Classification

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

P312, 2013

Classification

Introduction

- In our previous slide set on discriminant analysis, we saw how, with two groups, a *linear discriminant function* could, under certain circumstances, lead to an optimal rule for classifying observations into two groups on the basis of a set of measurements.
- In that slide set, we concentrated on the discrimination part of discriminant analysis, i.e., how to discover which dimension(s) in the data optimally discriminate between groups.
- We saw that there is, indeed, an intimate connection between discriminant analysis and MANOVA.

Introduction

- In our previous slide set on discriminant analysis, we saw how, with two groups, a *linear discriminant function* could, under certain circumstances, lead to an optimal rule for classifying observations into two groups on the basis of a set of measurements.
- In that slide set, we concentrated on the discrimination part of discriminant analysis, i.e., how to discover which dimension(s) in the data optimally discriminate between groups.
- We saw that there is, indeed, an intimate connection between discriminant analysis and MANOVA.

Introduction

- In our previous slide set on discriminant analysis, we saw how, with two groups, a *linear discriminant function* could, under certain circumstances, lead to an optimal rule for classifying observations into two groups on the basis of a set of measurements.
- In that slide set, we concentrated on the discrimination part of discriminant analysis, i.e., how to discover which dimension(s) in the data optimally discriminate between groups.
- We saw that there is, indeed, an intimate connection between discriminant analysis and MANOVA.

Introduction

- In this slide set, we concentrate on the *classification* side of discriminant analysis.
- We take a deeper look at how observations are classified into a group via a classification rule, how to evaluate the success of such a rule, and how to deal with a situation in which the rule works poorly.

Introduction

- In this slide set, we concentrate on the *classification* side of discriminant analysis.
- We take a deeper look at how observations are classified into a group via a classification rule, how to evaluate the success of such a rule, and how to deal with a situation in which the rule works poorly.

The Linear Classification Function

- The process of classification with linear discriminant functions can be viewed in several equivalent ways. In the *Discriminant Analysis* slides, we discussed one approach which involves comparing two groups by computing a difference of their discriminant scores from a cutoff value.
- An alternative approach that generalizes immediately to multiple groups is to classify the jth vector of observations $\boldsymbol{x}_j$ by computing for each group i a weighted (squared) distance score from $\boldsymbol{x}_j$ to the ith group centroid

$$D_i(\boldsymbol{x}_j) = (\boldsymbol{x}_j - \overline{\boldsymbol{x}}_i)' \boldsymbol{S}^{-1} (\boldsymbol{x}_j - \overline{\boldsymbol{x}}_i) \tag{1}$$

  and assign the jth observation to the group for which $D_i(\boldsymbol{x}_j)$ is a minimum. We can refer to $D_i(\boldsymbol{x}_j)$ as a *quadratic classification function* as it is a quadratic form.
- By expanding Equation 1 eliminating terms that do not involve i, and multiplying by $-1/2$, we can determine an equivalent *linear classification function*

$$L_i(\boldsymbol{x}_j) = \overline{\boldsymbol{x}}_i' \boldsymbol{S}^{-1} \boldsymbol{x} - \frac{1}{2} \overline{\boldsymbol{x}}_i' \boldsymbol{S}^{-1} \overline{\boldsymbol{x}}_i \tag{2}$$

- The j observation is assigned to the group for which $L_i(\boldsymbol{x}_j)$ is a maximum.

The Linear Classification Function

- The process of classification with linear discriminant functions can be viewed in several equivalent ways. In the *Discriminant Analysis* slides, we discussed one approach which involves comparing two groups by computing a difference of their discriminant scores from a cutoff value.
- An alternative approach that generalizes immediately to multiple groups is to classify the jth vector of observations $\boldsymbol{x}_j$ by computing for each group i a weighted (squared) distance score from $\boldsymbol{x}_j$ to the ith group centroid

$$D_i(\boldsymbol{x}_j) = (\boldsymbol{x}_j - \overline{\boldsymbol{x}}_i)'\boldsymbol{S}^{-1}(\boldsymbol{x}_j - \overline{\boldsymbol{x}}_i) \tag{1}$$

  and assign the jth observation to the group for which $D_i(\boldsymbol{x}_j)$ is a minimum. We can refer to $D_i(\boldsymbol{x}_j)$ as a *quadratic classification function* as it is a quadratic form.
- By expanding Equation 1 eliminating terms that do not involve i, and multiplying by $-1/2$, we can determine an equivalent *linear classification function*

$$L_i(\boldsymbol{x}_j) = \overline{\boldsymbol{x}}_i'\boldsymbol{S}^{-1}\boldsymbol{x} - \frac{1}{2}\overline{\boldsymbol{x}}_i'\boldsymbol{S}^{-1}\overline{\boldsymbol{x}}_i \tag{2}$$

- The j observation is assigned to the group for which $L_i(\boldsymbol{x}_j)$ is a maximum.

