Chapter 8

Discriminant Analysis

8.1 Introduction

Definition 8.1. In *supervised classification*, there are k known groups and m cases. Each case is assigned to exactly one group based on its measurements $\boldsymbol{w}_i$.

Suppose there are k populations or groups where $k \geq 2$. Assume that for each population there is a probability density function (pdf) $f_j(\boldsymbol{z})$ where $\boldsymbol{z}$ is a $p \times 1$ vector and $j = 1, ..., k$. Hence if the random vector $\boldsymbol{x}$ comes from population j, then $\boldsymbol{x}$ has pdf $f_j(\boldsymbol{z})$. Assume that there is a random sample of n_j cases $\boldsymbol{x}_{1,j}, ..., \boldsymbol{x}_{n_j,j}$ for each group. Let $(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ denote the sample mean and covariance matrix for each group. Let $\boldsymbol{w}_i$ be a new $p \times 1$ random vector from one of the k groups, but the group is unknown. Usually there are many $\boldsymbol{w}_i$, and *discriminant analysis* attempts to allocate the $\boldsymbol{w}_i$ to the correct groups.

Definition 8.2. The *maximum likelihood discriminant rule* allocates case $\boldsymbol{w}$ to group a if $\hat{f}_a(\boldsymbol{w})$ maximizes $\hat{f}_j(\boldsymbol{w})$ for $j = 1, ..., k$.

For the following rules, assume that costs of correct and incorrect allocation are unknown or equal, and assume that the probabilities $\rho_a(\boldsymbol{w}_i)$ that $\boldsymbol{w}_i$ is in group a are unknown or equal: $\rho_a(\boldsymbol{w}_i) = 1/k$ for $a = 1, ..., k$. Often it is assumed that the k groups have the same covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}$. Then the pooled covariance matrix estimator is

$$\boldsymbol{S}_{pool} = \frac{1}{n-k} \sum_{j=1}^{k} (n_j - 1) \boldsymbol{S}_j$$

where $n = \sum_{j=1}^k n_j$. Let $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)$ be the estimator of multivariate location and dispersion for the jth group, eg the sample mean and sample covariance matrix $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$.

Definition 8.3. Assume the population dispersion matrices are equal: $\boldsymbol{\Sigma}_j \equiv \boldsymbol{\Sigma}$ for $j = 1, ..., k$. Let $\hat{\boldsymbol{\Sigma}}_{pool}$ be an estimator of $\boldsymbol{\Sigma}$. Then the *linear discriminant rule* is allocate $\boldsymbol{w}$ to the group with the largest value of

$$d_j(\boldsymbol{w}) = \hat{\boldsymbol{\mu}}_j^T \hat{\boldsymbol{\Sigma}}_{pool}^{-1} \boldsymbol{w} - \frac{1}{2} \hat{\boldsymbol{\mu}}_j^T \hat{\boldsymbol{\Sigma}}_{pool}^{-1} \hat{\boldsymbol{\mu}}_j = \hat{\alpha}_j + \hat{\boldsymbol{\beta}}_j^T \boldsymbol{w}$$

where $j = 1, ..., k$. *Linear discriminant analysis* (LDA) uses $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_{pool}) = (\overline{\boldsymbol{x}}_j, \boldsymbol{S}_{pool})$.

Definition 8.4. The *quadratic discriminant rule* is allocate $\boldsymbol{w}$ to the group with the largest value of

$$Q_j(\boldsymbol{w}) = \frac{-1}{2} \log(|\hat{\boldsymbol{\Sigma}}_j|) - \frac{1}{2} (\boldsymbol{w} - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1} (\boldsymbol{w} - \hat{\boldsymbol{\mu}}_j)$$

where $j = 1, ..., k$. *Quadratic discriminant analysis* (QDA) uses $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$.

Definition 8.5. The *distance discriminant rule* allocates $\boldsymbol{w}$ to the group with the smallest squared distance $D^2_{\boldsymbol{w}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\boldsymbol{w} - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1} (\boldsymbol{w} - \hat{\boldsymbol{\mu}}_j)$ where $j = 1, ..., k$.

Definition 8.6. Assume that $k = 2$ and that there is a group 0 and a group 1. Let $\rho(\boldsymbol{w}) = P(\boldsymbol{w} \in \text{group } 1)$. Let $\hat{\rho}(\boldsymbol{w})$ be the logistic regression estimate of $\rho(\boldsymbol{w})$. The *logistic regression discriminant rule* allocates $\boldsymbol{w}$ to group 1 if $\hat{\rho}(\boldsymbol{w}) \geq 0.5$ and allocates $\boldsymbol{w}$ to group 0 if $\hat{\rho}(\boldsymbol{w}) < 0.5$. Logistic regression produces an estimated sufficient predictor $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{w}$. Then

$$\hat{\rho}(\boldsymbol{w}) = \frac{e^{ESP}}{1 + e^{ESP}} = \frac{\exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{w})}{1 + \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{w})}.$$

Let $Y_i = j$ if case i is in group j for $j = 0, 1$. Then a *response plot* is a plot of ESP versus Y_i (on the vertical axis) with $\hat{\rho}(\boldsymbol{x}_i) \equiv \hat{\rho}(ESP)$ added as a visual aid where $\boldsymbol{x}_i$ is the vector of predictors for case i. Also divide the ESP into J slices with approximately the same number of cases in each slice. Then compute the sample mean = sample proportion in slice s: $\hat{\rho}_s = \overline{Y}_s = \sum_s Y_i/m_s$ where m_s is the number of cases in slice s. Then plot the resulting step function as a visual aid. If n_0 and n_1 are the sample sizes of both groups and $n_i > 5p$, then the logistic regression model was useful if the step function of observed slice proportions scatter fairly closely about the logistic curve $\hat{\rho}(ESP)$.

