Statistical Data Mining - I

Homework 3

Priya Karadi - 50291880

Question 1

Using the Boston data set (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.

- The Boston dataset has 506 observations and 14 variables. Since I wanted to do a do a categorical classification, I have added a new column crim_class to categorize all crim values greater that the median to be of class 1, and below the median as 0.
- I have used 75% of the dataset for training different models and reserved the rest for testing.

Data Modelling

Logistic Regression Model:

- From the function glm(). Since logistic regressions works with binomial categorical variables only, I have set ‘family= “binomial” ’. This will fit the model showing the probability of maximum likelihood of an observation belonging to a particular class.
- The significant variable are as follows.

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.06472  -0.10884   0.00000   0.00046   2.53077

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -42.036063   8.465208  -4.966 6.84e-07 ***
zn           -0.123796   0.048706  -2.542 0.011031 *
indus        -0.077060   0.055505  -1.388 0.165034
chas          0.740302   0.951745   0.778 0.436665
nox          60.767050  10.460645   5.809 6.28e-09 ***
rm           -0.867565   0.870177  -0.997 0.318765
age           0.028763   0.015157   1.898 0.057734 .
dis           1.060536   0.308850   3.434 0.000595 ***
rad           0.793264   0.198353   3.999 6.35e-05 ***
tax          -0.006557   0.003033  -2.162 0.030615 *
ptratio       0.285592   0.143780   1.986 0.046998 *
black        -0.010174   0.005561  -1.830 0.067288 .
lstat         0.103237   0.063453   1.627 0.103740
medv          0.238356   0.087660   2.719 0.006546 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- I have then predicted, by setting ‘type = “response” ’, the class of the train dataset to observe how well the model does on already seen data. The predicted probabilities were rounded to be used with the ConfusionMatrix() of the caret package. This has yielded an accuracy of 91%.
- The observed test error was 0.1259843.
- Similar procedure was carried out for the test dataset that yielded the following result (accuracy = 87%):

```
> cm_logistic
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 54  6
         1 10 57

               Accuracy : 0.874
                 95% CI : (0.8035, 0.9262)
    No Information Rate : 0.5039
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7481
 Mcnemar's Test P-Value : 0.4533

            Sensitivity : 0.8438
            Specificity : 0.9048
         Pos Pred Value : 0.9000
         Neg Pred Value : 0.8507
             Prevalence : 0.5039
         Detection Rate : 0.4252
   Detection Prevalence : 0.4724
      Balanced Accuracy : 0.8743

       'Positive' Class : 0
```

- To observe whether a subset of the variables provide better accuracy, I used performed the same steps on the subset of most significant predictors viz., ("nox","dis","rad","medv","crim_class")
- I expected to see an increase in the accuracy, however, it decreased to 81%.
- The observed test error was 0.1811024.

```
Console   Terminal ×
~/Documents/MES Data Science/Statistical Data Mining/Assignments/Assignment_3/
> cm_logistic_2
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 53 12
         1 11 51

               Accuracy : 0.8189
                 95% CI : (0.7408, 0.8816)
    No Information Rate : 0.5039
    P-Value [Acc > NIR] : 1.606e-13

                  Kappa : 0.6377
 Mcnemar's Test P-Value : 1

            Sensitivity : 0.8281
            Specificity : 0.8095
         Pos Pred Value : 0.8154
         Neg Pred Value : 0.8226
             Prevalence : 0.5039
         Detection Rate : 0.4173
   Detection Prevalence : 0.5118
      Balanced Accuracy : 0.8188
```

Linear Discriminant Analysis (LDA):

- LDA can be used for classification of data when the number of classes are more than 2. It uses Baye's Method for finding the maximum likelihood of an observation to belong to a particular class.
- The function lda() was called to fit the model on the dataset with all predictors. Then predict() followed by ConfusionMatrix() revealed the following result with test error 0.9055118 and accuracy of 81%.

```
> cm_lda
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 58 18
         1  6 45

               Accuracy : 0.811
                 95% CI : (0.732, 0.875)
    No Information Rate : 0.5039
    P-Value [Acc > NIR] : 6.953e-13

                  Kappa : 0.6215
 Mcnemar's Test P-Value : 0.02474

            Sensitivity : 0.9062
            Specificity : 0.7143
         Pos Pred Value : 0.7632
         Neg Pred Value : 0.8824
             Prevalence : 0.5039
         Detection Rate : 0.4567
   Detection Prevalence : 0.5984
      Balanced Accuracy : 0.8103

       'Positive' Class : 0
```

