectively. For this station, $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} = (0.62)(74.7) + (2.20)(3.37) = 53.76 < 59.59$, which would result in Columbus being classified as belonging to Group 2. ◇

Example 15.1 was constructed with $K = 2$ variables in each data vector in order to allow the geometry of the problem to be easily represented in two dimensions. However, it is not necessary to restrict the use of discriminant analysis to situations with only bivariate observations. In fact, discriminant analysis is potentially most powerful when allowed to operate on higher-dimensional data. For example, it would be possible to extend Example 15.1 to classifying stations according to average temperature and precipitation for all 12 months. If this were done, each data vector $\boldsymbol{x}$ would consist of $K = 24$ values. The discriminant vector $\boldsymbol{a}$ would also consist of $K = 24$ elements, but the dot product $\delta_1 = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}$ would still be a single scalar that could be used to classify group memberships.

Usually high-dimensional vectors of atmospheric data exhibit substantial correlation among the K elements, and thus carry some redundant information. For example, the 12 monthly mean temperatures and 12 monthly mean precipitation values are not mutually independent. If only for computational economy, it can be a good idea to reduce the dimensionality of this kind of data before subjecting it to a discriminant analysis. This reduction in dimension is most commonly achieved through a principal component analysis (Chapter 13). When the groups in a discriminant analysis are assumed to have the same covariance structure, it seems natural to perform this PCA on the estimate of their common variance–covariance matrix, $[S_{\text{pool}}]$. However, if the shape of the dispersion of the group means (as characterized by Equation 15.18) is substantially different from $[S_{\text{pool}}]$, its leading principal components may not be good discriminators, and better results might be obtained from a discriminant analysis based on PCA of the overall covariance, $[S]$ (Jolliffe, 2002). If the data vectors are not of consistent units (some temperatures and some precipitation amounts, for example), it will make more sense to perform the PCA on the corresponding correlation matrix. The discriminant analysis can then be carried out using M-dimensional data vectors composed of elements that are the leading M principal components, rather than the original K-dimensional raw data vectors. The resulting discriminant function will then pertain to the principal components in the $(M \times 1)$ vector $\boldsymbol{u}$, rather than to the original $(K \times 1)$ data, $\boldsymbol{x}$. In addition, if the first two principal components account for a large fraction of the total variance, the data can effectively be visualized in a plot like Figure 15.1, in which the horizontal and vertical axes are the first two principal components.

15.2.2. Fisher's Linear Discriminant for Multivariate Normal Data

Use of Fisher's linear discriminant requires no assumptions about the specific nature of the distributions for the two groups, $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$, except that they have equal covariance matrices. If in addition these are two multivariate normal distributions, or they are sufficiently close to multivariate normal for the sampling distributions of their means to be essentially multivariate normal according to the Central Limit Theorem, there are connections to the Hotelling T^2 test (Section 12.5) regarding differences between the two means.

In particular, Fisher's linear discriminant vector (Equation 15.3) identifies a direction that is identical to the linear combination of the data that is most strongly significant (Equation 12.57b), under the null hypothesis that the two population mean vectors are equal. That is, the vector $\boldsymbol{a}$ defines the direction maximizing the separation of the two means for both a discriminant analysis and the T^2 test. Furthermore, the distance between the two means in this direction (Equation 15.5) is their Mahalanobis distance with respect to the pooled estimate $[S_{\text{pool}}]$ of the common covariance $[\Sigma_1] = [\Sigma_2]$, which is proportional (through the factor $n_1^{-1} + n_2^{-1}$, in Equation 12.42a) to the 2-sample T^2 statistic itself (Equation 12.40).

In light of these relationships, one way to look at Fisher's linear discriminant, when applied to multivariate normal data, is as an implied test relating to the null hypothesis that $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. Even if this null hypothesis is true, the corresponding sample means in general will be different, and the result of the T^2 test is an informative necessary condition regarding the reasonableness of conducting the discriminant analysis. A multivariate normal distribution is fully defined by its mean vector and covariance matrix. Since $[\Sigma_1] = [\Sigma_2]$ already has been assumed, if in addition the two multivariate normal data groups are consistent with the condition $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, then there is no basis upon which to discriminate between them. Note, however, that rejecting the null hypothesis of equal means in the corresponding T^2 test is not a sufficient condition for good discrimination: arbitrarily small mean differences can be detected by this test as sample sizes increase, even though the scatter of the two data groups may overlap to such a limited degree that discrimination is completely pointless.

15.2.3. Minimizing Expected Cost of Misclassification

The point $\hat{m}$ on Fisher's discriminant function halfway between the projections of the two sample means is not always the best point at which to make a separation between groups. One might have prior information that the probability of membership in Group 1 is larger than that for Group 2, perhaps because Group 2 members are rather rare overall. If this is so, it would usually be desirable to move the classification boundary toward the Group 2 mean, with the result that more future observations $\boldsymbol{x}$ would be classified as belonging to Group 1. Similarly, if misclassifying a Group 1 data value as belonging to Group 2 were to be a more serious error than misclassifying a Group 2 data value as belonging to Group 1, again we would want to move the boundary toward the Group 2 mean.

One rational way to accommodate these considerations is to define the classification boundary based on the *expected cost of misclassification* (ECM) of a future data vector. Let p_1 be the prior probability (the unconditional probability according to previous information) that a future observation $\boldsymbol{x}_0$ belongs to Group 1, and let p_2 be the prior probability that the observation $\boldsymbol{x}_0$ belongs to Group 2. Define $P(2\,|\,1)$ to be the conditional probability that a Group 1 object is misclassified as belonging to Group 2, and $P(1\,|\,2)$ as the conditional probability that a Group 2 object is misclassified as belonging to Group 1. These conditional probabilities will depend on the two PDFs $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$, respectively; and on the placement of the classification criterion, because these conditional probabilities will be given by the integrals of their respective PDFs over the regions in which classifications would be made to the other group. That is,

$$P(2|\,1) = \int_{R_2} f_1(\boldsymbol{x})\,d\boldsymbol{x} \tag{15.10a}$$

and

$$P(1|\,2) = \int_{R_1} f_2(\boldsymbol{x})\,d\boldsymbol{x}, \tag{15.10b}$$

where R_1 and R_2 denote the regions of the K-dimensional space of $\boldsymbol{x}$ in which classifications into Group 1 and Group 2, respectively, would be made. Unconditional probabilities of misclassification are given by the products of these conditional probabilities with the corresponding prior probabilities, that is, $P(2\,|\,1)$ p_1 and $P(1\,|\,2)$ p_2.

If $C(1\,|\,2)$ is the cost, or penalty, incurred when a Group 2 member is incorrectly classified as part of Group 1, and $C(2\,|\,1)$ is the cost incurred when a Group 1 member is incorrectly classified as part of Group 2, then the expected cost of misclassification will be

$$ECM = C(2|\,1)\,P(2|\,1)\,p_1 + C(1|\,2)\,P(1|\,2)\,p_2. \tag{15.11}$$

The classification boundary can be adjusted to minimize this expected cost of misclassification, through the effect of the boundary on the misclassification probabilities (Equations 15.10). The resulting classification rule is

$$\text{Assign } \boldsymbol{x}_0 \text{ to Group 1 if } \frac{f_1(\boldsymbol{x}_0)}{f_2(\boldsymbol{x}_0)} \geq \frac{C(1|\,2)\,p_2}{C(2|\,1)\,p_1}, \tag{15.12a}$$

or

$$\text{Assign } \boldsymbol{x}_0 \text{ to Group 2 if } \frac{f_1(\boldsymbol{x}_0)}{f_2(\boldsymbol{x}_0)} < \frac{C(1|2)\,p_2}{C(2|1)\,p_1}. \tag{15.12b}$$

That is, classification of $\boldsymbol{x}_0$ depends on the ratio of its likelihoods according to the PDFs for the two groups (i.e., the ratio of the two density functions, evaluated at $\boldsymbol{x}_0$), in relation to the ratios of the products of the misclassification costs and prior probabilities. Accordingly, it is not actually necessary to know the two misclassification costs specifically, but only their ratio, and likewise it is necessary only to know the ratio of the prior probabilities. If $C(1|2) \gg C(2|1)$—that is, if misclassifying a Group 2 member as belonging to Group 1 is especially undesirable—then the ratio of likelihoods on the left-hand side of Equation 15.12 must be quite large ($\boldsymbol{x}_0$ must be substantially more plausible according to $f_1(\boldsymbol{x})$]) in order to assign $\boldsymbol{x}_0$ to Group 1. Similarly, if Group 1 members are intrinsically rare, so that $p_1 \ll p_2$, a higher level of evidence must be met in order to classify $\boldsymbol{x}_0$ as a member of Group 1. If both misclassification costs and prior probabilities are equal, or of the costs and priors compensate to make the right-hand sides of Equation 15.12 equal to 1, then classification is made according to the larger of $f_1(\boldsymbol{x}_0)$ or $f_2(\boldsymbol{x}_0)$.