Examining some of the rules for $k = 2$ and one predictor w is informative. First, assume group 2 has a uniform(-10,10) distribution and group 1 has a uniform$(a-1, a+1)$ distribution. If $a = 0$ is known, then the maximum likelihood discriminant rule assigns w to group 1 if $-1 < w < 1$ and assigns w to group 2, otherwise. This occurs since $f_2(w) = 1/20$ for $-10 < w < 10$ and $f_2(w) = 0$, otherwise, while $f_1(w) = 1/2$ for $-1 < w < 1$ and $f_1(w) = 0$, otherwise. For the distance rule, the distances are basically the absolute value of the z-score. Hence $D_1(w) \approx 1.732|w - a|$ and $D_2(w) \approx 0.1732|w|$. If w is from group 1, then w will not be classified very well unless $|a| \geq 10$ or if w is very close to a. In particular, if $a = 0$ then expect nearly all w to be classified to group 2 if w is used to classify the groups. On the other hand, if $a = 0$, then $D_1(w)$ is small for w in group 1 but large for w in group 2. Hence using $z = D_1(w)$ in the distance rule would result in classification with low error rates.

Similarly if group 2 comes from a $N_p(\boldsymbol{0}, 10\boldsymbol{I}_p)$ distribution and group 1 comes from a $N_p(\boldsymbol{\mu}, \boldsymbol{I}_p)$ distribution, the maximum likelihood rule will tend to classify $\boldsymbol{w}$ in group 1 if $\boldsymbol{w}$ is close to $\boldsymbol{\mu}$ and to classify $\boldsymbol{w}$ in group 2 otherwise. The two misclassification error rates should both be low. For the distance rule, the distances D_i have an approximate χ^2_p distribution if $\boldsymbol{w}$ is from group i. If covering ellipsoids from the two groups have little overlap, then the distance rule does well. If $\boldsymbol{\mu} = \boldsymbol{0}$, then expect all $\boldsymbol{w}$ to be classified to group 2 with the distance rule, but $D_1(\boldsymbol{w})$ will be small for $\boldsymbol{w}$ from group 1 and large for $\boldsymbol{w}$ from group 2, so using the single predictor $z = D_1(\boldsymbol{w})$ in the distance rule would result in classification with low error rates. More generally, if group 1 has a covering ellipsoid that has little overlap with the observations from group 2, using the single predictor $z = D_1(\boldsymbol{w})$ in the distance rule should result in classification with low error rates even if the observations from group 2 do not fall in an ellipsoidal region.

Now suppose the k groups come from the same family of elliptically contoured $EC(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, g)$ distributions where g is a decreasing function that does not depend on j for $j = 1, ..., k$. For example, could have $\boldsymbol{w} \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. Using Equation (3.5), $\log(f_j(\boldsymbol{w})) =$

$$\log(k_p) - \frac{1}{2}\log(|\boldsymbol{\Sigma}_j)|) + \log(g[(\boldsymbol{w} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{w} - \boldsymbol{\mu}_j)]) =$$

$$\log(k_p) - \frac{1}{2}\log(|\boldsymbol{\Sigma}_j)|) + \log(g[D^2_{\boldsymbol{w}}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)]).$$

Hence the maximum likelihood rule leads to the quadratic rule if the k groups have $N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ distributions, and the maximum likelihood rule leads to the distance rule if the groups have dispersion matrices that have the same determinant: $\det(\boldsymbol{\Sigma}_j) = |\boldsymbol{\Sigma}_j| \equiv |\boldsymbol{\Sigma}|$ for $j = 1, ..., k$. This is a much weaker assumption that of equal dispersion matrices. For example, let $c_X \boldsymbol{\Sigma}_j$ be the covariance matrix of $\boldsymbol{x}$, and let $\boldsymbol{\Gamma}_j$ be an orthogonal matrix. Then $\boldsymbol{y} = \boldsymbol{\Gamma}_j \boldsymbol{x}$ corresponds to rotating $\boldsymbol{x}$, and $c_X \boldsymbol{\Gamma}_j \boldsymbol{\Sigma}_j \boldsymbol{\Gamma}_j^T$ is the covariance matrix of $\boldsymbol{y}$ with $|\mathrm{Cov}(\boldsymbol{x})| = |\mathrm{Cov}(\boldsymbol{y})|$.

Note that if the k groups come from the same family of elliptically contoured $EC(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, g)$ distributions with nonsingular covariance matrices $c_X \boldsymbol{\Sigma}_j$, then $D^2_{\boldsymbol{w}}(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ is a consistent estimator of $D^2_{\boldsymbol{w}}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)/c_X$. Hence the distance rule using $(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ is a maximum likelihood rule if the $\boldsymbol{\Sigma}_j$ have the same determinant.

Now $D^2_{\boldsymbol{w}}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = \boldsymbol{w}^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{w} - \boldsymbol{w}^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\mu}_j - \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{w} + \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\mu}_j = \boldsymbol{w}^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{w} - 2\boldsymbol{\mu}_j^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{w} + \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\mu}_j = \boldsymbol{w}^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{w} + \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}_j^{-1}(-2\boldsymbol{w} + \boldsymbol{\mu}_j)$. Hence if $\boldsymbol{\Sigma}_j \equiv \boldsymbol{\Sigma}$ for $j = 1, ..., k$, then want to minimize $\boldsymbol{\mu}_j^T \boldsymbol{\Sigma}_j^{-1}(-2\boldsymbol{w} + \boldsymbol{\mu}_j)$ or maximize $\boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1}(2\boldsymbol{w} - \boldsymbol{\mu}_j)$, which is leads to the linear discriminant rule.

The maximum likelihood rule is robust to nonnormality, but it is difficult to estimate $\hat{f}_j(\boldsymbol{w})$ if $p > 1$. The linear discriminant rule and distance rule are robust to nonnormality, as is the logistic regression discriminant rule if $k = 2$. Expect the distance rule to be best when the ellipsoidal covering regions of the k groups have little overlap.

