Multivariate Statistics
Supervised Learning  Lecture 7
Linear Discriminant Analysis
```
dist(cereal.pca$scores)
hclust(*, "ward")
```
“Given it is known there are at least two categories or object types, how do we optimally distinguish them?”
Classification Problems

Disease detection
Fault detection

Change point detection

Validation of Clustering

Classification at different levels of detail
Simpson's Paradox
Simpson’s Paradox
Berkeley Sex Bias Case

Example: 44% of male applicants are admitted by a university, but only 33% of female applicants

Does this mean there is unfair discrimination?

University investigates and breaks down figures for Engineering and English programmes

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accept</td>
<td>35</td>
<td>20</td>
</tr>
<tr>
<td>Refuse entry</td>
<td>45</td>
<td>40</td>
</tr>
<tr>
<td>Total</td>
<td>80</td>
<td>60</td>
</tr>
</tbody>
</table>
Simpson’s Paradox

- No relationship between sex and acceptance for either programme
  - So no evidence of discrimination
- Why?
  - More females apply for the English programme, but it hard to get into
  - More males applied to Engineering, which has a higher acceptance rate than English
- Must look deeper than single cross-tab to find this out

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engineering</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accept</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>Refuse entry</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>60</td>
<td>20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accept</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>Refuse entry</td>
<td>15</td>
<td>30</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>40</td>
</tr>
</tbody>
</table>
The patterns show a significant difference from each other in one of the transformed axes.

**FIGURE 8.4** A cloud of data points is shown in two dimensions, and the density plots formed by projecting this cloud onto each of two axes, 1 and 2, are indicated. The projection onto axis 1 has maximum variance, and clearly shows the bimodal, or clustered character of the data.
Principal Component Analysis (PCA)

- PCA and classification
**Linear Discriminant Analysis, two-classes (1)**

- The objective of LDA is to perform dimensionality reduction while preserving as much of the class discriminatory information as possible.
  - Assume we have a set of D-dimensional samples \( \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\} \), \( N_1 \) of which belong to class \( \omega_1 \), and \( N_2 \) to class \( \omega_2 \). We seek to obtain a scalar \( y \) by projecting the samples \( x \) onto a line:
    \[
    y = w^T x
    \]
  - Of all the possible lines we would like to select the one that maximizes the separability of the scalars.
    - This is illustrated for the two-dimensional case in the following figures.
LINEAR DISCRIMINANT ANALYSIS

$Z = b_1 x_1 + b_2 x_2$
Linear Discriminant Analysis, two-classes (2)

- In order to find a good projection vector, we need to define a measure of separation between the projections
  - The mean vector of each class in $x$ and $y$ feature space is
    \[
    \mu_i = \frac{1}{N_i} \sum_{x \in \omega_i} x \quad \text{and} \quad \mu = \frac{1}{N} \sum_{y \in \omega} y = \frac{1}{N} \sum_{x \in \omega_i} w^T x = w^T \mu_i
    \]
  - We could then choose the distance between the projected means as our objective function
    \[
    J(w) = \| \tilde{\mu}_1 - \tilde{\mu}_2 \| = \| w^T (\mu_1 - \mu_2) \|
    \]
  - However, the distance between the projected means is not a very good measure since it does not take into account the standard deviation within the classes

This axis yields better class separability

This axis has a larger distance between means
The solution proposed by Fisher is to maximize a function that represents the difference between the means, normalized by a measure of the within-class scatter.

- For each class we define the **scatter**, an equivalent of the variance, as

\[
\tilde{s}_i^2 = \sum_{y \in \omega_i} (y - \tilde{\mu}_i)^2
\]

- where the quantity \( \left( \tilde{s}_1^2 + \tilde{s}_2^2 \right) \) is called the **within-class scatter** of the projected examples.

