Statistical Data Mining I Project 3

Question 1

In this problem used the Boston data set from the ISLR package and we fit classification models in order to predict whether a given suburb has a crime rate above or below the median. First the median of the crime was determined and observations above was given value 1, otherwise they were given 0. This produced the response variable column. The old crime column was then removed. Afterwards, the data was split into a training set and a testing set.

The relationship between the predictors was determined using a correlation plot, as shown in figure 1. We can see that predictors tax and rad are the most strongly correlated predictors with value of 0.9 which is greater than 0.75. Other predictors such as nox and indus are strongly correlated with a 0.77 which is greater than 0.75. Thus, the tax predictor was removed as it is strongly correlated rad and other predictors.

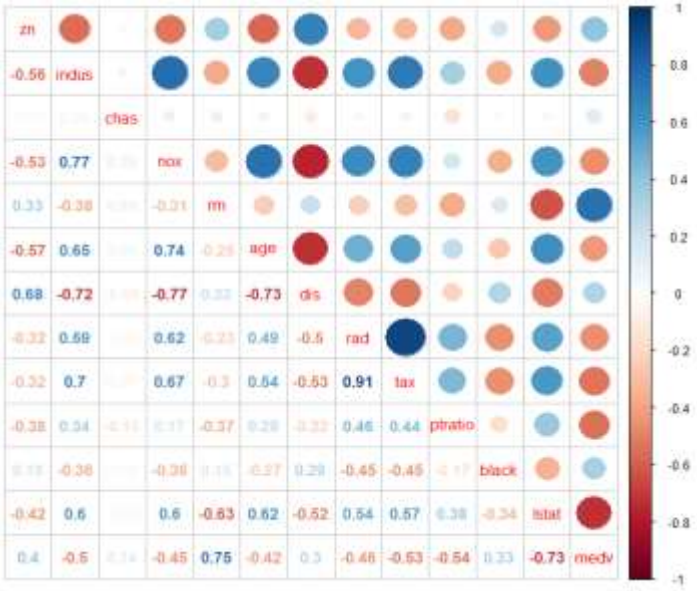

Figure 1: Correlation plot of the different predictors for the Boston data

Afterwards logistic regression, LDA and kNN models using various subsets of the predictors were explored and I decided to explore the predictor's nox, rad and dis.

Logistic regression

A logistic regression was fit and the testing error were determined to be 0.1496063 and training error was 0.1372032.
A confusion matrix and the corresponding statistic values were determined as shown in table 1.

Table 1: Confusion matrix and corresponding statistics for logistic regression

```
          Reference
Prediction  0  1
         0 48 10
         1  9 60

               Accuracy : 0.8504
                 95% CI : (0.7763, 0.9075)
    No Information Rate : 0.5512
    P-Value [Acc > NIR] : 6.662e-13

                  Kappa : 0.6981

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.8421
            Specificity : 0.8571
         Pos Pred Value : 0.8276
         Neg Pred Value : 0.8696
             Prevalence : 0.4488
         Detection Rate : 0.3780
   Detection Prevalence : 0.4567
      Balanced Accuracy : 0.8496

       'Positive' Class : 0
```

From the confusion table we can see that the

- Accuracy was 0.8504= 48+60/127
- Recall was 0.8421=48/48+9
- Precision was 0.8276= 48/58

Describe findings

The testing error and training error for this model are not to high being fairly close to zero. Furthermore, the accuracy which is true positive and true negatives combined, the precision which is the ratio of correctly predicted positive observations to the total predicted positive observations, and the recall (sensitivity) which is the ratio of correctly predicted positive observations to the all observations in the actual class are around 0.8. Since all these values are close to 1, it suggests that model is fairly accurate.

LDA model

A LDA model was fit and the testing error were determined to be 0.1653543 and training error was 0.1398417
A confusion matrix and the corresponding statistic values were determined as shown in table 2.

