Data Algorithms II

Chelsea Nowlin

Assignment 3

# Exercise 10, 11, 13

**10.** **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 from this chapter’s lab, except that it contains 1, 089 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?**

There don’t appear to be any substantial correlations between the lag variables and the year. The only substantial correlation observed was between volume and year at 0.842

**(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?**

The Lag2 variable appears to be the only statistically significant predictors based on its p-value of less than 0.05. All other predictors have p-values greater than 0.05.

**(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.**

The percentage of correct predictions based on the training data, calculated as (54+557) / (54 + 48 + 430 + 557) equals 56% with a training error rate of 44%. Alternatively, when the market is up, the model is correct 92% (557/ (48 + 557)). When the market is down the model is correct 11% (54 / (54 + 430)).

**(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).**

The percentage of correct predictions based on the test data, calculated as (9 + 56) / (9 + 5 + 34 + 56) equals 62.5% with a 37.5% test error rate. Alternatively, when the market is up the model is correct 92% (56 / (56 + 5)). When the market is down the model is correct 21% (9/ (9 + 34)).

**(e) Repeat (d) using LDA.**

The results are very similar to the results of the previous logistic regression model.

**(f) Repeat (d) using QDA.**

The percentage of correct predictions based on the test data, calculated as (0 + 61) / (0 + 0 + 43 + 61) equals 59% with a 41% test error rate. Alternatively, when the market is up the model is correct 100% (61 / (61 + 0)). When the market is down the model is correct 0% (0/ (0 + 43)).

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

The percentage of correct predictions based on the test data, calculated as (21 + 31) / (21 + 30 + 22 + 31) equals 59% with a 41% test error rate. Alternatively, when the market is up the model is correct 51% (31 / (31 + 30)). When the market is down the model is correct 49% (21/ (21 + 22)).

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

Based on comparison of error rates of the models, logistic regression and LDA had the smallest error rate percentages.

**(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.**

Based on the comparison of the models, logistic regression and LDA still have better performance based on lower error rate percentages.

**11. 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.**

The variables mpg, cylinders, weight, displacement, and horsepower appear to have an association.

**(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?**

The test error rate calculated is 12.64%.

**(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?**

The test error rate calculated is 13.19%.

**(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?**

The test error rate calculated is 12.09%.

**(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?**

The test error rate calculated is 15.38% for K = 1.

The test error rate calculated is 16.48% for K = 10.  
 The test error rate calculated is 14.29%. for K = 100. <- K value equal 100 performs the best based on lowest error rate.

**13. 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.**

The test error rate calculated is 18.18% for logistic regression.

The test error rate calculated is 15.81% for the logistic regression.

The test error rate calculated is 13.44% for LDA.

The test error rate calculated is 15.02% for the LDA.

For KNN k =1, the error rate is 45.85%

For KNN k =10, the error rate is 11.86%

For KNN k=100, the error rate is 49%