**Homework 3 (Due on October 25th midnight)**

**Total marks 40**

Chapter 4, Q7 on page 170 (with some changes)

1. Suppose that we wish to predict whether a given stock will issue a dividend this year (“Yes” or “No”) based on , last year’s percent proﬁt. We examine a large number of companies and discover that the mean value of for companies that issued a dividend was = 10, while the mean for those that didn’t was = 0. In addition, the variance of for these two sets of companies was = 36. Finally, 80% of companies issued dividends. Assuming that follows a normal distribution, **predict the probability that a company will issue a dividend this year given that its percentage proﬁt was = 4 last year.** To answer this, first answer a) to e).
2. Write down what is P( | Dividend = Yes) **(3 marks)**
3. Write down what is P( | Dividend = No) **(3 marks)**
4. Use *dnorm()* function in R to calculate conditional probabilities in a) and b) when = 4 **(4 marks, 2 marks each)**
5. What is the value of P( Dividend = Yes)? **(2 marks)**
6. What is the value of P( Dividend = No)? **(2 marks)**
7. Now predict the probability that a company will issue a dividend this year given that its percentage proﬁt was = 4 last year. **Hint:** Use Bayes’ rule as we discussed in the class. **(6 marks)**

Chapter 4, Q 11 on pp 171-172 (with some changes)

1. 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 in R. 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. **(2 marks)**

(b) Which of the continuous features seem most likely to be useful in predicting mpg? Use *cor()* function in R and consider features with correlation coefficients > 0.6 as useful in predicting. **(2 marks)**

(c) Split the data into a training set and a test set holding 30% of data for testing. Use sample.split() function in the library ‘caTools’ in R to split the data with the random seed 101. Use set.seed() function in R to assign the random seed. **(3 marks)**

(d) Perform LDA on the training data in order to predict mpg01 using the variables that seemed most associated with mpg in (b). You may not include mpg as it was used to derive mpg01. What is the test error of the model obtained? Use *lda()* function in the library ‘MASS’ in R.**(3 marks)**

(e) Perform QDA on the training data in order to predict mpg01 using the variables that seemed most associated with mpg in (b). You may not include mpg as it was used to derive mpg01. What is the test error of the model obtained? Use *qda()* function in the library ‘MASS’ in R **(3 marks)**

(f) Perform logistic regression on the training data in order to predict mpg01 using the variables that seemed most associated with mpg in (b). You may not include mpg as it was used to derive mpg01. What is the test error of the model obtained? **(3 marks)**

(g) Perform KNN on the training data, with several values of K (use K=1, K=5, K=10, K=15, K=20, K=30, K=50, K=100, K=150, K=200) in order to predict mpg01. Use only the variables that seemed most associated with mpg in (b). You may not include mpg as it was used to derive mpg01. Obtain test errors corresponds to each K. Which value of K seems to perform the best on this data set? Use *knn()* function in the library **(4 marks)**