MAS8383 Statistical Learning Methodology: Project

Semester 1, 2023/2024

Guidelines

- Submit your report (in pdf & Rmd) and the video of your oral presentation via the submission point on Canvas by **2023-11-24 (Fri) 16:00**.
- Report:
  - You must write a report which should not exceed twelve pages, written in Rmarkdown. Project reports exceeding this limit will be penalised.
  - You do not need to include an appendix for the R code as long as you submit the Rmd file, in which the R code is visible.
- Oral presentation:
  - The video of your oral presentation should not exceed five minutes. The format must be mp4 and the file must be zipped up as a Zip or Tar Archive file before uploading it to Canvas.
  - The oral presentation is pass / fail. This means it does not carry a mark, but must be passed in order to pass the module. Its main purpose is to encourage you to focus on explaining statistical ideas in your own words.

Data

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.

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

```
## Install the package
install.packages("mlbench")
```

The command above should be run at the console once, and **must not be included in the Rmd file**. Once the package is installed, 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
## 1 1000025            5         1          1             1            2
```

```
## 2 1002945            5         4          4            5            7
## 3 1015425            3         1          1            1            2
## 4 1016277            6         8          8            1            3
## 5 1017023            4         1          1            3            2
## 6 1017122            8        10         10            8            7
##   Bare.nuclei Bl.cromatin Normal.nucleoli Mitoses     Class
## 1           1           3                1       1    benign
## 2          10           3                2       1    benign
## 3           2           3                1       1    benign
## 4           4           3                7       1    benign
## 5           1           3                1       1    benign
## 6          10           9                7       1 malignant
```

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

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.

Before working on the parts below, you should begin by cleaning the data. There are a few points to consider:

- 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 and explain why this is a reasonable thing to do.
- 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
## 24 1057013            8         4          5            1            2
##    Bare.nuclei Bl.cromatin Normal.nucleoli Mitoses     Class
## 24        <NA>           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
## 24 FALSE        FALSE     FALSE      FALSE         FALSE        FALSE
##    Bare.nuclei Bl.cromatin Normal.nucleoli Mitoses Class
## 24        TRUE       FALSE           FALSE   FALSE FALSE
```

- While you should not drop any variables in the original data, you may want to create additional data frames with the necessary variables, depending on the method or model to be used. For example, for a supervised learning method, you will need to drop the `Id` variable:

```
## Create a new data frame without the Id variable
df0 <- BreastCancer[, names(BreastCancer) != "Id"]
```

  Similarly, for unsupervised learning methods, you will need to drop `Id` and `Class`.

Part 1: Exploratory data analysis (10%)

a. How might you summarise the data graphically and numerically?
b. What does this tell you about the relationships between the response variable (`Class`) and predictor variables (the cytological characteristics)?
c. What does this tell you about the relationships between predictor variables?

Part 2: Hierarchical clustering (16%)

a. Apply hierarchical clustering with single-linkage using correlation-based distance and plot the dendrogram. Do the variables separate the samples into the two groups? Explain the statistical reasoning behind your answer.
b. Repeat above using complete-linkage and average-linkage. Do your results depend on the type of linkage used?
c. Suppose you would like to know which variables differ the most across the two groups. Suggest a way to address this question, ideally using ideas from the chapter on PCA, and apply it here. Be aware that there is no single, "correct" answer; all that is required is a sensible solution.

Part 3: K-means clustering (24%)

a. Write your own function to implement the K-means algorithm. The function should take as arguments a matrix `X`, with n rows and p columns, representing the data matrix; and an integer `K` representing the number of clusters. It should return a list containing the cluster partition, cluster means and within-cluster sum-of-squares SS_W on termination. Your code *must* include comments to indicate how the function works.
b. Taking $K = 2$, apply your function to the breast cancer data. Run the function multiple times and find the one which leads to the smallest value of SS_W. Verify that this gives the same cluster partition as R's `kmeans` function when the latter is also run with multiple random starts.
c. For a given data set and choice of K, assuming the stopping conditions of R's `kmeans` function and your own K-means function are the same, will the procedure in the above part always give two equivalent cluster partitions? Explain your answer.

Part 4: Classification (40%)

Your goal in this part 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 there is no "correct" answer for this part. 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.

a. Split the data into a training and validation set. For the purpose of this part, take the data from (approximately) 80% of the patients as training data and the remaining as test data.

b. 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;

c. Present the coefficients of the fitted models, and any other useful graphical or numerical summaries. 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?

d. 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.

e. Select a final "best" classifier, justifying your choice. Does it include all the predictor variables? Why or why not? What is the nature of the misclassification errors it tends to make?

Form, presentation, and reproducibility (10%)

Overall presentation and writing text

- While this project brief contains several parts, each of which focuses on a chapter / method, you should view the whole project as a comprehensive analysis of one data set, and draw connections

between different parts in your report.
- Think about what you want to talk about the data. Include useful numerical and graphical summaries. Write some text that complements the numbers and plots in a coherent way.
- 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.
- Make sure there is sufficient text describing the background of the data, and the analysis.
- Do not copy a large part of the code in the text to explain what it does, as it wastes space.
- Check the punctuation and grammar e.g. don't forget to have a space between the full stop of last sentence and the beginning of next sentence.
- Write in complete sentences.

Rmarkdown

- The Rmd you submit has to match the pdf. This means that if I open the Rmd file on my computer and click "knit to pdf", I should be able to generate the same pdf. You should write everything using Rmarkdown, including tables and plots. In this way, your analysis can be reproduced, and the R code doesn't need to be included as an appendix.
- If you start writing your report from the Rmarkdown template, remember to remove the template section headings (e.g. "R Markdown", "Including Plots") and template text (such as "This is an R Markdown document ...") and enter text relevant to the assignment.
- Do not use absolute paths (that only exist on your computer) in the Rmd file, as others will not be able to reproduce your results. As a rule of thumb, you don't need absolute paths if the Rmd file and the data are in the same folder / directory.
- Do not hardcode the inline results. Look up "Rmarkdown inline code".
- For section headings, leave a space between the last # and the first character, so that the section headings will be rendered properly.
- Do not include the following commands (potentially with arguments within) in an Rmd file:
  - `install.packages()`
  - `View()`
  - `setwd()`
- To facilitate the flow of your analysis and save space, you may not want to show the R code (unless you need to) in the pdf version of the report. This can be done by the chunk option `echo = FALSE` e.g.

````
```{r, echo = FALSE}
1 + 1
```
````

- Leave an empty line after an R code chunk and before new text, so that the text starts on a new line in the generated document.

Plots

- Do not make plots for the sake of making them. Think about whether they are useful at showing the prominent aspects of the data. Also, think if the information can be alternatively summarised by numbers or tables.
- Make sure that the labels and the axes are large enough and informative.

Oral presentation (Pass / Fail)

Present a summary of the main findings from the project report. You can make your slides using whatever presentation software you like, for example Latex Beamer, PowerPoint or Keynote. The video of your oral presentation should not exceed five minutes.