Table 2: Confusion matrix and corresponding statistics for logistic regression

```
          Reference
Prediction  0  1
         0 51 15
         1  6 55

               Accuracy : 0.8346
                 95% CI : (0.7584, 0.8946)
    No Information Rate : 0.5512
    P-Value [Acc > NIR] : 1.257e-11

                  Kappa : 0.6706

 Mcnemar's Test P-Value : 0.08086

            Sensitivity : 0.8947
            Specificity : 0.7857
         Pos Pred Value : 0.7727
         Neg Pred Value : 0.9016
             Prevalence : 0.4488
         Detection Rate : 0.4016
   Detection Prevalence : 0.5197
      Balanced Accuracy : 0.8402

       'Positive' Class : 0
```

From the confusion table we can see that the

- Accuracy was 0.8346= 51+55/127
- Recall was 0.8947=51/51+6
- Precision was 0.7727= 51/51+6

Describe findings

The testing error and training error for this model are not too high being fairly close to zero. Furthermore, the accuracy which is true positive and true negatives combined, the precision which is the ratio of correctly predicted positive observations to the total predicted positive observations, and the recall (sensitivity) which is the ratio of correctly predicted positive observations to the all observations in the actual class are close to 1. This suggests that model is fairly accurate, which is approximately close to the logistic regression model fit.

KNN model

A KNN model was fit and the testing error was determined for 10 k values.
The testing errors for the ten k values are respectively:

```
0.09448819 0.10236220 0.06299213 0.05511811 0.07086614 0.07874016 0.07874016 0.10236220 0.09448819 0.08661417
```

Describe findings

Based on the KNN model the when k=4, the test error is 0.05511811 which is the least error. Using the KNN model with k=4 provides the least testing error compared to the logistic regression model and the compared to the LDA model. With the logistic regression model giving the next least testing error.

Question 2

In this problem we had to use the diabetes data set. We had to disregard the first three columns. The fourth column is the observation number, and the next five columns are the predictors (glucose. area, insulin.area, SSPG, relative. weight, and fasting.plasma.glucose). The final column is the class number.

We had to produce pairwise scatterplots for all five variables (glucose. area, insulin.area, SSPG, relative. weight, and fasting.plasma.glucose), with different colors representing the three different classes, shown in figure 2 and 3.