Rule of thumb 8.1. Use the distance rule if $n_j > 10p$ for $j = 1, ..., k$. Make the k DD plots using the $\boldsymbol{x}_{i,j}$ for each group to check for outliers, which could be cases that were incorrectly classified. If the distance rule error rates are very poor for some groups and very good for others, compute $z_j = D_j$, the distances for all n cases based on the jth group, where $j = 1, ..., k$. Since the

z_j may be highly correlated, use no more than $k-1$ of the z_j as predictors. The error rates computed using the data $\boldsymbol{x}_{i,j}$ with known groups give a lower bound on the error rates for the $\boldsymbol{w}_i$. That is, the error rates computed on the training data $\boldsymbol{x}_{i,j}$ are optimistic. When the discriminant rule is applied to the m $\boldsymbol{w}_i$ where the groups are unknown, the error rates will be higher. If equal covariance matrices are assumed, plot $D_i(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ versus $D_i(\overline{\boldsymbol{x}}_j, \boldsymbol{\Sigma}_{pool})$ for each of the k groups, where the $\boldsymbol{x}_{i,j}$ are used for $i = 1, ..., n_j$. The plotted points should cluster tightly about the identity line if n_j is large in each of the k plots if the assumption is reasonable. The linear discriminant rule has some robustness against the assumption of equal covariance matrices.

8.2 Two New Methods

Assume the k groups come from k distributions where the prediction regions from Section 5.2 are reasonable. For example, the jth group may have n_j cases that are iid $EC_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, g_j)$ for $j = 1, ..., k$. That is, there may be k different elliptically contoured distributions with different location vectors and dispersion matrices.

Two new methods of discriminant analysis will be considered. For each group, compute $D_i(j) \equiv D_i(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ and the maximum distance $D_{(n_j)}(j)$ where $i = 1, ..., n_j$ and $j = 1, ..., k$. Then $\{\boldsymbol{z} : D_{\boldsymbol{z}}(j) \leq D_{(n_j)}(j)\}$ is a covering region for the jth group since the hyperellipsoid contains all n_j cases $\boldsymbol{x}_{i,j}$ from the jth group.

Let $\boldsymbol{w}$ be a new case to be classified. If $D_{\boldsymbol{w}}(j) > D_{(n_j)}(j)$ for all $j = 1, ..., k$, then both Methods 1 and 2 allocate $\boldsymbol{w}$ to the group a with the smallest value of

$$\frac{D_{\boldsymbol{w}}(j)}{D_{(n_j)}(j)}. \tag{8.1}$$

Now consider the groups where $D_{\boldsymbol{w}}(j) \leq D_{(n_j)}(j)$ for at least one j. Hence $\boldsymbol{w}$ is in at least one of the k covering regions.

For Method 1, allocate $\boldsymbol{w}$ to group a with the smallest $D_{\boldsymbol{w}}(a)$ for the groups with $D_{\boldsymbol{w}}(j) \leq D_{(n_j)}(j)$. Method 1 is very similar to the distance rule, but when $\boldsymbol{w}$ is in at least one of the k covering regions, distances are only computed for the groups that have covering regions that contain $\boldsymbol{w}$. Also, Equation (8.1) is used instead of the smallest distance if $\boldsymbol{w}$ is not in any of the k covering regions.

Method 2 combines Method 1 with a maximum likelihood rule based on a kernel density estimator of $\hat{f}_j$. For Method 2, if there is only one group a where $D_{\boldsymbol{w}}(a) \leq D_{(n_a)}(a)$, allocate $\boldsymbol{w}$ to group a. Otherwise compute $\hat{f}_j(\boldsymbol{w})$ for the groups where $D_{\boldsymbol{w}}(j) \leq D_{(n_j)}(j)$ and allocate $\boldsymbol{w}$ to the group a with the largest $\hat{f}_a(\boldsymbol{w})$.

Note: To find the z_j of Rule of thumb 8.1, find $D_h(j)$ using all n of the $\boldsymbol{x}_{i,j}$, eg stack the $\boldsymbol{x}_{i,j}$ into an $n \times 1$ vector $\boldsymbol{x}$ and compute the $D_h(j)$ for $h = 1, ..., n$. These k new predictor variables still have known groups. Find $D_{\boldsymbol{w}_i}(j)$ for $i = 1, ..., m$ and $j = 1, ..., k$ to create k new predictor variables for the ith case to be classified. Then input up to $k - 1$ of these variables, with or without some of the p original predictor variables, into Method 1 or 2. Section 8.3 will give an example.

8.2.1 The Kernel Density Estimator

Definition 8.7. Let $K(\boldsymbol{z})$ be a multivariate probability density function. Then a *kernel density estimator* is

$$\hat{f}(\boldsymbol{z}) = \frac{1}{n} \ \frac{1}{h^p} \ \sum_{i=1}^{n} K\left(\frac{1}{h}(\boldsymbol{z} - \boldsymbol{x}_i)\right)$$

where there are n iid cases $\boldsymbol{x}_i$ that come from a population with unknown pdf $f(\boldsymbol{z})$.

For example, the uniform distribution on the unit hypersphere has

$$K(\boldsymbol{z}) = \frac{p\Gamma(p/2)}{2\pi^{p/2}} I(\boldsymbol{z}^T\boldsymbol{z} \leq 1)$$

so

$$\hat{f}(\boldsymbol{z}) = \frac{p\Gamma(p/2)}{2\pi^{p/2}} \ \frac{1}{n} \ \frac{1}{h^p} \ \sum_{i=1}^{n} I(\|\boldsymbol{z} - \boldsymbol{x}_i\|^2 \leq h^2).$$

Following Silverman (1986, p. 84), want the bias and variance of $\hat{f}$ to go to 0 as $n \to \infty$, and this will happen if $h \to 0$ and $nh^p \to \infty$. The asymptotically optimal value of h satisfies $h_{opt} \propto \dfrac{1}{n^{\frac{1}{p+4}}}$.

Now suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid from a multivariate distribution with pdf f, and consider a hypersphere of radius r centered at $\boldsymbol{w}$ where r is small enough so that if $\boldsymbol{z}$ is in the hypersphere, then $f(\boldsymbol{z}) \approx f(\boldsymbol{w})$. Then the probability that an observation $\boldsymbol{x}_i$ falls in the hypersphere $\approx f(\boldsymbol{w})$ (volume of the hypersphere) $= f(\boldsymbol{w}) \dfrac{2\pi^{p/2}}{p\Gamma(p/2)} r^p \propto r^p$. Hence the number of $\boldsymbol{x}_i$ in the hypersphere $\propto nr^p$. If $r = h_{opt}$ then this number is $\propto n^{\frac{4}{4+p}}$. If $r = h \propto n^{\frac{1}{2p}}$, then the number of cases that fall in the hypersphere is proportional to $\sqrt{n}$.

