Assignment 3: Classification and Resampling

BUAD 5072 – Fall 2017

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

Wednesday November 29th, 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. **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 for each question should be rm(list=ls())
2. Each question in the assignment should begin with the following three comment lines, where *n* is the question number:

   ```
   #######################
   #### QUESTION n ####
   #######################
   ```

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()
   c. Do not include install.packages() functions
   d. Include statements to load all necessary packages
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.
5. I suggest that you read the entire assignment before starting – there are sometimes notes and suggestions at the end of the assignment document.

7. Assignment Tasks:

Question 1: (25%)

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) 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 display 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.
d) **From the confusion matrix,** compute and display 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 display 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.

h) Repeat e) and f) using QDA.

i) Repeat e) and f) using KNN with k = 1.

j) Repeat e) and f) using KNN with k = 5.

k) Based on the confusion matrices, which of these methods appears to provide the best results on this data?

Question 2: (25%)

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. 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.

   i. You can compute the median using the median() function.

   ii. 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 (but not mpg).

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 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 the same five performance statistics for this new set of predictions as was requested in Question 1.

f. Repeat d) and e) using LDA.

g. Repeat d) and e) using QDA.

h. Repeat d) and e) using KNN with k = 1.

i. Repeat d) and e) using KNN, with various values of k. Choose the model that performs best.

j. Based on the confusion matrices, which of these methods appears to provide the best results on this data?

Question 3: (25%)

a) Using the Boston data set in the MASS package, create training and test sets in the ratio of 80/20, then 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 nox, rad and dis as predictors. Evaluate these models using confusion matrices as above and describe your findings.

Question 4: (25%)

In this question, we will perform cross-validation on a simulated data set. Do the following:

a) Generate a simulated data set as follows:

```
set.seed(5072)
x=rnorn(100)
y = x – 2 * x^2 + rnorm(100)
```

b) Create a data frame containing these x and y variables in named columns X and Y

c) Create a scatterplot of X against Y.

d) Set the random seed to 123, then compute the LOOCV errors (using the cv.glm() function) that result from fitting the following four models using least squares:

i. $Y = \beta_0 + \beta_1 X + \epsilon$

ii. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$

iii. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$

iv. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \epsilon.$

e) Repeat d) using random seed 456 and report your LOOCV errors as above. Are your results the same as what you got in d)? Why?

f) Which of the models in (d) had the smallest LOOCV error? Is this what you expected? Explain your answer.

g) Comment on the statistical significance of the coefficient estimates (citing p-values) that results from fitting each of the models in d) using least squares. Do these results agree with the conclusions drawn based on the cross-validation results?