Minimizing the ECM (Equation 15.11) does not require assuming that the distributions $f_1(\boldsymbol{x})$ or $f_2(\boldsymbol{x})$ have specific forms, or even that they are of the same parametric family. But it is necessary to know or assume a functional form for each of them in order to evaluate the left-hand side of Equation 15.12. Often it is assumed that both $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$ are multivariate normal (possibly after data transformations for some or all of the elements of $\boldsymbol{x}$), with equal covariance matrices that are estimated using $[S_{\text{pool}}]$. In this case Equation 15.12a, for the conditions under which $\boldsymbol{x}_0$ would be assigned to Group 1, becomes

$$\frac{2\pi^{-K/2}\left|[S_{\text{pool}}]\right|^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{x}_0-\overline{\boldsymbol{x}}_1)^T[S_{\text{pool}}]^{-1}(\boldsymbol{x}_0-\overline{\boldsymbol{x}}_1)\right)}{2\pi^{-K/2}\left|[S_{\text{pool}}]\right|^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{x}_0-\overline{\boldsymbol{x}}_2)^T[S_{\text{pool}}]^{-1}(\boldsymbol{x}_0-\overline{\boldsymbol{x}}_2)\right)} \geq \frac{C(1|2)\,p_2}{C(2|1)\,p_1}, \tag{15.13a}$$

which, after some rearrangement, is equivalent to

$$(\overline{\boldsymbol{x}}_1-\overline{\boldsymbol{x}}_2)^T[S_{\text{pool}}]^{-1}\boldsymbol{x}_0 - \frac{1}{2}(\overline{\boldsymbol{x}}_1-\overline{\boldsymbol{x}}_2)^T[S_{\text{pool}}]^{-1}(\overline{\boldsymbol{x}}_1+\overline{\boldsymbol{x}}_2) \geq \ln\left(\frac{C(1|2)\,p_2}{C(2|1)\,p_1}\right). \tag{15.13b}$$

The left-hand side of Equation 15.13b looks elaborate, but its elements are familiar. In particular, its first term is exactly the linear combination $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_0$ in Equation 15.7. The second term is the midpoint $\hat{m}$ between the two means when projected onto $\boldsymbol{a}$, defined in Equation 15.6. Therefore if $C(1|2) = C(2|1)$ and $p_1 = p_2$ (or if other combinations of these quantities yield $\ln(1) = 0$ on the right-hand side of Equation 15.13b), the minimum ECM classification criterion for two multivariate normal populations with equal covariance is exactly the same as Fisher's linear discriminant. To the extent that the costs and/or prior probabilities are not equal, Equation 15.13 results in movement of the classification boundary away from the midpoint defined in Equation 15.6, and toward the projection of one of the two means onto $\boldsymbol{a}$.

15.2.4. Unequal Covariances: Quadratic discrimination

Discrimination and classification are much more straightforward, both conceptually and mathematically, if equality of covariances for the G populations can be assumed. For example, it is the equality-of-covariance assumption that allows the simplification from Equation 15.13a to

Equation 15.13b for two multivariate normal populations. If it cannot be assumed that $[\Sigma_1]=[\Sigma_2]$, and instead these two covariance matrices are estimated separately by $[S_1]$ and $[S_2]$, respectively, minimum ECM classification for two multivariate populations leads to classification of $\boldsymbol{x}_0$ as belonging to Group 1 if

$$\frac{1}{2}\boldsymbol{x}_0^T\left([S_1]^{-1}-[S_2]^{-1}\right)\boldsymbol{x}_0+\left(\overline{\boldsymbol{x}}_1^T[S_1]^{-1}-\overline{\boldsymbol{x}}_2^T[S_2]^{-1}\right)\boldsymbol{x}_0-\text{const}\geq \ln\left(\frac{C(1|2)\,p_2}{C(2|1)\,p_1}\right), \tag{15.14a}$$

where

$$\text{const}=\frac{1}{2}\left(\ln\frac{|[S_1]|}{|[S_2]|}+\overline{\boldsymbol{x}}_1^T[S_1]^{-1}\overline{\boldsymbol{x}}_1-\overline{\boldsymbol{x}}_2^T[S_2]^{-1}\overline{\boldsymbol{x}}_2\right) \tag{15.14b}$$

contains scaling constants not involving $\boldsymbol{x}_0$.

The mathematical differences between Equations 15.13b and 15.14 result because cancellations and recombinations that are possible in Equation 15.13 when the covariance matrices are equal, whereas additional terms in Equation 15.14 result when they are not. Classification and discrimination using Equation 15.14 are more difficult conceptually because the regions R_1 and R_2 are no longer necessarily contiguous. Equation 15.14, for classification with unequal covariances, is also less robust to non-Gaussian data than classification with Equation 15.13, when equality of covariance structure can reasonably be assumed.

Figure 15.2 illustrates quadratic discrimination and classification with a simple, one-dimensional example. Here it has been assumed for simplicity that the right-hand side of Equation 15.14a is $\ln(1)=0$, so the classification criterion reduces to assigning $\boldsymbol{x}_0$ to whichever group yields the larger likelihood, $f_g(\boldsymbol{x}_0)$. Because the variance for Group 1 is so much smaller, both very large and very small $\boldsymbol{x}_0$ will be assigned to Group 2. Mathematically, this discontinuity for the region R_2 results from the first (i.e., the quadratic) term in Equation 15.14a, which in $K=1$ dimension is equal to ${x_0}^2(1/{s_1}^2-1/{s_2}^2)/2$. In higher dimensions the shapes of quadratic classification regions will usually be curved and more complicated.

Another approach to nonlinear discrimination within the straightforward framework of linear discriminant analysis is to extend the original data vector $\boldsymbol{x}$ to include non-linear-derived variables based on its elements, as is also done when computing support vector machine classifiers (Section 15.6.1). For example, if the original data vectors $\boldsymbol{x}=[x_1,x_2]^{\mathrm{T}}$ are $K=2$-dimensional, a quadratic discriminant analysis can be carried out in the $K^*=5$-dimensional space of the extended data vector $\boldsymbol{x}^*=[x_1,x_2,{x_1}^2,{x_2}^2,x_1x_2]^{\mathrm{T}}$. The resulting classification boundary can subsequently be mapped back to the original K-dimensional space of $\boldsymbol{x}$, where in general it will be nonlinear.

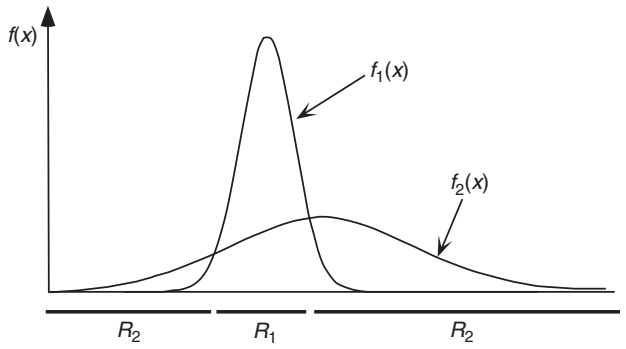

FIGURE 15.2 Discontinuous classification regions resulting from unequal variances for the populations described by two Gaussian PDFs $f_1(x)$ and $f_2(x)$.

15.3. MULTIPLE DISCRIMINANT ANALYSIS (MDA)

15.3.1. Fisher's Procedure for More Than Two Groups

Fisher's linear discriminant, described in Section 15.2.1, can be generalized for discrimination among $G = 3$ or more groups. This generalization is called *multiple discriminant analysis* (MDA). Here the basic problem is to allocate a K-dimensional data vector $\boldsymbol{x}$ to one of G > 2 groups on the basis $J = \min(G–1, K)$ discriminant vectors, $\boldsymbol{a}_j$, $j = 1, \ldots, J$. The projection of the data onto these vectors yield the J discriminant functions

$$\delta_j = \boldsymbol{a}_j^T \boldsymbol{x}, \quad j = 1, \ldots, J. \tag{15.15}$$

The discriminant functions are computed on the basis of a training set of G data matrices $[X_1]$, $[X_2]$, $[X_3]$, …, $[X_G]$, dimensioned, respectively, $(n_g \times K)$. A sample variance–covariance matrix can be computed from each of the G sets of data, $[S_1]$, $[S_2]$, $[S_3]$, …, $[S_G]$, according to Equation 11.30. Assuming that the G groups represent populations having the same covariance matrix, the pooled estimate of this common covariance matrix is estimated by the weighted average

$$[S_{\text{pool}}] = \frac{1}{n-G} \sum_{g=1}^{G} (n_g - 1)[S_g], \tag{15.16}$$

where there are n_g observations in each group, and the total sample size is

$$n = \sum_{g=1}^{G} n_g. \tag{15.17}$$

Equation 12.42b is a special case of Equation 15.16, with $G = 2$.

Computation of multiple discriminant functions also requires calculation of the *between-groups covariance matrix*

$$[S_B] = \frac{1}{G-1} \sum_{g=1}^{G} (\overline{\boldsymbol{x}}_g - \overline{\boldsymbol{x}}_\bullet)(\overline{\boldsymbol{x}}_g - \overline{\boldsymbol{x}}_\bullet)^T, \tag{15.18}$$

where the individual group means are calculated as in Equation 15.2, and.

$$\overline{\boldsymbol{x}}_\bullet = \frac{1}{n} \sum_{g=1}^{G} n_g \overline{\boldsymbol{x}}_g \tag{15.19}$$

is the grand, or overall vector mean of all n observations. The between-groups covariance matrix $[S_B]$ is essentially a covariance matrix describing the dispersion of the G sample means around the overall mean (compare Equation 11.35).

