STAT 318/462: Data Mining
Assignment 2
Due Date: 3pm, 21st September, 2018

Your printed assignment must be submitted in the STAT318/462 assignment box on the fourth floor of the Erskine building (by MATH/STAT reception).

You can work in pairs on this assignment and submit a single co-authored set of answers. Marks will be lost for unexplained, poorly presented and incomplete answers. Whenever you are asked to do computations with data, feel free to do them any way that is convenient. If you use *R* (recommended), please provide your code. All figures and plots must be clearly labelled.

1. **(2 marks)** Suppose we collect data for a group of students that have taken STAT318 with variables X_1 = hours spent studying per week, X_2 = number of classes attended and

$$Y = \begin{cases} 1 & \text{if the student received a GPA value} \geq 7 \text{ in STAT318} \\ 0 & \text{otherwise.} \end{cases}$$

We fit a logistic regression model and find the estimated coefficients to be $\hat{\beta}_0 = -16, \hat{\beta}_1 = 1.4$ and $\hat{\beta}_2 = 0.3$.

(a) Estimate the probability of a student getting a GPA value ≥ 7 in STAT318 if they study for 5 hours per week and attend all 36 classes.

(b) If a student attends 18 classes, how many hours do they need to study per week to have a 50% chance of getting a GPA value ≥ 7 in STAT318?

2. **(10 marks)** In this question, you will fit a logistic regression model to predict the probability of a banknote being forged using the Banknote data set. This data has been divided into training and testing sets: `BankTrain.csv` and `BankTest.csv` (download these sets from Learn). The response variable is y (the fifth column), where $y = 1$ denotes a forged banknote and $y = 0$ denotes a genuine banknote. Although this data set has four predictors, **you will be using x_1 and x_3 to fit your model**[1].

(a) Perform multiple logistic regression using the training data. **Comment on the model obtained**.

(b) Estimate the standard errors for $\hat{\beta}_1$ and $\hat{\beta}_2$ (regression coefficients for x_1 and x_3) using the bootstrap.

(c) Suppose we classify observations using

$$f(x) = \begin{cases} \textit{forged banknote} & \text{if } \Pr(Y = 1|X = x) > \theta \\ \textit{genuine banknote} & \text{otherwise.} \end{cases}$$

i. Plot the training data (using a different symbol for each class) and the decision boundary for $\theta = 0.5$ on the same figure.

[1] These predictors are features extracted from an image of a banknote: x_1 is the variance of a Wavelet Transformed image and x_3 is the kurtosis of a Wavelet Transformed image.

ii. Using $\theta = 0.5$, compute the confusion matrix for the testing set and **comment** on your output.

iii. Find a value of θ that reduces the training error as much as possible. Compute the confusion matrix for the testing set using your best θ value and **comment** on your output.

3. **(6 marks)** In this question, you will fit linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) models to the training set from question 2 of this assignment.

   (a) Fit an LDA model to predict the probability of a banknote being forged using the predictors x_1 and x_3. Compute the training and test error (using the testing set from question 2).

   (b) Repeat part (a) using QDA.

   (c) **Comment** on your results from parts (a) and (b). Compare these methods with the logistic regression model from question 2. **Which method would you recommend and why?**

4. **(2 mark)** Consider a binary classification problem $Y \in \{0, 1\}$ with one predictor X. Assume that X is normally distributed ($X \sim \mathrm{N}(\mu, \sigma^2)$) in each class with $X \sim$ N(0,4) in class 0 and $X \sim$ N(2,4) in class 1. Calculate Bayes error rate when the prior probability of being in class 0 is $\pi_0 = 0.4$. (*Bayes error rate is the test error rate using Bayes classifier.*)