Project

Submit your project report via Canvas by

***16:30 on Friday 18th November**.*

Please note that:

- *The report should not exceed 12 pages. Project reports exceeding this limit will be penalised. You may also like to include an Appendix (which does not count towards the page limit) containing supplementary tabular and graphical output.*
- *You should submit your work as a single electronic file in PDF format along with a single file containing the R code used to produce your report.*

1 Project brief

In this project, you will analyse the `BreastCancer` data set which concerns characteristics of breast tissue samples collected from 699 women in Wisconsin using fine needle aspiration cytology (FNAC). This is a type of biopsy procedure in which a thin needle is inserted into an area of abnormal-appearing breast tissue. Nine easily-assessed cytological characteristics, such as uniformity of cell size and shape, were measured for each tissue sample on a one to ten scale. Smaller numbers indicate cells that looked healthier in terms of that characteristic. Further histological examination established whether each of the samples was benign or malignant. The objective of the clinical experiment was to determine the extent to which a tissue sample could be classified as benign or malignant using only the nine cytological characteristics.

For the purposes of this project, you may assume that the patients can be regarded as a random sample from the population of women experiencing symptoms of breast cancer.

The data set is part of the `mlbench` package. The package can be installed by typing into the console

```
> install.packages("mlbench")
```

It can then be loaded into R and inspected as follows:

```
> ## Load mlbench package
> library(mlbench)
> ## Load the data
> data(BreastCancer)
> ## Check size
> dim(BreastCancer)

[1] 699  11

> ## Print first few rows
> head(BreastCancer)
```

```
       Id Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size Bare.nuclei
1 1000025            5         1          1             1            2           1
2 1002945            5         4          4             5            7          10
3 1015425            3         1          1             1            2           2
4 1016277            6         8          8             1            3           4
5 1017023            4         1          1             3            2           1
6 1017122            8        10         10             8            7          10
  Bl.cromatin Normal.nucleoli Mitoses     Class
1           3               1       1    benign
2           3               2       1    benign
3           3               1       1    benign
4           3               7       1    benign
5           3               1       1    benign
6           9               7       1 malignant
```

More information on the variables can be found by typing `?BreastCancer` in the console.

1.1 Task

Your goal is to build a classifier for the `Class` – benign or malignant – of a tissue sample based on (at least some of) the nine cytological characteristics. It should be stressed that this is a real data set and there is no "correct" answer. Instead, what is required is evidence of an understanding of the main statistical ideas, sound interpretation of results, sensible and reasoned comparisons of classifiers, and demonstration of competence in the use of R as a tool for data analysis.

This part of the project should be written up as a coherent report, giving consideration to the points detailed in Section 1.1.1 below. You may like to include R code in your report. Alternatively, you can simply place the code in an Appendix and refer to it as appropriate. You do not need to comprehensively describe everything you have done to explore and model the data. However, you should provide a narrative which details and justifies the salient features of your approach, in addition to reporting and interpreting your results.

1.1.1 Points to consider

- You should begin by cleaning the data:
  - Technically, the nine cytological characteristics are ordinal variables on a 1 – 10 scale. In the `BreastCancer` data, they are encoded as factors. For the purposes of this project, we will treat them as quantitative variables. You should *carefully* convert the factors to quantitative variables.
  - This data set contains some missing observations on predictors, encoded as `NA`. For the purposes of this project, you should remove all of the rows where there are missing values before carrying out any further analysis. To do this, you may find the `is.na` function helpful. For instance

```
> ## Print 24th row of Breast Cancer data and note there is a NA in the
> ## Bare.nuclei column:
> BreastCancer[24,]
```

```
       Id Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size Bare.nuclei
24 1057013            8         4          5             1            2        <NA>
```

```
   Bl.cromatin Normal.nucleoli Mitoses     Class
24           7               3       1 malignant
> ## Test whether each element on the 24th row is a NA:
> is.na(BreastCancer[24,])
      Id Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size Bare.nuclei
24 FALSE        FALSE     FALSE      FALSE         FALSE        FALSE        TRUE
   Bl.cromatin Normal.nucleoli Mitoses Class
24       FALSE           FALSE   FALSE FALSE
```

- Consider some exploratory data analysis. For example, how might you summarise the data graphically and numerically? What does this tell you about the relationships between the response variable and predictor variables and about the relationships between predictor variables?

- You should build classifiers using each of the following methods:

  – At least one method for subset selection in logistic regression;

  – At least one regularized form of logistic regression, i.e. with a ridge or LASSO penalty;

  – At least one discriminant analyis method, i.e. the Bayes classifier for linear disciminant analysis (LDA) or quadratic discriminant analysis (QDA).

  For the variants of logistic regression, you should present the coefficients of the fitted model, and any other useful graphical or numerical summaries. For LDA and QDA present estimates of the group means. In each case, discuss what your results show. For example, which variables drop out of the model when you use subset selection or the LASSO? What do the parameters tell you about the relationships between the response and predictor variables?

- Compare the performance of your models using cross-validation based on the test error. Think about how you might do this in a way that makes the comparison fair.

- Select a final "best" classifier, justifying your choice. Does it include all the predictor variables? Why or why not?