

## Exercise 4

**Tutorial Exercise.**

1. If the observations for class  $g$  are from a  $N_p(\boldsymbol{\mu}_g, \boldsymbol{\Sigma})$  distribution for  $g = 1, 2$  for some  $\boldsymbol{\mu}_g$  and positive definite  $\boldsymbol{\Sigma}$ , and each class is equally likely, show that the Bayes rule is given by

$$\text{allocate } x \text{ to class 1 if } \boldsymbol{\alpha}^T(\mathbf{x} - \boldsymbol{\mu}) > 0$$

where  $\boldsymbol{\alpha} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$  and  $\boldsymbol{\mu} = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$ .

2. The attached **R** output relates to a discriminant analysis for two species of iris.
- Test the hypothesis that the two species have the same mean vectors.
  - Determine the discriminant function from the output.
  - Using Fisher discriminant rule, determine how many of the 100 observations are misclassified (calculating the resubstitution error rate).
  - How would you classify the following plant with the measurement (Sepal.Length, Sepal.Width, Petal.Length, Petal.Width) = (7, 4.5, 5.1, 2) using Fisher discriminant rule.
3. The attached **R** output relates a classification analysis of 200 crab *Leptograpsus variegatus* specimens found on the shores of Western Australia.
- Using KNN classifier, determine how many of the 200 observations are misclassified by calculating:
    - the resubstitution error rate, and
    - the test-set error rate when samples are randomly divided into two groups.
  - How would you classify the following crab with the measurement (FL,RW,CL,CW) = (19.3, 14.5, 40, 30.9)
    - using a KNN classifier, and
    - binary tree classifier.
  - What is the 5-fold cross validation error rate when
    - using a KNN classifier, and
    - binary tree classifier.
  - Based on the CV error rate, which is the preferred classifier?

## Computer Exercise 1.

Data on 9 mandible measurements (in mm) for samples drawn from different species of dog are stored in Splus in **canine1** (modern Thai dogs) and **canine2** (golden jackals). These data were partially analysed in Computer Exercise Week 2. You can obtain the data using

```
canine1 = read.table(file =  
url("http://www.maths.usyd.edu.au/u/UG/SM/STAT3014/r/Data/canine1.dat"))  
and  
canine2 = read.table(file =  
url("http://www.maths.usyd.edu.au/u/UG/SM/STAT3014/r/Data/canine2.dat"))
```

1. Calculate Fisher's linear discriminant function based on the variables  $V_1, \dots, V_5$  and obtain the allocation rule for dogs based on this function.
2. Obtain a discriminant rule using the function `lda` and comment on the results. Determine how many of observations are misclassified (i.e. calculate the resubstitution error rate).
3. Use the function `knn` from the `class` package, calculate the resubstitution error rate for  $k = 3$ .
4. Randomly divide the samples into 4 groups and calculate the 4-fold cross-validated error rate for  $k = 3$ .
5. Repeat Questions 3 and 4 using a tree classifier via the function `rpart` from the `rpart` package. Which is the classifier has the lower 4-fold cross-validated error rate?