The Linear Classification Function

- The process of classification with linear discriminant functions can be viewed in several equivalent ways. In the *Discriminant Analysis* slides, we discussed one approach which involves comparing two groups by computing a difference of their discriminant scores from a cutoff value.
- An alternative approach that generalizes immediately to multiple groups is to classify the jth vector of observations $\boldsymbol{x}_j$ by computing for each group i a weighted (squared) distance score from $\boldsymbol{x}_j$ to the ith group centroid

$$D_i(\boldsymbol{x}_j) = (\boldsymbol{x}_j - \overline{\boldsymbol{x}}_i)'\boldsymbol{S}^{-1}(\boldsymbol{x}_j - \overline{\boldsymbol{x}}_i) \tag{1}$$

  and assign the jth observation to the group for which $D_i(\boldsymbol{x}_j)$ is a minimum. We can refer to $D_i(\boldsymbol{x}_j)$ as a *quadratic classification function* as it is a quadratic form.
- By expanding Equation 1 eliminating terms that do not involve i, and multiplying by $-1/2$, we can determine an equivalent *linear classification function*

$$L_i(\boldsymbol{x}_j) = \overline{\boldsymbol{x}}_i'\boldsymbol{S}^{-1}\boldsymbol{x} - \frac{1}{2}\overline{\boldsymbol{x}}_i'\boldsymbol{S}^{-1}\overline{\boldsymbol{x}}_i \tag{2}$$

- The j observation is assigned to the group for which $L_i(\boldsymbol{x}_j)$ is a maximum.

The Linear Classification Function

- The process of classification with linear discriminant functions can be viewed in several equivalent ways. In the *Discriminant Analysis* slides, we discussed one approach which involves comparing two groups by computing a difference of their discriminant scores from a cutoff value.
- An alternative approach that generalizes immediately to multiple groups is to classify the jth vector of observations $\boldsymbol{x}_j$ by computing for each group i a weighted (squared) distance score from $\boldsymbol{x}_j$ to the ith group centroid

$$D_i(\boldsymbol{x}_j) = (\boldsymbol{x}_j - \overline{\boldsymbol{x}}_i)'\boldsymbol{S}^{-1}(\boldsymbol{x}_j - \overline{\boldsymbol{x}}_i) \tag{1}$$

  and assign the jth observation to the group for which $D_i(\boldsymbol{x}_j)$ is a minimum. We can refer to $D_i(\boldsymbol{x}_j)$ as a *quadratic classification function* as it is a quadratic form.
- By expanding Equation 1 eliminating terms that do not involve i, and multiplying by $-1/2$, we can determine an equivalent *linear classification function*

$$L_i(\boldsymbol{x}_j) = \overline{\boldsymbol{x}}_i'\boldsymbol{S}^{-1}\boldsymbol{x} - \frac{1}{2}\overline{\boldsymbol{x}}_i'\boldsymbol{S}^{-1}\overline{\boldsymbol{x}}_i \tag{2}$$

- The j observation is assigned to the group for which $L_i(\boldsymbol{x}_j)$ is a maximum.

Incorporating Prior Probabilities

- If the probabilities of group membership are not equal, and group i occurs with probability p_i, then the linear classification function $L_i(\boldsymbol{x}_j)$ can be modified as follows to optimize the classification if the population distributions are multinormal with equal covariance matrices:

$$
\begin{aligned}
L_i^*(\boldsymbol{x}_j) &= \ln p_i + \overline{\boldsymbol{x}}_i' \boldsymbol{S}^{-1} \boldsymbol{x} - \frac{1}{2} \overline{\boldsymbol{x}}_i' \boldsymbol{S}^{-1} \overline{\boldsymbol{x}}_i && (3) \\
&= \ln p_i + L_i(\boldsymbol{x}_j) && (4)
\end{aligned}
$$

