DATA SCI 415 - Homework 5

Due Wednesday, February 26, 2025, 11:59 PM

1. For the one dimensional (training) data below, give the linear discriminant analysis and quadratic discriminant analysis classifiers.

x	-3	-2	0	1	-1	2	3	4	5
y	-1	-1	-1	-1	1	1	1	1	1

   (a) What are the parameters needed to specify the LDA and QDA models respectively? What are their estimated values using the training data? (Hint: Note that there is only one input variable, so you don't need to worry about covariance, but you do need to consider the effect of the variance of X.)

   (b) Write down the discriminant functions.

   (c) Compute the training errors using LDA and QDA respectively, i.e., the misclassification error when applying your classifier to the training data?

   (d) Given a test set of (x, y) pairs

x	-1.5	-1	0	1	0.5	1	2.5	5
y	-1	-1	-1	-1	1	1	1	1

   what are the test errors?

   (e) Which is more suitable for this (training) data set, LDA or QDA? Justify your answer.

2. In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the `Auto` data set.

(a) Create a binary variable, `mpg01`, that contains a 1 if `mpg` contains a value above its median, and a 0 if `mpg` contains a value below its median. You can compute the median using the `median()` function. Note you may find it helpful to use the `data.frame()` function to create a single data set containing both `mpg01` and the other `Auto` variables.

(b) Explore the data graphically in order to investigate the association between `mpg01` and the other features. Which of the other features seem most likely to be useful in predicting `mpg01`? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

(c) Split the data into a training set and a test set.

(d) Perform LDA on the training data in order to predict `mpg01` using the variables that seemed most associated with `mpg01` in (b). What is the test error of the model obtained?

(e) Perform QDA on the training data in order to predict `mpg01` using the variables that seemed most associated with `mpg01` in (b). What is the test error of the model obtained?

(f) Perform logistic regression on the training data in order to predict `mpg01` using the variables that seemed most associated with `mpg01` in (b). What is the test error of the model obtained?

(g) Perform KNN on the training data, with several values of K, in order to predict `mpg01`. Use only the variables that seemed most associated with `mpg01` in (b). What test errors do you obtain? Which value of K seems to perform the best on this data set?