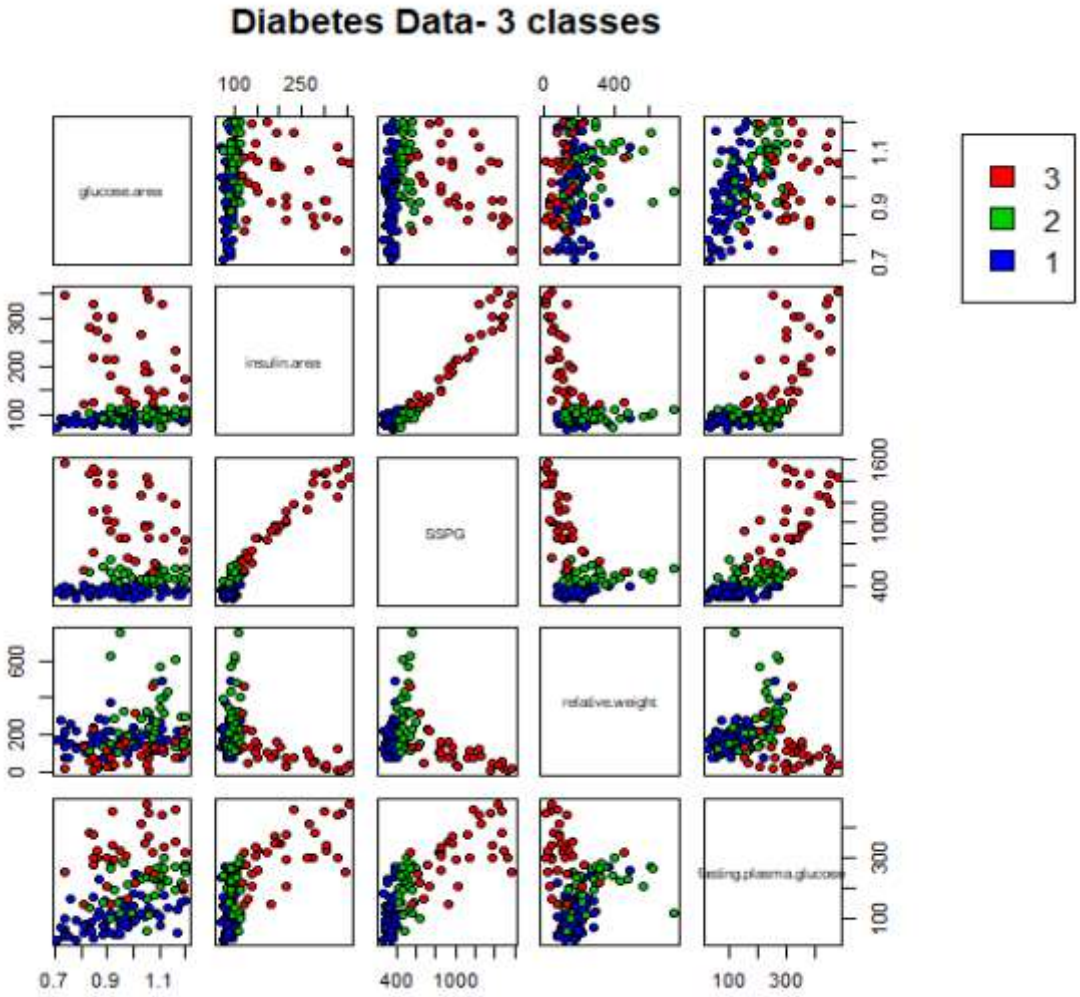

Figure 2: Pairwise scatterplots for all five variables (glucose. area, insulin.area, SSPG, relative. weight, and fasting.plasma.glucose), with different colors representing the three different classes.

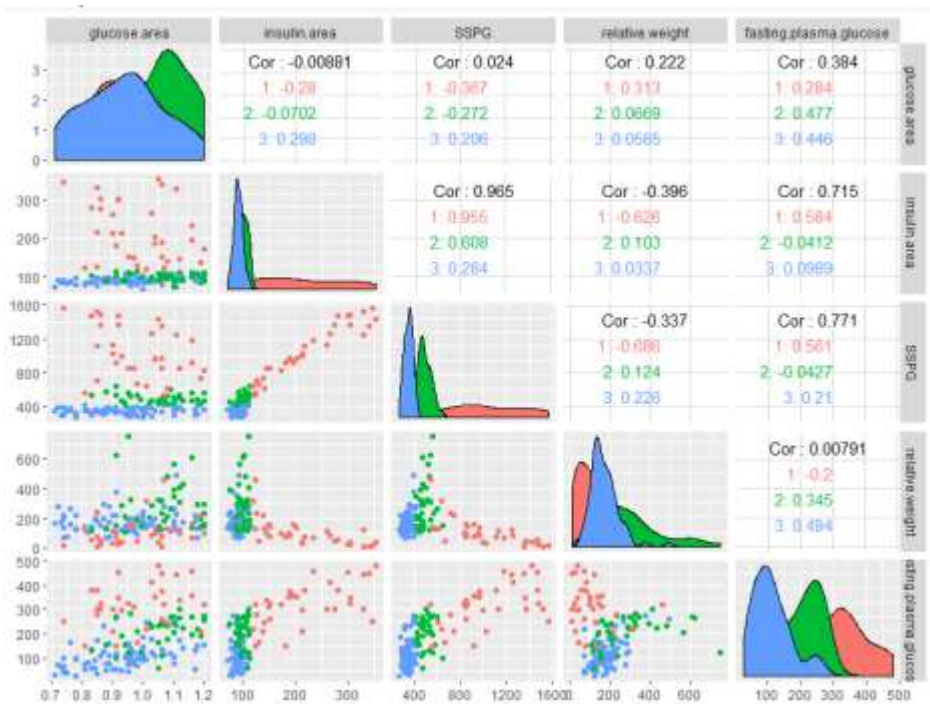

Figure 3: pairwise scatterplots for all five variables (glucose. area, insulin.area, SSPG, relative. weight, and fasting.plasma.glucose), with different colors representing the three different classes.

a) Do you see any evidence that the classes may have difference covariance matrices? That they may not be multivariate normal?

