Homework 3

Freya Genesis D Mello

November 2018

1 Problem Statement 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.

1.1 Preparing the data

The Boston data set is studied to identify the two classes of low and high crime rates. They're classified into low and high risk areas in Boston for housing choices. The median of the crime rate is used to split the data set into the two levels. The test and training data is set and prepared. The exhaustive subset selection is used to learn that the best subset model is with 4 predictors. 1

1.2 Logistic Regression

The training and test data is fit in the logistic regression model for various sizes of subsets from 1 to 13 predictors. The test and training errors are computed on every iteration to see that the least error rate is observed when the number of predictors were 4 for the subset being created for a minimum error of 0.09473256 in train set and 0.106242 in the test set. 2

1.3 LDA

The training and test data is fit in the LDA for various sizes of subsets from 1 to 13 predictors. The test and training errors are computed on every iteration to see that the least error rate is observed when the number of predictors were 4 for the subset being created for a minimum error of 0.1291291 in train set and 0.1271676 in the test set. 3

As compared to Logistic Regression it is seen that LDA shows a higher error rate especially for the training set.

As the trend is observed in the LDA and Logistic Regression graphs, the train error reduces as the subset size of the model increases. The train error reduces at first but then it rises to acceptable error values.

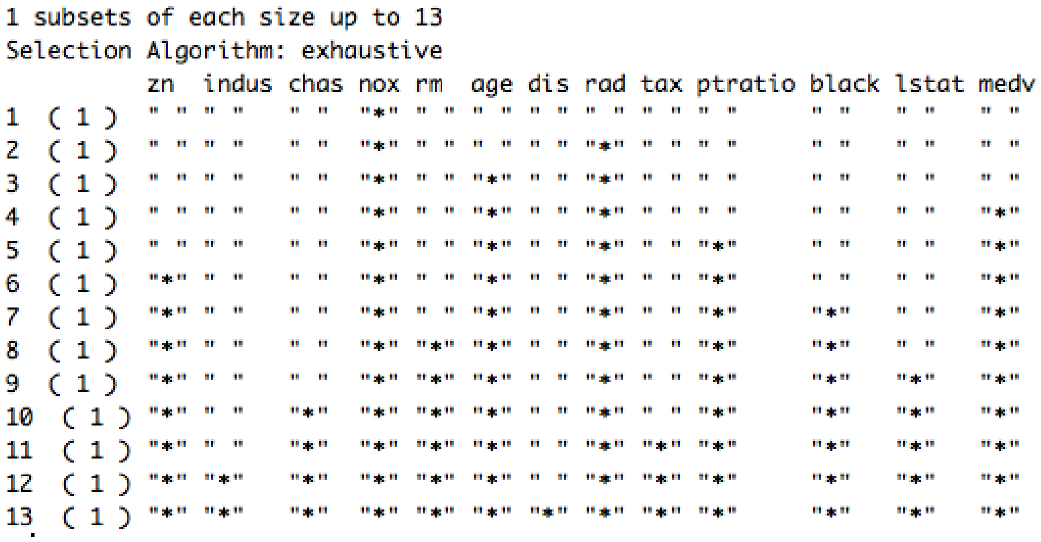
1.4 KNN

In comparison to the LDA, Logistic regression model performs better in predicting the crime rate in areas of Boston. According to the error rates observed in 4 and 5, the KNN with k-value = 2 has the lowest test and k-value = 4 has lowest training error but it's not an optimal model since we are trying to over fit the data by keeping value of k as low.