The number J of discriminant functions that can be computed is the smaller of $G - 1$ and K. Thus for the two-group case discussed in Section 15.2, there is only $G - 1 = 1$ discriminant function, regardless of the dimensionality K of the data vectors. In the more general case, the discriminant functions are derived from the first J eigenvectors (corresponding to the nonzero eigenvalues) of the matrix

$$[S_{pool}]^{-1}[S_B]. \tag{15.20}$$

This $(K \times K)$ matrix in general is not symmetric. The discriminant vectors $\boldsymbol{a}_j$ are aligned with these eigenvectors, but are often scaled to yield unit variances for the data projected onto them, that is,

$$\boldsymbol{a}_j^T \left[S_{pool}\right] \boldsymbol{a}_j = 1, \;\; j = 1, \ldots, J. \tag{15.21}$$

Usually computer routines for calculating eigenvectors will scale eigenvectors to unit length, that is, $||\boldsymbol{e}_j|| = 1$, but the condition in Equation 15.21 can be achieved by calculating

$$\boldsymbol{a}_j = \frac{\boldsymbol{e}_j}{\left(\boldsymbol{e}_j^T \left[S_{\text{pool}}\right] \boldsymbol{e}_j\right)^{1/2}}, \;\; j = 1, \ldots, J. \tag{15.22}$$

The first discriminant vector $\boldsymbol{a}_1$, which is associated with the largest eigenvalue of the matrix in Equation 15.20, makes the largest contribution to separating the G group means, in aggregate; and $\boldsymbol{a}_J$, which is associated with the smallest nonzero eigenvalue, makes the least contribution overall.

Many texts develop Fisher's method for MDA in terms of the eigenvectors of the product $[W]^{-1}[B]$, where $[W] = (n - G)[S_{\text{pool}}]$ is called the *within-groups covariance matrix*, and $[B] = (G - 1)[S_B]$ is called the sample between-groups matrix; rather than the eigenvectors of Equation 15.20. However, since the two matrix products are the same apart from a scalar multiple, their eigenvectors are also the same and so yield the same discriminant vectors $\boldsymbol{a}_j$ in Equation 15.22.

The J discriminant vectors $\boldsymbol{a}_j$ define a J-dimensional discriminant space, in which the G groups exhibit maximum separation. The projections δ_j (Equation 15.15) of the data onto these vectors are sometimes called the *discriminant coordinates* or canonical variates. This second name derives from the fact that equivalent results can be obtained through CCA using indicator variables for the groups (Hastie et al., 2009). As was also the case when distinguishing between $G = 2$ groups, observations $\boldsymbol{x}$ can be assigned to groups according to which of the G group means is closest in discriminant space. For the $G = 2$ case the discriminant space is one-dimensional, consisting only of a line. The group assignment rule (Equation 15.7) is then particularly simple. More generally, the Euclidean distances in discriminant space between the candidate vector $\boldsymbol{x}_0$ and each of the G group means are evaluated in order to find which is closest. It is actually easier to evaluate these in terms of squared distances, yielding the classification rule:

$$\text{Assign } \boldsymbol{x}_0 \text{ to group } g \text{ if } \sum_{j=1}^{J} \left[\boldsymbol{a}_j \left(\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_g\right)\right]^2 \le \sum_{j=1}^{J} \left[\boldsymbol{a}_j (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_h)\right]^2, \;\; \text{for all } h \ne g. \tag{15.23}$$

That is, the sum of the squared distances between $\boldsymbol{x}_0$ and each of the group means, along the directions defined by the vectors $\boldsymbol{a}_j$, is compared in order to find the closest group mean.

Computing the discriminant vectors $\boldsymbol{a}_j$ allows one to define the discriminant space, which can lead to informative graphical displays of the data in this space or subspaces of it. However, if only a classification rule is needed, it can be computed without the eigendecomposition of the matrix product in Equation 15.20. Specifically, a candidate vector $\boldsymbol{x}_0$ is assigned to the group g that maximizes

$$\overline{\boldsymbol{x}}_g^T \left[S_{\text{pool}}\right]^{-1} \boldsymbol{x}_0 - \frac{1}{2} \overline{\boldsymbol{x}}_g^T \left[S_{\text{pool}}\right]^{-1} \overline{\boldsymbol{x}}_g. \tag{15.24}$$

Example 15.2. Multiple Discriminant Analysis with G = 3 Groups

Consider discriminating among all three groups of data in Table 15.1. Using Equation 15.16 the pooled estimate of the assumed common covariance matrix is

$$\left[S_{\text{pool}}\right] = \frac{1}{28 - 3}\left(9 \begin{bmatrix} 1.47 & 0.65 \\ 0.65 & 1.45 \end{bmatrix} + 8 \begin{bmatrix} 2.08 & 0.06 \\ 0.06 & 0.17 \end{bmatrix} + 8 \begin{bmatrix} 4.85 & -0.17 \\ -0.17 & 0.10 \end{bmatrix}\right) = \begin{bmatrix} 2.75 & 0.20 \\ 0.20 & 0.61 \end{bmatrix}, \tag{15.25}$$

and using Equation 15.18 the between-groups covariance matrix is

$$[S_B] = \frac{1}{2}\left(\begin{bmatrix} 12.96 & 5.33 \\ 5.33 & 2.19 \end{bmatrix} + \begin{bmatrix} 2.89 & -1.05 \\ -1.05 & 0.38 \end{bmatrix} + \begin{bmatrix} 32.49 & 5.81 \\ 5.81 & 1.04 \end{bmatrix} \right) = \begin{bmatrix} 24.17 & 5.04 \\ 5.04 & 1.81 \end{bmatrix}. \tag{15.26}$$

The directions of the two discriminant functions are specified by the eigenvectors of the matrix

$$[S_{\text{pool}}]^{-1}[S_B] = \begin{bmatrix} 0.373 & -.122 \\ -.122 & 1.685 \end{bmatrix} \begin{bmatrix} 24.17 & 5.04 \\ 5.04 & 1.81 \end{bmatrix} = \begin{bmatrix} 8.40 & 1.65 \\ 5.54 & 2.43 \end{bmatrix}, \tag{15.27a}$$

which, when scaled according to Equation 15.22 are

$$\boldsymbol{a}_1 = \begin{bmatrix} 0.542 \\ 0.415 \end{bmatrix} \quad \text{and} \quad \boldsymbol{a}_2 = \begin{bmatrix} -0.282 \\ 1.230 \end{bmatrix}. \tag{15.27b}$$

The discriminant vectors $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ define the directions of the first discriminant function $\delta_1 = \boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{x}$ and the second discriminant function $\delta_2 = \boldsymbol{a}_2^{\mathrm{T}}\boldsymbol{x}$. Figure 15.3 shows the data for all three groups in Table 15.1 plotted in the discriminant space defined by these two functions. Points for Groups 1 and 2 are shown by circles and ×'s, as in Figure 15.1, and points for Group 3 are shown by +'s. The heavy symbols locate the respective vector means for the three groups. Note that the point clouds for Groups 1 and 2 appear to be stretched and distorted relative to their arrangement in Figure 15.1. This is because the matrix in Equation 15.27a is not symmetric so that its eigenvectors, and therefore the two discriminant vectors in Equation 15.27b, are not orthogonal.

The heavy dashed lines in Figure 15.3 divide the portions of the discriminant space that are assigned to each of the three groups by the classification criterion in Equation 15.23. These are the regions closest to each of the group means, and so define a Voronoi tessellation (or collection of Thiessen polygons) in the discriminant space. (Note, however, that the corresponding partition of the data space does not consist of

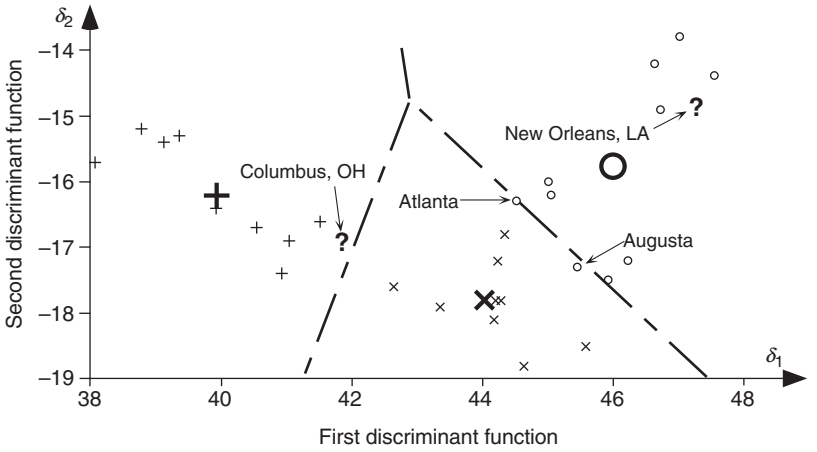

FIGURE 15.3 Illustration of the geometry of multiple discriminant analysis applied to the $G = 3$ groups of data in Table 15.1. Group 1 stations are plotted as circles, Group 2 stations are plotted as ×'s, and Group 3 stations are plotted as +'s. The three vector means are indicated by the corresponding heavy symbols. The two axes are the first and second discriminant functions, and the heavy dashed lines divide sections of this discriminant space allocated to each group. The data for Atlanta and Augusta are misclassified as belonging to Group 2. The two stations Columbus and New Orleans, which are not part of the training data in Table 15.1, are shown as question marks, and are allocated to Groups 3 and 1, respectively.

perpendicular bisectors of the line segments between group vector means, unless the pooled covariance matrix is proportional to $[I]$.) Here the data for Atlanta and Augusta have both been misclassified as belonging to Group 2 rather than Group 1. For Atlanta, for example, the squared distance to the Group 1 mean is $[0.542(78.6–80.6)+0.415(4.73–5.67)]^2+[-0.282(78.6–80.6)+1.230(4.73–5.67)]^2=2.52$, and the squared distance to the Group 2 mean is $[0.542\ (78.6–78.7)+0.415(4.73–3.57)]^2+[-0.282\ (78.6–78.7)+1.230(4.73–3.57)]^2=2.31$. A line in this discriminant space could be drawn by eye that would include these two stations in the Group 1 region. That the discriminant analysis has not specified this line is probably a consequence of the assumption of equal covariance matrices not being well satisfied. In particular, the points in Group 1 appear to be more positively correlated in this discriminant space than the members of the other two groups.

