PAUL M Girdler
CLASS 15
paulgird@buffalo.edu.au

Homework Assignment 3
EAS506 - Statistical Data Mining 1

2018-11-03

1 Exercise 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.

The Boston dataset contains 506 data points with 13 predictors and 1 response variable, **crim**, the per capita crime rate by town. The response variable, **crim**, was converted to a class variable, with two classes:

0 = below median per capita crime rate
1 = above median per capita crime rate

The data, excluding the class, was scaled. The data was randomly split into training and test with a ratio of 4:1 respectively.

PCA was performed on the full dataset to explore any relationships. The data appeared to be adequately explained by 2 to 4 principle components. However, those principle components were a linear combination of a number of predictors. As such they were not helpful for inference. This may also infer that the data has complex structural relationships.

1.1 Summary

Describe your findings.

Table: Summary of Results

Method	Parameter	Training Error	Test Error
Logistic Regression	$p = 12$	8.17%	11.76%
LDA	$p = 4$	12.87%	14.71%
kNN	$k = 1$	0%	3.92%

KNN outperformed the other methods in terms of predictive accuracy, as evident in the table above. KNN method, however, provides no interpret-ability.

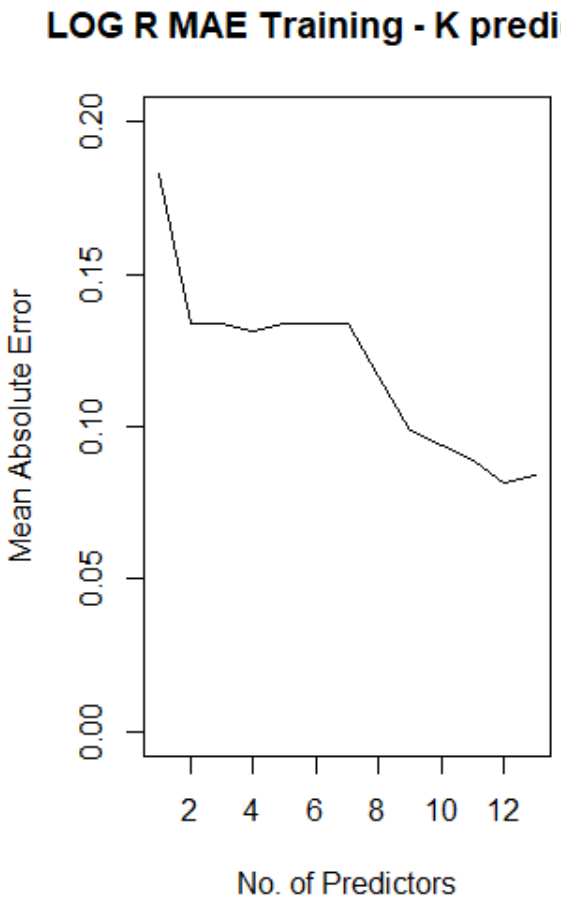
```
Call:
glm(formula = crim ~ ., family = "binomial", data = red.training)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.01342  -0.11393   0.00000   0.00061   2.42709

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.0414     0.8268   3.678 0.000235 ***
zn           -2.8861     1.0390  -2.778 0.005475 **
indus        -0.5157     0.3508  -1.470 0.141471
nox           6.7782     1.1225   6.038 1.56e-09 ***
rm           -0.7468     0.5779  -1.292 0.196287
age           0.8181     0.3988   2.052 0.040212 *
dis           2.2683     0.6246   3.632 0.000282 ***
rad           6.7170     1.5851   4.238 2.26e-05 ***
tax          -1.0918     0.4970  -2.197 0.028045 *
ptratio       0.7026     0.2974   2.363 0.018139 *
black        -1.0651     0.5353  -1.989 0.046649 *
lstat         0.7855     0.4151   1.892 0.058468 .
