Sheet 5

Hand in solution by May 2 in the lecture room

Discriminant Analysis

1. Load the ecosystem data of the alp in Grisons:

```
t.url <- "http://stat.ethz.ch/~stahel/courses/multivariate/datasets/vegenv.dat"
d.vegenv <- read.table(t.url,header=TRUE)
```

The variable `VegetationGroup` is an indicator for the type of vegetation present. We want to know whether the vegetation type can be characterized by the presence or absence of the plant species `Nardstri, Caluvulg, Festrubr`. You can perform linear discriminant analysis by the function `lda` from the package `MASS`. Load the package with the command `library(MASS)`.

a) Perform linear discriminant analysis. Use square root-transformed species counts and restrict the analysis to groups 3 and 4. Plot the result. Are these two groups well separated?
Hints: The groups can be selected via
`t.d <- d.vegenv[d.vegenv[,"VegetationGroup"]>=3,]`.
For the lda, use
`t.r <- lda(VegetationGroup ~ sqrt(Nardstri) + sqrt(Caluvulg)`
`+ sqrt(Festrubr), data=t.d)`.
Plot the results using `plot(t.r)`.

b) Now carry out the same analysis for all 4 groups. When viewed graphically, which groups appear to be well separated, and which groups are difficult to seperate?

c) The function `predict` assigns a group to each observation based on the results of the discriminant analysis. Compare the predictions with the true group labels to see how well the classifier performs. Are your findings in line with **a)** and **b)**.
Hint: `t.pr <- predict(t.r)` is a list containing the class-prediction `t.pr$class`, the values of the discrimant function and the posterior probabilities. The command
`table(t.pr$class,t.d$VegetationGroup)` produces a cross-table of the predicted and true labels.

2. Load the data set of the “1000 franc” bills:
`t.url <- "http://stat.ethz.ch/Teaching/Datasets/NDK/banknot.dat"`
`d.bankn <- read.table(t.url, header=TRUE)`.
Recall that the `CODE==1` bills are forged. We want to separate original and counterfeit bills by performing linear discriminant analysis using the variables `LEFT` and `RIGHT` as predictors.

a) Make a scatter plot of the variables `LEFT` and `RIGHT` using different colours to differentiate the two groups. If you like, use the function `jitter()` to see all points.
Hint: Some of the values are tied, `jitter` adds a small amount of noise to break up the ties. For plotting:
`plot(jitter(RIGHT) ∼ jitter(LEFT), data=d.bankn,`
`col=c("red", "blue")[d.bankn$CODE+1], xlab="LEFT", ylab="RIGHT")`.

b) Perform linear discriminant analysis and provide a cross-table for the predicted and the true classes.

c) Draw the decision boundary (i.e., the line which separates the two groups) of the lda into the scatter plot from **a)**.
Hint: The β-parameter of the linear discriminant function can be extracted via *lda-Objekt*`$scaling`. The α-parameter can be calculated as follows: The sample mean $\underline{\overline{x}}$ of all data satisfies the condition $\alpha + \underline{\beta}^T \cdot \underline{\overline{x}} = 0$, i.e., $\underline{\overline{x}}$ lies on the decision boundary, and hence $\alpha = -\underline{\beta}^T \cdot \underline{\overline{x}}$. The line is given by those points $\underline{x} = [x_1, x_2]^T$ for which $\alpha + \underline{\beta}^T \cdot \underline{x} = 0$ is true. Solve this equation for x_2 to obtain intercept and slope of the separating line. `abline(a,b)` adds a line with intercept `a` and slope `b` to an existing plot.

d) Calculate the sensitivity and the specificity using the cross-table from **b)**. Why are these values over-optimistic?

e) More realistic numbers for the error rates can be obtained by cross-validation: The discriminant function is calculated without using the point to be classified. Perform cross-validation by setting the argument `lda(..., CV=TRUE)`. Extract the predicted labels via *lda-Objekt*`$class`, i.e., you do not require the command `predict()` in this case. Again, provide the cross-table, and compute the sensitivity and the specificity. Compare with the numbers from **d)**.

3. We still consider the paper money data from the previous exercise, and we use the variables `LEFT` and `RIGHT` as predictors to seperate original and forged bills via *quadratic* discriminant analysis (qda) and logistic regression.

a) In principle, what is the difference between linear and quadratic discriminant analysis? Looking at the covariance matrices for the two groups in this example, do you think that qda gives smaller error rates than lda?

b) Perform quadratic discriminant analysis and provide a cross-table for the predicted and the true classes. Use cross-validation for the class predictions. Calculate the sensitivity and the specificity using the cross-table. Compare with the corresponding numbers obtained from lda.
Hint: You can perform qda by the function `qda` from the package `MASS`. `qda` is used in the very same way as `lda`.

c) (*) Repeat **b)** using logistic regression instead.
Hint: Use `t.r <- glm(CODE ∼ LEFT + RIGHT, data=?, familiy=binomial)` to fit the model. The posterior probabilities can be obtained via

```
t.pp <- predict(t.r, newdata=?, type="response").
```

Use `ifelse(t.pp>=0.5,1,0)` to predict the class. `glm` does not provide automatic cross-validation, i.e., it must be implemented “by hand”.