cc

Confusion Matrix

	0	**1**	**total**
0	64	24	
1	3	82	
total	P	N	



```
1 subsets of each size up to 13
Selection Algorithm: exhaustive
          zn  indus chas nox rm  age dis rad tax ptratio black lstat medv
1  ( 1 )  " " " "   " "  "*" " " " " " " " " " " " "     " "   " "   " "
2  ( 1 )  " " " "   " "  "*" " " " " " " "*" " " " "     " "   " "   " "
3  ( 1 )  " " " "   " "  "*" " " "*" " " "*" " " " "     " "   " "   " "
4  ( 1 )  " " " "   " "  "*" " " "*" " " "*" " " " "     " "   " "   "*"
5  ( 1 )  " " " "   " "  "*" " " "*" " " "*" " " "*"     " "   " "   "*"
6  ( 1 )  "*" " "   " "  "*" " " "*" " " "*" " " "*"     " "   " "   "*"
7  ( 1 )  "*" " "   " "  "*" " " "*" " " "*" " " "*"     "*"   " "   "*"
8  ( 1 )  "*" " "   " "  "*" "*" "*" " " "*" " " "*"     "*"   " "   "*"
9  ( 1 )  "*" " "   " "  "*" "*" "*" " " "*" " " "*"     "*"   "*"   "*"
10  ( 1 ) "*" " "   "*"  "*" "*" "*" " " "*" " " "*"     "*"   "*"   "*"
11  ( 1 ) "*" " "   "*"  "*" "*" "*" " " "*" "*" "*"     "*"   "*"   "*"
12  ( 1 ) "*" "*"   "*"  "*" "*" "*" " " "*" "*" "*"     "*"   "*"   "*"
13  ( 1 ) "*" "*"   "*"  "*" "*" "*" "*" "*" "*" "*"     "*"   "*"   "*"
```

Figure 1: Exhaustive Subset selection giving the best subsets

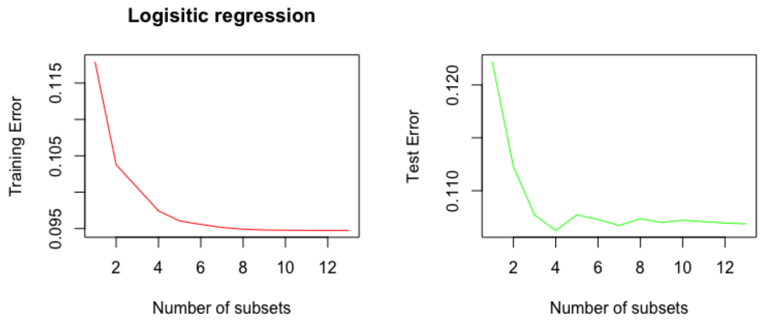


Figure 2: Logistic Regression observing test and train data errors for various subset sizes.

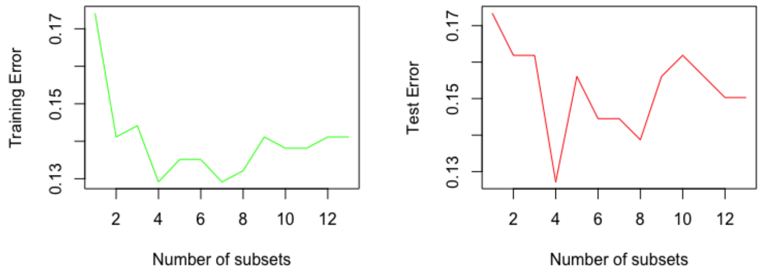
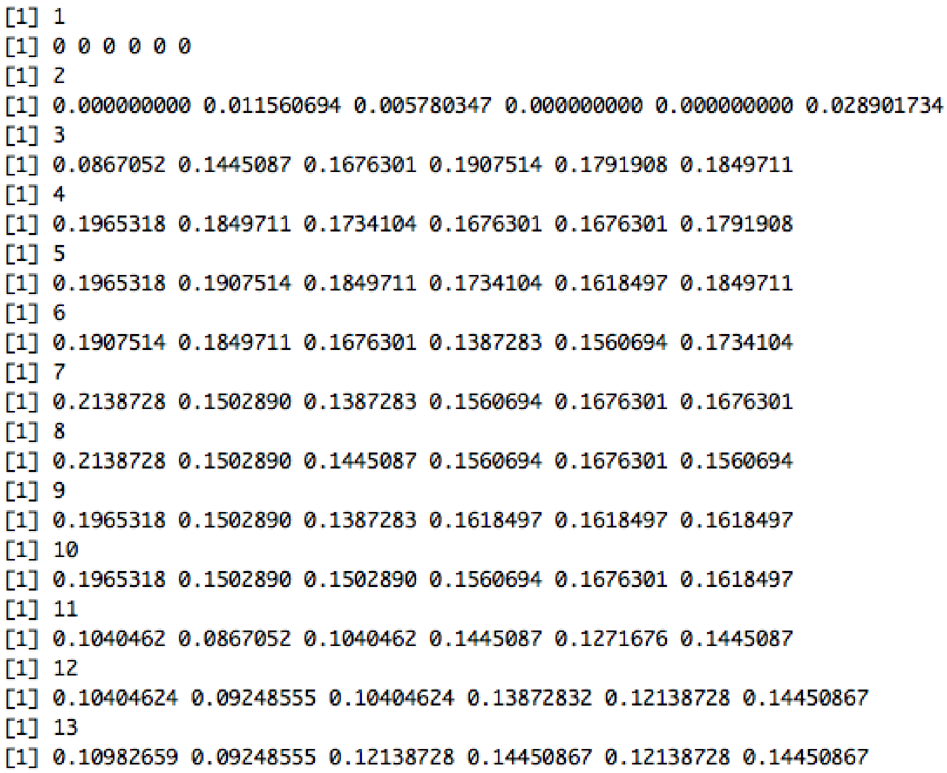