The data points for the two stations Columbus and New Orleans, which are not part of the training data in Table 15.1, are shown by the question marks in Figure 15.3. The location in the discriminant space of the point for New Orleans, for which $\boldsymbol{x}=[82.1\ 6.73]^{\mathrm{T}}$, is $\delta_1=(0.542)(82.1)+(0.415)(6.73)=47.3$ and $\delta_2=(-0.282)(82.1)+(1.230)(6.73)=-14.9$, which is within the region assigned to Group 1. The coordinates in discriminant space for the Columbus data, $\boldsymbol{x}=[74.7\ \ 3.37]^{\mathrm{T}}$, are $\delta_1=(0.542)(74.7)+(0.415)(3.37)=41.9$ and $\delta_2=(-0.282)(74.7)+(1.230)(3.37)=-16.9$, which is within the region assigned to Group 3.

Drawing the discriminant space in Figure 15.3 required computation of the eigenvectors of the matrix product in Equation 15.20. If this diagram had not been of interest, the same group assignments could have been made instead using the computationally simpler Equation 15.24. Evaluating the data for New Orleans with respect to the three group means using Equation 15.24 yields

$$[80.6\,5.67]\begin{bmatrix}.373 & -.122\\ -.122 & 1.685\end{bmatrix}\begin{bmatrix}82.1\\ 6.73\end{bmatrix}-\frac{1}{2}[80.6\,5.67]\begin{bmatrix}.373 & -.122\\ -.122 & 1.685\end{bmatrix}\begin{bmatrix}80.6\\ 5.67\end{bmatrix}=1226.7 \qquad (15.28a)$$

$$[78.7\,3.57]\begin{bmatrix}.373 & -.122\\ -.122 & 1.685\end{bmatrix}\begin{bmatrix}82.1\\ 6.73\end{bmatrix}-\frac{1}{2}[78.7\,3.57]\begin{bmatrix}.373 & -.122\\ -.122 & 1.685\end{bmatrix}\begin{bmatrix}78.7\\ 3.57\end{bmatrix}=1218.6 \qquad (15.28b)$$

and

$$[71.3\,3.17]\begin{bmatrix}.373 & -.122\\ -.122 & 1.685\end{bmatrix}\begin{bmatrix}82.1\\ 6.73\end{bmatrix}-\frac{1}{2}[71.3\,3.17]\begin{bmatrix}.373 & -.122\\ -.122 & 1.685\end{bmatrix}\begin{bmatrix}71.3\\ 3.17\end{bmatrix}=1200.1 \qquad (15.28c)$$

for Groups 1, 2, and 3, respectively, so that membership in Group 1 is chosen because the result in Equation 15.28a is largest. Similarly, the calculations for Columbus are

$$[80.6\,5.67]\begin{bmatrix}.373 & -.122\\ -.122 & 1.685\end{bmatrix}\begin{bmatrix}74.7\\ 3.37\end{bmatrix}-\frac{1}{2}[80.6\,5.67]\begin{bmatrix}.373 & -.122\\ -.122 & 1.685\end{bmatrix}\begin{bmatrix}80.6\\ 5.67\end{bmatrix}=1010.2 \qquad (15.29a)$$

$$[78.7\,3.57]\begin{bmatrix}.373 & -.122\\ -.122 & 1.685\end{bmatrix}\begin{bmatrix}74.7\\ 3.37\end{bmatrix}-\frac{1}{2}[78.7\,3.57]\begin{bmatrix}.373 & -.122\\ -.122 & 1.685\end{bmatrix}\begin{bmatrix}78.7\\ 3.57\end{bmatrix}=1016.6 \qquad (15.29b)$$

and

$$[71.3\,3.17]\begin{bmatrix}.373 & -.122\\ -.122 & 1.685\end{bmatrix}\begin{bmatrix}74.7\\ 3.37\end{bmatrix}-\frac{1}{2}[71.3\,3.17]\begin{bmatrix}.373 & -.122\\ -.122 & 1.685\end{bmatrix}\begin{bmatrix}71.3\\ 3.17\end{bmatrix}=1219.2 \qquad (15.29c)$$

the largest of which is Equation 15.28c, so that the assignment is made to Group 3. ◇

Graphical displays of the discriminant space such as that in Figure 15.3 can be quite useful for visualizing the separation of data groups. If $J = \min(G–1, K) > 2$, we cannot plot the full discriminant space in only two dimensions, but it is still possible to calculate and plot its first two components, δ_1 and δ_2. The relationships among the data groups rendered in this reduced discriminant space will be a good approximation to those in the full J-dimensional discriminant space, if the corresponding eigenvalues of Equation 15.20 are large relative to the eigenvalues of the omitted dimensions. Similarly to the idea expressed in Equation 13.4 for PCA, the reduced discriminant space will be a good approximation to the full discriminant space, to the extent that $(\lambda_1 + \lambda_2)/\sum_j \lambda_j \approx 1$.

15.3.2. Minimizing Expected Cost of Misclassification

The procedure described in Section 15.2.3, accounting for misclassification costs and prior probabilities of group memberships, generalizes easily for MDA. Again, if equal covariances for each of the G populations can be assumed, there are no other restrictions on the PDFs $f_g(\boldsymbol{x})$ for each of the populations except that that these PDFs can be evaluated explicitly. The main additional complication is to specify cost functions for all possible $G(G-1)$ misclassifications of Group g members into Group h,

$$C(h|\,g); \;\; g = 1, \ldots, G; \; h = 1, \ldots, G; \; g \neq h. \tag{15.30}$$

The resulting classification rule is to assign an observation $\boldsymbol{x}_0$ to the group g for which

$$\sum_{\substack{h=1 \\ h \neq g}}^{G} C(g|\,h)\, p_h f_h(\boldsymbol{x}_0) \tag{15.31}$$

is minimized. That is, the candidate Group g is selected for which the probability-weighted sum of misclassification costs, considering each of the other $G-1$ groups h as the potential true home of $\boldsymbol{x}_0$, is smallest. Equation 15.31 is the generalization of Equation 15.12 to $G \geq 3$ groups.

If all the misclassification costs are equal, minimizing Equation 15.31 simplifies to classifying $\boldsymbol{x}_0$ as belonging to that group g for which

$$p_g f_g(\boldsymbol{x}_0) \geq p_h f_h(\boldsymbol{x}_0), \quad \text{for all } h \neq g. \tag{15.32}$$

If in addition the PDFs $f_g(\boldsymbol{x})$ are all multivariate normal distributions, with possibly different covariance matrices $[\Sigma_g]$,(the logs of) the terms in Equation 15.32 take on the form

$$\ln(p_g) - \frac{1}{2} \ln\left|[S_g]\right| - \frac{1}{2}(\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_g)^T [S_g]^{-1} (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_g). \tag{15.33}$$

The observation $\boldsymbol{x}_0$ would be allocated to the group whose multinormal PDF $f_g(\boldsymbol{x})$ maximizes Equation 15.33. The unequal covariances $[S_g]$ result in this classification rule being quadratic. If all the covariance matrices $[\Sigma_g]$ are assumed equal and are estimated by $[S_{\text{pool}}]$, the classification rule in Equation 15.33 simplifies to choosing that Group g maximizing the linear discriminant score

$$\ln(p_g) + \overline{\boldsymbol{x}}_g^T [S_{pool}]^{-1} \boldsymbol{x}_0 - \frac{1}{2} \overline{\boldsymbol{x}}_g^T [S_{pool}]^{-1} \overline{\boldsymbol{x}}_g. \tag{15.34}$$

This rule minimizes the total probability of misclassification. If the prior probabilities p_g are all equal, Equation 15.43 reduces to Equation 15.24.

15.3.3. Probabilistic Classification

The classification rules presented so far choose one and only one of the G groups in which to place a new observation $\boldsymbol{x}_0$. Except for very easy cases, in which group means are well separated relative to the data scatter, these rules rarely will yield perfect results. Accordingly, probability information describing classification uncertainties is often useful.

