Statistical Data Mining

Homework 3

Executive Summary

By Priyanka Sariya

*With the stock market information, we want to predict when a day might be bad (<-1%), poor (-1%<x<0), good(0<x<1%), and great (>1%). Thus, using the data from HW2, instead of predicting up or down, or predicting the value, see if we can use the lag variables to predict the classifications above.*

*Use Bayes and LDA to predict the classes above and see if either does better. Train your data up through the end of 2017 and test it on 2018 data using the same data as in HW 2.*

The assignment is to predict the lag from yesterday’s stock and today’s stock value using Bayes model and Linear Discriminant Analysis and then compare and contrasts the predictions by the two models.

I used the Bayes model first to do the predictions. The data for first divided into training and testing sets and the bayes model was run on the training set. The model thus generated was tested on the testing set. This testing was done to predict the classification, that is, if the day is bad, poor, good or great.

Next I used the LDA model for predictions. Again, the model was trained using training set and then the resulted model was used to predict the classifications in the testing set.

The accuracy of the two models was as below –

**Bayes Model**

> table(pred.classify$class, my.test$classify)

bad good great poor

bad 1 0 0 2

good 9 56 11 27

great 1 1 6 0

poor 5 31 3 37

> mean(pred.classify$class == my.test$classify)

[1] 0.5263158

**LDA Model**

> table(my.test$lda.pred.classify, my.test$classify)

bad good great poor

bad 3 1 0 3

good 10 74 13 46

great 0 0 5 0

poor 3 13 2 17

> mean(my.test$lda.pred.classify == my.test$classify)

[1] 0.5210526

So we see that Bayes and LDA models gave about the same level of accuracy of above 52%, with Bayes giving slightly better accuracy by 0.5%