TMA4300 Computer Intensive Statistical Methods
Exercise 3, Spring 2017

Note: The solution to ALL problems must be handed in no later than **May 4th 2017, 12:00**.
Hint: In almost all exercises you will need to use the R-function `sample`.

Problem A: Classification and cross validation

We will consider three different classification approaches: linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and k-nearest neighbours (KNN), and apply them to two different simulated data scenarios. The scenarios are as follows:

Scenario 1: There are 20 observations in each of three classes. The observations within each class are bivariate normal random variables, i.e. there are two predictors, with the following mean vectors:

$$\boldsymbol{\mu}_1 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \boldsymbol{\mu}_3 = \begin{pmatrix} 0 \\ -2 \end{pmatrix}.$$

The covariance matrix, which is the same for all three classes, has the form $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 0.7 \end{pmatrix}$.

Scenario 2: There are 20 observations in each of three classes. The observations within each class are bivariate normal random variables, i.e. there are two predictors, with the following mean vectors:

$$\boldsymbol{\mu}_1 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \boldsymbol{\mu}_3 = \begin{pmatrix} 0 \\ -2 \end{pmatrix}.$$

The covariance matrix for the first class is denoted by $\boldsymbol{\Sigma}_1$, for the second class by $\boldsymbol{\Sigma}_2$, and for the third class by $\boldsymbol{\Sigma}_3$, and defined as follows:

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_3 = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 0.25 \end{pmatrix},$$

a) Write a function `compareClassifiers()` that takes as input one simulated data set and returns the missclassification rate obtained by each of the three classification approaches based on 10-fold cross-validation. For KNN use $k = 1, \ldots, 10$, so that you obtain in total 12 estimated miss-classification rates.

b) Apply the function implemented in a) to 100 random data sets simulated from Scenario 1. To get an idea on how the data behave, you might want to plot some of them. Generate one figure containing boxplots for all classification approaches showing their respective missclassification rates. (Hint: See `example(boxplot)` in R on how to generate multiple boxplots in one figure). Interpret the result.

c) Do the same described in b) for Scenario 2.

d) Do the results coincide with your intuition? Why might they differ?

NOTE: You are allowed to use the R-function `lda` and `qda` available in the library `MASS`, and the R-function `knn` available in the library `class`.

Problem B: Comparing $AR(2)$ parameter estimators using resampling of residuals

The data files and pre-programmed R-code can be downloaded from the course webpage. Look in the `probBhelp.R`-file and read the documentation to see how the code works. Load the code and data into `R` with

```
source("probBhelp.R")
source("probBdata.R")
```

In this exercise you should analyse the data in `data3A$x`, which contains a sequence of length $T = 100$ of a non-Gaussian time-series, and compare two different parameter estimators.
We consider an AR(2) model which is specified by the relation

$$x_t = \beta_1 x_{t-1} + \beta_2 x_{t-2} + e_t,$$

where e_t are iid random variable with zero mean and constant variance.
The least sum of squared residuals (LS) and least sum of absolute residuals (LA) are obtained by minimising the following loss functions with respect to $\boldsymbol{\beta}$:

$$Q_{LS}(\boldsymbol{x}) = \sum_{t=3}^{T} \left(x_t - \beta_1 x_{t-1} - \beta_2 x_{t-2}\right)^2$$
$$Q_{LA}(\boldsymbol{x}) = \sum_{t=3}^{T} \left|x_t - \beta_1 x_{t-1} - \beta_2 x_{t-2}\right|$$

Denote the minimisers by $\widehat{\boldsymbol{\beta}}_{LS}$ and $\widehat{\boldsymbol{\beta}}_{LA}$ (calculated by `ARp.beta.est`), and define the estimated residuals to be $\hat{e}_t = x_t - \hat{\beta}_1 x_{t-1} - \hat{\beta}_2 x_{t-2}$ for $t = 3, \ldots, T$, and let $\bar{e}$ be the mean of these. The $\hat{e}_t$ can be re-centered to have mean zero by defining $\hat{\epsilon}_t = \hat{e}_t - \bar{e}$. (Results for $\hat{\epsilon}_t$ obtained by LS and LA can be calculated with `ARp.resid`).

