EAS596, Homework_3

Abhishek Kumar, Class#1

2/10/2018

SOLUTION 1

To predict whether a given suburb has a crime rate above or below the median, we make a new varible "crime_above_median" that is 0 if the crime in the suburb is less than the median else will be 1. We then divide the dataset into training and test in the ratio of 75:25 respectively. To analyze the correlation among variable let's look at the correlation plot:

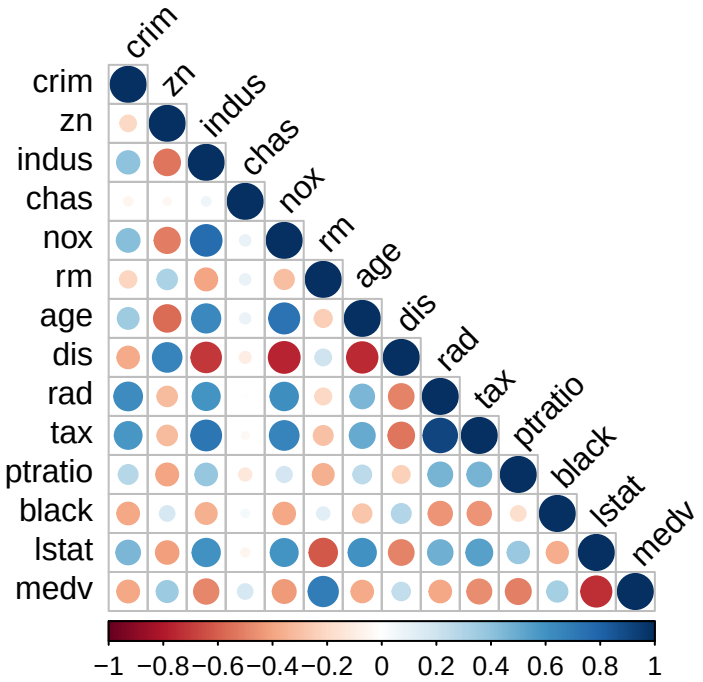


Figure 1: Correlation Matrix Intensity

From the above correlation plot we can infer that crime is very loosely related to 'chas','zn' and 'rm'. We can also deduce that 'chas' has negligible magnitude while 'zn' and 'rm' has very less but more magnitude than 'chas'. So we will work on three subsets: (a)With all the variables (b) Excluding 'chas' (c) Excluding 'chas', 'zn' and 'rm'

USING LOGISTIC REGRESSION

Including all variables

```
## Confusion Matrix for test data:

##
## logi.test.pred_a  0  1
##                0 62  7
##                1  3 54

## Train Accuracy:  0.9105263

## Test Accuracy:  0.9206349
```

Excluding 'chas'

```
## Confusion Matrix for test data:

##
## logi.test.pred_b  0  1
##                0 61  7
##                1  4 54

## Train Accuracy:  0.9105263

## Test Accuracy:  0.9126984
```

Excluding 'chas', 'rm', 'tax'

```
## Confusion Matrix:

##
## logi.test.pred_c  0  1
##                0 61  8
##                1  4 53

## Train Accuracy:  0.9131579

## Test Accuracy:  0.9047619
```

USING LDA

Including all variables

```
## Confusion Matrix for test data:
```

```
##
##      0  1
##   0 62 15
##   1  3 46
```

```
## Training Accuracy:  0.8571429
```

```
## Test Accuracy:  0.8571429
```

Excluding 'chas'

```
## Confusion Matrix for test data:
```

```
##
##      0  1
##   0 62 15
##   1  3 46
```

```
## Training Accuracy:  0.8736842
```

```
## Test Accuracy:  0.8571429
```

Excluding 'chas', 'rm', 'tax'

```
## Confusion Matrix for test data :
```

```
##
##      0  1
##   0 64 15
##   1  1 46
```

```
## Training Accuracy : 0.8631579
```

```
## Test Accuracy, excluding chas, tax, rm:  0.8730159
```

USING kNN

Including all variables

```
## #kNN MODEL; including all variables
```

```
## Confusion Matrix for test data :
```

```
##
## knn.pred_a  0  1
##          0 64  9
##          1  1 52
```

```
## kNN Test Accuracy, including all variables:  0.9206349
```

Excluding 'chas'

```
## #kNN MODEL; excluding 'chas'
```

```
## Confusion Matrix for test data :
```

```
##
## knn.pred_b  0  1
##          0 64  9
##          1  1 52
```

```
## Test Accuracy:  0.9206349
```

Excluding 'chas', 'rm', 'tax'

```
## #kNN MODEL; excluding 'chas', 'rm', 'tax'
```

```
## Confusion Matrix for test data :
```

```
##
## knn.pred_c  0  1
##          0 57 12
##          1  8 49
```

```
## Test Accuracy:  0.8412698
```