- The Fisher linear discriminant is defined as the linear function \( \mathbf{w}^\intercal \mathbf{x} \) that maximizes the criterion function

\[
J(w) = \frac{|\tilde{\mu}_1 - \tilde{\mu}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}
\]

- Therefore, we will be looking for a projection where examples from the same class are projected very close to each other and, at the same time, the projected means are as farther apart as possible.
Limitations of LDA

- LDA produces at most C-1 feature projections
  - If the classification error estimates establish that more features are needed, some other method must be employed to provide those additional features

- LDA is a parametric method since it assumes unimodal Gaussian likelihoods
  - If the distributions are significantly non-Gaussian, the LDA projections will not be able to preserve any complex structure of the data, which may be needed for classification

- LDA will fail when the discriminatory information is not in the mean but rather in the variance of the data
LDA Vs. PCA: Coffee discrimination with a gas sensor array

- These figures show the performance of PCA and LDA on an odor recognition problem
  - Five types of coffee beans were presented to an array of chemical gas sensors
  - For each coffee type, 45 “sniffs” were performed and the response of the gas sensor array was processed in order to obtain a 60-dimensional feature vector
- Results
  - From the 3D scatter plots it is clear that LDA outperforms PCA in terms of class discrimination
  - This is one example where the discriminatory information is not aligned with the direction of maximum variance
Quadratic Discriminant Analysis

Is very similar to linear discriminant analysis, except that every group is allowed to have its own variance matrix, allowing the ellipses to have a different orientation.
FIGURE 4.1. The left plot shows some data from three classes, with linear decision boundaries found by linear discriminant analysis. The right plot shows quadratic decision boundaries. These were obtained by finding linear boundaries in the five-dimensional space $X_1, X_2, X_{12}, X_1^2, X_2^2$. Linear inequalities in this space are quadratic inequalities in the original space.
Problem description

- We wish to predict group membership for a number of subjects from a set of predictor variables.
- The criterion variable (also called grouping variable) is the object of classification. This is ALWAYS a categorical variable!!
- Simple case: two groups and $p$ predictor variables.
Example

- We want to know whether somebody has lung cancer. Hence, we wish to predict a yes or no outcome.
- Possible predictor variables: number of cigarettes smoked a day, coughing frequency and intensity etc.
Approach (1)

- Linear discriminant analysis constructs one or more **discriminant equations** $D_i$ (linear combinations of the predictor variables $X_k$) such that the different groups differ as much as possible on $D$.
- Discriminant function:

$$D_i = b_0 + \sum_{k=1}^{p} b_k X_k$$
Approach (2)

- More precisely, the weights of the discriminant function are calculated in such a way, that the ratio (between groups SS)/(within groups SS) is as large as possible.
- Number of discriminant functions = \( \min(\text{number of groups} - 1, p) \).
Interpretation

- First discriminant function $D_1$ distinguishes first group from groups 2, 3, ..., N.
- Second discriminant function $D_2$ distinguishes second group from groups 3, 4, ..., N.
- etc
Visualization (two outcomes)
Visualization (3 outcomes)
Approach (3)

- To calculate the optimal weights, a training set is used containing the correct classification for a group of subjects.

- EXAMPLE (lung cancer): We need data about persons for whom we know for sure that they had lung cancer (e.g. established by means of an operation, scan, or x-rays)!
Approach (4)

- For a new group of subjects for whom we do not yet know the group they belong to, we can use the previously calculated discriminant weights to obtain their discriminant scores.
- We call this “classification”.
Example (1)

- The famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables **sepal length** and **width** and **petal length** and **width**, respectively, for 50 flowers from each of 3 species of iris.

- The species are Iris **setosa**, **versicolor**, and **virginica**.
<table>
<thead>
<tr>
<th></th>
<th>Obs</th>
<th>S.Length</th>
<th>S.Width</th>
<th>P.Length</th>
<th>P.Width</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.1</td>
<td>3.5</td>
<td>1.4</td>
<td>0.2</td>
<td>setosa</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>4.9</td>
<td>3.0</td>
<td>1.4</td>
<td>0.2</td>
<td>setosa</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>4.7</td>
<td>3.2</td>
<td>1.3</td>
<td>0.2</td>
<td>setosa</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4.6</td>
<td>3.1</td>
<td>1.5</td>
<td>0.2</td>
<td>setosa</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5.0</td>
<td>3.6</td>
<td>1.4</td>
<td>0.2</td>
<td>setosa</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>5.4</td>
<td>3.9</td>
<td>1.7</td>
<td>0.4</td>
<td>setosa</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>4.6</td>
<td>3.4</td>
<td>1.4</td>
<td>0.3</td>
<td>setosa</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>5.0</td>
<td>3.4</td>
<td>1.5</td>
<td>0.2</td>
<td>setosa</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>4.4</td>
<td>2.9</td>
<td>1.4</td>
<td>0.2</td>
<td>setosa</td>
<td></td>
</tr>
</tbody>
</table>
Example (2)

