

## 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

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

**b) Linear Discriminant Analysis (LDA)**

Use the result from a) and replace  $\hat{\Sigma}_j$  by  $\hat{\Sigma}$  to get

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

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

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

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  $B = \{x \mid \text{??}\}$ .

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  $i \in \{1, \dots, 1000\}$ ) 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 bootstrap 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 the generalization error which model is the preferred method? Use a boxplot to determine if there's a significant difference between the methods for the given data.
- e\*) Additionally, calculate the out-of-bootstrap (OOB) estimate of the generalization error for both methods, using the same loss function as in d). Compare these estimates with the estimates of the generalization error from d).

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

Use different colours when plotting different classes.

**Preliminary discussion:** Friday, April 17.

**Deadline:** Friday, April 24.