To define the kernel density estimator used in Method 2, let $v_j = \lceil 2\sqrt{n_j} \rceil$ and let $r_j^2 = \|\boldsymbol{x}_{i,j} - \overline{\boldsymbol{x}}_j\|^2_{(v_j)} = D^2_{(v_j)}(\overline{\boldsymbol{x}}_j, \boldsymbol{I}_p)$ where the n_h $\boldsymbol{x}_{i,j}$ are in group j. Hence the hypersphere centered at $\overline{\boldsymbol{x}}_j$ with radius r_j contains $\approx 2\sqrt{n}$ of the $\boldsymbol{x}_{i,j}$ in group j. Then the kernel density estimator used in Method 2 is

$$\hat{f}_j(\boldsymbol{w}) = \frac{p\Gamma(p/2)}{2\pi^{p/2}} \; \frac{1}{n_j} \; \frac{1}{(r_j)^p} \; \sum_{i=1}^{n_j} I(\|\boldsymbol{w} - \boldsymbol{x}_{i,j}\|^2 \leq r_j^2)$$

which is equal to the number of the $\boldsymbol{x}_{i,j}$ in the hypersphere of radius r_j centered at $\boldsymbol{w}$ divided by $n_j V_{r_j}$ where V_{r_j} is the volume of the hypersphere.

The main reasons for using this kernel density estimator are that it is simple to explain, fast to compute and does not use too few observations when $p > 4$. Since kernel density estimators do not work well for $p > 1$, speed is more important than asymptotic optimality. Also only need a crude estimator since if $f_a(\boldsymbol{w})$ is the pdf that maximizes $f_j(\boldsymbol{w})$, only need $\hat{f}_a(\boldsymbol{w})$ to maximize the $\hat{f}_j(\boldsymbol{w})$: hence extremely accurate estimators of the $f_j(\boldsymbol{w})$ are not needed. Using good predictors with p small is important since the performance of kernel density estimators decreases very rapidly as the number of predictors increases. See Silverman (1986, p. 94).

8.3 Some Examples

The *mpack* functions `ddiscr` and `ddiscr2` do discriminant analysis using Methods 1 and 2. The functions need x: the training data that has been classified into k groups, w: the data to be classified, *group*: a vector of integers where the ith element is j if the ith element of x is in group j, and $xwflag$ which is set equal to T if $w = x$ and to F if $w \neq x$. Each row of w and x corresponds to a case. The functions return the distances of the $\boldsymbol{x}$ and $\boldsymbol{w}$ computed for the k groups, the classifications for the $\boldsymbol{x}$ and $\boldsymbol{w}$, the error rates for the $\boldsymbol{x}$ classifications for each group, and the total error rate.

Example 8.1. Generated n random $N_p(\mathbf{0}, \boldsymbol{I}_p)$ random variables $\boldsymbol{x}_i$. Then $\boldsymbol{x}$ was put in group 1 if $D^2_{\boldsymbol{x}_i} \le \chi^2_{p,0.5}$ and in group 2 otherwise. Expect group 2 to have smaller distances than group 1 so error rate will be near 1 for group 1 and near 0 for group 2. Output is shown below with $p = 2$ and shows that this was the case. Then the predictor $D_i(1)$ was used in *out2*, reducing the dimension from $p = 2$ to 1. The error rates were low since group 1 falls in an ellipsoidal region so the distances are a good predictor. Method 2 worked much better on the raw data and about the same as Method 1 when the predictor $D_i(1)$ was used.

```
n <- 100
p <- 2
x <- matrix(rnorm(n*p),nrow=n,ncol=p)
group <- 1 + 0*1:n
covv <- diag(p)
mns<- apply(x, 2, mean)
md2 <- mahalanobis(x, center = mns, covv)
group[md2>qchisq(0.5,p)] <- 2

out1 <- ddiscr(x,w=x,group,xwflag=T)
out2<-ddiscr(x=out1$mdx[,1],w=out1$mdw[,1],group,xwflag=T)
out3 <- ddiscr2(x,w=x,group,xwflag=T)
out4<-ddiscr2(x=out1$mdx[,1],w=out1$mdw[,1],group,xwflag=T)

out1$err
[1] 0.9787234 0.0000000
out2$err
[1] 0.08510638 0.01886792
out3$err
[1] 0.0000000 0.1320755
out4$err
[1] 0.04255319 0.05660377

out1$toterr
[1] 0.46
out2$toterr
[1] 0.05
out3$toterr
```

```
[1] 0.07
out4$toterr
[1] 0.05
```

