DAR/IDAR Coursework 1

- Please submit ONE file to the coursework 1 answer submission portal on moodle: one of the following (.pdf/.html/.doc) files created by RStudio. Please include any R code, plots or results.
- Your files should be named as follows:
  MSc/BSc_CW1_xxxxxxxx_initial_lastname.pdf (.html/.doc)
  where xxxxxxxx is your student ID. For instance, MSc/BSc_CW1_12345678_Wan.pdf.
- Don't forget to write down your programme (MSc or BSc), name and student ID on the first page of your answer sheets as well.
- Each question below has two weightings. The first weighting is for MSc students and the second weighting is for BSc students.

1. Statistical learning methods (12% | 12%)

Marking scheme:
- MSc: 3% each
- BSc: 3% each.

For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

(a) The number of predictors p is extremely large, and the number of observations n is small.

(b) The sample size n is extremely large, and the number of predictors p is small.

(c) The relationship between the predictors and response is highly non-linear.

(d) The standard deviation of the error terms, i.e. $\sigma = \text{sd}(\varepsilon)$, is extremely high.

2. Bayes' rule (12% | 12%)

Marking scheme:
- MSc: 1% for each probability.
- BSc: 1% for each probability.

Given a dataset including 20 observations (S_1, ..., S_20) about the temperature (i.e. hot or cool) for playing golf (i.e. yes or no), you are required to use the Bayes' rule to calculate by hand the probability of playing golf according to the temperature, i.e. P(Play Golf | Temperature).

	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_10
Temperature	cool	hot	hot	hot	cool	cool	hot	cool	hot	hot
Play Golf	yes	no	yes	no	yes	yes	no	yes	no	yes

	S_11	S_12	S_13	S_14	S_15	S_16	S_17	S_18	S_19	S_20
Temperature	hot	hot	hot	cool	hot	hot	cool	cool	cool	hot
Play Golf	no	no	yes	no	no	no	yes	no	no	no

3. Descriptive analysis (22% | 22%)

Marking scheme: - MSc: (a) - (d) 3% each, the rest 5% each. - BSc: (b)(c) 3% each, the rest 4% each.

This question involves the `Auto` dataset included in the "ISLR" package.

(a) Which of the predictors are quantitative, and which are qualitative?

(b) What is the range of each quantitative predictor? You can answer this using the `range()` function.

(c) What is the median and variance of each quantitative predictor?

(d) Now remove the 11th through 79th observations (inclusive) in the dataset. What is the range, median, and variance of each predictor in the subset of the data that remains?

(e) Using the full data set, investigate the relationship between individual predictors with the target response gas mileage (`mpg`) graphically. Comment on your findings.

(f) Suppose that we wish to predict `mpg` on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting `mpg`? Justify your answer.

4. Linear regression (24% | 24%)

Marking scheme: - MSc: (a) i - iv 3% each, (b) 4% (c) 8%. - BSc: (a) i - iv 3% each, (b)(c) 6% each.

This question involves the use of simple linear regression on the `Auto` dataset.

(a) Use the `lm()` function to perform a simple linear regression with `mpg` as the response and `acceleration` as the predictor. Use the `summary()` function to print the results. Comment on the output. For example:

i. Is there a relationship between the predictor and the response?

ii. How strong is the relationship between the predictor and the response?

iii. Is the relationship between the predictor and the response positive or negative?

iv. What is the predicted `mpg` associated with an `acceleration` of 14.50? What are the associated 97% confidence and prediction intervals?

(b) Plot the response and the predictor. Use the `abline()` function to display the least squares regression line.

(c) Plot the 97% confidence interval and prediction interval in the same plot as (b) using different colours and legends.

5. Logistic regression and cross validation (30% | 30%)

Marking scheme: - MSc: (a)-(d) 5% each, (e) 10%. - BSc: 6% each.

A recent study has shown that the accurate prediction of the office room occupancy leads to potential energy savings of 30%. In this question, you are required to build logistic regression models by using different environmental measurements as predictors (features), such as temperature, humidity, light, CO_2 and humidity ratio, to predict the office room occupancy. The provided training dataset consists of 2,000 observations, whilst the testing dataset consists of 300 observations.

(a) Load the training and testing datasets from corresponding files, and display the statistics about different predictors in the training dataset.

(b) Conduct a 10-fold cross validation to evaluate the predictive accuracy of a logistic regression model that uses Temperature as the only predictor. Report the average accuracy and AUROC value obtained over the 10-fold cross validation. Set the value of random seed as "100" when generating fold indices. Consider the predictive label equals to 1, if the predictive probability is greater than 0.5.

(c) Conduct a 10-fold cross validation to evaluate the predictive accuracy of a logistic regression model that uses HumidityRatio as the only predictor. Report the average accuracy and AUROC value obtained over the 10-fold cross validation. Set the value of random seed as "100" when generating fold indices. Consider the predictive label equals to 1, if the predictive probability is greater than 0.5.

(d) Conduct a 10-fold cross validation to evaluate the predictive accuracy of a logistic regression model that uses Temperature and HumidityRatio in the training dataset. Report the average accuracy and AUROC value obtained over the 10-fold cross validation. Set the value of random seed as "100" when generating fold indices. Consider the predictive label equals to 1, if the predictive probability is greater than 0.5.

(e) Compare the performance of those three different models on predicting the testing dataset. Draw ROC curves for all individual models and calculating the corresponding AUROC values. Discuss the comparison results obtained by the 10-fold cross validation and the hold-out testing.