Homework 3 : Statistical Data Mining (EAS 506)

Shubham Sharma, Person No.: 50290293, Class No.: 44

November 3, 2018

Question 1: 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 Boston Data Set has been taken from the MASS package. The variable crim has been converted to a qualitative variable with values 1 if the value is greater than the median value and 0 elsewhere. The training and test set data is split in the ratio 3:1.

The prior probabilities while fitting the LDA model indicate that 51.45% of the suburbs have crime rate above the median while 48.55% have it below the median in the training set. Considering all the predictors, we get a misclassification error rate of 0.165 on the test set and 0.133 on the training set. The sensitivity from the confusion matrix comes out to be 94.83% thus we can say that the model is able to predict the suburbs where the crime rate is below than the median 94.83% of the time in the test set while the specificity, i.e. the percentage of the test set on which the model is able to correctly classify a suburb as having the crime rate more than the median of the distribution is 73.91%.

The confusion matrix for the LDA model is shown in Figure 1.

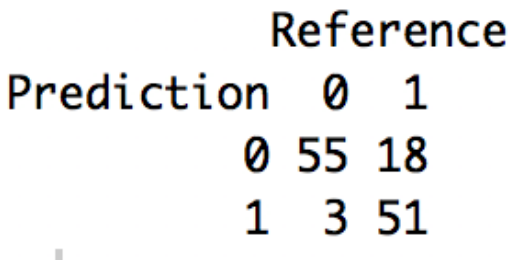

Figure 1: Confusion Matrix for LDA model

For the logistic regression model the misclassification error rate is 0.0945 on the test set and 0.0844 on the training set as should be the case. The sensitivity and specificity values are 94.83% and 86.96% respectively.

The confusion matrix for the logistic regression model is shown in Figure 2.

For the KNN model with k=1 the misclassification error rate is 0.0945 on the test set. The sensitivity and specificity values are 89.66% and 91.30% respectively. For k = 5 the values are 0.1024, 89.66% and 89.86% respectively.

The confusion matrix for the 1-KNN model is shown in Figure 3.

Logistic regression performs better than other methods if we see the trend of the misclassification error rate on the test data. On examining the p-values from the logistic regression model as in Figure 4, we take the most significant predictors, i.e. zn, nox, age, dis, rad, ptratio, black, medv in case 2.

In case 2, we get the following results as in Table 1.