- Dependent variable?
- Predictor variables?
- Number of discriminant functions?
Step 1: Analyze data

- The idea is to start with analyzing the data.
- We start with linear discriminant analysis.
- Do the predictors vary sufficiently over the different groups?
- If not, they will be bad predictors.
- Formal test for this: Wilks’ test
- This test assesses whether the predictors vary enough to distinguish different groups.
**Step 1a: Sample statistics**

**Call:**
```r
call <- lda(Species ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width, data = iris)
```

**Prior probabilities of groups:**
- **setosa**: 0.3333333
- **versicolor**: 0.3333333
- **virginica**: 0.3333333

**Group means:**

<table>
<thead>
<tr>
<th></th>
<th>Sepal.Length</th>
<th>Sepal.Width</th>
<th>Petal.Length</th>
<th>Petal.Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>setosa</td>
<td>5.006</td>
<td>3.428</td>
<td>1.462</td>
<td>0.246</td>
</tr>
<tr>
<td>versicolor</td>
<td>5.936</td>
<td>2.770</td>
<td>4.260</td>
<td>1.326</td>
</tr>
<tr>
<td>virginica</td>
<td>6.588</td>
<td>2.974</td>
<td>5.552</td>
<td>2.026</td>
</tr>
</tbody>
</table>

**Coefficients of linear discriminants:**

<table>
<thead>
<tr>
<th></th>
<th>LD1</th>
<th>LD2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sepal.Length</td>
<td>0.8293776</td>
<td>0.02410215</td>
</tr>
<tr>
<td>Sepal.Width</td>
<td>1.5344731</td>
<td>2.16452123</td>
</tr>
<tr>
<td>Petal.Length</td>
<td>-2.2012117</td>
<td>-0.93192121</td>
</tr>
<tr>
<td>Petal.Width</td>
<td>-2.8104603</td>
<td>2.83918785</td>
</tr>
</tbody>
</table>

**Proportion of trace:**

<table>
<thead>
<tr>
<th></th>
<th>LD1</th>
<th>LD2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trace</td>
<td>0.9912</td>
<td>0.0088</td>
</tr>
</tbody>
</table>
Step 1b: Formal test

- X<-as.matrix(iris[-5])
- iris.manova<-manova(X~iris$Species)
- iris.wilks<-summary(iris.manova,test=“Wilks”)
- Relevant output: Wilks’ lamba equals 0.023, with p-value 2.2e-16. Thus, at a 0.001 significance level, we do not reject the discriminant model. (yes!, we are happy!)
Step 1c: Canonical correlation

- \( \sqrt{1 - \text{Wilks' lambda}} = \text{canonical correlation} \)
- Refers to the amount of variance in the grouping variable that is explained by the predictor variables.
- The higher this value, the better!!!!
- (The smaller Wilks' lambda, the better!!!)
Step 2: Discriminant function (1)

- Look at the coefficients of the standardized (!) discriminant functions to see what predictors play an important role.
- The larger the coefficient of a predictor in the standardized discriminant function, the more important its role in the discriminant function.
Step 2: Discriminant function (2)

- The coefficients represent partial correlations:

  *the contribution of a variable to the discriminant function in the context of the other predictor variables in the model.*

- Limitations: with more than two outcomes more difficult to interpret.
Step 2: Getting discr. functions

Call:
iris.lda<-lda(Species ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width, data = iris)

Prior probabilities of groups:
setosa versicolor virginica
0.3333333 0.3333333 0.3333333

Group means:

<table>
<thead>
<tr>
<th></th>
<th>Sepal.Length</th>
<th>Sepal.Width</th>
<th>Petal.Length</th>
<th>Petal.Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>setosa</td>
<td>5.006</td>
<td>3.428</td>
<td>1.462</td>
<td>0.246</td>
</tr>
<tr>
<td>versicolor</td>
<td>5.936</td>
<td>2.770</td>
<td>4.260</td>
<td>1.326</td>
</tr>
<tr>
<td>virginica</td>
<td>6.588</td>
<td>2.974</td>
<td>5.552</td>
<td>2.026</td>
</tr>
</tbody>
</table>

Coefficients of linear discriminants:

<table>
<thead>
<tr>
<th></th>
<th>LD1</th>
<th>LD2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sepal.Length</td>
<td>0.8293776</td>
<td>0.02410215</td>
</tr>
<tr>
<td>Sepal.Width</td>
<td>1.5344731</td>
<td>2.16452123</td>
</tr>
<tr>
<td>Petal.Length</td>
<td>-2.2012117</td>
<td>-0.93192121</td>
</tr>
<tr>
<td>Petal.Width</td>
<td>-2.8104603</td>
<td>2.83918785</td>
</tr>
</tbody>
</table>

Proportion of trace:  
LD1  0.9912  0.0088
Step 3: Comparing discrim. funcs

- Which discriminant function has most discriminating power?
- Look at the “eigenvalues”, also called the “singular values” or “characteristic roots”. Each discriminant function has such a value. They reflect the amount of variance explained in the grouping variable by the predictors in a discriminant function.
- Always look at the ratio of the eigenvalues to assess the relative importance of a discriminant function.
Step 3: Getting eigenvalues

iris.lda$svd

> iris.lda$svd

svd: the singular values, which give the ratio of the between- and within-group standard deviations on the linear discriminant variables.
Step 4: More interpretation

- Trace
- Useful plots
- Group centroids
Step 4a: Trace

Call:
iris.lda<-lda(Species ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width, data = iris)

Prior probabilities of groups:
setosa versicolor virginica
0.3333333 0.3333333 0.3333333

Group means:

<table>
<thead>
<tr>
<th></th>
<th>Sepal.Length</th>
<th>Sepal.Width</th>
<th>Petal.Length</th>
<th>Petal.Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>setosa</td>
<td>5.006</td>
<td>3.428</td>
<td>1.462</td>
<td>0.246</td>
</tr>
<tr>
<td>versicolor</td>
<td>5.936</td>
<td>2.770</td>
<td>4.260</td>
<td>1.326</td>
</tr>
<tr>
<td>virginica</td>
<td>6.588</td>
<td>2.974</td>
<td>5.552</td>
<td>2.026</td>
</tr>
</tbody>
</table>

Coefficients of linear discriminants:

<table>
<thead>
<tr>
<th></th>
<th>LD1</th>
<th>LD2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sepal.Length</td>
<td>0.8293776</td>
<td>0.02410215</td>
</tr>
<tr>
<td>Sepal.Width</td>
<td>1.5344731</td>
<td>2.16452123</td>
</tr>
<tr>
<td>Petal.Length</td>
<td>-2.2012117</td>
<td>-0.93192121</td>
</tr>
<tr>
<td>Petal.Width</td>
<td>-2.8104603</td>
<td>2.83918785</td>
</tr>
</tbody>
</table>

Proportion of trace:

<table>
<thead>
<tr>
<th></th>
<th>LD1</th>
<th>LD2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.9912</td>
<td>0.0088</td>
</tr>
</tbody>
</table>
Step 4a: Trace interpretation

- The first trace number indicates the percentage of between-group variance that the first discriminant function is able to explain from the total amount of between-group variance.

- High trace number = discriminant function plays an important role!
Step 4b: Useful plots

- Take e.g. first and second discriminant function. Plot discriminant function values of objects in scatter plot, with predicted groups. Does the discriminant function discriminate well between the different groups?

- Combine plot with “group centroids”. (Average values of discriminant functions for each group)
Step 4c: R code for plot

# Plot
LD1<-predict(iris.lda)$x[,1]
LD2<-predict(iris.lda)$x[,2]
plot(LD1,LD2,xlab="first linear discriminant",ylab="second linear discriminant",type="n")
text(cbind(LD1,LD2),labels=unclass(iris$Species))

# 1="setosa"
# 2="versicolor"
# 3="virginica"

# Group centroids
sum(LD1*(iris$Species=="setosa"))/sum(iris$Species=="setosa")
sum(LD2*(iris$Species=="setosa"))/sum(iris$Species=="setosa")

sum(LD1*(iris$Species=="versicolor"))/sum(iris$Species=="versicolor")
sum(LD2*(iris$Species=="versicolor"))/sum(iris$Species=="versicolor")

sum(LD1*(iris$Species=="virginica"))/sum(iris$Species=="virginica")
sum(LD2*(iris$Species=="virginica"))/sum(iris$Species=="virginica")
Step 5: Prediction (1)
Cross Validation

- Using the estimated discriminant model, classify new subjects.
- We consider the following approach:
  - Calculate the probability that a subject belongs to a certain group using the estimated discriminant model.
  - Do this for all groups.
  - Classification rule: subject is assigned to group it has the highest probability to fall into.
Step 5: Prediction (3) Bootstrap

- Prediction on training set: to assess how well the discriminant model predicts.
- Prediction on a new data set: to predict the group new object belongs to.
- Best way: N-k 75%-25% (must sample from this set—potentially huge number of splits)
  - Average over all splits
- N-1-- most conservative
Step 5: Prediction in R

- `iris.predict<- predict(iris.lda,iris[,1:4])`
  Predict class for all objects.