Example 8.2. Now groups 1 and 2 had $n_i = 50$, and group 1 used $\boldsymbol{x} \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ while group 2 used $\boldsymbol{x} \sim N_p(2 \ \ \boldsymbol{1},\ \boldsymbol{I}_p)$. Output is shown below for $p = 2$. Now the single predictor $D_i^2(1)$ was slightly worse than using the raw data, and Method 1 was about as good as Method 2, which is not surprising since both methods approximate the maximum likelihood discriminant rule when the groups are multivariate normal with the same covariance matrix.

```
 n <- 100
 p <- 2
 x <- matrix(rnorm(n*p),nrow=n,ncol=p)
 group <- 1 + 0*1:n
 group[1:50] <- 1
 group[51:100] <- 2
 x[51:100,] <- x[51:100,] + c(2,2)
 out1 <- ddiscr(x,w=x,group,xwflag=T)
 out2<-ddiscr(x=out1$mdx[,1],w=out1$mdw[,1],group,xwflag=T)
 out3 <- ddiscr2(x,w=x,group,xwflag=T)
 out4<-ddiscr2(x=out1$mdx[,1],w=out1$mdw[,1],group,xwflag=T)

out1$err
[1] 0.12 0.08
out2$err
[1] 0.14 0.10
out3$err
[1] 0.08 0.12
out4$err
[1] 0.14 0.10

library(MASS)
group <- pottery[pottery[,1]!=5,1]
group <- (as.integer(group!=1)) + 1
x <- pottery[pottery[,1]!=5,-1]
```

```
out<-lda(x,group)
1-mean(predict(out,x)$class==group)
[1] 0.03571429
out<-lda(x[,-c(1)],group)
1-mean(predict(out,x[,-c(1)])$class==group)
out<-lda(x[,-c(1,2)],group)
1-mean(predict(out,x[,-c(1,2)])$class==group)
out<-lda(x[,-c(1,2,3)],group)
1-mean(predict(out,x[,-c(1,2,3)])$class==group)
out<-lda(x[,-c(1,2,3,4)],group)
1-mean(predict(out,x[,-c(1,2,3,4)])$class==group)
out<-lda(x[,-c(1,2,3,4,5)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5)])$class==group)
[1] 0.03571429
out<-lda(x[,-c(1,2,3,4,5,6)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,6)])$class==group)
[1] 0.07142857
out<-lda(x[,-c(1,2,3,4,5,7)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7)])$class==group)
out<-lda(x[,-c(1,2,3,4,5,7,8)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8)])$class==group)
out<-lda(x[,-c(1,2,3,4,5,7,8,9)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9)])$class==group)
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10)])$class==group)
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,11)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,11)])$class==group)
[1] 0.07142857
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,12)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,12)])$class==group)
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,12,13)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,12,13)])$class==group)
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,12,13,14)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,12,13,14)])$class==
group)
[1] 0.07142857
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,12,13,15)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,12,13,15)])$class==
group)
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,12,13,15,16)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,12,13,15,16)])$
class==group)
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,12,13,15,16,17)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,12,13,15,16,17)])$
class==group)
[1] 0.03571429
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,12,13,15,16,17,18)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,12,13,15,16,17,18)])
$class==group)
[1] 0.07142857
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,12,13,15,16,17,19)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,12,13,15,16,17,19)])
$class==group)
[1] 0.03571429
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,12,13,15,16,17,19,20)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,12,13,15,16,17,19,
20)])$class==group)
[1] 0

#x6,x11,x14,x18 seem good for LDA
```

8.4 Summary

1) In *supervised classification*, there are k known groups or populations and m cases. Each case is assigned to exactly one group based on its measurements $\boldsymbol{w}_i$. Assume that for each population there is a probability density function (pdf) $f_j(\boldsymbol{z})$ where $\boldsymbol{z}$ is a $p \times 1$ vector and $j = 1, ..., k$. Hence if the random vector $\boldsymbol{x}$ comes from population j, then $\boldsymbol{x}$ has pdf $f_j(\boldsymbol{z})$. Assume that there is a random sample of n_j cases $\boldsymbol{x}_{1,j}, ..., \boldsymbol{x}_{n_j,j}$ for each group. Let $(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ denote the sample mean and covariance matrix for each group. Let $\boldsymbol{w}_i$ be a new $p \times 1$ random vector from one of the k groups, but the group is unknown. Usually there are many $\boldsymbol{w}_i$, and *discriminant analysis* attempts to allocate the $\boldsymbol{w}_i$ to the correct groups.

2) The *maximum likelihood discriminant rule* allocates case $\boldsymbol{w}$ to group a if $\hat{f}_a(\boldsymbol{w})$ maximizes $\hat{f}_j(\boldsymbol{w})$ for $j = 1, ..., k$. This rule is robust to nonnormality

and the assumption of equal population dispersion matrices, but $\hat{f}_j$ is hard to compute for $p > 1$.

3) Given the $\hat{f}_j(\boldsymbol{w})$ or a plot of the $\hat{f}_j(\boldsymbol{w})$, determine the maximum likelihood discriminant rule.

For the following rules, assume that costs of correct and incorrect allocation are unknown or equal, and assume that the probabilities $\rho_a(\boldsymbol{w}_i)$ that $\boldsymbol{w}_i$ is in group a are unknown or equal: $\rho_a(\boldsymbol{w}_i) = 1/k$ for $a = 1, ..., k$. Often it is assumed that the k groups have the same covariance matrix $\boldsymbol{\Sigma_x}$. Then the pooled covariance matrix estimator is

$$\boldsymbol{S}_{pool} = \frac{1}{n-k} \sum_{j=1}^{k} (n_j - 1) \boldsymbol{S}_j$$

where $n = \sum_{j=1}^{k} n_j$. Let $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)$ be the estimator of multivariate location and dispersion for the jth group, eg the sample mean and sample covariance matrix $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$.

4) Assume the population dispersion matrices are equal: $\boldsymbol{\Sigma}_j \equiv \boldsymbol{\Sigma}$ for $j = 1, ..., k$. Let $\hat{\boldsymbol{\Sigma}}_{pool}$ be an estimator of $\boldsymbol{\Sigma}$. Then the *linear discriminant rule* is allocate $\boldsymbol{w}$ to the group with the largest value of

$$d_j(\boldsymbol{w}) = \hat{\boldsymbol{\mu}}_j^T \hat{\boldsymbol{\Sigma}}_{pool}^{-1} \boldsymbol{w} - \frac{1}{2} \hat{\boldsymbol{\mu}}_j^T \hat{\boldsymbol{\Sigma}}_{pool}^{-1} \hat{\boldsymbol{\mu}}_j = \hat{\alpha}_j + \hat{\boldsymbol{\beta}}_j^T \boldsymbol{w}$$

where $j = 1, ..., k$. *Linear discriminant analysis* (LDA) uses $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_{pool}) = (\overline{\boldsymbol{x}}_j, \boldsymbol{S}_{pool})$. LDA is robust to nonnormality and somewhat robust to the assumption of equal population covariance matrices.

5) The *quadratic discriminant rule* is allocate $\boldsymbol{w}$ to the group with the largest value of

$$Q_j(\boldsymbol{w}) = \frac{-1}{2} \log(|\hat{\boldsymbol{\Sigma}}_j|) - \frac{1}{2} (\boldsymbol{w} - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1} (\boldsymbol{w} - \hat{\boldsymbol{\mu}}_j)$$

where $j = 1, ..., k$. *Quadratic discriminant analysis* (QDA) uses $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$. QDA has some robustness to nonnormality.