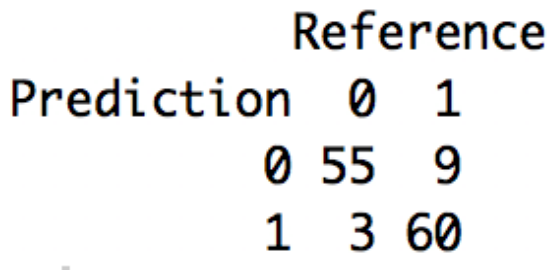

Figure 2: Confusion Matrix for Logistic Regression model

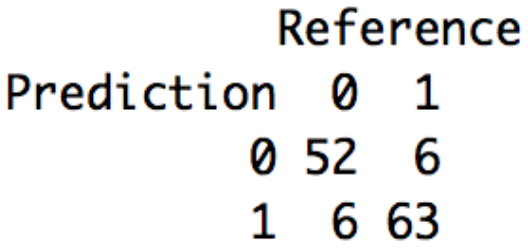

Figure 3: Confusion Matrix for 1-KNN model

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -34.469088   7.835599  -4.399 1.09e-05 ***
zn           -0.082031   0.039305  -2.087 0.036884 *
indus        -0.048569   0.050363  -0.964 0.334854
chas          1.102206   0.895928   1.230 0.218608
nox          50.722959   8.998760   5.637 1.73e-08 ***
rm           -1.163472   0.872421  -1.334 0.182331
age           0.031942   0.015578   2.050 0.040327 *
dis           0.791210   0.260688   3.035 0.002405 **
rad           0.632259   0.183566   3.444 0.000573 ***
tax          -0.003831   0.003434  -1.115 0.264647
ptratio       0.391024   0.153108   2.554 0.010652 *
black        -0.013614   0.006522  -2.087 0.036859 *
lstat         0.004631   0.061110   0.076 0.939595
medv          0.250144   0.090822   2.754 0.005883 **
```

Figure 4: Coefficients from the Logistic Regression model

Model	Misclassification Test Set error rate	Sensitivity(%)	Specificity(%)
LDA	0.1570	96.55	73.91
Logistic Regression	0.1339	91.38	82.61
1-KNN	0.1811	86.21	78.26
5-KNN	0.1654	87.93	79.71

Table 1: Results for the predictors: zn, nox, age, dis, rad, ptratio, black, medv

We observe that the performance of the logistic regression model in this case is better than the other models although we observe an increase in all the errors as compared to the previous case which indicate that relevant information have been excluded in the predictors which were excluded from the model. On examining the p-values from the logistic regression model of the previous case as in Figure 6, we take the most significant predictors, i.e. zn, nox, age, dis, rad, medv in case 3.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -32.538227   6.998322  -4.649 3.33e-06 ***
zn           -0.091710   0.036963  -2.481 0.013096 *
nox          41.190638   7.451627   5.528 3.24e-08 ***
age           0.024417   0.011889   2.054 0.040000 *
dis           0.750413   0.237970   3.153 0.001614 **
rad           0.531417   0.139686   3.804 0.000142 ***
ptratio       0.229812   0.125336   1.834 0.066717 .
black        -0.011794   0.006235  -1.892 0.058547 .
medv          0.155582   0.041130   3.783 0.000155 ***
---
```

Figure 5: Coefficients from the Logistic Regression model

In the case 3 we get the following results as in Table 2.

Model	Misclassification Test Set error rate	Sensitivity(%)	Specificity(%)
LDA	0.157	96.55	73.91
Logistic Regression	0.134	89.66	84.06
1-KNN	0.149	86.21	84.06
5-KNN	0.157	89.66	79.71

Table 2: Results for the predictors: zn, nox, age, dis, rad, medv

In this case we observe that there is a slight improvement in the performance of the KNN models although LDA and logistic regression models perform the same. Thus, we see that overall, logistic regression and KNN models are performing better than the LDA model for this data which indicates that the data distribution is considerably far from a linear relationship and also that the observations do not belong to a multivariate Guassian distribution.

Question 2: Disregard the first three columns. 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. Assume the population prior probabilities are estimated using the relative frequencies of the classes in the data.

(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?

The given data set was converted to a table which has the following column names:

OBSNO - Observation Number

G.AREA - glucose.area

I.AREA - insulin.area

SSPG - SSPG

RWT - relative.weight

FPG - fasting.plasma.glucose

CLASS - class number

The pairwise scatter plot is shown in the figure 6.