- Next I used the subset of highly correlated variables which were the same as the ones used in logistic model. This gave results with test error as 0.8897638 and accuracy 85%.

```
Console   Terminal ×
~/Documents/MES Data Science/Statistical Data Mining/Assignments/Assignment_3/
> cm_logistic_2
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 53 12
         1 11 51

               Accuracy : 0.8189
                 95% CI : (0.7408, 0.8816)
    No Information Rate : 0.5039
    P-Value [Acc > NIR] : 1.606e-13

                  Kappa : 0.6377
 Mcnemar's Test P-Value : 1

            Sensitivity : 0.8281
            Specificity : 0.8095
         Pos Pred Value : 0.8154
         Neg Pred Value : 0.8226
             Prevalence : 0.5039
         Detection Rate : 0.4173
   Detection Prevalence : 0.5118
      Balanced Accuracy : 0.8188
```

k-Nearest Neighbor (kNN):

- I have used the knn() to fit the model for k-values of 3, 5, 10.
- For K = 3, including all predictors, error observed was 0.03937008 and accuracy 91%.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 61  8
         1  3 55

               Accuracy : 0.9134
                 95% CI : (0.8503, 0.956)
```

- For K = 5, including all predictors, error observed was 0.02362205 and accuracy 89%.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 59  8
         1  5 55

               Accuracy : 0.8976
                 95% CI : (0.8313, 0.9444)
```

- For K = 10, including all predictors, error observed was 0.007874016 and accuracy 88%.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 57  8
         1  7 55

               Accuracy : 0.8819
                 95% CI : (0.8127, 0.9324)
```

- Same procedure was carried out for highly correlated variable as in logistic model. The results were as follows:

  K=3, error = 0.007874016, accuracy = 91%.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 59  6
         1  5 57

               Accuracy : 0.9134
                 95% CI : (0.8503, 0.956)
```

  K=5, error = 0.01574803, accuracy = 88%.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 58  8
         1  6 55

               Accuracy : 0.8898
                 95% CI : (0.822, 0.9384)
```

K=10, error = 0.03149606, accuracy = 84%.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 56 12
         1  8 51

               Accuracy : 0.8425
                 95% CI : (0.7673, 0.9011)
```

Comparison:

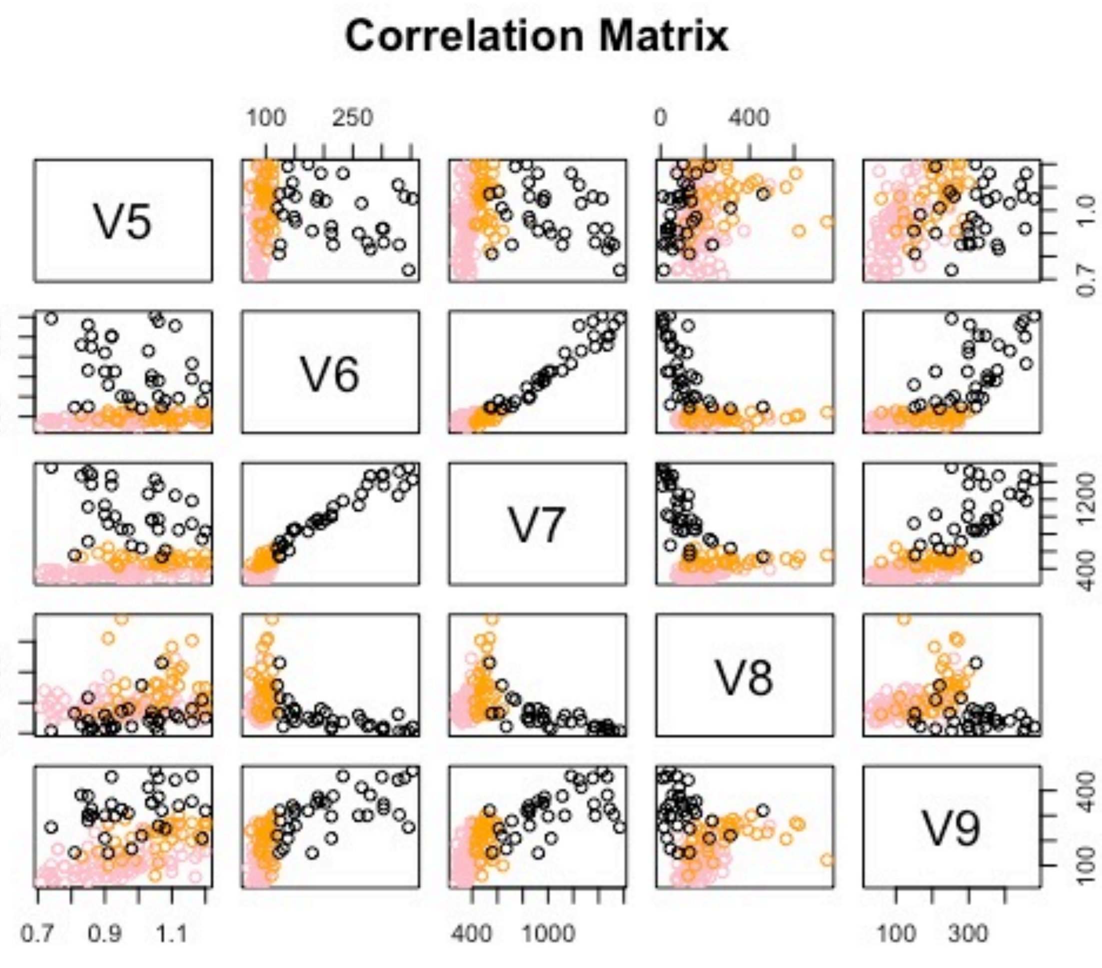
Model	Accuracy (%)
Logistic Regression	91
Linear Discriminant Analysis	81
3-Nearest Neighbor	91
5- Nearest Neighbor	89
10- Nearest Neighbor	88

Question 2

Download the diabetes data set (http://astro.temple.edu/~alan/DiabetesAndrews36_1.txt). 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?

