Series 8

1. The dataset `heart.dat` contains data for 99 people sorted by age. In each age group the total number of individuals (m_i) is known, as well the number of those with symptoms of heart disease (y_i).

a) Estimate the parameters of a simple logistic regression which relates the probability of having symptoms to the age of the individual. Does age influence this probability in a significant way? How do you interpret the sign of the coefficient of `age`?

R hint:
The data is located at `http://stat.ethz.ch/Teaching/Datasets/heart.dat`.
The logistic regression model can be fitted by using the command

```
fit <- glm(cbind(y, m - y) ~ age, family = binomial, data = heart).
```

Binomial responses $Y_i \sim \text{Bin}(m_i, \pi_i)$ for $m_i > 1$ should be entered as a (two-column) matrix, with the number of "successes" (Y_i) in the first column and the number of "failures" ($m_i - Y_i$) in the second.

b) Plot the probability estimate against age. At what age would you expect 10%, 20%, ..., 90% of people to have symptoms of heart disease? Discuss your results.

R hint:
You can obtain probability estimates at arbitrary ages `new.age` by using the command

```
predict(fit, newdata = data.frame(age = new.age), type = "response")
```

2. **a)** **Quadratic Discriminant Analysis (QDA)**
Assume the normal model $X|Y = j \sim \mathcal{N}_p(\mu_j, \Sigma_j)$, $\mathbb{P}[Y = j] = p_j$, $\sum_{j=0}^{J-1} p_j = 1$.
Show that (6.2) and (6.4) lead to

$$\hat{\delta}_j^{QDA}(x) = -\log(\det(\hat{\Sigma}_j))/2 - (x - \hat{\mu}_j)^\intercal \hat{\Sigma}_j^{-1}(x - \hat{\mu}_j)/2 + \log(\hat{p}_j).$$

b) **Linear Discriminant Analysis (LDA)**
Use the result from a) and replace $\hat{\Sigma}_j$ by $\hat{\Sigma}$ to get

$$\begin{aligned}\hat{\delta}_j^{LDA}(x) &= x^\intercal \hat{\Sigma}^{-1}\hat{\mu}_j - \hat{\mu}_j{}^\intercal \hat{\Sigma}^{-1}\hat{\mu}_j/2 + \log(\hat{p}_j) \qquad (1)\\ &= (x - \hat{\mu}_j/2)^\intercal \hat{\Sigma}^{-1}\hat{\mu}_j + \log(\hat{p}_j).\end{aligned}$$

c) The LDA decision function can be written as (see (1) above)

$$\hat{\delta}_j(x) = x^\intercal b_j + c_j,$$

where $b_j \in \mathbb{R}^p$ and $c_j \in \mathbb{R}$. Assume that we only have two classes ($j = 0, 1$). Use the equation above to characterize the decision boundary.

d) **Small Simulation**
Use the R-code below to generate data samples from three groups of normal distributions; change the covariance matrix and mean vectors if you like:

```
library(mvtnorm) ## Needed for rmvnorm
library(MASS)    ## Needed for lda/qda
## Read in a function that plots LDA/QDA decision boundaries
source("http://stat.ethz.ch/teaching/lectures/FS_2010/CompStat/predplot.R")
## Covariance Matrix
sigma <- cbind(c(0.5, 0.3), c(0.3, 0.5))
## Mean vectors
mu1 <- c(3, 1.5)
mu2 <- c(4, 4)
```

```
mu3 <- c(8.5, 2)
m <- matrix(0, nrow = 300, ncol = 3)
## Grouping vector
m[,3] <- rep(1:3, each = 100)
## Simulate data
m[1:100,1:2] <- rmvnorm(n = 100, mean = mu1, sigma = sigma)
m[101:200,1:2] <- rmvnorm(n = 100, mean = mu2, sigma = sigma)
m[201:300,1:2] <- rmvnorm(n = 100, mean = mu3, sigma = sigma)
m <- data.frame(m)
```

Perform LDA and plot the results:

```
fit <- lda(x = m[,1:2], grouping = m[,3])
predplot(fit, m)
```

Manually calculate (see c)) the boundary between group 1 and 2. Add your solution to the plot with `abline()`

Hint:

If `A <- fit$scaling`, it holds (in the case of $p+1$ groups in $\mathbb{R}^p$) that $\hat{\Sigma}^{-1} = AA^\intercal$. The means and prior probabilites can also be found in the `lda`-object. However, you may also want to do everything on your own, i.e., without using the result of `lda`; in this case, you can use the estimators for $\hat{\mu}_j$ and $\hat{\boldsymbol{\Sigma}}$ given in Chapter 6.3.1 of the lecture notes, just above Formula (6.5).

Preliminary discussion: Friday, May 06.

Deadline: Friday, May 13.