UNIVERSITY OF CAPE TOWN
DEPARTMENT OF STATISTICAL SCIENCES
Data Science – Supervised Learning 2019

ASSIGNMENT 1

Due date: Tuesday, 26 March 2019 at 12:00
Late submissions will be penalised at 10% per day (pro rata)

INSTRUCTIONS:

- Present your final report as a pdf document. You may use any typesetting software you wish, but I would encourage you to use R-Markdown or LaTeX.
- Provide complete code for both questions under separate headings as an appendix to your write-up. Start each question on a new page.
- You may NOT provide R output interspersed between your answers! Please typeset relevant elements in the output either in-line, or tabulate results formally. Plots can be very useful, but use them sparingly – make sure that a given plot is relevant to the question and pertains to text in your answer. Figures are meant to enrich your analysis, don't leave it to the reader to analyse. Provide captions for all figures and tables. Square figures only!
- When you typeset R code use `courier` or an equivalent 'typewriter'-like font.
- You are expected to work on this **on your own**. Please attach a plagiarism declaration to your report – a template is provided on Vula.

QUESTION 1 – PREDICTING HOUSE VALUES IN BOSTON SUBURBS

The goal of this exercise is to predict the median value of owner occupied homes (in \$1000s) for suburbs in Boston, based on 12 explanatory variables.

The full dataset, available on Vula, is given in the file `boston.csv`. Each of the 506 rows in the dataset corresponds to a different suburb. The descriptions of the variables are provided in the file `boston_vars.txt`.

You will not be working with the full dataset; each student must first create thier own unique sub-sample of 400 observations **which is to be used for this problem**. To do this, run the `boston_data_splitter.R` file, after setting the seed to your unique project number.

(a) After splitting your dataset into separate training and testing sets, fit a multiple linear regression model to your training set, regressing `medv` on all the explanatory variables. Discuss the model in terms of its fit and variable significance. Use your model to predict `medv` for the testing set, and calculate the corresponding mean squared error (MSE).

(b) Attempt to improve on the linear model fitted in (a), where improvement is measured by the testing set MSE. You may apply any of the variable selection techniques covered in the course, including LASSO regularization. Your final model must also include an interaction term (you may have more than one, but one is sufficient). Be sure to motivate the reasoning behind the choice of interaction. Briefly investigate and discuss the residual diagnostics of your final model.

(c) Fit a large regression tree to your training data by relaxing the stopping criteria, and use this tree to predict on the test set. Prune your tree down, motivating the choice of tree size. How does your final tree compare to your final regression model in (b) in terms of prediction?

Question 2 – Classifying Email as Spam or Legitimate

Email spam, or junk mail, has been plaguing inboxes and threatening users since the mid-1990's. Apart from being an annoyance, modern unsolicited emails often contain disguised links which lead to viruses, malware or phishing web sites. When recipients of junk mail label incoming emails as spam, we can look for features in the content of the mail that distinguishes it from legitimate email, thereby presenting us with a supervised classification problem where the goal is to correctly classify incoming mail based on these features.

The dataset provided in `spam_data.csv` contains 3,601 observations that were classified either as spam or email. For each email 57 attributes were measured, the descriptions of which can be found here. This dataset can also be loaded from the R package `ElemStatLearn`; after loading the library, the command `?ElemStatLearn::spam` will also provide information on the data.

(a) Fit a logistic regression model to your training set, applying all of the variable selection techniques and choose one of the techniques. Interpret your results and make sure your findings make sense.

(b) Do the same as in (a) using discriminant analysis and in addition investigate the validity of the assumptions of your model choice.

(c) Use the `h2o` R package to fit random forests to the aforementioned data. Make use of the `h2o.grid` function to try various combinations of the different possible model parameters. Identify and discuss your best model, including variable importance.

(d) Repeat the process in (c) for boosted trees.

(e) Test all the above models on the unseen data contained in `spam_test.csv`. Report on each model's performance by at least investigating the misclassification rate and interpreting this in context of the problem.