Master of Science (Business Analytics)

—

MIS40970: Data Mining for Business Analytics

—

Assignment 1

Due date and time: 9 pm, Monday 8th April, 2019

Assessment weight: 15%

Introduction

This assignment is take-home and done individually, so it will measure your fundamental understanding of the coursework without depending on recall. It involves applying methods from this course and analysing the results according to various criteria.

There are two main purposes in doing this assignment:

- to develop further your ability to investigate a dataset using R in order to answer questions
- to develop your analysis and reporting skills

Assignment submission

Assignment submission should consist of two parts – (a) a written report on your work and (b) a zip file of all the R scripts you used in the assignment.

The report should include:

(a) A standard cover page stating that this is all your own work

(b) A title page, containing

- title and the actual date of the assignment submission
- your full name (as per UCD records) and student number

(c) The main body of the report containing your analysis and conclusions (no more than 10 pages)

(d) Appendices with graphs and tables at the end of the document (if needed)

The cover page, title page and appendices do not count towards the page limit.

The R scripts should run in RStudio. The code should be commented and formatted.

Name files with the report and the zip folder with R scripts:

- Surname_Student Number

Submit your assignments through Brightspace. As a backup, email report and code files to gborlikova@ucd.ie with the subject line Assignment, Surname_Student Number.

The report should be written in a reader friendly form (imagine presenting it to a business). The report does not need to contain all code and outputs, only snippets of code/output that are required to facilitate the flow of the report. Results should be consolidated into groupings that help comparison and analysis.

Create separate code files for each question that requires coding (file name – Surname_StudentNumber_QuestionNumber, relevant extention - .R). All stages of data manipulation should be documented in code. Comment code as needed. Incomplete or non-working code will affect the marks. However, even if you encounter problems and your code does not work, but you have a good grasp of the theory behind the questions, a good report answering the questions will still earn you some marks. Put notes/comments in the code indicating where you run into problems.

The final 5% of marks will be allocated based on the quality (readability) of the report and the code.

If the assignment is late, it will be penalised according to UCD policies. Also please check the University policy on plagiarism.

Assignment

Apply what you have learned about Data Mining to date. Include plain English interpretation of your analyses.

Q1 Compare and contrast supervised and unsupervised learning **[10 marks]**

Q2 Exercise 3.2 (Applied predictive Modeling M. Kuhn, K. Johnson) **[15 marks]**

Data were collected to predict disease in 683 soybeans. The 35 predictors are mostly categorical and include information on the environmental conditions (e.g., temperature, precipitation) and plant conditions (e.g., left spots, mold growth). The outcome labels consist of 19 distinct classes.

The data can be loaded via:

```
> library(mlbench)
> data(Soybean)
> ## See ?Soybean for details
```

(a) Investigate the frequency distributions for the categorical predictors. Are any of the distributions degenerate in the ways discussed in the relevant book chapter?
(b) Roughly 18% of the data are missing. Are there particular predictors that are more likely to be missing? Is the pattern of missing data related to the classes?
(c) Develop a strategy for handling missing data, either by eliminating predictors or imputation.

Q3 Exercise 4.4 (parts a-c) (Applied predictive Modeling M. Kuhn, K. Johnson) **[15 marks]**

96 samples of seven types of oils were used to develop a methodology for food laboratories to determine the type of oil from a sample. The data can be found in the caret package using data(oil). The oil types are contained in a factor variable called oilType. The types are pumpkin (coded as A), sunflower (B), peanut (C), olive (D), soybean (E), rapeseed (F) and corn (G).

```
> library(caret)
> data(oil)
> ## See ?oil for details
> str(oilType)
Factor w/ 7 levels "A","B","C","D",..: 1 1 1 1 1 1 1 1 1 1 ...
> table(oilType)
oilType
A  B  C  D  E  F  G
37 26 3  7  11 10 2
```

(a) Use the sample function in base R to create a completely random sample of 60 oils. How closely do the frequencies of the random sample match the original samples? Repeat this procedure several times of understand the variation in the sampling process.

(b) Use the caret package function createDataPartition to create a stratified random sample. How does this compare to the completely random samples?
(c) With such a small samples size, what are the options for determining performance of the model? Should a test set be used?

Q4 Exercise 16.1 (parts a-e) (Applied predictive Modeling M. Kuhn, K. Johnson) **[30 marks]**

The "adult" data set at the UCI Machine Learning Repository is derived from census records. In these data, the goal is to predict whether a person's income was large (defined in 1994 as more than $50K) or small. The predictors include educational level, type of job (e.g., never worked, and local government), capital gains/losses, work hours per week, native country, and so on (exclude variable fnlwgt from analysis). After filtering out data where the outcome class is unknown, there were 48842 records remaining. The majority of the data were associated with a small income level (75.9%).

The data are contained in the arules package and the appropriate version can be loaded using

```
> library(arules)
> data(AdultUCI)
> ## See ? AdultUCI for details
```

(a) Load the data and investigate the predictors in terms of their distributions and potential correlations.
(b) Determine an appropriate split of the data.
(c) Build several classification models for these data (at least one algorithm of the type decision tree). Output relevant summaries and confusion matrices. Do the results favour the small income class?
(d) Is there a good trade-off that can be made between the sensitivity and specificity?
(e) Use sampling methods to improve the model fit. Pay attention to the models' generalisation and ability to correctly predict both classes.

Q5 Exersise 9 (with modifications) chapter 10 (An Introduction to Statistical Learning G. James, D. Witten, T. Hastie, R. Tibshirani) **[25 marks]**

Load USArrests dataset uploaded to Brightspace. Perform hierarchical and K-Means clustering on the states.
(a) Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states.
(b) Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?
(c) Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one.
(d) What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed? Provide a justification for your answer.

(e) Cluster the states using K-Means with k = 3 and 20 restarts. Try K-Means on both raw data and scaled data. Compare the results on raw and scaled data with the results of hierarchical clustering.

Additional notes

The final **[5 marks]** will be allocated based on the quality (readability) of the report and the code.

You might need to first install relevant package (install.packages()) and then load it (library()) in order to load the data.

Description of USArrests dataset

Violent Crime Rates by US State

This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.

[, 1] State character
[, 2] Murder numeric Murder arrests (per 100,000)
[, 3] Assault numeric Assault arrests (per 100,000)
[, 4] UrbanPop numeric Percent urban population
[, 5] Rape numeric Rape arrests (per 100,000)