**STATISTICAL DATA MINING 1**

**Homework 3**

**Ashwin Vijayakumar (50249042)**

**Class Number 3**

1. **Boston Data Set :**



The following analysis is for the Boston data set , where we try to build a model to predict whether a given suburb has a high or low crime rate (above or below median) . We explore Logistic regression , Linear Discriminant Analysis and KNN for the same .

First we split the data into training and test set . We then perform the training of the model on the train set and ultimately calculate the prediction (test) error by running the model against our test set .

1. **Logistic Regression :**

By fitting a logistic regression model over the training set and predicting the test error on the predicted values for the test set , we obtain the logistic regression test error as :

lr.test.error

[1] 0.09448819

1. **LDA :**

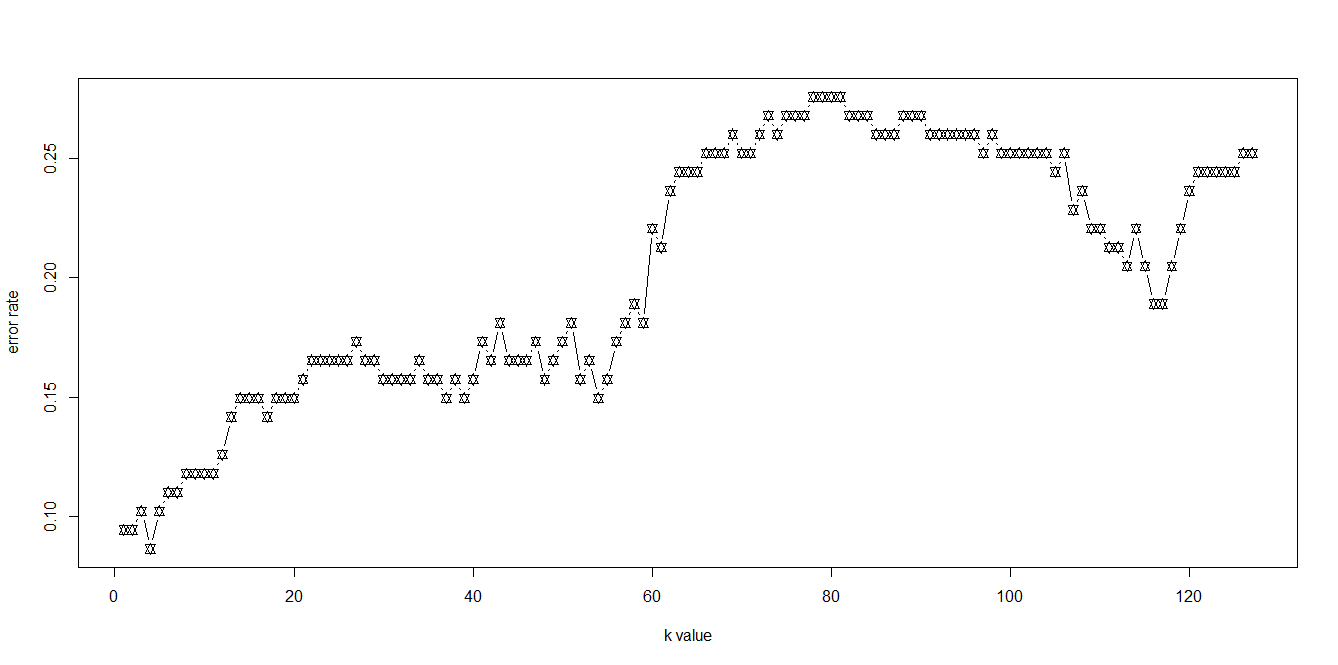
By fitting a linear discriminant analysis model over the training set and predicting the test error on the predicted values for the test set , we obtain the LDA test error as :

lda.test.error

[1] 0.1653543

1. **KNN :**

To find the test error on a Knn model , we first train our model over the training set for all values of K starting from 1 to Number of test samples and evaluate it over the test set .The following plot shows how the error values vary for each selection of K .