Probabilistic classification, that is, specification of probabilities for $\boldsymbol{x}_0$ belonging to each of the G groups, can be achieved through an application of Bayes' Theorem:

$$\Pr\{\text{Group } g \mid \boldsymbol{x}_0\} = \frac{p_g f_g(\boldsymbol{x}_0)}{\sum_{h=1}^{G} p_h f_h(\boldsymbol{x}_0)}. \tag{15.35}$$

Here the p_g are the prior probabilities for group membership, which often will be the relative frequencies with which each of the groups is represented in the training data. The PDFs $f_g(\boldsymbol{x})$ for each of the groups can be of any form, so long as they can be evaluated explicitly for particular values of $\boldsymbol{x}_0$.

Often it is assumed that each of the $f_g(\boldsymbol{x})$ is multivariate normal distribution. In this case, Equation 15.35 becomes

$$\Pr\{\text{Group } g \mid \boldsymbol{x}_0\} = \frac{p_g \left(|[S_g]|^{-1/2} \exp\left(-\frac{1}{2} (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_g)^T [S_g]^{-1} (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_g) \right) \right)}{\sum_{h=1}^{G} p_h \left(|[S_h]|^{-1/2} \exp\left(-\frac{1}{2} (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_h)^T [S_h]^{-1} (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_h) \right) \right)}. \tag{15.36}$$

Equation 15.36 simplifies if all G of the covariance matrices are assumed to be equal, because in that case the factors involving determinants cancel. This equation also simplifies if all the prior probabilities are equal (i.e., $p_g = 1/G$, $g = 1, \ldots, G$), because these probabilities then cancel.

Example 15.3. Probabilistic Classification with $G = 3$ Groups

Consider probabilistic classification for Columbus, Ohio, into the three climate region groups of Example 15.2. The July mean vector for Columbus is $\boldsymbol{x}_0 = [74.7^\circ\text{F}, 3.37 \text{ in.}]^T$. Figure 15.3 shows that this point is near the boundary between the (nonprobabilistic) classification regions for Groups 2 (Central United States) and 3 (Northeastern United States) in the two-dimensional discriminant space, but the calculations in Example 15.2 do not quantify the certainty with which Columbus has been placed in Group 3.

Assume for simplicity that the three prior probabilities are equal, and that the three groups are all samples from multivariate normal distributions with a common covariance matrix. The pooled estimate for the common covariance is given in Equation 15.25, and its inverse is indicated in the middle equality of Equation 15.27a. The groups are then distinguished by their mean vectors, indicated in Table 15.1.

The differences between $\boldsymbol{x}_0$ and the three group means are

$$\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_1 = \begin{bmatrix} -5.90 \\ -2.30 \end{bmatrix}, \quad \boldsymbol{x}_0 - \overline{\boldsymbol{x}}_2 = \begin{bmatrix} -4.00 \\ -0.20 \end{bmatrix}, \text{and} \quad \boldsymbol{x}_0 - \overline{\boldsymbol{x}}_3 = \begin{bmatrix} 3.40 \\ 0.20 \end{bmatrix}; \tag{15.37a}$$

yielding the likelihoods

$$f_1(\boldsymbol{x}_0) \propto \exp\left(-\frac{1}{2} [-5.90 \;\; -2.30] \begin{bmatrix} .373 & -.122 \\ -.122 & 1.679 \end{bmatrix} \begin{bmatrix} -5.90 \\ -2.30 \end{bmatrix} \right) = .000094, \tag{15.37b}$$

$$f_2(\boldsymbol{x}_0) \propto \exp\left(-\frac{1}{2}[-4.00 \;\; -0.20]\begin{bmatrix} .373 & -.122 \\ -.122 & 1.679 \end{bmatrix}\begin{bmatrix} -4.00 \\ -0.20 \end{bmatrix}\right) = .054, \tag{15.37c}$$

and

$$f_3(\boldsymbol{x}_0) \propto \exp\left(-\frac{1}{2}[3.40 \;\; 0.20]\begin{bmatrix} .373 & -.122 \\ -.122 & 1.679 \end{bmatrix}\begin{bmatrix} 3.40 \\ 0.20 \end{bmatrix}\right) = .122. \tag{15.37d}$$

Substituting these likelihoods into Equation 15.36 yields the three classification probabilities

$$\Pr(\text{Group 1}\,|\,\boldsymbol{x}_0) = .000094/(.000094 + .054 + .122) = .0005, \tag{15.38a}$$

$$\Pr(\text{Group 2}\,|\,\boldsymbol{x}_0) = .054/(.000094 + .054 + .122) = 0.31, \tag{15.38b}$$

and

$$\Pr(\text{Group 3}\,|\,\boldsymbol{x}_0) = .122/(.000094 + .054 + .122) = 0.69. \tag{15.38c}$$

Even though the group into which Columbus was classified in Example 15.2 is the most likely, there is still a substantial probability that it might belong to Group 2 instead. The possibility that Columbus is really a Group 1 station appears to be remote. ◇

15.4. FORECASTING WITH DISCRIMINANT ANALYSIS

Discriminant analysis is a natural tool to use in forecasting when the predictand consists of a finite set of discrete categories (groups), and vectors of predictors $\boldsymbol{x}$ are known sufficiently far in advance of the discrete observation that will be predicted. Apparently the first use of discriminant analysis for forecasting in meteorology was described by Miller (1962), who forecast airfield ceiling in five MECE categories at a lead time of 0–2 h, and also made five-group forecasts of precipitation type (none, rain/freezing rain, snow/sleet) and amount ($\leq$0.05 in., and $>$0.05 in., if nonzero). Both of these applications today would be called *nowcasting*, because of the very short lead time. Some other examples of the use of discriminant analysis for forecasting can be found in Drosdowsky and Chambers (2001) and Ward and Folland (1991).

An informative case study in the use of discriminant analysis for forecasting is provided by Lehmiller et al. (1997). They consider the problem of forecasting hurricane occurrence (i.e., whether or not at least one hurricane will occur) during summer and autumn, within subbasins of the northwestern Atlantic Ocean, so that $G = 2$ for forecasts in each subbasin. They began with a quite large list of potential predictors and so needed to protect against overfitting in their $n = 43$-year training sample, 1950–1992. Their approach to predictor selection was computationally intensive, but statistically sound: different discriminant analyses were calculated for all possible subsets of predictors, and for each of these subsets the calculations were repeated 43 times, in order to produce leave-one-out cross-validations. The chosen predictor sets were those with the smallest number of predictors that minimized the number of cross-validated incorrect classifications.

Figure 15.4 shows one of the resulting discriminant analyses, for occurrence or nonoccurrence of hurricanes in the Caribbean Sea, using standardized African rainfall predictors that would be known as of 1 December in the preceding year. Because this is a binary forecast (two groups), there is only a single linear discriminant function, which would be perpendicular to the dividing line labeled "discriminant partition line" in Figure 15.4. This line compares to the long-short dashed dividing line in Figure 15.1. (The discriminant vector $\boldsymbol{a}$ would be perpendicular to this line and pass through the origin.)

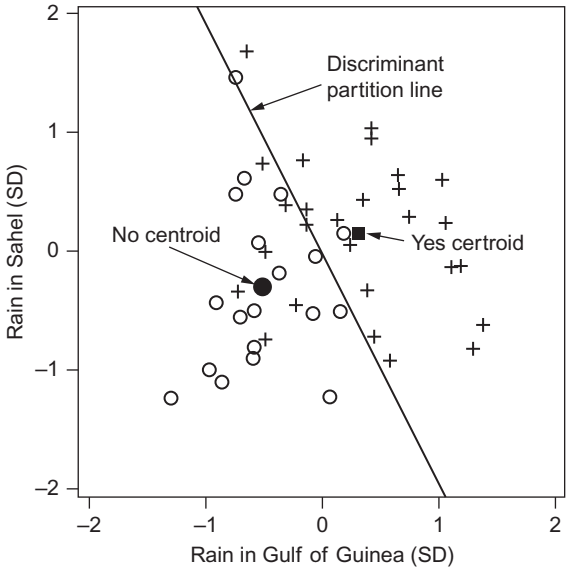

FIGURE 15.4 Binary (yes/no) forecasts for occurrence of at least one hurricane in the Caribbean Sea during summer and autumn, using two standardized predictors observed as of 1 December of the previous year to define a single linear discriminant function. Circles and plusses show the training data, and the two solid symbols locate the two group means (centroids). *From Lehmiller et al. (1997). © American Meteorological Society. Used with permission.*

The $n = 43$-year training sample is indicated by the open circles and plusses. Seven of the 18 hurricane years have been misclassified as "no" years, and only two of 25 nonhurricane years have been misclassified as "yes" years. Since there are more "yes" years, accounting for unequal prior probabilities would have moved the dividing line down and to the left, toward the "no" group mean (solid circle). Similarly, for some purposes it might be reasonable to assume that the cost of an incorrect "no" forecast would be greater than that of an incorrect "yes" forecast, and incorporating this asymmetry would also move the partition down and to the left, producing more "yes" forecasts.

