Homework 2

Big Data Analytics

Due time: November 22, 2018, 11:59 PM

Instructions:
1. Please turn in your homework in a pdf format generated by R Markdown.
2. Discussion are encouraged, but you need to work on the problems by yourself.

1 (35 points)
Problem 1 (Problem 10, Chapter 4)

This question should be answered using the *Weekly* data set, which is part of the *ISLR* package.

(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) Experiments 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 (35 points)
Problem 2 (Problem 11, Chapter 4)

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 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 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? Which value of K seems to perform the best on this data set?

3 (30 points)
Problem 3 (Problem 13, Chapter 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 subset of the predictors. Describe your findings.