- `iris.classify<-iris.predict$class`
  Get predicted class for all objects.

- `iris.classperc<- sum(iris.classify==iris[,5])/150`
  Calculate % correctly classified objects.

- Priors are set automatically, but you can set them manually as well if you want.
Step 5: Quality of prediction (1)

- To assess the quality of a prediction, make a prediction table.

- Rows with observed categories of dependent variable, columns with forecasted categories.

- Ideally, the off-diagonal elements should be zero.
Step 5: Quality of prediction (2)

- The percentage correctly classified objects is usually compared to
  - the “random” classification
    \((100/N)\%\) probability in group \(i=1,\ldots,N\).
  - the “probability matching” classification
    Probability of assigning group \(i=1,\ldots,N\) to an object is equal to the fraction of objects in class \(i\).
Step 5: Get table in R

```r
table(Original = iris$Species, Predicted = predict(iris.lda)$class)
```

<table>
<thead>
<tr>
<th>Original</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>setosa</td>
<td>50 0 0</td>
</tr>
<tr>
<td>versicolor</td>
<td>0 48 2</td>
</tr>
<tr>
<td>virginica</td>
<td>0 1 49</td>
</tr>
</tbody>
</table>

Grouping variable

Predicted classes
Step 6: Structure coefficients

- Correlations between predictors and discriminant values indicate which predictor is most related to discriminant function (not corrected for the other variables)

- Example: cor(iris[,1],LD1)

  (Note difference with discriminant coefficients!!!)

  - cor(iris[,1],LD1)
  - [1] -0.7918878
  - > cor(iris[,2],LD1)
  - [1] 0.530759
  - > cor(iris[,3],LD1)
  - [1] -0.9849513
  - > cor(iris[,4],LD1)
  - [1] -0.972812
Assumptions underlying LDA

- Independent subjects.
- Normality: the variance-covariance matrix of the predictors is the same in all groups.
- If the latter assumption is violated: use quadratic discriminant analysis in the same manner as linear discriminant analysis.

**ALWAYS CHECK YOUR ASSUMPTIONS.......**
Assignment (1)

- Find an appropriate data set of at least 3 groups (in the area your area of interest).
  ADVICE: do not take too many groups.
- Think of a research question.
- Explain why discriminant analysis is a suitable method for your problem.
- Do a rigorous discriminant analysis, containing at least: sample statistics, formal testing, interpretation, prediction, quality assessment of prediction.
- **INTERPRET YOUR RESULTS!!!!**
Download R

- http://www.r-project.org/
- MASS library needed (automatically downloaded and installed)
- Load MASS library in R before you start!
- Various R manuals available at the website (and elsewhere on the web)
- Short questions: pass by my office
- Longer questions: make an appointment by phone
Data sets in R (1)

- Make an ASCII file of your data (e.g. in Notepad with a .txt extension)
- Give appropriate names to the various columns (a “header”)
- Round numbers in the same amount of digits!!!!!!!!!!!!!!!!! (you can do this in a earlier stage in Excel)
- Use the command
  
  ```r
  naam<-read.table("filenaam.txt", header=TRUE)
  ```
Before you start, load the right library (MASS), using Packages \rightarrow Load Packages \rightarrow MASS

Also set your working directory to the right one (where your files are located), using the command Files \rightarrow Change directory \rightarrow Browse
Logistic Regression
identical to 2-Class Discriminant

Let $\pi_t$ be the probability of membership in group $t$.

Use maximum likelihood to fit

$$\log\left(\frac{\pi_t}{1 - \pi_t}\right) = \beta_0 + \sum \beta_i Y_{it}$$

Classify a sample into group $t$ if the predicted $\log\left(\frac{\pi_t}{1 - \pi_t}\right)$ is the maximum over all groups.

Again, we can weight by prior probability.
Nonlinear and NonParametric Classifiers: Neural Networks SVM