The multivariate normal (MVN), is used as the joint probability density function for continuous variables. The overall shape of the data defines the covariance matrix which is used to capture the spread of data. From figure 2 and 3 we can see that is clear that green and blue are closer in terms of their trends. The red is really different, this is the diabetic's class 3. The class covariance matrices are quite different, as this tells us how the variables vary with respect to each other, even though we can see they are correlated, but the spread of points is different for even the blue and green class.

b) Linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) were fitted.

LDA testing error was determined to be: 5.232518

QDA testing error was determined to be: 3.549844

We can see that the performance of QDA compare to that of LDA that qda is better. In this case the test prediction error is better but almost the close to LDA as there is not much difference of about 1.7. In addition, QDA is more flexible than LDA and has higher variance. The QDA is used when the training set is very large, so that the variance of the classifier is not an issue.

(c) We considered an individual which has (glucose area = 0.98, insulin area =122, SSPG = 544. Relative weight = 186, fasting plasma glucose = 184).

Using the fitted LDA model we determined to which class LDA assigns this individual, which as we can see from the posterior (probability of class) it is class 3

```
Levels: 1 2 3

$posterior
            1         2         3
1 0.001544973 0.4444845 0.5539705

$x
         LD1        LD2
1 -0.2159991 0.07535272
```

Using the fitted QDA model we determined to which class QDA assign this individual, which as we can see from the posterior (probability of class) it is class 2

```
Levels: 1 2 3

$posterior
            1         2         3
1 0.001544973 0.4444845 0.5539705
```

Question 3

Problem 3

a) Logistic regression model - the sum of posterior probabilities of classes is equal to 1. Show this holds for $k=K$.

The posterior probabilities are given by

$$P(G=k|X=x) = \frac{\exp(\beta_{k0}+\beta_k^T x)}{1+\sum_{l=1}^{K-1}\exp(\beta_{l0}+\beta_l^T x)} \quad \text{for } k=1,\dots,K-1$$

$$P(G=k|X=x) = \frac{1}{1+\sum_{l=1}^{K-1}\exp(\beta_{l0}+\beta_l^T x)}$$

Thus show: $\sum_{k=1}^{K} P(G=k|X=x)=1$

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

$$= \sum_{k=1}^{K-1} \frac{\exp(\beta_{k0}+\beta_k^T x)}{1+\sum_{l=1}^{K-1}\exp(\beta_{l0}+\beta_l^T x)} + \frac{1}{1+\sum_{l=1}^{K-1}\exp(\beta_{l0}+\beta_l^T x)}$$

$$= \frac{\sum_{k=1}^{K-1}\exp(\beta_{k0}+\beta_k^T x)+1}{1+\sum_{l=1}^{K-1}\exp(\beta_{l0}+\beta_l^T x)} = 1.$$

Problem 3b

$$P(x) = \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}$$

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

$$1-P(x) = \frac{1}{1+\exp(\beta_0+\beta_1 x)}.$$

$$\frac{P(x)}{1-P(x)} = \frac{\exp(\beta_0+\beta_1 x)/1+\exp(1+\exp(\beta_0+\beta_1 x)}{\frac{1}{(\exp(\beta_0+\beta_1 x)+1)}}$$

$$\therefore \quad \frac{P(x)}{1-P(x)} = \left(\frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}\right)\left(\frac{1+\exp(\beta_0+\beta_1 x)}{1}\right)$$

$$= \boxed{\exp(\beta_0+\beta_1 x).}$$

Question 4

In this problem we has to perform cross-validation on a simulated data set. We has to generate simulated data as follows and has to compute the LOOCV errors that result from fitting the following four models using least squares:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$$
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \varepsilon$$

a) The LOOCV error are as follows:

   Model 1: 6.975212

   Model 2: 0.9664678

   Model 3: 1.000017

   Model 4: 0.9993215

b) From the four models the models had the smallest LOOCV error is model 2 i.e. the second degree polynomial.
   This is what we expect because the true data is of quadratic form. The relation between x and y is quadratic. This can be seen from the scatterplot in figure it is clear that there is a quadratic relationship between x and y and that a least squares regression line (1st polynomial) would not be the best fit for the data. Furthermore, an increase the power (maybe overfitting) for a small decrease the test error.