medv          2.3395     0.7762   3.014 0.002579 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The logistic regression model was the next best in terms of predictive accuracy. However, unlike the other two methods, the logistic regression model has good inter-pretability. The significant predictors are shown in the model summary above.

1.2 Part A

Logistic Regression

Logistic Regression was performed, and a model fit, for the best subset for every **k** number of predictors. The Mean Absolute Error was calculated for each best subset model.



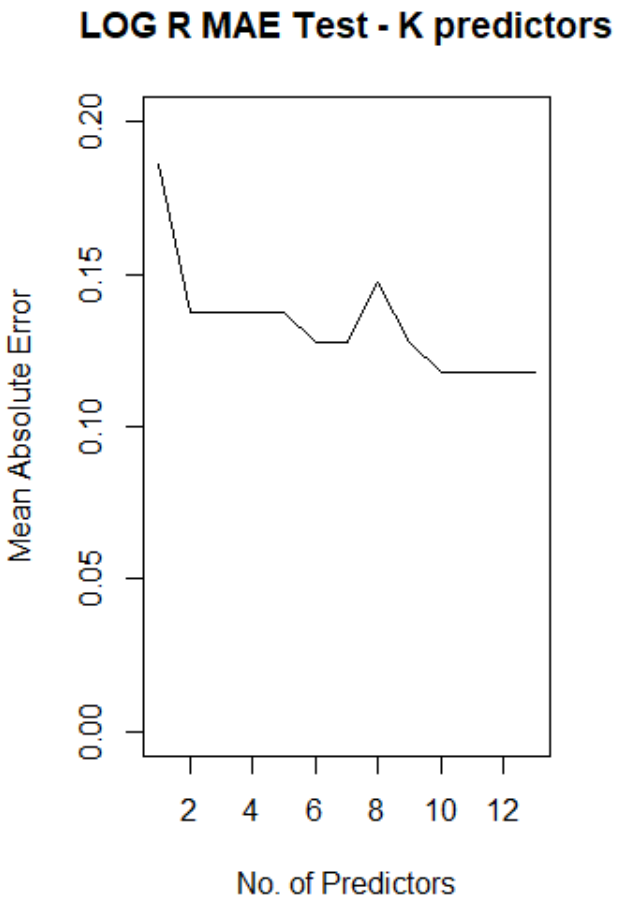


Table: Confusion Matrix Logistic Regression Training Prediction

	Actual BELOW	Actual ABOVE
Predicted BELOW	TN = 186	FN = 18
Predicted ABOVE	FN = 15	TP = 185

Table: Confusion Matrix Logistic Regression Test Prediction

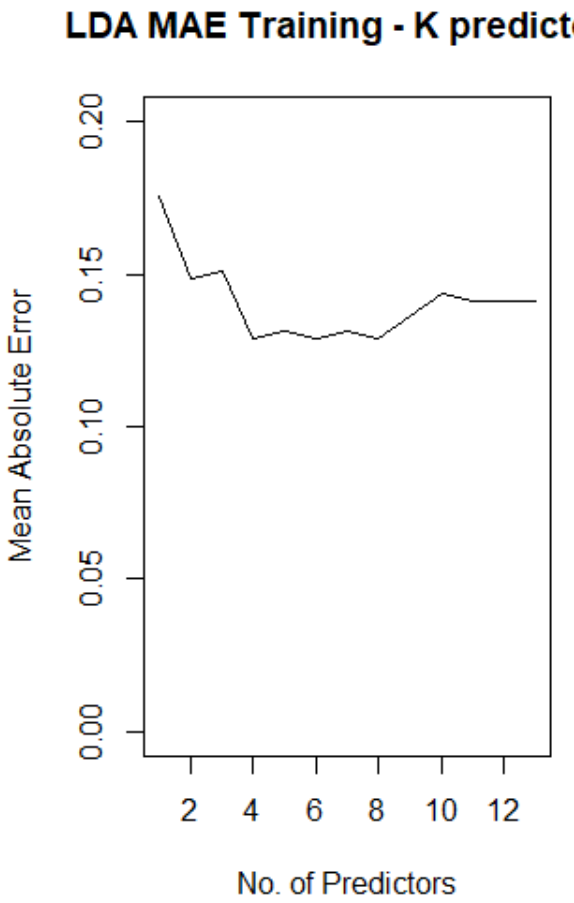
	Actual BELOW	Actual ABOVE
Predicted BELOW	TN = 45	FN = 5
Predicted ABOVE	FN = 7	TP = 45

BELOW = below median per capita crime rate
ABOVE = above median per capita crime rate

1.3 Part B

LDA

LDA was performed, and a model fit, for the best subset for every **k** number of predictors. The Mean Absolute Error was calculated for each best subset model.



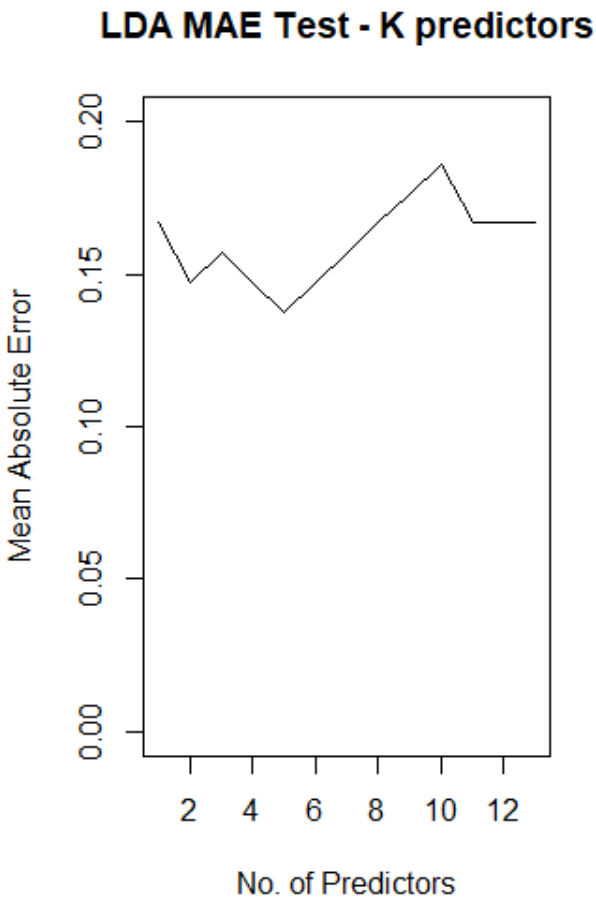


Table: Confusion Matrix LDA Training Prediction