Quadratic Classification Functions

- If population covariance matrices differ across groups, then the linear classification approach discussed in the previous section is, in general, no longer optimal.
- A modified approach minimizes the (squared) distance function
$$D_i(\boldsymbol{x}_j) = (\boldsymbol{x}_j - \overline{\boldsymbol{x}}_i)' \boldsymbol{S}_i^{-1} (\boldsymbol{x}_j - \overline{\boldsymbol{x}}_i) \tag{5}$$
where $\boldsymbol{S}_i$ is the sample covariance matrix for the ith group.
- Note that unless n_i is greater than p, the number of predictors in $\boldsymbol{x}$, then $\boldsymbol{S}_i$ will be singular and the quadratic method cannot be used.
- If we assume multivariate normality, with prior probabilities for the groups of $p_i, i = 1, \ldots k$, then the optimal rule can be written as follows: Assign vector of scores $\boldsymbol{x}_j$ to the group for which
$$Q_i(\boldsymbol{x}_j) = \ln p_i - \frac{1}{2} \ln |\boldsymbol{S}_i| - \frac{1}{2} (\boldsymbol{x}_j - \overline{\boldsymbol{x}}_i)' \boldsymbol{S}_i^{-1} (\boldsymbol{x}_j - \overline{\boldsymbol{x}}_i) \tag{6}$$
is a maximum.

Quadratic Classification Functions

- If population covariance matrices differ across groups, then the linear classification approach discussed in the previous section is, in general, no longer optimal.
- A modified approach minimizes the (squared) distance function
$$D_i(\boldsymbol{x}_j) = (\boldsymbol{x}_j - \overline{\boldsymbol{x}}_i)' \boldsymbol{S}_i^{-1} (\boldsymbol{x}_j - \overline{\boldsymbol{x}}_i) \tag{5}$$
where $\boldsymbol{S}_i$ is the sample covariance matrix for the ith group.
- Note that unless n_i is greater than p, the number of predictors in $\boldsymbol{x}$, then $\boldsymbol{S}_i$ will be singular and the quadratic method cannot be used.
- If we assume multivariate normality, with prior probabilities for the groups of $p_i, i = 1, \ldots k$, then the optimal rule can be written as follows: Assign vector of scores $\boldsymbol{x}_j$ to the group for which
$$Q_i(\boldsymbol{x}_j) = \ln p_i - \frac{1}{2} \ln |\boldsymbol{S}_i| - \frac{1}{2} (\boldsymbol{x}_j - \overline{\boldsymbol{x}}_i)' \boldsymbol{S}_i^{-1} (\boldsymbol{x}_j - \overline{\boldsymbol{x}}_i) \tag{6}$$
is a maximum.

Quadratic Classification Functions

- If population covariance matrices differ across groups, then the linear classification approach discussed in the previous section is, in general, no longer optimal.
- A modified approach minimizes the (squared) distance function
$$D_i(\boldsymbol{x}_j) = (\boldsymbol{x}_j - \overline{\boldsymbol{x}}_i)'\boldsymbol{S}_i^{-1}(\boldsymbol{x}_j - \overline{\boldsymbol{x}}_i) \tag{5}$$
where $\boldsymbol{S}_i$ is the sample covariance matrix for the ith group.
- Note that unless n_i is greater than p, the number of predictors in $\boldsymbol{x}$, then $\boldsymbol{S}_i$ will be singular and the quadratic method cannot be used.
- If we assume multivariate normality, with prior probabilities for the groups of $p_i, i = 1, \ldots k$, then the optimal rule can be written as follows: Assign vector of scores $\boldsymbol{x}_j$ to the group for which
$$Q_i(\boldsymbol{x}_j) = \ln p_i - \frac{1}{2}\ln|\boldsymbol{S}_i| - \frac{1}{2}(\boldsymbol{x}_j - \overline{\boldsymbol{x}}_i)'\boldsymbol{S}_i^{-1}(\boldsymbol{x}_j - \overline{\boldsymbol{x}}_i) \tag{6}$$
is a maximum.

