STAT 4533/5533 - Homework 2 - Due February 14, 2024

Instructions: Everyone must answer Questions 1 and 2(a) - (f). Graduate students are also required to answer Questions 2(g), 2(h), and 3. Questions 4 and 5 are optional and will not be graded. Solutions must be created using R Markdown, and you will upload your HTML file to D2L to submit your answers. Your markdown document must be logically organized and easy to read. Do not print entire data sets in your output.

1. Chapter 4 Exercise 6

2. This question uses the `iris` data set which contains information related to flower measurements for three different species of iris. You do not need to create training and test sets for this problem.

   (a) Create a new data frame called `iris2` by removing all of the setosa species flowers from the original data set.

   (b) Fit a logistic regression model on the data to predict species using the other variables.

   (c) Repeat part (b) using LDA.

   (d) Repeat part (b) using QDA.

   (e) Repeat part (b) using naive Bayes.

   (f) Which model performs best? Justify your conclusion with appropriate statistics.

   (g) GRADUATE STUDENTS ONLY: Which model had the highest sensitivity?

   (h) GRADUATE STUDENTS ONLY: Does it seem reasonable to use LDA and QDA with this data. Do some investigating and explain why or why not.

3. GRADUATE STUDENTS ONLY: A statistician fits a logistic regression model to classify a binary response variable using one predictor variable. You know that $P(Y = 1|X = 4) = 0.88877$ and $P(Y = 1|X = 6) = 0.96562$. Use this information to find the estimated values of β_0 and β_1.

   The following questions are recommended, but will not be graded:

4. Chapter 4 Exercise 14 - For parts (d) - (g) use cylinders, displacement, horsepower, weight, and acceleration, and also remove the original mpg variable from the dataset.

5. Chapter 4 Exercise 15