RMI 8300
Assignment 2
Please show your work clearly to get full credit

1. This question should be answered using the Weekly data set, which is part of the ISLR package. This data is similar in nature to the Smarket data, except that it contains 1089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

(a) Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?

(b) Use the full data set to perform a logistic regression with Direction as the response and the five lag variables plus Volume as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?

(c) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

(d) Now fit the logistic regression model using a training data period from 1990 to 2008, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held-out data (that is, the data from 2009 and 2010).

(e) Repeat (d) using LDA.

(f) Repeat (d) using QDA.

(g) Repeat (d) using KNN with K =1.

(h) Which of these methods appears to provide the best results on this data?

(i) Experiment with different combinations of predictors, including possible transformations and interactions, for each of the methods. Report the variables, method, and associated confusion matrix that appears to provide the best results on the held-out data. Note that you should also experiment with values for K in the KNN classifier.

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?

3. This problem involves writing functions.

(a) Write a function, **Power(),** that prints out the result of raising 4 to the 5th power. In other words, your function should compute 4^5 and print out the results.
*Hint: Recall that x^a raises x to the power a. Use the **print()** function to output the result.*

(b) Create a new function, **Power2(),** that allows you to pass any two numbers, x and a, and prints out the value of x^a. You can do this by beginning your function with the line

```
>Power2 = function (x,a){
```

You should be able to call your function by entering, for instance,

```
>Power2(2,5)
```

on the command line. This should output the value of $2^5 = 32$.

(c) Using the **Power2()** function that you just wrote, compute 8^7, 11^{23}, and 65^{45}.

(d) Now create a new function, **Power3(),** that actually returns the result x^a as an R object, rather than simply printing it to the screen. That is, if you store the value x^a in an object called result within your function, then you can simply **return()** this result, using:

```
>return(result)
```

The line above should be the last line in your function, before the } symbol.

(e) Now using the **Power3()** function, create a plot of $f(x) = x^4$. The x-axis should display a range of integers from 1 to 10, and the y-axis should display x^4. Label the axes appropriately, and use an appropriate title for the figure. Consider displaying either the x-axis, the y-axis, or both on the log-scale. You can do this by using log=''x'', log=''y'',or log=''xy'' as arguments to the **plot()** function.

(f) Create a function, **PlotPower(),** that allows you to create a plot of x against x^a for a fixed a and for a range of values of x. For instance, if you call

```
>PlotPower(1:5,4)
```

then a plot should be created with an x-axis taking on values 1, 2,3,4,5 and a y-axis taking on values 1^4, $2^4, 3^4, 4^4, 5^4$.

4. Using the Boston data set, fit classification models in order to predict whether a given suburb has a crime rate above or below the median. Explore logistic regression, LDA, and KNN models using various subsets of the predictors. Describe your findings.