Assignment 3: Classification

BUAD 5072 – Fall 2018

1. Objectives

The purpose of this assignment is to provide you with some experience working with the classification methods and resampling approaches we have been discussing in class.

2. What You Will Need

- Access to a Windows computer with R

3. What You Will Hand In

Submit your script file as Assignment3.R via Blackboard - Assignment 3.

4. Due Date

Friday December 7th, just before midnight.

5. Note on Collaboration

This is a Category C assignment. Specifically, you may work with others or receive help from the instructor on this assignment. You must, however, turn in your own paper. You may not divide the work with others or copy another student's work. **In particular, you may not copy another person's code. The code you submit must have been written by you alone. It would be an honor code offense to do so**.

6. Preliminaries:

To get set up for the assignment, follow these steps:

1. As the first statement in your script file, enter rm(list=ls())
2. Each part of each question in the assignment should begin with the following three comment lines, where *n* is the question number:

   ```
   ############################
   #### QUESTION n Part m ####
   ############################
   ```

3. I should be able to run your script on my computer without errors or interruptions. For this to happen, you must:
   a. Avoid entering file path information…my files will be located in a different location that yours, and so your code will fail on my machine. Instead, always refer only to files in your working directory.
   b. Do not use functions like file.choose(), fix(),edit(), or q()
4. Do not create console output other than what is asked of you explicitly. For example, in your final script, remove any statements that you used to verify the contents or structure of data.

7. Assignment Tasks:

Note: Please do not round off your calculations in this assignment.

Question 1: (33%)

In this problem you will use the Weekly data set, which is part of the ISLR package. This data is similar in nature to the Smarket data from Chapter 4's lab, except that it contains 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010. Do the following:

a) rm(list=ls()) and Set the random seed to 5072.
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. On a single comment line, identify predictors that are statistically significant, if any.
c) Create and (using the print() function) a confusion matrix in the form we have been using in class. Since there's no obvious choice here for the null hypothesis, assume that Down is the null hypothesis.
   a. Recall the contrasts() function.
d) **From the confusion matrix,** compute and display (using the print() function) the following performance statistics:
   - The overall fraction of correct predictions.
   - The overall error rate
   - Type I and Type II error rates

- The Power of the model
- The Precision of the model

e) Fit a logistic regression model using a training data period from 1990 to 2008, with Lag2 as the only predictor.

f) For the held out data (that is, the data from 2009 and 2010), use the model just created to construct and print (using the print() function) a confusion matrix in the format outlined in c) above, and from this table, compute the same five performance statistics for this new set of predictions.

g) Repeat e) and f) using LDA. In this and the following items h) through j), do not reset the random seed. The random seed should be set only once, in step a).

h) Repeat e) and f) using QDA.

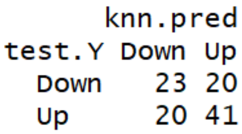
i) Fit a KNN model using the training data period from 1990 to 2008, with Lag2 as the only predictor, as above. Evaluate odd-number values of k from 1 to 99, and choose the k that produces the lowest overall error rate on the held out test data (that is, the data from 2009 and 2010).

   a. Note that since there is only one predictor in this question, there is no need to scale the x-values. This may not be the case in later questions, however.

j) Use the best k to construct and print (using the print() function) a confusion matrix <u>for the test</u> set in the format outlined in c) above, and from this table, compute and print (using the print() function) the same five performance statistics as above for this new set of test predictions.

k) Use the print() function to indicate which of the methods used in steps e) through j) appears to provide the lowest overall error rate on the test data.

My confusion matrix for step j) is as follows:



```
      knn.pred
test.Y Down Up
  Down   23 20
  Up     20 41
```

<u>Question 2</u>: (33%)

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. Do the following:

a. rm(list=ls()) and Set the random seed to 5072.

b. Create a binary variable, mpg01 that contains a 1 if mpg contains a value above its median, and a 0 otherwise.

c. Split the data into a training set and a test set using the sample() function as usual. The training set should be approximately 80% of the total number of rows.

d. Perform logistic regression on the training data in order to predict mpg01 using the variables cylinders, displacement and weight

e. For the test set, use the model just created to construct and display (using the print() function) a confusion matrix in the format outlined in Question 1 (assuming below-median mpg to be the null hypothesis), and from this table, compute and display (using the print() function) the same five performance statistics for this new set of predictions that were requested in Question 1.

f. Repeat d) and e) using LDA. In this and the following items g) through i), do not reset the random seed. The random seed should be set only once, in step a).
g. Repeat d) and e) using QDA.
h. Fit a KNN model on the training data in order to predict mpg01 using the variables cylinders, displacement and weight, as above. Evaluate odd-number values of k from 1 to 11, and choose the k that produces the lowest overall error rate on the test set.
i. Use the best k to construct and print (using the print() function) a confusion matrix for the test set in the format outlined in e) above, and from this table, compute and print (using the print() function) the same five performance statistics as above for this new set of test predictions.
j. Use the print() function to indicate which of these methods appears to provide the lowest overall error rate on this data.

The success rate of my best KNN model was 0.9367089

Question 3: (34%)

a) rm(list=ls()) and Set the random seed to 5072.
b) Using the Boston data set in the MASS package, create training and test sets in the ratio of 80/20.
c) Fit a logistic regression model on the training data using nox, rad and dis as predictors in order to predict whether a given suburb has a crime rate above or below the median.
d) Using the test data, produce the following:
    i. The ROC curve (with an appropriate title identifying the method)
    ii. The AUC (use a print() function)
    iii. A confusion matrix with a cutoff of 0.5 (use a print() function)
    iv. A confusion matrix with a cutoff of 0.4 (use a print() function)
e. Repeat steps c) and d) using LDA
f. Repeat steps c) and d) using QDA
g. Fit a KNN model on the training data using nox, rad and dis as predictors in order to predict whether a given suburb has a crime rate above or below the median, as above. Evaluate odd-number values of k from 1 to 49, and choose the k that produces the lowest overall error rate on the test set.
h. Use the model with the best k to print (using the print() function) the lowest overall error rate on the test set.
i. Use the print() function to indicate which of these four methods appears to provide the lowest overall error rate on this data. Use the test error rate associated with a cutoff of 0.5 for the first three models.
j. Use the print() function to indicate which of the first three methods provides the highest power at cutoff=0.4.

The power of my logistic regression model at cutoff=0.4 was the highest at 0.8035714.