STAT3406 and STAT4064
Semester 1, 2023
Assignment 1

The question, section and page numbers refer to the textbook:
James, G, Witten, D, Hastie, T and Tibshirani, R (2021). *An Introduction to Statistical Learning: with Applications in R*, Second Edition, Springer-Verlag, New York.

Assignments may be completed by teams of 2 students or individually. If two students work together, submit one assignment with names and student IDs and clearly state each person's contribution. Your answers should be in a report format. It is not expected any `R` code, and this will not be marked. Keep in mind that the presentation of your findings is essential. Printing out the software output is not a valid answer.

Due date: Thursday 06 April 2023 at 5 pm

Assignment Questions

1. Consider the confusion matrix of a 2-class problem.

   (a) [2 marks] Explain the two ways an incorrect classification can be made and why it may be important to make a distinction between the two types of error.

   (b) [2 marks] Describe briefly a scenario in which it does matter which of the two types of error has been made. Explain why the distinction matters for your scenario.

   (c) [3 marks] Assume we have the confusion matrix of a 2-class problem obtained from the training data. Describe and give reasons for the change (including the direction of change – positive or negative) you would expect to see when you calculate a confusion matrix for the testing data.

2. Use the `Auto` data available at the `R` package `ISLR2`, which is described in Section 2.3.4, p. 48.

   (a) [1 marks] Divide the data in two groups of equal size (first and second half, do not randomise). Consider the first record of the second half of the data and list the value of `acceleration` of this record.

   (b) [2 marks] Use all observations of the data. Use `acceleration` as the response (Y), and `displacement`, `horsepower`, `weight`, and `mpg` as predictor variables (features, X). Conduct a linear regression with pairwise interactions between the predictor variables (`horsepower`, `weight`), and (`mpg`, `weight`). Considering the hypothesis tests for $H_0 : \beta_j = 0$, for each $j = 1, \ldots, k$ (k is the number of parameters in your model), state which predictors variables you would keep in your model. Consider a 5% significance level.

   (c) [2 marks] Write down an expression for the final model you derive in part (2b) and comment.

3. Consider the `Auto` data and the variables `acceleration`, `displacement`, `horsepower`, `weight`, and `mpg`. In items (3b) – (3e) we use `lda` with the default proportions for classification.

(a) [1 marks] Create a new qualitative variable `mpgclass` with categories 'low', 'medium' and 'high' as follows:
`mpgclass` is 'low' if `mpg` < 20
`mpgclass` is 'medium' if $20 \leq$ `mpg` < 27
`mpgclass` is 'high' if `mpg` ≥ 27.
Present the proportion of each category.

(b) [2 marks] Use `acceleration, displacement, horsepower` and `weight` as predictors (X), and `mpgclass` as the class labels (Y). Perform an LDA and determine the classification error on the data and show its confusion matrix.

(c) [2 marks] Extract a data set considering only the cars from the year 75, and consider it as a test data. Apply the rule constructed in part (3b) to the this test data. Calculate and show the test error and display your results in a confusion matrix.

(d) [2 marks] Consider the test data from item (3c), cars from the year 75, and construct a training data with cars from all other year. Perform classification on the training data and calculate the error on the training data. Show the training error and the the confusion matrix for the training error.

(e) [2 marks] Consider the training and test data from item (3d), and use the classification rule obtained there (3d) to predict the class membership of the cars in the test data. State your test error and display the results obtained on the test data in a confusion matrix.

(f) [3 marks] Compare the results of items (3b)–(3e) and comment on the various errors and confusion matrices. Explain why we expect the test error in item (3c) to be smaller than that obtained in part (3e).