Homework2 - Classification

Lu Haibo

Mondy, Mar 11, 2019

Deadline: Mar 18, 2019

1. This question should be answered using the `Weekly` data set, which is part of the `ISLR` package. This data contains 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010. For each date, we have recorded the percentage returns for each of the five previous trading days, `Lag1` through `Lag5`. We have also recorded `Volume` (the number of shares traded on the previous day, in billions), `Today` (the percentage return on the date in question) and `Direction` (whether the market was Up or Down on this date).

```
library(ISLR)
data(Weekly)
```

a. Produce some numerical and graphical summaries of the `Weekly` data. Do there appear to be any patterns? (scatter plots, correlation, …)

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 `stargazer` 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 the held out 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. Using the `Boston` data set, in the library `MASS`, 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.