6) The *distance discriminant rule* allocates $\boldsymbol{w}$ to the group with the smallest squared distance $D^2_{\boldsymbol{w}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\boldsymbol{w} - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1} (\boldsymbol{w} - \hat{\boldsymbol{\mu}}_j)$ where $j = 1, ..., k$. This rule is robust to nonnormality and the assumption of equal $\boldsymbol{\Sigma}_j$, but needs $n_j > 10p$ for $j = 1, ..., k$.

7) Assume that $k = 2$ and that there is a group 0 and a group 1. Let $\rho(\boldsymbol{w}) = P(\boldsymbol{w} \in \text{group } 1)$. Let $\hat{\rho}(\boldsymbol{w})$ be the logistic regression (LR) estimate of $\rho(\boldsymbol{w})$. Logistic regression produces an estimated sufficient predictor $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{w}$. Then

$$\hat{\rho}(\boldsymbol{w}) = \frac{e^{ESP}}{1 + e^{ESP}} = \frac{\exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{w})}{1 + \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{w})}.$$

The *logistic regression discriminant rule* allocates $\boldsymbol{w}$ to group 1 if $\hat{\rho}(\boldsymbol{w}) \geq 0.5$ and allocates $\boldsymbol{w}$ to group 0 if $\hat{\rho}(\boldsymbol{w}) < 0.5$. Equivalently, the LR rule allocates $\boldsymbol{w}$ to group 1 if $ESP > 0$ and allocates $\boldsymbol{w}$ to group 0 if $ESP < 0$.

8) Let $Y_i = j$ if case i is in group j for $j = 0, 1$. Then a *response plot* is a plot of ESP versus Y_i (on the vertical axis) with $\hat{\rho}(\boldsymbol{x}_i) \equiv \hat{\rho}(ESP)$ added as a visual aid where $\boldsymbol{x}_i$ is the vector of predictors for case i. Also divide the ESP into J slices with approximately the same number of cases in each slice. Then compute the sample mean = sample proportion in slice s: $\hat{\rho}_s = \overline{Y}_s = \sum_s Y_i / m_s$ where m_s is the number of cases in slice s. Then plot the resulting step function as a visual aid. If n_0 and n_1 are the sample sizes of both groups and $n_i > 5p$, then the logistic regression model was useful if the step function of observed slice proportions scatter fairly closely about the logistic curve $\hat{\rho}(ESP)$. If the LR response plot is good, $n_0 > 5p$ and $n_1 > 5p$, then the LR rule is robust to nonnormality and the assumption of equal population dispersion matrices. Know how to tell a good LR response plot from a bad one.

9) Given LR output, as shown below in symbols and for a real data set, and given $\boldsymbol{x}$ to classify, be able to a) compute ESP, b) classify $\boldsymbol{x}$ in group 0 or group 1, c) compute $\hat{\rho}(\boldsymbol{x})$.

Label	Estimate	Std. Error	Est/SE	p-value
Constant	$\hat{\alpha}$	$se(\hat{\alpha})$	$z_{o,0}$	for Ho: $\alpha = 0$
x_1	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1} = \hat{\beta}_1 / se(\hat{\beta}_1)$	for Ho: $\beta_1 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	$\hat{\beta}_p$	$se(\hat{\beta}_p)$	$z_{o,p} = \hat{\beta}_p / se(\hat{\beta}_p)$	for Ho: $\beta_p = 0$

```
Binomial Regression Kernel mean function = Logistic
Response = Status  Terms = (Bottom Left) Trials = Ones
Coefficient Estimates
```

```
Label      Estimate        Std. Error     Est/SE    p-value
Constant  -389.806         104.224        -3.740     0.0002
Bottom     2.26423         0.333233        6.795     0.0000
Left       2.83356         0.795601        3.562     0.0004
```

10) Suppose there is training data $\boldsymbol{x}_{ij}$ for $i = 1, ..., n_j$ for group j. Hence it is known that $\boldsymbol{x}_{ij}$ came from group j where there are $k \geq 2$ groups. Use the discriminant analysis method to classify the training data. If m_j of the n_j group j cases are correctly classified, then the *apparent error rate for group* j is $1 - m_j/n_j$. If $m_A = \sum_{j=1}^k m_j$ of the $n = \sum_{j=1}^k n_j$ cases were correctly classified. Then the *apparent error rate* AER $= 1 - m_A/n$.

11) For the `ddiscr` method, get the apparent error rate for each of the k groups with the following commands. Replace `ddiscr` by `ddiscr2` for the `ddiscr2` method.

```
out1 <- ddiscr(x,w=x,group,xwflag=T)
out1$err
```

Get apparent error rates for `ddiscr,` `LDA` and `QDA` with the following commands.

```
out1 <- ddiscr(x,w=x,group,xwflag=T)
out1$toterr

out2  <- lda(x,group)
1-mean(predict(out2,x)$class==group)

out3  <- qda(x,group)
1-mean(predict(out3,x)$class==group)
```

Get the AERs for the methods that use variables x_1, x_3 and x_7 with the following commands.

```
out <- ddiscr(x[,c(1,3,7)],w=x[,c(1,3,7)],group,xwflag=T)
out$toterr

out <- lda(x[,c(1,3,7)],group)
1-mean(predict(out,x[,c(1,3,7)])$class==group)
```

```
out <- qda(x[,c(1,3,7)],group)
1-mean(predict(out,x[,c(1,3,7)])$class==group)
```

Get the AERs for the methods that leave out variables x_1, x_4 and x_5 with the following commands.

```
out <- ddiscr(x[,-c(1,4,5)],w=x[,-c(1,4,5)],group,xwflag=T)
out$toterr

out <- lda(x[,-c(1,4,5)],group)
1-mean(predict(out,x[,-c(1,4,5)])$class==group)

out <- qda(x[,-c(1,4,5)],group)
1-mean(predict(out,x[,-c(1,4,5)])$class==group)
```

12) Expect the apparent error rate to be too low: the method works better on the training data than on the new data to be classified.

13) Cross validation (CV): for $i = 1, ..., n$ where the training data has n cases, compute the discriminant rule with case i left out and see if the rule correctly classifies case i. Let m_C be the number of cases correctly classified. Then the CV error rate is $1 - m_C/n$.