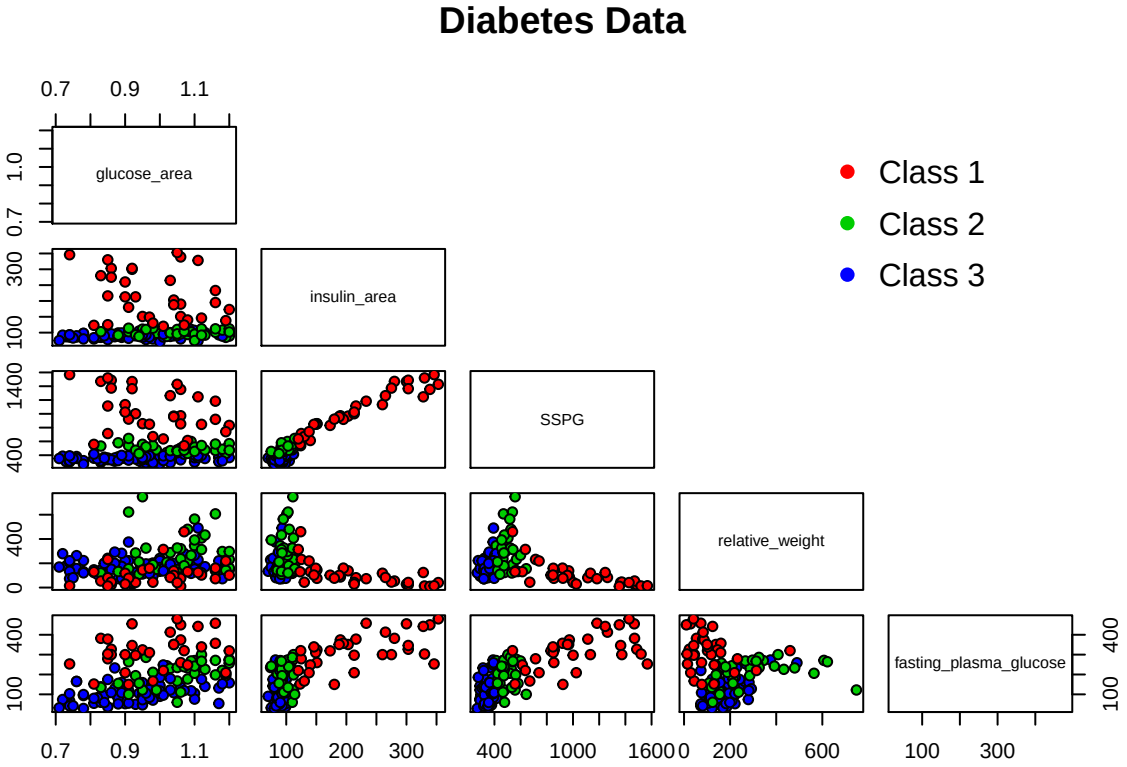
INFERENCES :

From the above computations, we see that the performance of logistic regression and knn is better than the LDA classification method. This maybe because the predictors are not normally distributed and also because the classes not linearly seperable. Although there is some difference in the performance of the LDA, it is not very pronounced.

SOLUTION 2

(a)

When we look at the pairplot below we can see that most of the distributions are different for each of the plot. Hence each class will have different covariance with respect to the pairwise variables and eventually will have different covariance matrices.



Its hard to determine if each class have a multivariate distribution. So, we'll do a multivariate test using MVN library to check if the classes have multivariate distribution.

```
## sROC 0.1-2 loaded
```

```
## Multivariate test for class 1
```

```
##              Test         Statistic              p value Result
## 1 Mardia Skewness  72.1130242128528 0.000223018134108012     NO
## 2 Mardia Kurtosis 0.759602021039537    0.447492510707829    YES
## 3             MVN              <NA>                 <NA>     NO
```

```
## Multivariate test for class 2
```

```
##              Test          Statistic            p value Result
## 1 Mardia Skewness   50.8434201071473 0.0406859141683021     NO
## 2 Mardia Kurtosis 0.0757723206589703  0.939600237713607    YES
## 3             MVN               <NA>                <NA>     NO
```

```
## Multivariate test for class 3
```

```
##              Test        Statistic             p value Result
## 1 Mardia Skewness 68.2997291313893 0.00063808969256816     NO
## 2 Mardia Kurtosis 2.08698632017701  0.0368893711399998     NO
## 3             MVN             <NA>                <NA>     NO
```

From the above test we can see that class 1 and 2 passes just Mardia Kurtosis test but fails the Mardia Skewness test. While class 3 fails both the test. So we conclude that the classes does not have multivariate distribution. But as we have very less data points, our conclusion are not with confidence.

(b)

```
## FITTING LDA MODEL:

## Confusion Matrix for test data:

##
##      1  2  3
##   1  8  0  0
##   2  2 10  0
##   3  1  0 15

## Training Accuracy : 0.9082569

## Test Accuracy :  0.9166667

## FITTING QDA MODEL:

## QDA Confusion Matrix for test data:

##
##      1  2  3
##   1  9  0  0
##   2  2  9  0
##   3  0  1 15

## QDA Training Accuracy : 0.9541284

## QDA Test Accuracy :  0.9166667
```

(c)

```
## The posterior probabilities for the data is :

##              1         2         3
## 1 4.841209e-05 0.5462818 0.4536698

## And the class predicted by QDA is :  2

## The posterior probabilities for the data is :

##           1         2            3
## 1 0.2214092 0.7785877 3.117562e-06

## And the class predicted by QDA is :  2
```