Write up for homework 3

Hyungkyu Lim

1) When I generate correlation of each variable easepcialy for crim, there are several variables that have strong relationship with crim. Among them, I chose Indus, nox, rad,tax, dis and age variables to do logistic regression, LDA and Knn. For logistic regression, the train error is about 0.122 and test error is about 0.157 based on the confusion matrix(Confusion matrix function doesn't work in my R environment so, I used table function insted of confusion matrix. And table function works exactly same with the confusion matrix function). In table matrix, I calulated the test error for logistic regression using False Positive and False Negative.

```
## [1] 0.1219512
```

```
##    glm.y_hat_test
##      0  1
##   0 77 13
##   1 15 73
```

```
## [1] 0.157
```

For LDA, train error is about 0.14 and test error is about 0.174. I also used the table matrix to get the test error of LDA.

```
## [1] 0.1402439
```

```
##    lda.y_hat_test
##      0  1
##   0 85  5
##   1 26 62
```

```
## [1] 0.174
```

For Knn, the test error of K=1 is about 0.457. K=5 is 0.464 K=10 is 0.462. For K=20, the test error is 0.472. Based on the test error of Knn, I found when K increases, the test error is also incrases (except for K=5). As a result, the minimum test error among three models is logistic regression model. and the maximum test error among threee models is Knn model with K=20.

```
##    knn_1
##      0  1
##   0 78 12
##   1  8 80
```

```
## Warning in `!=.default`(knn_1, subset.variable[, 1]): longer object length
## is not a multiple of shorter object length
```

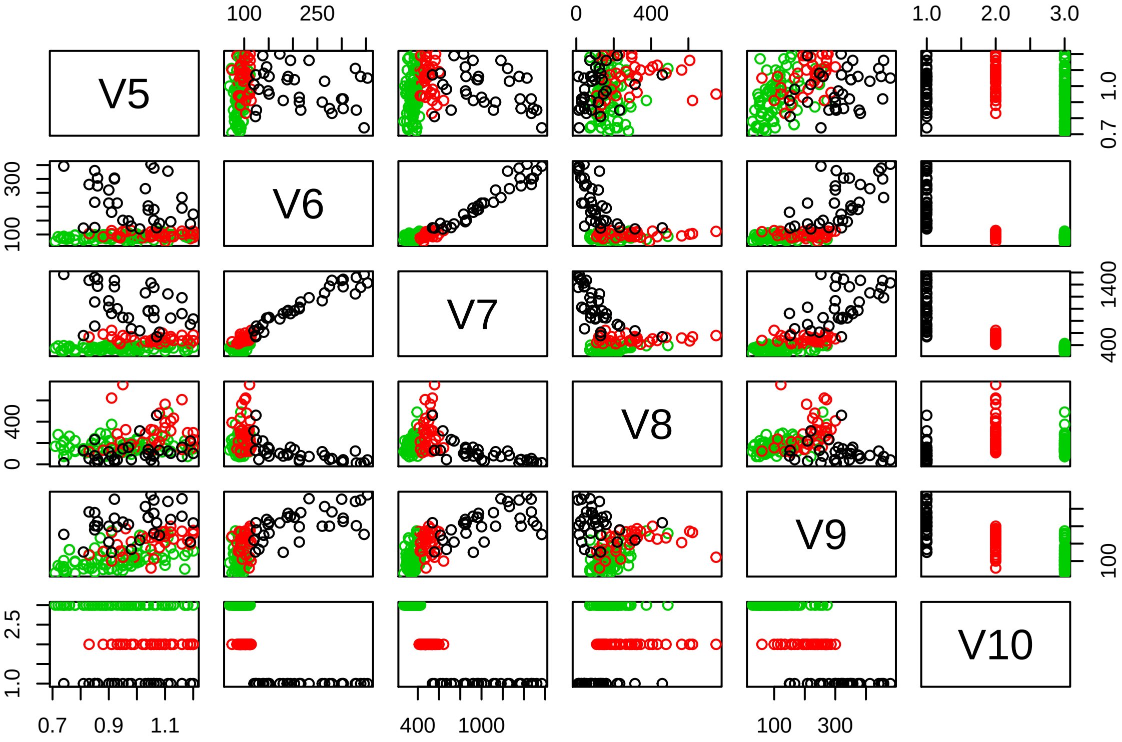
```
## Warning in is.na(e1) | is.na(e2): longer object length is not a multiple of
## shorter object length
```

```
## [1] 0.457
```

```
##    knn_5
##      0  1
##   0 82  8
##   1  8 80
```

```
## Warning in `!=.default`(knn_5, subset.variable[, 1]): longer object length
## is not a multiple of shorter object length

## Warning in `!=.default`(knn_5, subset.variable[, 1]): longer object length
## is not a multiple of shorter object length

## [1] 0.464

##    knn_10
##      0  1
##   0 79 11
##   1  8 80

## Warning in `!=.default`(knn_10, subset.variable[, 1]): longer object length
## is not a multiple of shorter object length

## Warning in `!=.default`(knn_10, subset.variable[, 1]): longer object length
## is not a multiple of shorter object length

## [1] 0.462

##    knn_20
##      0  1
##   0 77 13
##   1 12 76

## Warning in `!=.default`(knn_20, subset.variable[, 1]): longer object length
## is not a multiple of shorter object length

## Warning in `!=.default`(knn_20, subset.variable[, 1]): longer object length
## is not a multiple of shorter object length

## [1] 0.472
```

2) Based on scatter plot, I can see that 3 classes have different covariance matrices for some variables. However it looks that 3 classes have almost same covariance matrix for glucose.area variable.



To compare performance of LDA and QDA, I got the train and test error for those of two. Train error of LDA is about 0.096 and train error of QDA is about 0.032. Test error of LDA is about 0.12 and test error of QDA is about 0.14. So I guess performance of LDA is better than the performance of QDA.

```
## [1] 0.09574468
```

```
##                     diabetes_true_test
## diabetes_y_hat_test  1  2  3
##                   1  6  0  0
##                   2  1 12  1
##                   3  1  3 27
```

```
## [1] 0.118
```

```
## [1] 0.03191489
```

```
##                    diabetes_qda_y_hat_test
## diabetes_true_test  1  2  3
##                  1  7  1  0
##                  2  2 10  3
##                  3  0  1 27
```

```
## [1] 0.137
```

With the indivisual dataset, LDA assigns individual dataset class #3 and QDA assigns individual dataset #1.

```
## [1] 3
```

```
## [1] 1
```