EAS506LEC000 : STATISTICAL DATA MINING

HW ASSIGNMENT – 3

SUBMITTED BY : DITHYA SRIDHARAN

CLASS NUMBER : 48

QUESTION 1

PROBLEM STATEMENT

Consider the Boston Data Set in the ISLR Package. 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.

1. DATA SUMMARY:

The Boston data frame has 506 rows and 14 columns. This data frame contains the following columns: crim, zn, indus, chas, nox, rm, age, dis, rad, tax, ptratio, black, lstat, medv

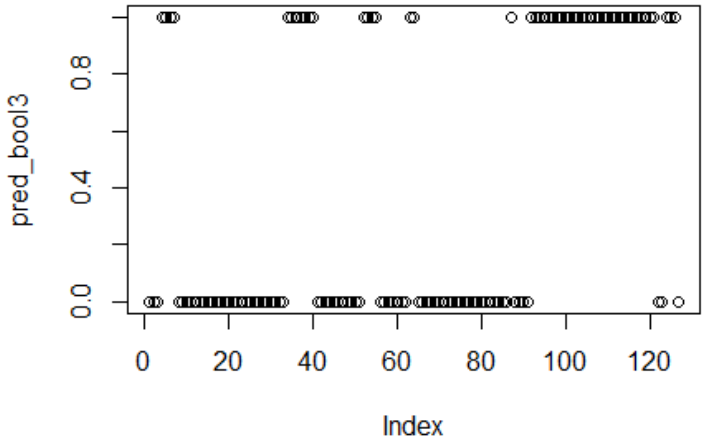
2. SUBSET SELECTION TO PERFORM MODELING ON:

The regsubsets() function (part of the leaps library) performs selection. Here we use the regsubsets function but specify the `method="exhaustive" option. The summary() command outputs the best set of variables for each model size. Selection based on Cp gives 5 variables - nox, medv, rad, zn, age. Selection based on BIC gives 4 variables - nox, medv, rad, age. Selection based on RSS gives all 13 variables. So we perform on the training data.

3.LOGISTIC REGRESSION

A logistic regression model on the training set, subset using 5 variables, subset using 4 variables, subset using 13 variables was fit.. It is found that the test MSE for different subsets is as follows :

SUBSET	TEST MSE
Subset 1 – Using 5 variables	0.1417323
Subset 2 – Using 4 variables	0.1338583
Subset 3 – Using 13 variables(except crim)	0.08661417

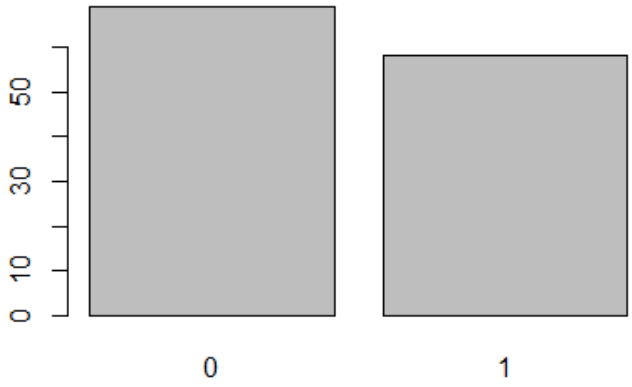


This gives a good enough model.The accuracy of the model is 0. 9133858. That is, the model is 91% is accurate.

4. KNN MODEL

A KNN model on the training set, subset using 5 variables, subset using 4 variables, subset using 13 variables was fit. We have considered for k=10. K=10 seems to be the best K value found. It is found that the test MSE for different subsets is as follows :

SUBSETS	TEST MSE
Subset 1 – Using 5 variables	0.1811024
Subset 2 – Using 4 variables	0.2047244
Subset 3 – Using 13 variables(except crim)	0.1102362

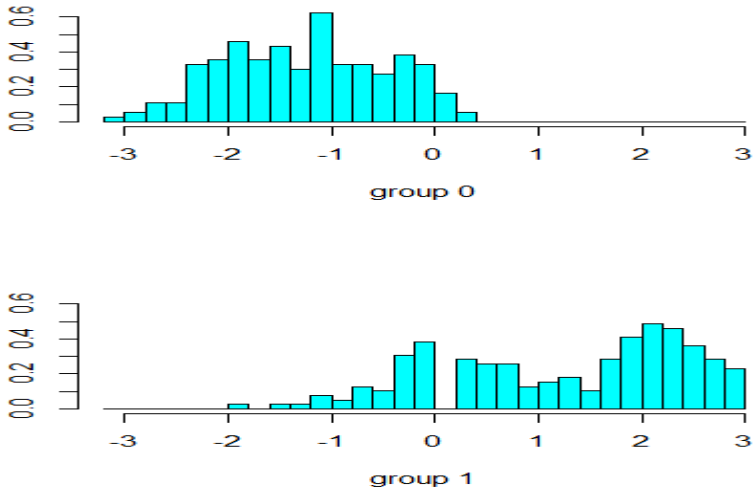


This gives a good model. The accuracy of the model is 0.8997638 ie, the model is 89% accurate. This is a good model but is slightly less accurate than logistic regression.