14) Suppose the training data has n cases. Randomly select a subset L of m cases to be left out when computing the discriminant rule. Hence $n - m$ cases are used to compute the discriminant rule. Let m_L be the number of cases from subset L that are correctly classified. Then the "leave a subset out" error rate is $1 - m_L/m$. Here m should be large enough to get a good rate. Often m uses between $0.1n$ and $0.5n$.

15) Variable selection is the search for a subset of variables that does a good job of classification.

16) Forward selection: suppose $X_1, ..., X_p$ are variables.

Step 1) Choose variable $W_1 = X_1$ that minimizes the AER.

Step 2) Keep W_1 in the model, and add variable W_2 that minimizes the AER. So W_1 and W_2 are in the model at the end of Step 2).

Step k) Have $W_1, ..., W_{k-1}$ in the model. Add variable W_k that minimizes the AER. So $W_1, ..., W_k$ are in the model at the end of Step k).

Step p) $W_1, ..., W_p = X_1, ..., X_p$, so all p variables are in the model.

17) Backward elimination: suppose $X_1, ..., X_p$ are variables.

Step 1) $W_1, ..., W_p = X_1, ..., X_p$, so all p variables are in the model.

Step 2) Delete variable $W_p = X_j$ such that the model with $p-1$ variables $W_1, ..., W_{p-1}$ minimizes the AER.

Step 3) Delete variable $W_{p-1} = X_j$ such that the model with $p-2$ variables $W_1, ..., W_{p-2}$ minimizes the AER.

Step k) $W_1, ..., W_{p-k+2}$ are in the model. Delete variable $W_{p-k+2} = X_j$ such that the model with $p-k+1$ variables $W_1, ..., W_{p-k+1}$ minimizes the AER.

Step p) Have W_1 and W_2 in the model. Delete variable W_2 such that the model with 1 variable W_1 minimizes the AER.

18) Other criterion can be used and `proc stepdisc` in *SAS* does variable selection.

19) In *R*, using LDA, leave one variable out at a time as long as the AER does not increase much, to find a good subset quickly.

8.5 Complements

For $k = 2$, an alternative to the logistic regression model is the discriminant function model. See Hosmer and Lemeshow (2000, p. 43–44). Assume that $\rho_j = P(Y = j)$ and that $\boldsymbol{x}|Y = j \sim N_k(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ for $j = 0, 1$. That is, the conditional distribution of $\boldsymbol{x}$ given $Y = j$ follows a multivariate normal distribution with mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}$ which does not depend on j. Notice that $\boldsymbol{\Sigma} = \text{Cov}(\boldsymbol{x}|Y) \neq \text{Cov}(\boldsymbol{x})$. Then as for the logistic regression model,

$$P(Y = 1|\boldsymbol{x}) = \rho(\boldsymbol{x}) = \frac{\exp(\alpha + \boldsymbol{\beta}^T \boldsymbol{x})}{1 + \exp(\alpha + \boldsymbol{\beta}^T \boldsymbol{x})}.$$

Definition 8.8. Under the conditions above, the **discriminant function** parameters are given by

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \tag{8.2}$$

and

$$\alpha = \log\left(\frac{\rho_1}{\rho_0}\right) - 0.5(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0).$$

To use Definition 8.8 to simulate logistic regression data, set $\rho_0 = \rho_1 = 0.5$, $\boldsymbol{\Sigma} = \boldsymbol{I}$, and $\boldsymbol{\mu}_0 = \boldsymbol{0}$. Then $\alpha = -0.5\boldsymbol{\mu}_1^T\boldsymbol{\mu}_1$ and $\boldsymbol{\beta} = \boldsymbol{\mu}_1$. The discriminant function estimators $\hat{\alpha}_D$ and $\hat{\boldsymbol{\beta}}_D$ are found by replacing the population quantities ρ_1, ρ_0, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}$ by sample quantities. Alternatively, generate n values of the $SP_i = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}_i$, then generate a binomial$(1, \rho(SP_i))$ case for $i = 1, ..., n$. This alternative method is useful since the $\boldsymbol{x}_i$ need not be from a multivariate normal distribution.

See Olive (2010: ch. 10, 2013) for more information about logistic regression and response plots for logistic regression.

Huberty and Olejnik (2006) and McLachlan (2004) are useful references for discriminant analysis. Silverman (1986, $\oint$ 6.1) and Raveh (1989) are good references for nonparametric discriminant analysis. Discrimination when $p > n$ is interesting. See Cai and Liu (2011) and Mai, Zou and Yuan (2012).

Logistic regression is a useful alternative to discriminant analysis when there are two groups. The distance rule and Methods 1 and 2 can use RFCH or RMVN to compute $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)$.

Hand (2006) notes that supervised classification is a research area in statistics, machine learning, pattern recognition, computational learning theory and data mining. Hand (2006) argues that simple classification methods, such as linear discriminant analysis, are almost as good as more sophisticated methods such as neural networks and support vector machines.

8.6 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

8.1*. Assume the cases in each of the k groups are iid from a population with covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}(j)$ Find $E(\boldsymbol{S}_{pool})$ assuming that the k groups have the same covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}(j) \equiv \boldsymbol{\Sigma}_{\boldsymbol{x}}$ for $j = 1, ..., k$.

```
Logistic Regression Output,
Response = nodal involvement, Terms = (acid size xray)
Label      Estimate        Std. Error     Est/SE    p-value
Constant  -3.57564         1.18002        -3.030     0.0024
acid       2.06294         1.26441         1.632     0.1028
size       1.75556         0.738348        2.378     0.0174
```

```
xray       2.06178          0.777103         2.653     0.0080

Number of cases: 53, Degrees of freedom: 49, Deviance: 50.660
```

8.2. Following Collett (1999, p. 11), treatment for prostate cancer depends on whether the cancer has spread to the surrounding lymph nodes. Let the response variable = group y = *nodal involvement* (0 for absence, 1 for presence). Let $x_1 = acid$ (serum acid phosphatase level), $x_2 = size$ (= tumor size: 0 for small, 1 for large) and $x_3 = xray$ (xray result: 0 for negative, 1 for positive). Assume the case to be classified has $\boldsymbol{x}$ with $x_1 = acid = 0.65$, $x_2 = 0$ and $x3 = 0$.

a) Find ESP for $\boldsymbol{x}$.

b) Is $\boldsymbol{x}$ classified in group 0 or group 1?

c) Find $\hat{\rho}(\boldsymbol{x})$.