15.5. CONVENTIONAL ALTERNATIVES TO CLASSICAL DISCRIMINANT ANALYSIS

Traditional discriminant analysis, as described in the first sections of this chapter, continues to be widely employed and extremely useful. Newer alternative approaches to discrimination and classification are also available. Two of these, based on conventional statistical methods, are described in this section. Two more modern "machine learning" alternatives are presented in Section 15.6.

15.5.1. Discrimination and Classification Using Logistic Regression

Section 7.6.2 described logistic regression, in which the nonlinear logistic function (Equation 7.36) is used to relate a linear combination of predictors, $\boldsymbol{x}$, to the probability of one of the elements of a dichotomous outcome. Figure 7.20 shows a simple example of logistic regression, in which the probability of occurrence of precipitation at Ithaca has been specified as a logistic function of the minimum temperature on the same day.

Figure 7.20 could also be interpreted as portraying classification into $G = 2$ groups, with $g = 1$ indicating precipitation days, and $g = 2$ indicating dry days. The number densities (points per unit length) of the dots along the top and bottom of the figure suggest the locations and shapes of the two underlying PDFs, $f_1(x)$ and $f_2(x)$, respectively, as functions of the minimum temperature, x. The medians of these two

conditional distributions for minimum temperature are near 23°F and 3°F, respectively. However, the classification function in this case is the logistic curve (solid), the equation for which is also given in the figure. Simply evaluating the function using the minimum temperature for a particular day provides an estimate of the probability that that day belonged to Group 1 (nonzero precipitation). A nonprobabilistic classifier could be constructed at the point of equal probability for the two groups, by setting the classification probability (= y in Figure 7.20) to 1/2. This probability is achieved when the argument of the exponential function is zero, implying a nonprobabilistic classification boundary of 15°F. In this case days could be classified as belonging to Group 1(wet) if the minimum temperature is warmer, and classified as belonging to Group 2 (dry) if the minimum temperature is colder. Seven days (the five warmest dry days, and the two coolest wet days) in the training data are misclassified by this rule. In this example the relative frequencies of the two groups are nearly equal, but logistic regression automatically accounts for relative frequencies of group memberships in the training sample (which estimate the prior probabilities) in the fitting process.

Figure 15.5 shows a forecasting example of two-group discrimination using logistic regression, with a $K = 2$-dimensional predictor vector $\boldsymbol{x}$. The two groups are years with (solid dots) and without (open circles) landfalling hurricanes on the southeastern U.S. coast from August onward, and the two elements of $\boldsymbol{x}$ are July average values of sea-level pressure at Cape Hatteras, and 200–700 mb wind shear over southern Florida. The contour lines indicate the shape of the logistic function, which in this case is a surface deformed into an S shape, analogously to the logistic function in Figure 7.20 being a line deformed in the same way. High surface pressures and wind shears simultaneously result in large probabilities for hurricane landfalls, whereas low values for both predictors yield small probabilities. This surface could be calculated as indicated in Equation 7.42, except that the vectors would be dimensioned (3×1) and the matrix of second derivatives would be dimensioned (3×3).

Hastie et al. (2009, Section 4.4.5) compare logistic regression and linear discriminant analyses for the $G = 2$-group situation, concluding that logistic regression may be more robust, but that the two generally give very similar results.

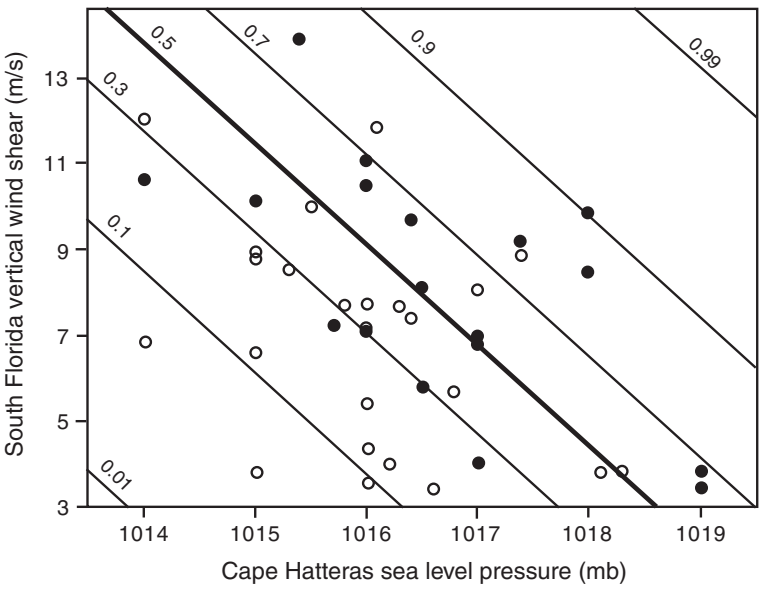

FIGURE 15.5 Two-dimensional logistic regression surface, estimating the probability of at least one landfalling hurricane on the southeastern U.S. coastline from August onward, on the basis of July sea-level pressure at Cape Hatteras and 200–700 mb wind shear over south Florida. Solid dots indicate hurricane years, and open dots indicate nonhurricane years, in the training data. *Adapted from Lehmiller et al. (1997). © American Meteorological Society. Used with permission.*

15.5.2. Discrimination and Classification Using Kernel Density Estimates

It was pointed out in Sections 15.2 and 15.3 that the G PDFs $f_g(\boldsymbol{x})$ need not be of particular parametric forms in order to implement Equations 15.12, 15.32, and 15.35, but rather it is necessary only that they can be evaluated explicitly. Gaussian or multivariate normal distributions often are assumed, but these and other parametric distributions may be poor approximations to data in some circumstances. Kernel density estimates (Section 3.3.6), which are nonparametric PDF estimates, provide viable alternatives. Indeed, nonparametric discrimination and classification motivated much of the early work on kernel density estimation (Silverman, 1986).

Nonparametric discrimination and classification are straightforward conceptually, but may be computationally demanding. The basic idea is to separately estimate the PDFs $f_g(\boldsymbol{x})$ for each of the G groups, using the methods described in Section 3.3.6. Somewhat subjective choices for appropriate kernel form and (especially) bandwidth are necessary. But having estimated these PDFs, they can be evaluated for any candidate $\boldsymbol{x}_0$, and thus lead to specific classification results. Nearest-neighbor classification methods, where membership probabilities are derived from relative frequencies of group memberships for the data vectors nearest the datum $\boldsymbol{x}_0$ to be classified, can be regarded as belonging to the class of kernel methods, where (hyper-) rectangular kernels are used to define the averaging.

Figure 15.6 illustrates the discrimination procedure for the same June Guayaquil temperature data (Table A.3) used in Figures 3.6 and 3.8. The distribution of these data is bimodal, as a consequence of four of the five El Niño years being warmer than 26°C whereas the warmest of the 15 non-El Niño years is 25.2°C. Discriminant analysis could be used to diagnose the presence or absence of El Niño, based on the June Guayaquil temperature, by specifying the two PDFs $f_1(x)$ for El Niño years and $f_2(x)$ for non-El Niño years. Parametric assumptions about the mathematical forms for these PDFs can be avoided through the use of kernel density estimates. The gray curves in Figure 15.6 show these two estimated PDFs. They exhibit fairly good separation, although $f_1(x)$, for El Niño years, is bimodal because the fifth El Niño year in the data set has a temperature of 24.8°C.

The posterior probability of an El Niño year as a function of the June temperature is calculated using Equation 15.35. The dashed curve is the result when equal prior probabilities $p_1 = p_2 = 1/2$ are assumed. Of course, El Niño occurs in fewer than half of all years, so it would be more reasonable to estimate the two prior probabilities as $p_1 = 1/4$ and $p_2 = 3/4$, which are the relative frequencies in the training sample. The resulting posterior probabilities are shown by the solid black curve in Figure 15.6.

Nonprobabilistic classification regions could be constructed using either Equation 15.12 or Equation 15.32, which would be equivalent if the two misclassification costs in Equation 15.12 were equal. If the two prior probabilities were also equal, the boundary between the two classification region

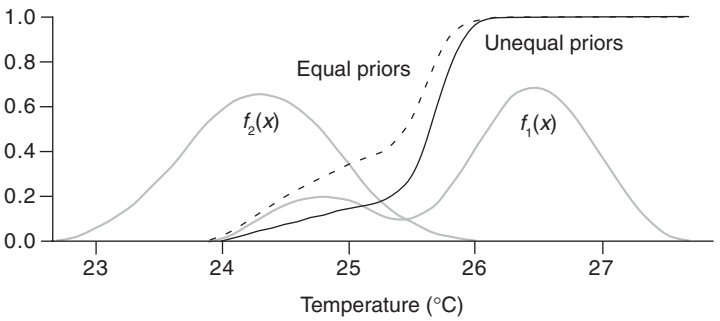

FIGURE 15.6 Separate kernel density estimates (quartic kernel, bandwidth = 0.92) for Guayaquil June temperatures during El Nino $f_1(x)$ and non-El Niño years $f_2(x)$, 1951–1970 (gray PDFs); and posterior probabilities for an El Niño year according to Equation 15.32, assuming equal prior probabilities (dashed), and prior probabilities estimated by the training-sample relative frequencies (solid).

would occur at the point where $f_1(x) = f_2(x)$, or $x \approx 25.45°C$. This temperature corresponds to a posterior probability of 1/2, according to the dashed curve. For unequal prior probabilities the classification boundary would shift toward the less likely group (i.e., requiring a warmer temperature to classify as an El Niño year), occurring at the point where $f_1(x) = (p_2/p_1)\, f_2(x) = 3\, f_2(x)$, or $x \approx 25.65$. Not coincidentally, this temperature corresponds to a posterior probability of 1/2 according to the solid black curve.