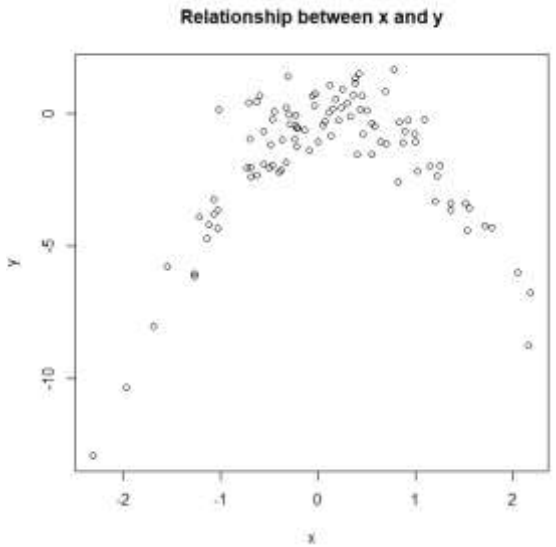

Figure 4: Relationship between x and y in the true data set

c) To comment on the statistical significance of the coefficient estimates we had to fit each of the four models using least squares.

Model 1:

```
Call:
lm(formula = y ~ x, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-10.083  -1.048   0.712   1.740   3.288

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.7263     0.2561  -6.741 1.09e-09 ***
x             0.4760     0.2806   1.696    0.093 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.548 on 98 degrees of freedom
Multiple R-squared:  0.02853,   Adjusted R-squared:  0.01862
F-statistic: 2.878 on 1 and 98 DF,  p-value: 0.09297
```

Model 2:

```
Call:
lm(formula = y ~ poly(x, 2), data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9368 -0.7291 -0.1191  0.6544  3.3320

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.68329    0.09704 -17.346  < 2e-16 ***
poly(x, 2)1   4.32330    0.97044   4.455 2.25e-05 ***
poly(x, 2)2 -23.34713    0.97044 -24.058  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9704 on 97 degrees of freedom
Multiple R-squared:  0.8606,    Adjusted R-squared:  0.8577
F-statistic: 299.3 on 2 and 97 DF,  p-value: < 2.2e-16
```

Model 3:

```
Call:
lm(formula = y ~ poly(x, 3), data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9078 -0.7189 -0.1174  0.6575  3.3062

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.6833     0.0975 -17.264  < 2e-16 ***
poly(x, 3)1   4.3233     0.9750   4.434 2.46e-05 ***
poly(x, 3)2 -23.3471     0.9750 -23.945  < 2e-16 ***
poly(x, 3)3   0.2971     0.9750   0.305    0.761
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.975 on 96 degrees of freedom
Multiple R-squared:  0.8607,    Adjusted R-squared:  0.8563
F-statistic: 197.7 on 3 and 96 DF,  p-value: < 2.2e-16
```

Model 4:

```
Call:
lm(formula = y ~ poly(x, 4), data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8630 -0.7018 -0.1688  0.6240  3.4230

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.68329    0.09742 -17.278  < 2e-16 ***
poly(x, 4)1   4.32330    0.97424   4.438 2.45e-05 ***
poly(x, 4)2 -23.34713    0.97424 -23.964  < 2e-16 ***
poly(x, 4)3   0.29713    0.97424   0.305    0.761
poly(x, 4)4   1.04583    0.97424   1.073    0.286
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9742 on 95 degrees of freedom
Multiple R-squared:  0.8624,    Adjusted R-squared:  0.8566
F-statistic: 148.8 on 4 and 95 DF,  p-value: < 2.2e-16
```

From the least squares models we can see that the first and second degree i.e. the linear and quadratic terms have statistically significant coefficient estimates as they have p-value <0.001 i.e.*** while the third and fourth degree terms do not and are thus not statistically significant.

d) This agrees strongly with our cross-validation results as the LOOCV error was minimum for the quadratic model, which has both the quadratic term and the linear term. Furthermore, this supports the conclusion that it is not worth it to fit a third or fourth degree polynomial as they would both be examples of overfitting and the correct model to fit the data is a second degree polynomial.