8.3. Recall that X comes from a uniform(a,b) distribution, written $x \sim U(a, b)$, if the pdf of x is $f(x) = \dfrac{1}{b-a}$ for $a < x < b$ and $f(x) = 0$, otherwise. Suppose group 1 has $X \sim U(-3, 3)$, group 2 has $X \sim U(-5, 5)$, and group 3 has $X \sim U(-1, 1)$. Find the maximum likelihood discriminant rule for classifying a new observation x.

```
out<-prcomp(state[,1:4],scale=T)
summary(out)
Importance of components: PC1    PC2    PC3     PC4
Standard deviation     1.6040 0.8803 0.6879 0.42318
Proportion of Variance 0.6432 0.1937 0.1183 0.04477
Cumulative Proportion  0.6432 0.8369 0.9552 1.00000

> out<-rprcomp(state[,1:4])
summary(out$out)
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.6705 0.8216 0.59362 0.42645
Proportion of Variance 0.6977 0.1688 0.08809 0.04546
Cumulative Proportion  0.6977 0.8664 0.95454 1.00000

Rotation:PC1          PC2         PC3         PC4
gdp     0.4525021  0.688328888 -0.5429877 -0.1631243
```

```
povrt  -0.5563898 -0.016929402 -0.2468286 -0.7932335
unins  -0.4442238  0.725197372  0.5076082  0.1381588
lifexp  0.5369706  0.002347129  0.6217506 -0.5701607

out <- lda(state[,1:4],state[,5])
1-mean(predict(out,state[,1:4])$class==state[,5])
[1] 0.3
```

8.4. The PCA and LDA output above is for the Minor (2012) state data where gdp = GDP per capita, povrt = poverty rate, unins = 3 year average uninsured rate 2007-9, and lifexp = life expectancy for the 50 states.

a) How many principal components are needed? Use a 0.9 threshold.

b) Which principal component corresponds to 9 gdp −9 unins −11 povrt +11 lifeexp?

c) The fifth variable was a 1 if the state was not worker friendly and a 2 if the state was worker friendly. With these two groups, what was the apparent error rate (AER) for LDA?

```
> out <- lda(x,group)
> 1-mean(predict(out,x)$class==group)
[1] 0.02
>
> out<-lda(x[,-c(1)],group)
> 1-mean(predict(out,x[,-c(1)])$class==group)
[1] 0.02
> out<-lda(x[,-c(1,2)],group)
> 1-mean(predict(out,x[,-c(1,2)])$class==group)
[1] 0.04
> out<-lda(x[,-c(1,3)],group)
> 1-mean(predict(out,x[,-c(1,3)])$class==group)
[1] 0.03333333
> out<-lda(x[,-c(1,4)],group)
> 1-mean(predict(out,x[,-c(1,4)])$class==group)
[1] 0.04666667
>
> out<-lda(x[,c(2,3,4)],group)
> 1-mean(predict(out,x[,c(2,3,4)])$class==group)
[1] 0.02
```

8.5. The above output is for LDA on the famous iris data set. the variables are x_1 = sepal length, x_2 = sepal width, x_3 = petal length and x_4 = petal width. These four predictors are in the x data matrix. There are three groups corresponding to types of iris: setosa versicolor virginica.

a) What is the AER using all 4 predictors?

b) Which variables, if any, can be deleted without increasing the AER in a)?

R/Splus Problems

Warning: Use the command *source("G:/mpack.txt")* **to download the programs. See Preface or Section 15.2.** Typing the name of the `mpack` function, eg *ddplot*, will display the code for the function. Use the `args` command, eg *args(ddplot)*, to display the needed arguments for the function.

8.5. Wisseman, Hopke and Schindler-Kaudelka (1987) pottery data has 36 pottery shards of Roman earthware produced between second century B.C. and fourth century A.D. Often the pottery was stamped by the manufacturer. A chemical analysis was done for 20 chemicals (variables), and 28 cases were classified as Arrentine (group 1) or nonArrentine (group 2), while 8 cases were of questionable origin. So the training data has $n = 28$ and $p = 20$.

a) Copy and paste the *R* commands for this part into *R* to make the data set.

b) Because of the small sample size, LDA should be used instead of QDA, as in the handout. Nonetheless, variable selection using QDA will be done. Copy and paste the *R* commands for this part into *R*. The first 9 variables result in no misclassification errors.

c) Now use commands like those shown in this section to delete variables whose deletion does not result in a classification error. Should get four variables are needed for perfect classification. What are they (eg X1, X2, X3 and X4)?

8.6. The distance discriminant rule is attractive theoretically as a maximum likelihood discriminant rule, but the distance rule does not work well for groups that have similar means. The `ddiscr` rule is a modification of the distance rule, and the `ddiscr2` rule tries to use the maximum likelihood rule where the $\hat{f}_j$ are estimated with a kernel density estimator.

The *R* code for this problem generates $N_2(\mathbf{0}, \boldsymbol{I})$ data where group 1 consists of the half set of cases closes to $\mathbf{0}$ in Mahalanobis distance (an ellipse about the origin), and group 2 consists of the remaining cases (the covering ellipse with inner ellipse removed).

a) Copy and paste the commands to make the data.

b) The commands for this part give the error rate for the ddiscr method that uses $\boldsymbol{x}$ as the two predictors. Put this output in *Word*.

c) The commands for this part give the error rate for the ddiscr method that uses the distances based on group 1 applied to all of the cases as the predictor. Put this output in *Word*.

d) The commands for this part give the error rate for the ddiscr2 method that uses $\boldsymbol{x}$ as the two predictors. Put this output in *Word*.

e) The commands for this part give the error rate for the ddiscr2 method that uses the distances based on group 1 applied to all of the cases as the predictor. Put this output in *Word*.

f) The commands for this part get the error rate for LDA using $\boldsymbol{x}$ as the two predictors.

g) The commands for this part get the error rate for QDA using $\boldsymbol{x}$ as the two predictors.

h) Which method worked the best?