In this assignment, we will use the mathematical notations from the following reference books:

- **Reference books:**

  1. *An Introduction to Statistical Learning with Applications in R* **(ISLR)**. Authors: James, G., Witten, D., Hastie, T., Tibshirani, R.

  2. *The Elements of Statistical Learning Data Mining, Inference, and Prediction* **(ESL)**. Authors: Hastie, T., Tibshirani, R., Friedman, J.

1 Regression

(a) **(3 points)** Explain the differences between the KNN classifier and KNN regression methods.

(b) **(7 points)** Suppose we have a data set with five predictors, X_1 =GPA, X_2 = IQ, X_3 = Gender (1 for Female and 0 for Male), X_4 = Interaction between GPA and IQ, and X_5 = Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50, \hat{\beta}_1 = 10, \hat{\beta}_2 = 1, \hat{\beta}_3 = 30, \hat{\beta}_4 = 0.01, \hat{\beta}_5 = -10$,.

- Which answer is correct, and why? i. For a fixed value of IQ and GPA, males earn more on average than females. ii. For a fixed value of IQ and GPA, females earn more on average than males. iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough. iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.
- Predict the salary of a female with IQ of 110 and a GPA of 4.0.
- True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

2 Classification

(a) **(6 points)** Using linear algebra, prove equation (1) and (2) are equivalent:

$$p(X) = \frac{e^{(w_0+w_1X)}}{1+e^{(w_0+w_1X)}} \tag{1}$$

$$\frac{p(X)}{1-p(X)} = e^{(w_0+w_1X)} \tag{2}$$

(b) **(8 points)** In this question, examine the difference between LDA and QDA.

- If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set? On the test set?
- If the Bayes decision boundary is non-linear, do we expect LDA or QDA to perform better on the training set? On the test set?

- In general, as the sample size n increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or be unchanged? Why?
- True or False: Even if the Bayes decision boundary for a given problem is linear, we will probably achieve a superior test error rate using QDA rather than LDA because QDA is flexible enough to model a linear decision boundary. Justify your answer.

(c) **(6 points)** Suppose we collect data for a group of students in a data analytics class with variables X_1 = hours studied, X_2 =undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, $\hat{\beta}_0 = -6, \hat{\beta}_1 = 0.1, \hat{\beta}_2 = 2$.

- Estimate the probability that a student who studies for 40 hours and has an undergrad GPA of 3.8 gets an A in the class.
- How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?

3 Decision Trees

(a) **(5 points)** Suppose you want to grow a decision tree to predict the *accident rate* based on the following accident data which provides the rate of accidents in 100 observations. Which predictor variable (weather or traffic) will you choose to split in the first step to maximize the information gain?

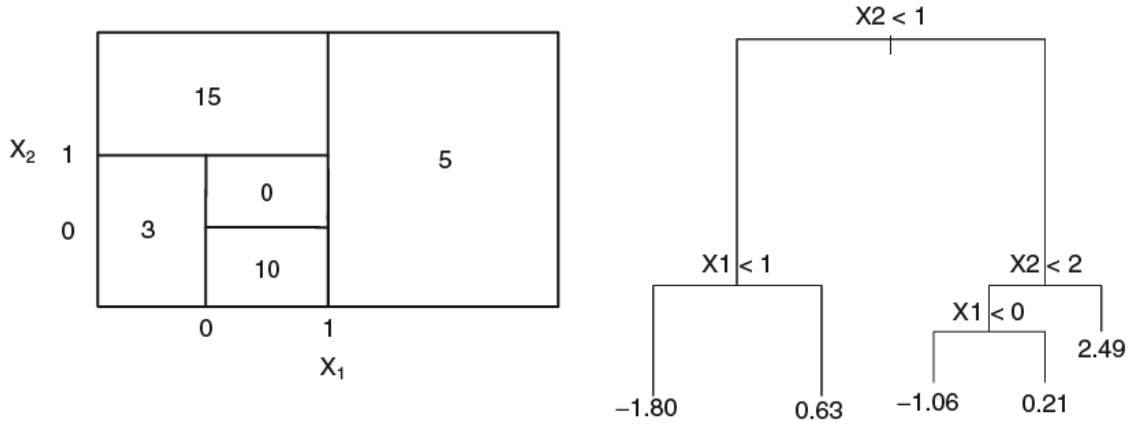
Weather	Traffic	Accident Rate	Number of observations
Sunny	Heavy	High	23
Sunny	Light	Low	5
Rainy	Heavy	High	50
Rainy	Light	Low	22

(b) **(5 points)** Suppose in another dataset, two students experiment with decision trees. The first student runs the decision tree learning algorithm on the raw data and obtains a tree T_1. The second student, normalizes the data by subtracting the mean and dividing by the variance of the features. Then, he runs the same decision tree algorithm with the same parameters and obtains a tree T_2. How are the trees T_1 and T_2 related?

(c) This question relates to the plots shown below in Figure 1.

- **(5 points)** (i) Sketch the tree corresponding to the partition of the predictor space illustrated in the left-hand panel of Figure 1. The numbers inside the boxes indicate the mean of Y within each region.
- **(5 points)** (ii) Create a diagram similar to the left-hand panel of Figure 1, using the tree illustrated in the right-hand panel of the same figure. You should divide up the predictor space into the correct regions, and indicate the mean for each region.

Figure 1: Left: A partition of the predictor space corresponding to (i). Right: A tree corresponding to (ii)



Submission Instruction: You need to provide the followings:

- Provide your answers to the problems in a PDF file, named as `IS777_hw#_fa19_LastName.pdf`. You need to submit the homework in electronic version as pdf file on Blackboard. If you choose handwriting instead of typing all the answers, you will get 40% points deducted.

Collaboration: You may collaborate. However, collaboration has to be limited to discussion only and you need to write your own solution and submit separately. You also need to list with whom you have discussed.