Figure 3: LDA observing test and train data errors for various subset sizes.



```
[1] 1
[1] 0 0 0 0 0 0
[1] 2
[1] 0.000000000 0.011560694 0.005780347 0.000000000 0.000000000 0.028901734
[1] 3
[1] 0.0867052 0.1445087 0.1676301 0.1907514 0.1791908 0.1849711
[1] 4
[1] 0.1965318 0.1849711 0.1734104 0.1676301 0.1676301 0.1791908
[1] 5
[1] 0.1965318 0.1907514 0.1849711 0.1734104 0.1618497 0.1849711
[1] 6
[1] 0.1907514 0.1849711 0.1676301 0.1387283 0.1560694 0.1734104
[1] 7
[1] 0.2138728 0.1502890 0.1387283 0.1560694 0.1676301 0.1676301
[1] 8
[1] 0.2138728 0.1502890 0.1445087 0.1560694 0.1676301 0.1560694
[1] 9
[1] 0.1965318 0.1502890 0.1387283 0.1618497 0.1618497 0.1618497
[1] 10
[1] 0.1965318 0.1502890 0.1502890 0.1560694 0.1676301 0.1618497
[1] 11
[1] 0.1040462 0.0867052 0.1040462 0.1445087 0.1271676 0.1445087
[1] 12
[1] 0.10404624 0.09248555 0.10404624 0.13872832 0.12138728 0.14450867
[1] 13
[1] 0.10982659 0.09248555 0.12138728 0.14450867 0.12138728 0.14450867
```