5. LDA MODEL:

A LDA model on the training set, subset using 5 variables, subset using 4, subset using 13 variables was fit. We have considered for k=10. K=10 seems to be the best K value found. It is found that the test MSE for different subsets is as follows :

SUBSETS	TEST MSE
Subset 1 – Using 5 variables	0.1574803
Subset 2 – Using 4 variables	0.1574803
Subset 3 – Using 13 variables(except crim)	0.1653543



This gives a good model. The accuracy of the model is 0.8425197 ie, the model is 84% accurate. This is a good model but is less accurate than logistic regression and KNN.

6. CONCLUSION

- It is found that the accuracy for logistic regression model is 0.9133858, KNN is 0.89976358,LDA is 0.8425197

- The test errors obtained by each model is comparable. Since LDA model has the highest MSE, LDA does not predict whether a given suburb has crime rate above or below the median with as much accuracy as the other two models for the given data.

- One of the benefits of is that we can begin to understand which variables are more important to the model, therefore we gain insight into the behavior of the real word phenomena.

QUESTION 2

PROBLEM STATEMENT

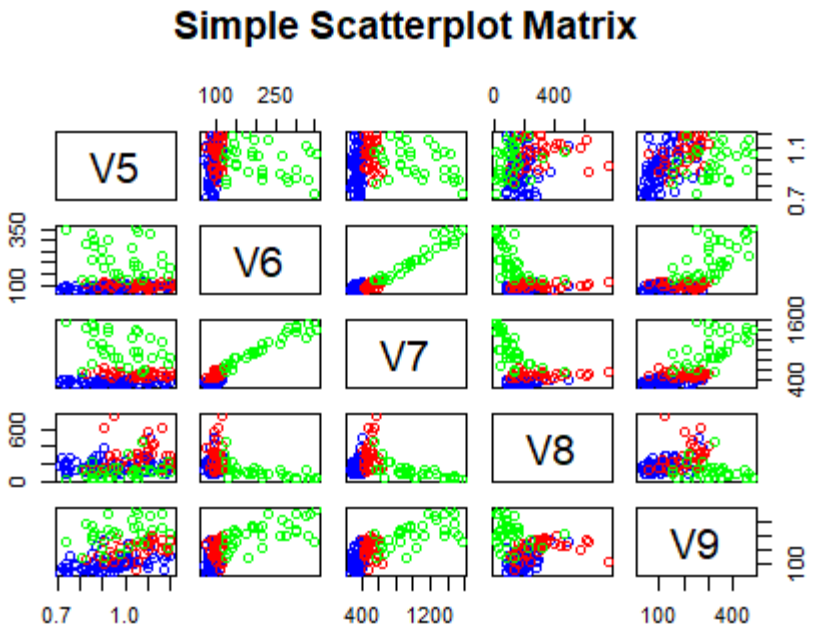
Consider the Diabetes Data Set in the MMST Package.
a)Produce pairwise scatterplots for all five variables, with different symbols or colors representing the three different classes. Do you see any evidence that the classes may have difference covariance matrices? That they may not be multivariate normal?
b) Apply linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). How does the performance of QDA compare to that of LDA in this case?
c) Suppose an individual has (glucose area = 0.98, insulin area =122, SSPG = 544. Relative weight = 186, fasting plasma glucose = 184). To which class does LDA assign this individual? To which class does QDA?

1. DATA SUMMARY

The dataset consists of 10 variables and 145 observations. The fourth column is the observation number, and the next five columns are the variables(glucose.area, insulin.area, SSPG, relative.weight, and fasting.plasma.glucose). The final column is the class number.

2. SCATTER PLOTS

Pairwise-scatter plots are plotted for the five variables – glucose area, insulin area, SSPG, rel ative weight, fasting plasma glucose for the three different classes.
Green indicates class 1, Red is class 2 and Blue indicates class 3.



a)Class 3 has different comparisons to class 1 and 2 as observed from plots. From this, we can say that the classes are not multivariate normal as covariance is different for each class. We can also say that they are linearly dependant since covariance is not 0.

3. LDA AND QDA TRAIN AND TEST ERRORS

The LDA and QDA model was fit for the data set and the training and test MSE was obtained for the respective models.

MODEL	**TRAINING/TEST**	**MSE**
LDA	Training	0.09259259
LDA	Test	0.05405405
QDA	Training	0.06481481
QDA	Test	0.02702703

It can be observed that the MSE obtained in QDA is lesser than that obtained by LDA. Hence for the given data set, QDA performs better than LDA in prediction

4.PREDICTING CLASS USING QDA AND LDA WITH GIVEN VALUES

It is observed that for an individual with glucose area = 0.98, insulin area =122, SSPG = 544. Relative weight = 186, fasting plasma glucose = 184, the QDA assigns the individual to class 2 and LDA assigns the individual to class 3.

5. CONCLUSION

- For the given data set, pairwise scatter plots reveal that class 3 compares differently to class 1 and class 2 with respect to it's relationship with the five variables.
- It can also be seen that the classes are not multivariate normal as covariance is different and they are found to have dependence.
- LDA and QDA MSEs are obtained and it reveals that for given data, QDA performs better.
- The individual with given variable values is found to fit into class 2 using QDA and class 3 using LDA.