Series 6

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 (N_i). The goal of this exercise is to estimate the probability of having such symptoms as a function of age using logistic regression.

The data is located at `http://stat.ethz.ch/Teaching/Datasets/heart.dat`.

a) In contrast to the binary classification example in the lecture notes (page 56), the response variable N has not a Bernoulli, but a binomial distribution: $N_1, \dots, N_n$ independent, $N_i \sim \text{Binomial}(m_i, \pi(x_i))$.
Show that the log-likelihood is in this case

$$\ell(\beta; (x_1, m_1, N_1), \dots, (x_n, m_n, N_n)) = \sum_{i=1}^{n} \left[\log \binom{m_i}{N_i} + N_i g(\beta; x_i) - m_i \log\left(1 + e^{g(\beta; x_i)}\right) \right],$$

where $g(\beta; x) = \beta_0 + \beta_1 x$ is the model function for the logistic transform of $\pi(x)$ (see Formula (6.6) in the lecture notes).

b) Write an R function `neg.ll(beta, data)` that calculates the *negative* log-likelihood

$$-\ell(\beta; (x_1, m_1, N_1), \dots, (x_n, m_n, N_n))$$

that you derived in task a). `beta` is a vector with two entries β_0 and β_1, and `data` is a data frame with columns `age`, `m` and `N` (as in `heart.dat`).
Make a contour plot of the negative log-likelihood of the `heart` dataset in the range $-10 \leq \beta_0 \leq 10$, $-1 \leq \beta_1 \leq 1$.
R hint: use a command like

```
> contour(beta0.grid, beta1.grid, neg.ll.values)
```

`beta0.grid` and `beta1.grid` are equidistant grids of values of β_0 and β_1 in the region of interest; use, e.g.

```
> beta0.grid <- seq(-10, 10, length = 51)
```

`neg.ll.values` is a matrix of negative log-likelihood values for the different values of β_0 and β_1.

c) Estimate the parameters β_0 and β_1 of the model function (see task a)) using the R function `glm`. Does age influence this probability in a significant way? How do you interpret the sign of the coefficient of `age`?
Compare the estimates from `glm` with estimates you get when minimizing the negative log-likelihood function you implemented in task b).
R hint: the logistic regression model can be fitted by using the command

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

Binomial responses $N_i \sim \text{Bin}(m_i, \pi_i)$ for $m_i > 1$ should be entered as a (two-column) matrix, with the number of "successes" (N_i) in the first column and the number of "failures" ($m_i - N_i$) in the second.
To minimize your function `neg.ll` from task b), use

```
> optim(c(0, 0), neg.ll, data = heart)
```

The first argument is the start value used for numerical optimization.

d) 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(MASS)    ## Needed for lda/qda and mvrnorm
## Read in a function that plots LDA/QDA decision boundaries
source("http://stat.ethz.ch/education/semesters/ss2012/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] <- mvrnorm(n = 100, mu = mu1, Sigma = sigma)
m[101:200,1:2] <- mvrnorm(n = 100, mu = mu2, Sigma = sigma)
m[201:300,1:2] <- mvrnorm(n = 100, mu = 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 $\hat{\mathbf{\Sigma}}$ given in Chapter 6.3.1 of the lecture notes, just above Formula (6.5).

Preliminary discussion: Friday, April 20.

Deadline: Friday, April 27.