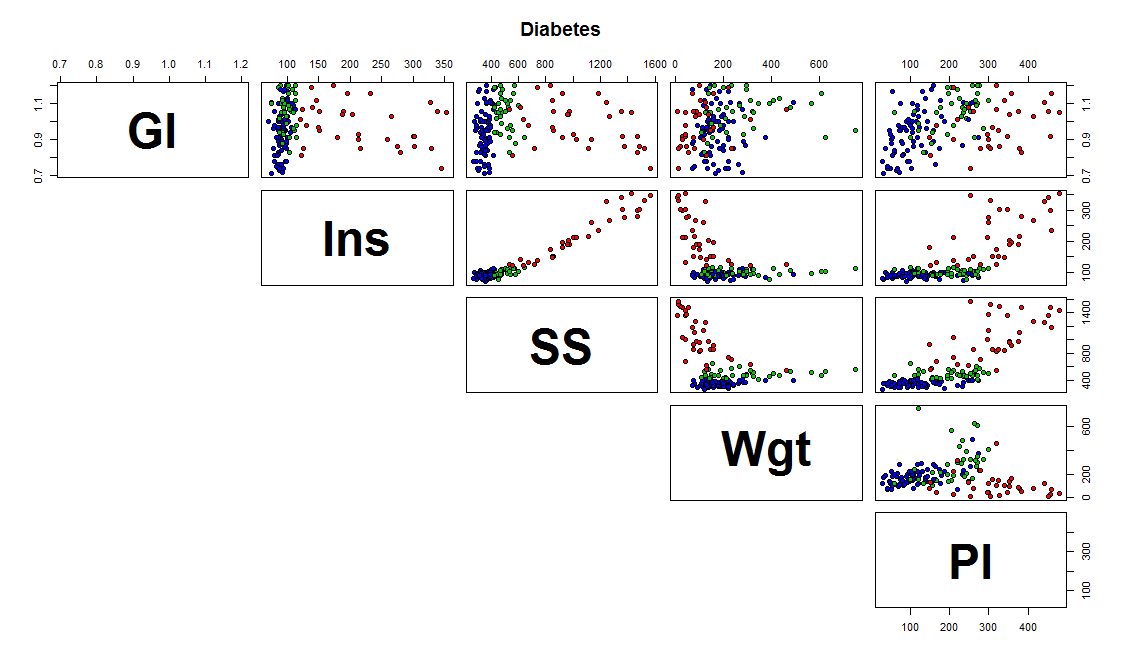


We find that a KNN model with k=4 has the least testing error (0.866) . Hence the prediction used by this model would be accurate in determining whether a suburb belongs to a region with high crime rate or low crime rate . We obtain the Knn test error as :

min.knn.error

[1] 0.08661417

Therefore , we see that the KNN model performs the best (Has a minimum error) compared to Logistic regression model and Linear Discriminant Model .



2)

a) When the data is a multivariate normal distribution and the covariance matrix is common, then based on the assumptions of the LDA model , prediction may be accurate (Provided the model is able to fit the data with a low error) . But if the assumption that the covariance matrix is same throughout does not hold , then the QDA may be a better model to be used as a predictor (Again , provided the model is able to fit the data with a low error) . In this case , we see that for each class , the covariance between each pair of predictor variables are not the same . Only certain predictor pairs , such as (Glucose,SSPG ) , (Insulin,SSPG) , (SSPG, Weight ) show common correlations for atmost 2 classes . But other pairs of predictors such as (Weight, Fasting Plasma Glucose) (Glucose , Insulin) and (Glucose and Fasting Plasma Glucose ) have data points belonging to different classes interspersed with each other , hence the assumption of a common covariance matrix may not lead to an optimal decision boundary in our LDA model. Also from the plots, its easy to see that most pairs of predictor variables are normally distributed , although in some of the plots not all classes seem to be normally distributed . Overall its safe to assume normal distribution of priors .

