BDA/IDAR Coursework 1

- Please submit TWO files to Dropbox 1 on moodle: a .rmd file and knit it to one of the following (.doc/.pdf/.html) files. Please include any R code, plots or results.
- Your files should be named as follows:
  MSc/BSc_CW1_xxxxxxxx_initial_lastname.rmd (.pdf/.html/.doc)
  where xxxxxxxx is your student ID. For instance, MSc_CW1_12345678_T_Han.rmd.
- 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. For instance, (10% | 0%) means that the question is worth 10% for MSc students and 0% for BSc students (optional).

1. Statistical learning methods **(8% | 12%)**

Marking scheme:
- MSc: 2% 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: 1% for each probability.

Given a dataset including 20 samples (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	hot	hot	hot	cool	hot	hot	cool	cool	cool	hot
Play Golf	no	no	yes	no	no	no	yes	no	no	yes

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

3. Descriptive analysis (12% | 22%)

Marking scheme:
- MSc: 2% each.
- BSc: (b)(c) 3% each, the rest 4% each.

This exercise involves the `Auto` data set studied in the class.

(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 predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

(f) Suppose that we wish to predict gas mileage (`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 (16% | 24%)

Marking scheme:
- MSc: (a) i - iv 2% each, (b)(c) 4% each.
- BSc: (a) i - iv 3% each, (b)(c) 6% each.

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

(a) Use the `lm()` function to perform a simple linear regression with mpg as the response and horsepower 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 a horsepower of 89? What are the associated 99% 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 99% confidence interval and prediction interval in the same plot as (b) using different colours and legends.

5. Logistic regression (24% | 30%)

Marking scheme:
- MSc: (a)-(d) 4% each, (e) 8%.
- 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 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 samples, whilst the testing dataset consists of 300 samples.

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

(b) Build a logistic regression model by only using the Temperature feature to predict the room occupancy. Display the confusion matrix and the predictive accuracy obtained on the testing dataset.

(c) Build a logistic regression model by only using the Humidity feature to predict the room occupancy. Display the confusion matrix and the predictive accuracy obtained on the testing dataset.

(d) Build a logistic regression model by using all features to predict the room occupancy. Display the confusion matrix and the predictive accuracy obtained on the testing dataset.

(e) Compare the predictive performance of three different models by drawing ROC curves and calculating the AUROC values. Discuss the comparison results.

6. Resampling methods **(28% | 0%)**

Marking scheme: MSc only. 4% each.

We are trying to learn regression parameters for a dataset which we know was generated from a polynomial of a certain degree, but we do not know what this degree is.

Assume the data was actually generated from a polynomial of degree 3 with some added noise, that is

$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon, \quad \varepsilon \sim N(0, 1)$

For training we have 400 (x, y)-pairs and for testing we are using an additional set of 400 (x, y)-pairs. Since we do not know the degree of the polynomial we learn two models from the data.

- Model A learns parameters for a polynomial of degree 2, and
- Model B learns parameters for a polynomial of degree 4.

(a) Which of these two models is likely to fit the test data better? Justify your answer.

We will now perform cross-validation on this simulated data set.

(b) Generate the simulated data set as follows:

```
set.seed(235)
x = 12 + rnorm(400)
y = 1 - x + 4*x^2 - 5*x^3 + rnorm(400)
```

Create a scatterplot of X against Y. Comment on what you find.

(c) Set the seed to be 34, and then compute the LOOCV and 10-fold CV errors that result from fitting the following two models using least squares:

i. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$
ii. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \varepsilon.$

Note you may find it helpful to use the `data.frame()` function to create a single data set containing both X and Y.

(d) Repeat (c) using random seed 68, and report your results. Are your results the same as what you got in (c)? Why?

(e) Which of the models in (c) had the smallest LOOCV and 10-fold CV error? Is this what you expected? Explain your answer.

(f) Comment on the coefficient estimates and the statistical significance of the coefficient estimates that results from fitting the preferred model in (a).

(g) Fit a cubic model and compute its LOOCV error, 10-fold CV error under seed 34, and comment on the coefficient estimates and the statistical significance of the coefficient estimates. Compare the LOOCV and 10-fold CV error with the preferred model in (a).