Quadratic Classification Functions

- If population covariance matrices differ across groups, then the linear classification approach discussed in the previous section is, in general, no longer optimal.
- A modified approach minimizes the (squared) distance function

$$D_i(\boldsymbol{x}_j) = (\boldsymbol{x}_j - \overline{\boldsymbol{x}}_i)'\boldsymbol{S}_i^{-1}(\boldsymbol{x}_j - \overline{\boldsymbol{x}}_i) \tag{5}$$

  where $\boldsymbol{S}_i$ is the sample covariance matrix for the ith group.
- Note that unless n_i is greater than p, the number of predictors in $\boldsymbol{x}$, then $\boldsymbol{S}_i$ will be singular and the quadratic method cannot be used.
- If we assume multivariate normality, with prior probabilities for the groups of $p_i, i = 1, \ldots k$, then the optimal rule can be written as follows: Assign vector of scores $\boldsymbol{x}_j$ to the group for which

$$Q_i(\boldsymbol{x}_j) = \ln p_i - \frac{1}{2}\ln|\boldsymbol{S}_i| - \frac{1}{2}(\boldsymbol{x}_j - \overline{\boldsymbol{x}}_i)'\boldsymbol{S}_i^{-1}(\boldsymbol{x}_j - \overline{\boldsymbol{x}}_i) \tag{6}$$

  is a maximum.

Estimating Misclassification Rates

- Once observations in the sample are classified, one may examine the accuracy of the rule created from the sample in classifying the observations in that sample.
- The result is a *Classification Table* that allows one to estimate both the proportion of observations correctly classified and the proportion of observations misclassified.
- We return to the football data set for an example.

```
> library(car)
> library(MASS)
> source(
+   "http://www.statpower.net/Content/312/R Stuff/Steiger R Library Functions.txt")
> fb.data <- read.table(
+   "http://www.statpower.net/Content/312/Lecture Slides/football.txt",header=T,sep=",")
> x <- as.matrix(fb.data[,2:7])
> Group <- as.matrix(fb.data[,1:1])
> source(
+   "http://www.statpower.net/Content/312/R Stuff/ClassifyCode.r")
```

Estimating Misclassification Rates

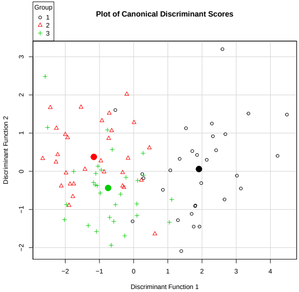
- The function `Classify` classifies observations according to either a linear or quadratic rule, and computes the Classification Table and error rates as well.

```
> out <- Classify(x,Group)
> head(out$Results)

  Group Classified WDIM CIRCUM FBEYE EYEHD EARHD JAW       L1       L2       L3
1     1          1 13.5  57.15  19.5  12.5  14.0  11 581.4637 577.4368 578.0970
2     1          1 15.5  58.42  21.0  12.0  16.0  12 657.9577 655.0008 655.6722
3     1          2 14.5  55.88  19.0  10.0  13.0  12 566.7910 570.0252 568.8719
4     1          1 15.5  58.42  20.0  13.5  15.0  12 637.3580 632.5968 634.4554
5     1          1 14.5  58.42  20.0  13.0  15.5  12 637.5758 630.7129 631.4172
6     1          1 14.0  60.96  21.0  12.0  14.0  13 659.0424 653.1503 652.0075

> out$Classification.Table

     Classified
Group  1  2  3
    1 26  1  3
    2  1 20  9
    3  2  8 20

> out$Proportion.Correct

[1] 0.7333333

> out$Error.Rate

[1] 0.2666667
```

Estimating Misclassification Rates

- From the Classification Table, it is clear that it is easy to classify members of Group 1, while there are plenty of misclassifications that result from confusing Groups 2 and 3.
- Looking back at the plot of canonical discriminant scores, it is easy to see why this is true.

```
> D <- Make.D(Group)
> H <- Make.H(Group)
> Plot.Discriminant.Scores(x,D,H,Group)
```



Estimating Misclassification Rates

- Using a quadratic rule improves the classification rates a bit.

```
> out <- Classify(x,Group,quadratic=TRUE)
> out$Classification.Table

     Classified
Group  1  2  3
    1 27  1  2
    2  2 21  7
    3  1  4 25

> out$Proportion.Correct

[1] 0.8111111

> out$Error.Rate

[1] 0.1888889
```

Bias in Error Rate Estimation

- Just as with R^2 in multiple regression, error rates obtained by applying a sample-based classification function to the same sample will be optimistic.
- One approach to de-biasing the error rate estimates is classical *cross-validation*, i.e., splitting the sample into a training sample and a test sample, and applying classification functions from one sample to the data in the other.