**Covariance Matrix :**

V1 V2 V3 V4 V5

V1 1.000000000 -0.008813193 0.0239843 0.222237813 0.384319804

V2 -0.008813193 1.000000000 0.9646281 -0.396234858 0.715480192

V3 0.023984304 0.964628091 1.0000000 -0.337020435 0.770942459

V4 0.222237813 -0.396234858 -0.3370204 1.000000000 0.007914263

V5 0.384319804 0.715480192 0.7709425 0.007914263 1.000000000

b)

LDA\_train\_error

[1] 0.0862069

LDA\_test\_error

[1] 0.1724138

QDA\_train\_error

[1] 0.02586207

QDA\_test\_error

[1] 0.1724138

The LDA training error as expected is less than the LDA test error . Similarly , the QDA training error is lesser than the QDA test error , when tested on the same data split . Also , the Overall training error in the QDA case seems to be lesser than the test error , but overall test error in both the cases are the same . This indicates that ,provided the assumptions that each model takes for granted holds , they both can be used for similar predictions . But we’ll see later why QDA would be a better selection.

c) The LDA model predicts the new data to be in class 3 whereas the QDA model predicts it to be in class 2 . But since the assumption needed for an Optimal LDA classifier may not withhold here , the QDA offers a better classification accuracy . In other words , the data [0.98,122,544,186,184] , assuming our assumptions hold , belongs to class 2 .

3) Under the logistic regression model , p(X) = exp(β0 + β1X)/ ( 1+ exp(β0 + β1X))

a) **Posterior probability for k=K is given by** :

P1 = P(C=k/X=x) = 1/( 1+∑[1 to K-1] exp (βl0+βTlx))

**Posterior probability for k=1…K-1 is given by** :

P2= P(C = [1..K-1] / X=x) = exp (βk0+βTkx)/(1+∑[1 to K-1]exp (βl0+βTlx))

P3 = Sum(P2) over all k from 1 to K-1

= ∑[1 to K-1]exp (βl0+βTlx) / (1+∑[1 to K-1]exp (βl0+βTlx) )

We see that P1 + P3 = (1 + ∑[1 to K-1]exp (βl0+βTlx) ) /(1+∑[1 to K-1]exp (βl0+βTlx) )

= 1 .

b)The logistic function is given by exp(β0 + β1X)/ (1+ exp(β0 + β1X)) and the Logit is given by the log of odds ration , i.e , log(p(X)/ (1-p(X))) .

Therefore , P(X) = exp(β0 + β1X)/ ( 1+ exp(β0 + β1X))

And 1 – P(X) = 1 - exp(β0 + β1X)/ ( 1+ exp(β0 + β1X))

= ( 1+ exp(β0 + β1X) - exp(β0 + β1X) )/( 1+ exp(β0 + β1X)

= (1)/( 1+ exp(β0 + β1X))

Which means P(X)/(1-P(X)) = exp(β0 + β1X)\*(1+ exp(β0 + β1X))/ ( 1+ exp(β0 + β1X))