15.6. "MACHINE LEARNING" ALTERNATIVES TO CONVENTIONAL DISCRIMINANT ANALYSIS

15.6.1. Support Vector Machines (SVM)

Support Vector Machines (SVM) provide an approach to discrimination and classification that is similar to LDA, in that both define a separating hyperplane that partition a data space into regions assigned to one or the other of two groups. In SVM, this data space is generally expanded to include nonlinear transformations of the underlying data variables, which yields nonlinear classification boundaries in the original data space. (This strategy can also be used with MDA, as mentioned in Section 15.2.4.) The two methods differ primarily with respect to the criterion used to define the linear separator in the expanded data space. In addition, SVM is nonparametric in that it does not require specification or assumption of forms for the probability distribution(s) of the data, and so is an attractive alternative if the usual parametric assumptions in LDA seem doubtful. On the other hand, SVM does not provide a mechanism for probabilistic classification.

The idea behind SVM discrimination is most easily approached by considering the case of two linearly separable groups, as exemplified for $K = 2$ dimensions in Figure 15.7a. In such cases, any number of linear functions

$$f(\boldsymbol{x}) = b_0 + \boldsymbol{b}^T\boldsymbol{x} = 0 \tag{15.39}$$

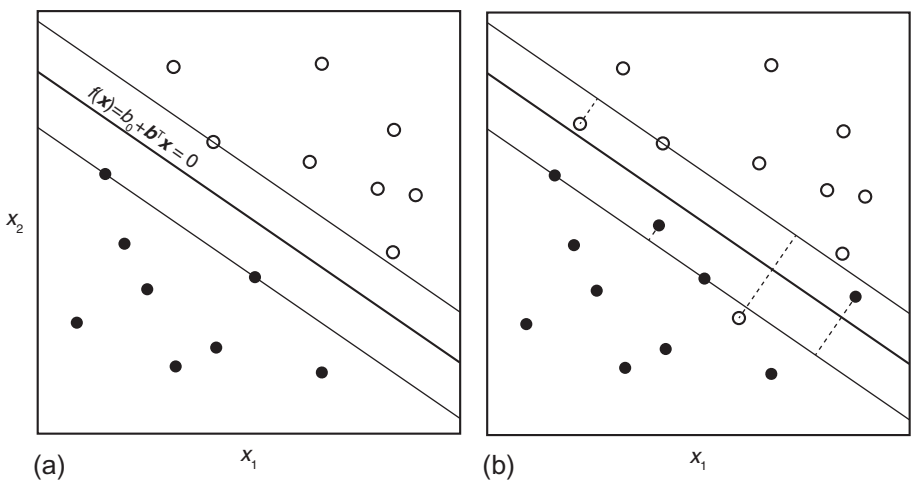

FIGURE 15.7 Illustration of (a) the maximum margin classifier, which is the support vector classifier when the two groups are linearly separable, and (b) the support vector classifier for two groups that are not linearly separable. The lengths of the light dashed lines in (b) indicate the magnitudes of the nonzero slack variables ε_i.

can be found to define lines (for $K=2$ dimensions, or surfaces for $K=3$, or hyperplanes for $K>3$) that fully separate the two groups, and thus can be used as classifiers. Classification of a data point $\boldsymbol{x}_i$ into one group or the other is coded as $y_i=\pm 1$, depending on which side of the partition it occurs:

$$y_i = \text{sign}\left(b_0 + \boldsymbol{b}^T\boldsymbol{x}_i\right) = \begin{cases} 1 \text{ if } f(\boldsymbol{x}_i) > 0 \\ -1 \text{ if } f(\boldsymbol{x}_i) < 0 \end{cases}. \tag{15.40}$$

The *maximum margin classifier* differs from the LDA separator in the case of linearly separable groups, in that the parameters defining Equation 15.39 are chosen to maximize the width of the margin around it. In other words, the maximum margin classifier finds the function in Equation 15.39 that yields the broadest possible buffer zone between the two groups in the training data. The margins bound the symmetrical zone around the classification function in Equation 15.39 (lighter lines in Figure 15.7 that are parallel to the separating function) that contains data points only on its boundary. These defining data vectors are called the *support vectors*.

Although only the values of the support vectors define the parameters in Equation 15.39, all the data are used in determining which points are the support vectors. The criterion for finding these parameters can be expressed as

$$\min_{b_0,\boldsymbol{b}} \|\boldsymbol{b}\|, \text{ subject to } y_i\left(b_0 + \boldsymbol{b}^T\boldsymbol{x}_i\right) \geq 1,\ i=1,\ldots,n, \tag{15.41}$$

where $||\bullet||$ denotes Euclidean length and n is the training sample size. Because all points in a linearly separable case can be correctly classified, necessarily $y_i(b_0+\boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}_i)>0$. Support vectors exactly on a margin edge will have $y_i(b_0+\boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}_i)=1$, and $y_i(b_0+\boldsymbol{b}^{\mathrm{T}}\mathbf{x}_i)>1$ for the other correctly classified points.

When two groups are not linearly separable, some points will necessarily lie on the wrong side of "their" margin, regardless of the parameters in Equation 15.39 and regardless of how thin the margin may be. In such cases the maximum margin classifier for linearly separable groups is relaxed to the *support vector classifier*, which allows some points in the training data to be misclassified, and also allows some correctly classified points to be inside their margin. These points are also support vectors, in addition to points that are exactly on their margin.

The degree of the intrusion into the margin of a correctly classified point, or degree of misclassification of a point on the wrong side of the separating hyperplane defined by Equation 15.39, is quantified by its corresponding *slack variable*, $\varepsilon_i \geq 0$. Any point on the correct side of its margin (i.e., those that are not support vectors), and support vectors that are exactly on their margin, have $\varepsilon_i=0$. Points within the margin but not misclassified have $0<\varepsilon_i<1$. Points exactly on the separating hyperplane have $\varepsilon_i=1$, and misclassified points have $\varepsilon_i>1$. The slack variables thus quantify the proportional amount by which $f(\boldsymbol{x}_i)$ falls on the wrong side of its margin. Figure 15.7b illustrates the idea, where four of the seven support vectors have nonzero slack variables, the magnitudes of which are indicated by the light dashed lines.

Extending Equation 15.41, the support vector classifier optimizes the parameters in Equation 15.39 according to

$$\min_{b_0,\boldsymbol{b}} \|\boldsymbol{b}\|, \text{ subject to } \begin{cases} y_i\left(b_0 + \boldsymbol{b}^T\boldsymbol{x}_i\right) \geq 1-\varepsilon_i,\ i=1,\ldots,n \\ \text{and} \qquad\qquad \sum_{i=1}^{n}\varepsilon_i \leq \text{constant} \end{cases}, \tag{15.42}$$

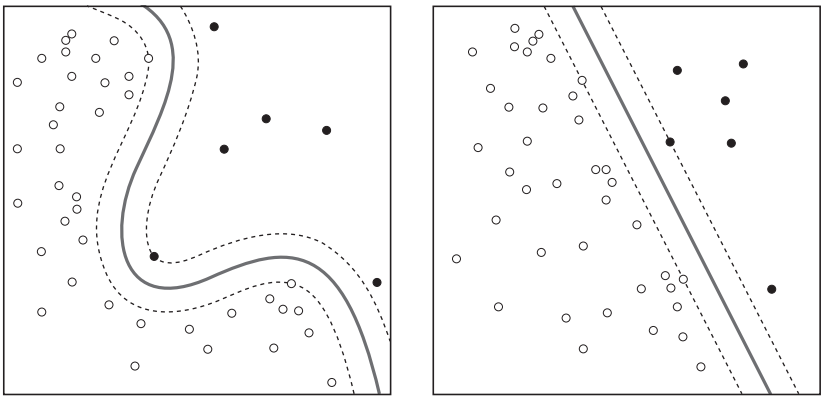

FIGURE 15.8 Illustration of (a) nonlinear classification boundary and associated margins, derived from (b) a linear maximum margin classifier in a higher-dimensional space that is rendered schematically in two dimensions. *From https://commons.wikimedia.org/wiki/File:Kernel_Machine.svg.*

where the magnitude of the adjustable constant limiting the sum of the slack variables is typically optimized by minimizing the number of misclassified training points through cross-validation. The width of the margin and the number of support vectors increase as this constant increases. Correctly classified points that are far from the separating hyperplane have little influence in defining it, which is another difference relative to LDA.

