Series 7

1. In section 6.3 (The view of discriminant analysis) of the lecture notes two discriminant classifiers are given. In this exercise we want to derive them.

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) in the lecture notes lead to finding the arg max of the following quantity:

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

b) Linear Discriminant Analysis (LDA)
Use the result from a) and replace $\hat{\Sigma}_j$ by $\hat{\Sigma}$ to show that the following quantity is maximized over j for LDA:

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

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 (the set of x where $\hat{\delta}_0^{LDA}(x) = \hat{\delta}_1^{LDA}(x)$), $B = \{\, x \,|\, ??? \,\}$.

2. The data frame `iris` gives the measurements in centimeters of the length and width of the sepal and petal (4 measurements in total) for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.

In this exercise we want to use a bootstrap with LDA and QDA on the iris data, by just using the petal information:

```
Iris <- iris[,c("Petal.Length","Petal.Width","Species")]
```

a) Fit the data with both the LDA and QDA methods. Then plot the classification boundaries for both methods while using the `predplot` function provided in the R-skeleton.
b) Use a bootstrap to generate $B = 1000$ bootstrap samples, then fit the bootstrap sample with both the LDA and QDA methods. Plot the bootstrap estimates $\hat{\mu}_j^{*i}$ ($j \in \{0, 1, 2\}$ and $i \in \{1, ..., 1000\}$) of the LDA method in a single plot with different colours for each class.
c) Plot the classification boundaries for both methods provided by the fits of the bootstrap samples in two separate plots. Once again use the function `predplot` provided in the R-skeleton.
d) Calculate the LOOCV estimate of the generalization error for both methods, where the loss function is defined as: $\rho(x, x') = \begin{cases} 0 & \text{if } x = x' \\ 1 & \text{else} \end{cases}$.
Based on this estimate which model is the preferred method?

R-Hints: Use the R-skeleton provided in the Exercises section of the website of the course.

Use different colours when plotting different classes.

3. The dataset `heart.dat` contains data for 99 people grouped 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`. You can download it using

```
> heart <- read.table("http://stat.ethz.ch/Teaching/Datasets/heart.dat", header = TRUE)
```

a) In contrast to the binary classification example in the lecture notes (page 57), the response variable N has not a Bernoulli, but a binomial distribution: $N_1, \ldots, N_n$ independent, $N_i \sim \text{Binomial}(m_i, \pi(x_i))$.
Show that the log-likelihood is in this case

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

where $g(\beta; x) = \beta_0 + \beta_1 x$ is the model function for the logistic transform of $\pi(x)$ (see section 6.4 of the lecture notes).

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

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

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`).
R hint: Have a look at `choose` to calculate the required binomial coefficient.
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 function call 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 = 101)
```

`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 function call

```
> 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: From a `glm()` fit, you can obtain probability estimates at arbitrary ages `new.age` by using the function call

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

Preliminary discussion: Friday, April 29.

Deadline: Friday, May 06.