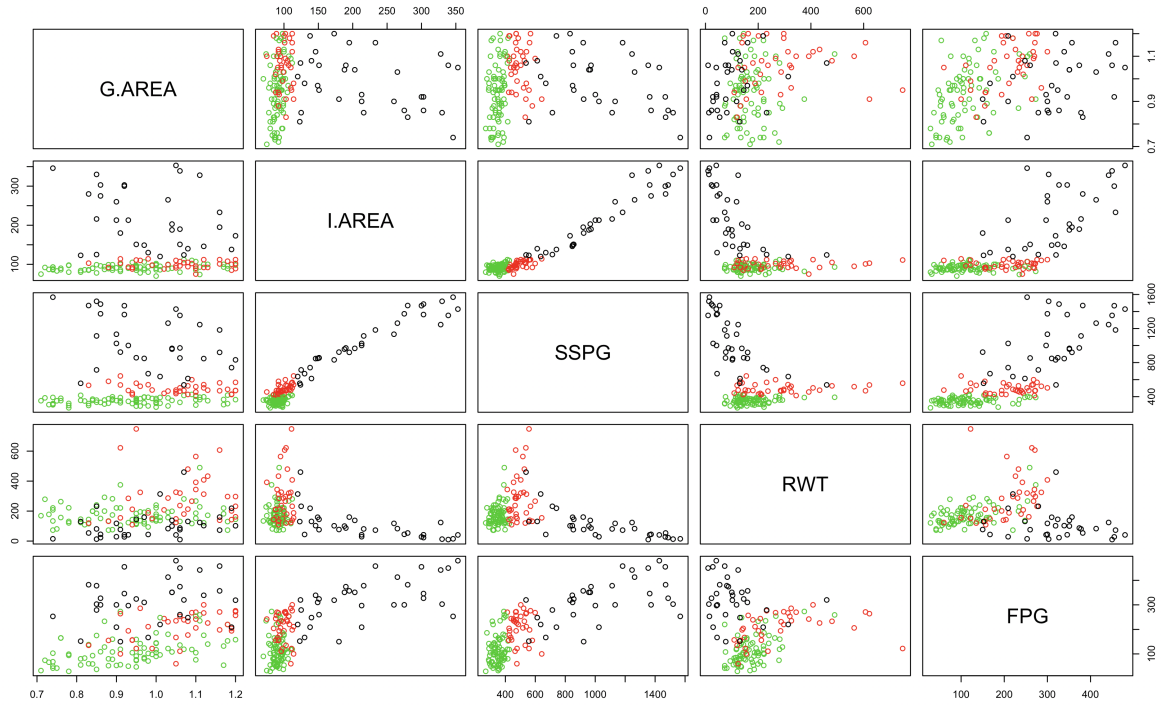

Figure 6: Pairwise Scatter Plot (1-Black, 2-Red, 3-Green)

It is clear from the plot that the classes have different covariance matrices and hence are not multivariate normal. This is because in multivariate Gaussian distribution each individual predictor will have a one-dimensional normal distribution with some correlation between each pair of predictors. Majority of the data should be concentrated in the middle portion for a particular class which is not seen in the plot obtained.

Following Misclassification error rates were observed (Table 3):

Model	Training Set	Test Set
LDA	0.074	0.270
QDA	0.037	0.108

Table 3: Misclassification error rates

Clearly, the test set error is higher for both the cases and the performance of QDA model is better than LDA. QDA predictions are accurate 89% of the time as compared to LDA which stands at 73%, for new unseen test set data. This suggests that the quadratic form assumed by QDA may capture the true relationship more accurately than the linear forms assumed by LDA.

The LDA model assigns the individual to class 3 and the QDA model assigns it to class 2.

Question 3: a) Under the assumptions in the logistic regression model, the sum of posterior probabilities of classes is equal to one. Show that this holds for k=K. b) Using a little bit of algebra, show that the logistic function representation and the logit representation for the logistic regression model are equivalent. In other words, show that the logistic function:

$$p(X) = \frac{exp(\beta_0 + \beta_1 X)}{1 + exp(\beta_0 + \beta_1 X)} \tag{1}$$

is equivalent to:

$$\frac{p(X)}{1 - p(X)} = exp(\beta_0 + \beta_1 X) \tag{2}$$

Solution: We know that the logistic regression models the posterior probabilities of the K class in the linear functions in x.

$$Pr(G = k|X = x) = \frac{e^{\beta_{k0} + \beta^{Tx} k}}{1 + \sum l = 1^{K-1} e^{\beta_{l0} + \beta_l^{Tx}}} \tag{3}$$

$$Pr(G = K|X = x) = \frac{1}{1 + \sum_{l=1}^{K-1} e^{\beta_{l0} + \beta_l^{Tx}}} \tag{4}$$

Dividing (1) by (2), we get

$$\frac{Pr(G = k|X = x)}{Pr(G = K|X = x)} = e^{\beta_{k0} + \beta_k^{Tx}} \tag{5}$$

$$Pr(G = k|X = x) = \frac{\frac{Pr(G=k|X=x)}{Pr(G=K|X=x)}}{1 + \frac{\sum_{l=1}^{K-1} Pr(G=l|X=x)}{Pr(G=k|X=x)}}$$

$$Pr(G = k|X = x) = \frac{\frac{Pr(G=k|X=x)}{Pr(G=K|X=x)}}{\frac{\sum_{l=1}^{K} Pr(G=l|X=x)}{Pr(G=k|X=x)}}$$

$$=> Pr(G = k|X = x) = \frac{Pr(G=k|X=x)}{\sum_{l=1}^{K} Pr(G=l|X=x)}$$

Therefore,

$$\sum_{l=1}^{K} Pr(G = l|X = x) = \frac{\sum_{l=1}^{K} Pr(G=l|X=x)}{\sum_{l=1}^{K} Pr(G=l|X=x)} = 1$$

Hence, proved.

(b) $P(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$

$$=> P(X) + P(X).e^{\beta_0 + \beta_1 X} = e^{\beta_0 + \beta_1 X}$$

$$=> P(X) = (1 - P(X)).e^{\beta_0 + \beta_1 X}$$

$=> \frac{P(X)}{1-P(X)} = e^{\beta_0+\beta_1 X}$

Thus, we can see that the logistic function representation and the logit representation for the logistic regression model are equivalent.