Learning Objectives: Data Science and Big Data Analytics

Working in R

- Be able to use R to perform mathematical calculations.
- Be able to use R to perform element-wise calculations to columns, (e.g. adding columns together, averaging columns, scaling columns).
- Be able to write a for loop in R e.g. to print a simple sequence 2,4,6,8 or perform a certain task 100 times.
- Be able to create empty data frames / vector columns that can be used to store the results calculated in a for loop.
  (Be able to use the ifelse function in R on a column to generate a new column based on those values.)
  (Be able to write simple R functions that take some input arguments and return a calculated result.)

Working with datasets in R

- Be able to load data in R using a variety of formats (csv, xls, Rdata)
- Be able to manipulate data:
  make selections of rows/columns,
  delete rows /columns,
  filter rows based on conditions,
  use column data to create new columns, c
  convert columns types (factor/character/numerical),
  be able to split rows into factors based on cuts on a numerical column,
- Be able to rename row and column labels
- Be able to sort a dataframe by column values
- Be able to use dplyr functions:

  select filter mutate arrange group_by / summarise

  (Be able to normalise columns (by shifting and scaling) so that it has mean = 0 and standard deviation = 1.)

Exploring datasets in R

- Be able to find and interpret summary statistics of a dataset column. (mean, median, standard deviation, variance, range, quantiles.
- Be able to make and interpret the following types of plots in R for any variable / pairs of variables: scatter plot, histogram (with defined cut/bin positions), boxplot
- Be able to identify outliers and take appropriate actions (include, filter out, or correct)
- Be able to use the table command to build a frequency table and interpret the results.
  (Be able to style plots: set x and y axis limits, set x and y axis labels, set title, set point marker types, set line styles, and set point colour based on a criteria or factor class.)
  (Be able to add annotations to plots: add text to a plot, add vertical / horizontal lines at given values, smoothed line onto scatter plot using scatter.smooth)
  (Be able to takes appropriate steps to identify where variables have missing data, and take appropriate measures (removal from data frame or exclusion from an analysis).)

Learning Objectives: Data Science and Big Data Analytics

Hypothesis testing in R

- Understand the assumptions made in performing a t-test, how to carry it out in R and interpret and explain the results (t-value, p-value, confidence interval).
- Be able to carry out a t-test on two groups of values to find evidence for a difference in the means, or test a hypothesis that one mean is greater/less than the other.
- Be able to carry out a t-test on two columns to find evidence to test the hypothesis that the columns are correlated.

Bivariate Regression in R

- Understand and be able to explain the principle of least-squares regression.
- Be able to carry out simple linear regression in R to make a linear model that predicts a response based on a single independent variable, and plot the line of best fit onto a scatter plot of the data.
- Be able to display and understand the results of the fit, and the associated measures returned.
- Be able to recall and explain the assumptions of linear regression, use R to explore if these assumptions are met, and carry out appropriate steps to identify and manage these issues.
- Be able to explore how simple variable transformations improve/do not improve linear fitting.
- Be able to make predictions using the fit, access the residuals, and calculate values of RSS, MSE, and RMSE.

Multivariate Regression in R

- Understand how to perform a multivariate fit in R.
- Understand how to interpret the fit coefficients in a multivariate fit, and their associated p-values.
- Understand how to interpret the f-test value returned from a multivariate fit.
- Understand how multivariate fits can include categorical variables, the use of dummy variables, and the interpretation of the resulting fit coefficients and their associated p-values.
- Be able to plot the result of multivariate fits appropriately.
- Be able to include interaction terms in a multivariate fit, and test for their significance.
- Understand the problem arising with highly correlated predictors, and be able to detect and suggest suitable steps to deal with these issues.

Learning Objectives: Data Science and Big Data Analytics

Model selection

- Be able to interpret fit results to evaluate the important of the different predictors used, and suggest model improvements.
- Understand how to perform an ANOVA test on a multivariate model to compare the performance to a model containing only a subset of the predictors, and interpret the result.
- Be able to use the stepAIC function to perform stepwise model selection to optimise a model, understanding the arguments that can be used.
- Be able to use the leaps function to test all possible predictor subsets and interpret the results and plots produced.

Cross Validation

- Understand why the performance of a model on its training data may differ significantly in comparison to a test / validation / new data.
- Be able to describe the approaches of validation set (aka hold out), LOOCV, k-fold validation, their relative strengths/disadvantages, and be able to carry them out in R using suitable library functions.
- Understand how to interpret the results of validation testing to select optimum models.
- Understand the method of resampling to generate boostrapped datasets that can be used when evaluating model performance.
- Understand how the process of bootstrapping can be used to investigate the distribution of fit parameter results.

Classification

- Be able to use R to calculate simple probabilities associated with sampling a dataset.
- Understand how the logistic function can be fitted to data to predict the probabilities associated with a binomial (0 or 1) variable in relation to the values of a set of predictors. Know why it is used in preference to a linear function.
- Be able to carry out a logistic regression fit in R and interpret the results.
- Understand how to display the results of a classification model using table to display the a confusion matrix.
- Be able to use the results of a classification model to make predictions, and measure the performance according to the misclassification rate.
- Be able to describe the assumptions which Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) make to build a classification model.
- Be able to use R to perform a LDA (Linear Discriminant Analysis) fit for classification.
- Be able to use R to perform a QDA (Quadratic Discriminant Analysis) fit for classification.
- Be able to apply methods of cross validation and model selection to classification methods.

Learning Objectives: Data Science and Big Data Analytics

K- Nearest Neighbours algorithm

- Be able to describe the principle of the K-nearest neighbours algorithm for classification and regression problems.
- Understand KNN requires consideration of the scale of each variable used as a predictor, and how to apply normalisation when appropriate.
- Be able to explore how the performance of the KNN method varies with K for a particular dataset and select an optimal K value using a cross validation method.

Comparisons of Fitting methods

- Understand that models can be used for both inference and prediction.
- Be able to describe the fitting methods considered in terms of their assumptions, and relative usefulness for making predictions/inferences.
- Be able to carry out suitable cross validation testing to measure the performance of different fit methods and evaluate the results.

Tree-based methods

(no exam questions will be based on this topic, but you may wish to utilise them in the exam where you have the option to apply an analysis of your own choice.)

- Understand how a decision tree can be used to build a model that can be used for regression or classification purposes.
- Understand the meaning of "top-down greedy" as applied to the method by which decision trees are constructed, and the measures used to select optimal split conditions at each branch.
- Understand that decision trees can overfit training data, and the use of cross-validated pruning to select an optimal tree size.
- Be able to use R to construct and prune trees to an optimal size using cross validation.
- Understand how the approach of bagging can be used to generate a set (also known as an ensemble) of decision trees who's predictions can be combined.
- Understand how Random Forest method modifies bagging, and the meaning of parameters *m* and *p* to the method.
- Be able to describe how Out-Of-Bag data points can be used to cross validate bagging and Random Forest model performance.
- Understand how to use the Random Forest library in R to carry out both bagging and Random forest methods, and evaluate model performance.
- Be able to explore how the number bagged datasets used (B) affects the performance of the bagging and Random Forest method and use this to select an optimum value.
- Be able to explore how different values of *m* (`mpar` in R) effect the performance of the Random Forest method and be able to use the result to select an optimal value.
- Be able to carry out the method of boosted decision trees in R to build a model, and assess performance based on the value of λ and number of trees used.