Bias in Error Rate Estimation

- Just as with R^2 in multiple regression, error rates obtained by applying a sample-based classification function to the same sample will be optimistic.
- One approach to de-biasing the error rate estimates is classical *cross-validation*, i.e., splitting the sample into a training sample and a test sample, and applying classification functions from one sample to the data in the other.

Bias in Error Rate Estimation

The Holdout Method

- An alternative approach is the *leave-one-out* or *holdout* method.
- With this approach, each observation vector is classified using classification functions *calculated from all the data but that observation.*
- Error rates are then estimated from the classification table.
- This method is, of course, more computationally intensive than the standard approach.

Bias in Error Rate Estimation

The Holdout Method

- An alternative approach is the *leave-one-out* or *holdout* method.
- With this approach, each observation vector is classified using classification functions *calculated from all the data but that observation.*
- Error rates are then estimated from the classification table.
- This method is, of course, more computationally intensive than the standard approach.

Bias in Error Rate Estimation
The Holdout Method

- An alternative approach is the *leave-one-out* or *holdout* method.
- With this approach, each observation vector is classified using classification functions *calculated from all the data but that observation.*
- Error rates are then estimated from the classification table.
- This method is, of course, more computationally intensive than the standard approach.

Bias in Error Rate Estimation

The Holdout Method

- An alternative approach is the *leave-one-out* or *holdout* method.
- With this approach, each observation vector is classified using classification functions *calculated from all the data but that observation.*
- Error rates are then estimated from the classification table.
- This method is, of course, more computationally intensive than the standard approach.

Bias in Error Rate Estimation

The Holdout Method

- The holdout method can be employed by using the function `Leave.One.Out`.
- This function repeatedly employs a service function `Make.Classification.Function` which returns the classification functions for any input data set, and thus can be immediately employed to predict the class of a new input vector.

```
> out <- Leave.One.Out(x,Group)
> out$Classification.Table

     Classified
Group  1  2  3
    1 26  1  3
    2  1 18 11
    3  2  9 19

> out$Proportion.Correct

[1] 0.7

> out$Error.Rate

[1] 0.3
```

Error Rates in Variable Selection

- Some authors, such as Rencher (in Chapter 9 of the second edition of his text) suggest combining error rate information with Wilks' Λ in assessing which variables to employ by means of a stepwise discriminant analysis.
- That is, a small improvements in Λ from adding a variable that is not accompanied by improvements in error rate might be considered illusory.

Error Rates in Variable Selection

- Some authors, such as Rencher (in Chapter 9 of the second edition of his text) suggest combining error rate information with Wilks' Λ in assessing which variables to employ by means of a stepwise discriminant analysis.
- That is, a small improvements in Λ from adding a variable that is not accompanied by improvements in error rate might be considered illusory.

Classification via the k Nearest Neighbor Rule

- Linear and Quadratic discriminant analysis are based on the supposition of a multivariate normal distribution.
- Other methods are available that do not make that assumption.
- Fix and Hodges (1951) proposed the k *nearest neighbor rule*.
- In this approach, we calculate the distance matrix between all observations using the function

$$D_{ij} = (\boldsymbol{x}_i - \boldsymbol{x}_j)'\boldsymbol{S}^{-1}(\boldsymbol{x}_i - \boldsymbol{x}_j)$$

- If sample sizes are equal, we then assign observation $\boldsymbol{x}_j$ to the class occupied by the majority of its k nearest neighbors. That is, for each of the k nearest neighbors, we compute k_i, the number that are in class i, and the class with the largest k_i is chosen.
- If sample sizes are unequal, we assign to the class i for which k_i/n_i is a maximum.
- If prior probabilities are incorporated, assign observation $\boldsymbol{x}_j$ to the class i for which $p_i k_i/n_i$ is a maximum.
- Of course, k must be chosen judiciously. Some authors suggest setting $k = \sqrt{n}$ for a "typical" group size n, while others suggest trying several values of k and settling on the one that produces the smallest error rate.
- The k-nearest neighbor method is implemented in the `class` library.

Classification via the k Nearest Neighbor Rule

```
> library(class)
> Classify <- rep(NA,90)
> for(i in 1:90)Classify[i] <- knn(x[-i,],
+                              x[i,],Group[-i],k=5)
> table(Group,Classify)
```

```
     Classify
Group  1  2  3
    1 26  2  2
    2  0 13 17
    3  3 11 16
```