The support vector classifier is most often implemented by extending the dimension of the predictor vectors $\boldsymbol{x}$ through the addition of (possibly very many) nonlinear transformations of the elements of $\boldsymbol{x}$. This extended implementation is the SVM. A simple and straightforward possibility is to expand the predictor space using a polynomial basis. For example, in $K = 2$ elements the quadratic polynomial expansion basis would consist of the $K^* = 5$-dimensional vector $\boldsymbol{x}^* = [x_1, x_2, {x_1}^2, {x_2}^2, x_1 x_2]^T$. A linear support vector classifier in this extended space then maps to a nonlinear classifier in the original space. When $K^* + 1 \geq n$, a hyperplane fully separating two groups can always be found in the extended space, in which case the maximum margin classifier would be implemented. Figs. 15.8 illustrates the result, in which panel (a) shows a nonlinear classification boundary and associated margins in the original 2-dimensional data space, which has been derived from a linear maximum margin classifier in a higher-dimensional space that is shown schematically in two dimensions in panel (b).

Extension of SVM to multiclass ($G > 2$ groups) can be achieved by computing binary SVM classifiers for all G $(G–1)/2$ possible group pairs. Each datum $\boldsymbol{x}$ to be assigned to a group is classified by each of these, and the datum is assigned to the group for which it received the largest number of "votes."

Computational aspects for SVM can be elaborate. Details can be found in such references as Efron and Hastie (2016), Hastie et al. (2009), and Hsieh (2009).

15.6.2. Classification Trees and Neural Network Classifiers

Classification Trees

Classification trees can be viewed as an application of the regression trees described in Section 7.8.1, for which the predictands are discrete categorical variables rather than continuous real numbers. That is, the outcomes to be predicted are the integer-valued group indices $g = 1, 2, \ldots, G$. The two methods were originally proposed together, under the name Classification And Regression Trees, or CART

(Breiman et al., 1984). As is also the case for a regression tree, a classification tree is built by recursive binary partitions based on the K elements of the predictor variables, where at each stage the split is made that best separates the groups to be discriminated.

For regression trees, each binary split is chosen as the one that minimizes the combined sum of squares in Equation 7.56. This metric is not appropriate in the case of classification trees because the group indices are arbitrarily chosen as consecutive integers, and in addition there may be no natural ordering among the groups. Instead, the splitting criteria are based on proportion of predictand values in a given group g at a current or potential terminal node ℓ,

$$p_{\ell,g} = \frac{1}{n_\ell} \sum_{\boldsymbol{x}_i \in \ell} I(y_i \in g). \tag{15.43}$$

The next binary split is then chosen which minimizes the "node impurity," or lack of homogeneity of group membership within the resulting branches, according to either the overall cross-entropy

$$E[L] = -\sum_{\ell=1}^{L} \sum_{g=1}^{G} p_{\ell,g} \ln\left(p_{\ell,g}\right) \tag{15.44}$$

or the *Gini Index*

$$E[L] = \sum_{\ell=1}^{L} \sum_{g=1}^{G} p_{\ell,g}\left(1 - p_{\ell,g}\right), \tag{15.45}$$

where L is the number of branches or nodes after the currently contemplated binary split. The cross-entropy measure fails if one or more groups are absent from one or more of the nodes.

Nonprobabilistic classification is decided according to which group has the largest proportional membership in each terminal node or branch. Figure 15.9 shows a portion of an example classification

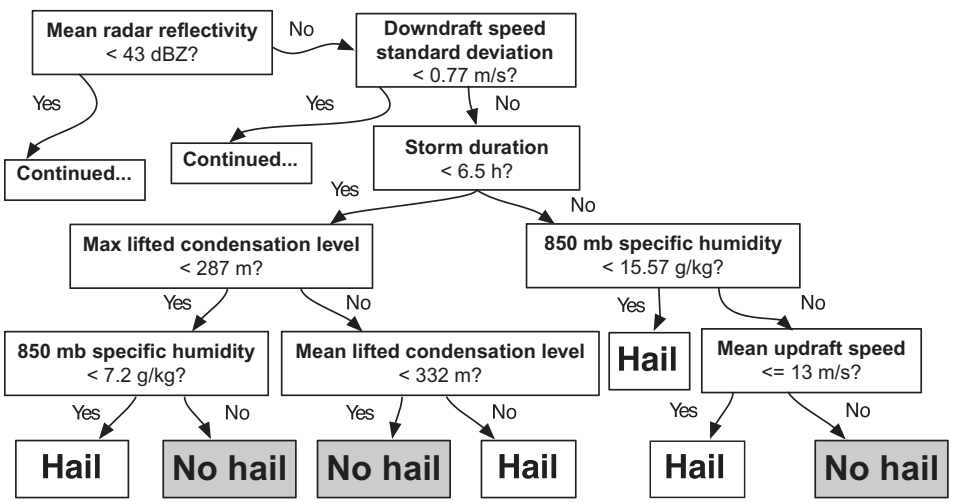

FIGURE 15.9 A portion of a classification tree yielding nonprobabilistic forecasts of hail occurrence or nonoccurrence. *From McGovern et al., 2017. © American Meteorological Society. Used with permission.*

tree, where the goal is to forecast one or the other of $G = 2$ future groups, pertaining to occurrence or nonoccurrence of hail. Alternatively, classification probabilities can be estimated using Equation 15.43 for each group $g = 1, 2, \ldots, G$, when the terminal node ℓ has been reached.

Because an individual classification tree is unstable, in the sense that small changes in the training data may yield substantial differences in the resulting tree structure, predictive performance can typically be improved by bootstrap aggregation, or bagging, of multiple trees, together with random predictor selection at each binary split. The result is a random forest, as described for regression trees in Section 7.8.1.

Neural Network Classifiers

Just as the framework of regression trees described in Section 7.8.1 can be used for classification by a discrete coding of the predictand variable, so also can neural networks (Section 7.8.2) be extended in an analogous way. As has already been noted in Section 7.8.2, use of a neural network for a classification problem usually involves choice of a suitable function for the output layer. In multicategory ($G > 2$) settings, often Equation 7.60 is used for this purpose. When only $G = 2$ groups are to be distinguished Equation 7.58 is another reasonable choice for the output function.

15.7. EXERCISES

15.1. Consider the two univariate PDFs $f_1(x) = 1 - |x|$, for $|x| \le 1$; and $f_2(x) = 1 - |x - 0.5|$, for $-0.5 \le x \le 1.5$.
   a. Sketch the two PDFs.
   b. Identify the classification regions when $p_1 = p_2$ and $C(1|2) = C(2|1)$.
   c. Identify the classification regions when $p_1 = 0.2$ and $C(1|2) = C(2|1)$.

15.2. Use Fisher's linear discriminant to classify years in Table A.3 as either El Niño or non-El Niño, on the basis of the corresponding temperature and pressure data.
   a. What is the discriminant vector, scaled to have unit length?
   b. Which, if any, of the El Niño years have been misclassified?
   c. Assuming bivariate normal distributions, repeat part (b) accounting for unequal prior probabilities.

15.3. Figure 15.4 illustrates a discriminant analysis relating two standardized African precipitation variables, Gulf of Guinea rainfall (x_1) and Sahel rainfall (x_2), to the occurrence of at least one hurricane in the Caribbean sea in the following year (Y or N). The sample mean vectors and pooled variance estimates are

$$\overline{\boldsymbol{x}}_Y = [0.29 \;\; 0.16]^T, \; \overline{\boldsymbol{x}}_N = [-0.50 \;\; -0.32]^T, \text{ and } [S_{pool}] = \begin{bmatrix} 1.000 & 0.154 \\ 0.154 & 1.000 \end{bmatrix}.$$

   a. Find the discriminant vector, using Fisher's procedure.
   b. Use Anderson's classification statistic to allocate the new observation $\boldsymbol{x}_0 = [1/2 \;\; 1/2]^T$, assuming equal misclassification costs and equal priors.
   c. Assuming that both $f_Y(\boldsymbol{x})$ and $f_N(\boldsymbol{x})$ are bivariate normal with equal covariance matrices, evaluate the posterior probability of at least one Caribbean hurricane (i.e., find Pr{"yes"}) in the following year, given $\boldsymbol{x}_0 = [1/2 \;\; 1/2]^T$.

TABLE 15.2 Likelihoods Calculated from the Forecast Verification Data for Subjective 12–24 h Projection Probability-of-Precipitation Forecasts for the United States During October 1980–March 1981, in Table 9.2

y_i	0.00	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
$p(y_i \mid o_1)$	0.0152	0.0079	0.0668	0.0913	0.1054	0.0852	0.0956	0.0997	0.1094	0.1086	0.0980	0.1169
$p(y_i \mid o_2)$	0.4877	0.0786	0.2058	0.1000	0.0531	0.0272	0.0177	0.0136	0.0081	0.0053	0.0013	0.0016

15.4. Average July temperature and precipitation at Ithaca, New York, are 68.6°F and 3.54 in.
   a. Classify Ithaca as belonging to one of the three groups in Example 14.2.
   b. Calculate probabilities that Ithaca is a member of each of the three groups, assuming bivariate normal distributions with common covariance matrix.

15.5. Using the forecast verification data in Table 9.2, we can calculate the likelihoods (i.e., conditional probabilities for each of the 12 possible forecasts, given either precipitation or no precipitation) in Table 15.2. The unconditional probability of precipitation is $p(o_1) = 0.162$. Considering the two precipitation outcomes as two groups to be discriminated between, calculate the posterior probabilities of precipitation if the forecast probability, y_i, is.
   a. 0.00.
   b. 0.10.
   c. 1.00.