1. Use the residual resampling bootstrap method to evaluate the relative performance of the two parameter estimators. Specifically, estimate the variance and bias of the two estimators.

   You may use `ARp.filter` as a helper function in your resampling code. Use at least $B = 1500$ bootstrap samples, each as long as the original data sequence ($T = 100$). To do a resampling, initialise values for x_1 and x_2 by picking a random consecutive subsequence from the data.

   The LS estimator is optimal for Gaussian AR(p) processes. Explain if it is also optimal for this problem?

2. Compute a 95% prediction interval for x_{101} based on both estimators. That means using the corresponding parameter estimates obtained in part 1), predict for each bootstrap iteration a value for x_{101}. From these B predictions derive a 95% quantile-based confidence interval.

Problem C: Permutation test

Bilirubin (see `http://en.wikipedia.org/wiki/Bilirubin`) is a breakdown product of haemoglobin, which is a principal component of red blood cells. If the liver has suffered degeneration, if the decomposition of haemoglobin is elevated, or if the gall bladder has been destroyed, large amounts of bilirubin can accumulate in the blood, leading to jaundice. The following data (taken from Jørgensen (1993)) contain measurements of the concentration of bilirubin (mg/dL) in blood samples taken from three young men.

Individual	Concentration (mg/dL)										
1	0.14	0.20	0.23	0.27	0.27	0.34	0.41	0.41	0.55	0.61	0.66
2	0.20	0.27	0.32	0.34	0.34	0.38	0.41	0.41	0.48	0.55	
3	0.32	0.41	0.41	0.55	0.55	0.62	0.71	0.91			

We will use the F-statistic to perform a permutation test.

Download the data file `bilirubin.txt` from the course webpage and read it into R using

```
bilirubin <- read.table("bilirubin.txt",header=T)
```

```
> head(bilirubin)
  meas pers
1 0.14   p1
2 0.20   p1
3 0.23   p1
4 0.27   p1
5 0.27   p1
6 0.34   p1
```

The first column, labelled `meas`, contains the concentrations (mg/dL) as shown in the table. The second column, `pers`, is an indicator for the individual.

1. Use a boxplot to inspect the logarithms of the concentrations for each individual. Be careful to use the same y-axis to make the plots comparable. Use the function `lm` in `R` to fit the regression model

$$\log Y_{ij} = \beta_i + \epsilon_{ij}, \qquad \text{with } i = 1, 2, 3 \text{ and } j = 1, \ldots, n_i \tag{1}$$

   where $n_1 = 11$, $n_2 = 10$ and $n_3 = 8$, and $\epsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. Use the F-test to test the hypothesis that $\beta_1 = \beta_2 = \beta_3$ and save the value of the F-statistic as `Fval`. Is the hypothesis accepted?

   (Hint: Define the model as `lm(log(meas)~pers,data=bilirubin)`. Then, the F-statistic of the test of interest is contained in the default output of `summary.lm`)

2. Write a function `permTest()` which generates a permutation of the data between the three individuals, consequently fits the model given in (1) and finally returns the value of the F-statistic for testing $\beta_1 = \beta_2 = \beta_3$.

3. Perform a permutation test using the function `permTest` to generate a sample of size 999 for the F-statistic. Compute the p-value for `Fval` using this sample. What do you observe?

Literature

Jørgensen, B. (1993). The Theory of Linear Models. Chapman and Hall

Oral presentations

Date	Problem	Team
24.04.2017	3: Problems A a) and b)	Kim Roger Kristiansen and Esten Nicolai WØien
	3: Problems A c) and d)	Katri Ailus and Carl Lakos
	3: Problem B	Håvard Heitlo Holm and Heidi Elisabeth
27.04.2017	3: Problem C	Markus Mortensen and Kristian Ruud