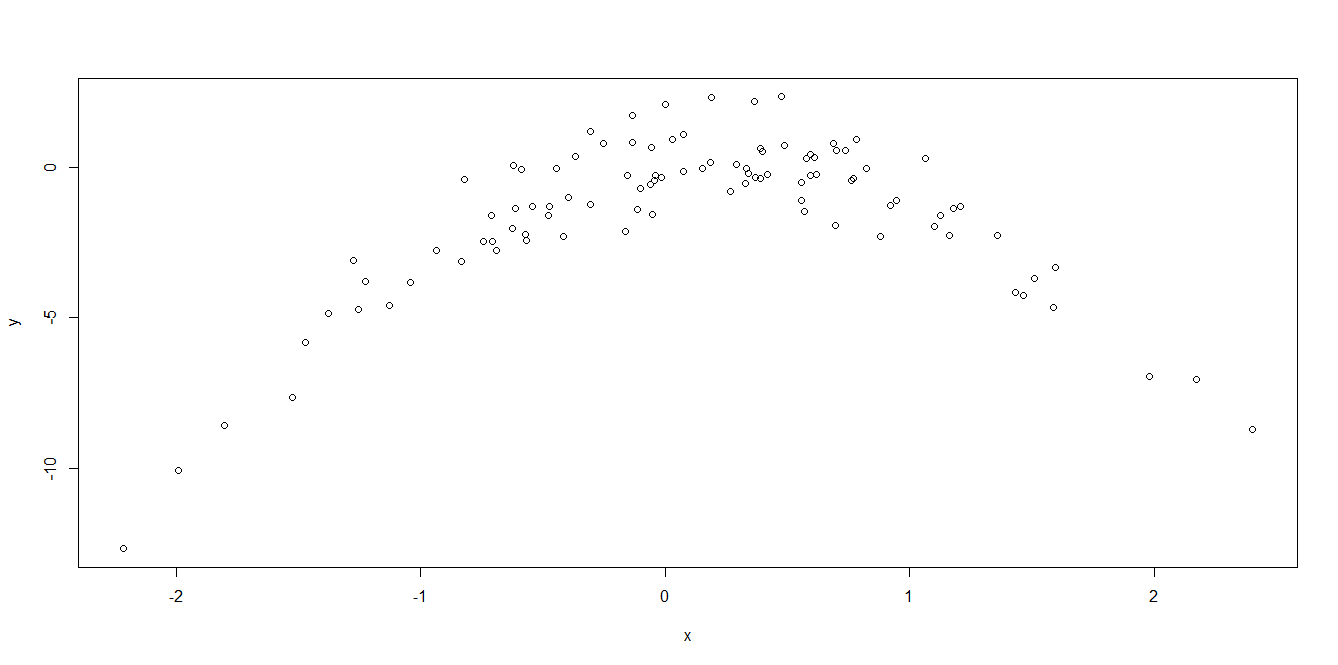
= exp(B0 + B1X)

Which is equivalent to the Logit function

Therefore the logistic function and the logit function representations are equivalent .

4 )

We generate an arbitrary data set using a random seed value .



The following analysis is for studying the effect LOOCV has on different models generated arbitrarily from random samples .

We use the following 4 models

Y = β0 + β1X + ε ---------------------1

Y = β0 + β1X + β2X2 + ε -------------------------2

Y = β0 + β1X + β2X2 + β3X3 + ε -----------------------------3

Y = β0 + β1X + β2X2 + β3X3 + β4X4 + ε----------------------------4

Using LOOCV , the following are the errors that we obtained :

Model 1 : [1] 7.288162

Model 2 : [1] 0.9374236

Model 3: [1] 0.9566218

Model 4: [1] 0.9539049

b) The second model is the one with the lowest LOOCV error . This is expected because the relation between x and y is quadratic in nature . This can be seen from the figure that the relation between y and x is parabolic or quadratic with degree = 2 , similar to the second model .

c)

This is the summary of our 4 models .

**Coefficients:**

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.6254 0.2619 -6.205 1.31e-08 \*\*\*

x 0.6925 0.2909 2.380 0.0192 \*

------------------------------------------------------------------------------------------------------------------------------------------

**Coefficients:**

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.5500 0.0958 -16.18 < 2e-16 \*\*\*

poly(x, 2)1 6.1888 0.9580 6.46 4.18e-09 \*\*\*

poly(x, 2)2 -23.9483 0.9580 -25.00 < 2e-16 \*\*\*

---

------------------------------------------------------------------------------------------------------------------------------------------

**Coefficients:**

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.55002 0.09626 -16.102 < 2e-16 \*\*\*

poly(x, 3)1 6.18883 0.96263 6.429 4.97e-09 \*\*\*

poly(x, 3)2 -23.94830 0.96263 -24.878 < 2e-16 \*\*\*

poly(x, 3)3 0.26411 0.96263 0.274 0.784

--------------------------------------------------------------------------------------------------------------------

**Coefficients:**

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.55002 0.09591 -16.162 < 2e-16 \*\*\*

poly(x, 4)1 6.18883 0.95905 6.453 4.59e-09 \*\*\*

poly(x, 4)2 -23.94830 0.95905 -24.971 < 2e-16 \*\*\*

poly(x, 4)3 0.26411 0.95905 0.275 0.784

poly(x, 4)4 1.25710 0.95905 1.311 0.193

As expected , according to the summary of the fits of each model , based on the p values and other coefficients , the quadratic model appears to be more statistically significant than the linear , cubic and polynomial with degree 4 .