Figure 4: Subsets models produced test errors for k values = 1,5,10,20,30,50

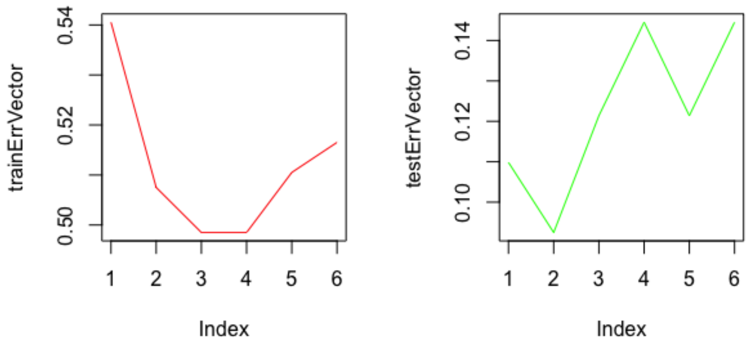


Figure 5: Train and test errors for the KNN values

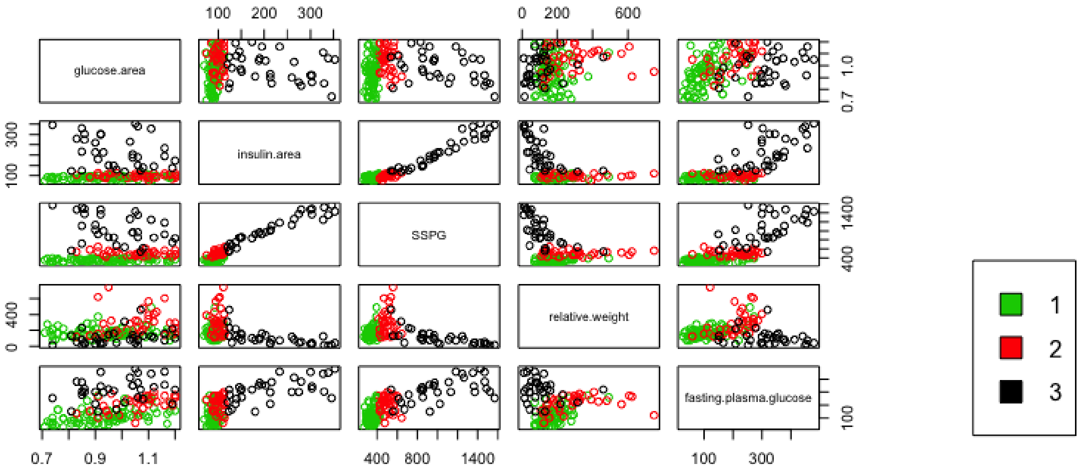


Figure 6: Scatter plot for spread of the data belonging to classes 1,2,3 among all 5 variables.

2 Problem Statement 2

Download the diabetes data set. 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. (Note: this data can also be found in the MMST library)
(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?
(b) Apply linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). How does the performance of QDA compare to that of LDA in this case?
(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?

2.1 Preparing the data set

The diabetes data set was read and learned. The first four variables of the data set were disregarded. The rest of the variables had their columns named to help explore them better.

The pairwise scatter plots were plotted for all five variables, with different colors representing the three different classes.

The classes do have a different covariance matrix and are not multivariate normal.

2.2 Linear Discriminant Analysis

The class number is the factor used to classify whether the person belongs either one of the classes 1,2 or 3. Using the class number, the LDA model helps obtain predicted values of train and test data and compare it with the actual values of the class number. The misclassification error is calculated as 0.08421053 for the train set and 0.14 for the test set. Here it is seen that the test error thus computed is more than the train and so the study is extended to QDA.

2.3 Quadratic Discriminant Analysis

In QDA, the model produces predicted values that are compared against the actual values and the misclassification error is noted. The misclassification error is calculated as 0.05263158 for the train set and 0.06 for the test set. The test and train error don't have much of a difference as compared the values from LDA model. The error values show that QDA worked better to predict than LDA for the right degree of classification.

2.4 Classifying the test data given

The data is used to test against the already created LDA and QDA models and the predicted value of the class number was observed. By both models, the class number predicted is 3.

3 Problem Statement 3

a) Under the assumptions in the logistic regression model, the sum of posterior probabilities of classes is equal to one. Show that this holds for k=K.
b) Using a little bit of algebra, show that the logistic function representation and the logit representation for the logistic regression model are equivalent.

4 Question 3

4.1 (a)

Posterior probabilities : Probability of an observation belonging to a class(G) given a particular observation(X)

The log ratios of posterior probabilities are called log-odds which gives a particular set of linear regression curve according to different classes considered.

For predicting the probability of the observation belonging to the 1st class given the observation:

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

For predicting the probability of the observation belonging to the 2nd class given the observation:

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

.

.

And so on till we reach the (K-1) class:

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

From above formulas, it can be clearly seen that

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

For k = 1, ... , K - 1

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

For the last class when we have exhausted all the classes, the log ratios of the posterior probabilities is given as:

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

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

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

Hence, under the assumptions in the logistic regression model, the sum of posterior probabilities of classes is equal to one. This holds for values from k=1 to k=K. i.e sum of probabilities of a particular observation belonging to all of the classes is 1.

4.2 (b)

The logistic function: $p(X) = \frac{\exp{(\beta_0+\beta_1 x)}}{1+\exp{(\beta_0+\beta_1 x)}}$

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

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

Dividing the above equations we get the logit representation

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

Hence, the logistic function representation and the logit representation for the logistic regression model are equivalent.