   Plotted the pairwise scatter plot for all five variables. The different colors indicate the different classes viz., Black=1, Orange= 2, Pink=3.



- It can be observed that there is an obvious negative covariance between V5 and V6, V6 and V8, and V7 and V8.

- Since the covariance is varied, there cannot be a multivariate normal distribution for this dataset.

b) Apply LDA and QDA. Compare their performance.

- LDA gives an accuracy of 88%

```
Confusion Matrix and Statistics

          Reference
Prediction  1  2  3
         1 20  4  1
         2  0 20  4
         3  0  3 56

Overall Statistics

               Accuracy : 0.8889
                 95% CI : (0.814, 0.9413)
```

- QDA gives an accuracy of 95%.

```
Confusion Matrix and Statistics

          Reference
Prediction  1  2  3
         1 23  2  0
         2  1 22  1
         3  0  1 58

Overall Statistics

               Accuracy : 0.9537
                 95% CI : (0.8953, 0.9848)
```

<u>Performance Comparison:</u>

- QDA has higher accuracy compared to LDA.
- QDA has lesser number of misclassifications compared to LDA.

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?

I put the test point in a new data frame and passed it to the prediction models of LDA and QDA. Following were the class predictions:

LDA: Class 3

QDA: Class 1

```
> tes_point_lda$class
[1] 3
Levels: 1 2 3
> tes_point_qda$class
[1] 1
Levels: 1 2 3
>
```

STATISTICAL DATA MINING - HOMEWORK-3

olution:3

(a) Assumption: sum of posterior probabilities of classes in logistic regression is equal to on.
Show that: This holds for $k = K$.

Proof:-

Let the observation be X and its probability of belonging to class $= G$.

For $G=1$;

According to log-odds ratio:-

$$\log\left(\frac{Pr(G=1 \mid X=x)}{Pr(G=K \mid X=x)}\right) = \beta_{10} + \beta_1^T x$$

for $G=2$;

$$\log\left(\frac{Pr(G=2 \mid X=x)}{Pr(G=K \mid X=x)}\right) = \beta_{20} + \beta_2^T x$$

Similarly for $G=K-1$

$$\log\left(\frac{Pr(G=K-1 \mid X=x)}{Pr(G=K \mid X=x)}\right) = \beta_{(K-1)0} + \beta_{(K-1)}^T x$$

$$\Rightarrow \frac{Pr(G=K-1 \mid X=x)}{Pr(G=K \mid X=x)} = \exp\left(\beta_{(K-1)0} + \beta_{(K-1)}^T x\right)$$

for $k = 1, 2, \ldots K-1$

$$Pr(G=k \mid X=x) = \frac{\exp(\beta_{(K-1)0} + \beta_{(K-1)}^T x)}{1 + \sum_{i=1}^{K-1} \exp(\beta_{i0} + \beta_i^T x)}$$

$$Pr(G=K \mid X=x) = \frac{1}{1 + \sum_{i=1}^{K-1} \exp(\beta_{i0} + \beta_i^T x)}$$

$$\therefore \sum_{k=1}^{K} Pr(G=k \mid X=x) = \frac{\sum_{k=1}^{K-1} \exp(\beta_{k0} + \beta_k^T x)}{1 + \sum_{i=1}^{K-1} \exp(\beta_{i0} + \beta_i^T x)}$$

$$+ \frac{1}{1 + \sum_{i=1}^{K-1} \exp(\beta_{i0} + \beta_i^T x)}$$

$$= \underline{\underline{1}}$$

$$\therefore \sum_{k=1}^{K} Pr(G=k \mid X=x) = 1$$

Hence, proved.

(b) TO show:- Logistic function representation and logit representation for the logistic regression model are equivalent.

i.e.

$$p(x) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)} \quad \text{— (1)}$$

and

$$\frac{p(x)}{1 - p(x)} = \exp(\beta_0 + \beta_1 X) \text{ are equivalent.} \quad \text{— (2)}$$

Solution:

$$p(x) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)} \quad \text{— (1)}$$

Dividing LHS by $1 - p(x)$

$$\therefore \frac{p(x)}{1 - p(x)} = \frac{\dfrac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}}{1 - \left[\dfrac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}\right]}$$

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

$$= \exp(\beta_0 + \beta_1 X)$$

= RHS of equation (2).

Hence, proved.