	Actual BELOW	Actual ABOVE
Predicted BELOW	TN = 195	FN = 46
Predicted ABOVE	FN = 6	TP = 157

Table: Confusion Matrix LDA Test Prediction

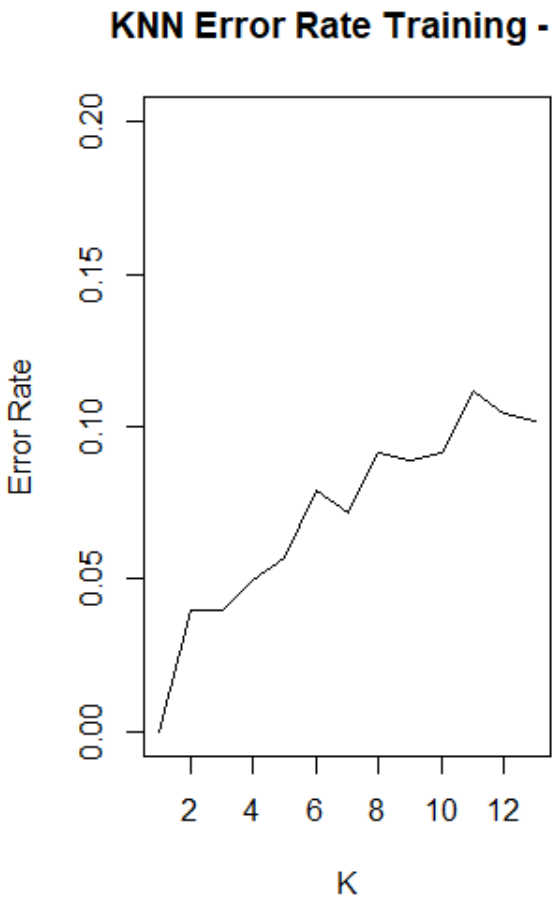
	Actual BELOW	Actual ABOVE
Predicted BELOW	TN = 51	FN = 14
Predicted ABOVE	FN = 1	TP = 36

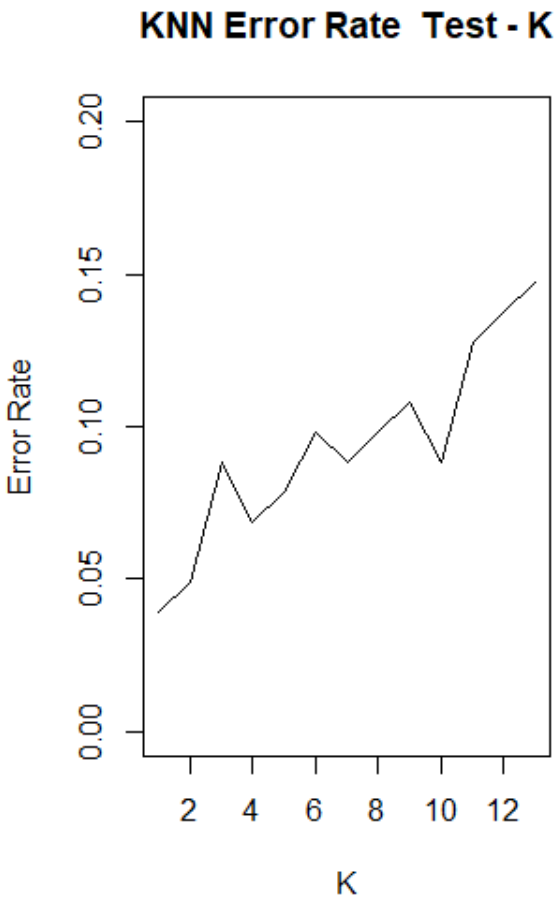
BELOW = below median per capita crime rate
ABOVE = above median per capita crime rate

1.4 Part C

KNN

KNN was performed, and a model fit, for a range of **k** nearest neighbours. The the mis-classifciation rate was calculated for each model.





$K = 1$, was selected as the complexity value.

Table: Confusion Matrix KNN Train Prediction

	Actual BELOW	Actual ABOVE
Predicted BELOW	TN = 201	FN = 0
Predicted ABOVE	FN = 0	TP = 203

Table: Confusion Matrix KNN Test Prediction

	Actual BELOW	Actual ABOVE
Predicted BELOW	TN = 52	FN = 4
Predicted ABOVE	FN = 0	TP = 46

BELOW = below median per capita crime rate
ABOVE = above median per capita crime rate

2 Exercise 2

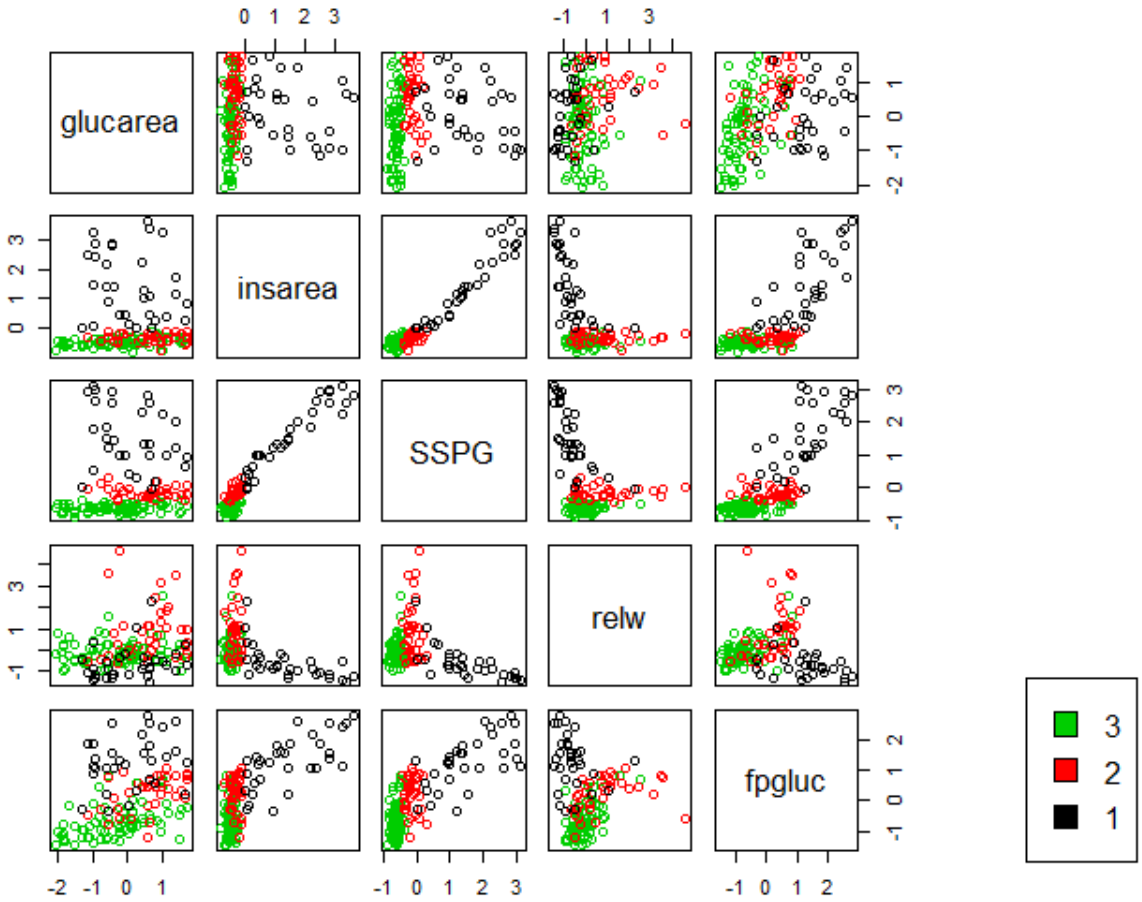
Download the diabetes data set. Disregard the first three columns. The fourth column is the observation number, and the next five columns are the variables (glucose.area, insulin.area, SSPG, relative.weight, and fasting.plasma.glucose). The final column is the class number. Assume the population prior probabilities are estimated using the relative frequencies of the classes in the data.

The Andrews Diabetes dataset contains 144 data points with 6 predictors and 1 target, **class**, with three classes: 1, 2, and 3. The data, excluding the class, was scaled.

2.1 Part 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?

Pairwise Scatter Plot: Andrews Diabetes Dataset



The plot above shows some differences in the scatter between classes. It is evident that the classes may have different covariance matrices. In particular, **Class 1** appears to have a different covariance matrix, to **Class 2** and **Class 3**. **Class 2** and **Class 3** appear to be related, but only moderately.

In addition, Mardia Multivariate Normality Test was performed and to confirm the observations above. The test indicated that data was NOT multivariate normal.

2.2 Part 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?

Table: Confusion Matrix LDA

	Actual 1	Actual 2	Actual 3
Predicted 1	25	6	1
Predicted 2	0	30	6
Predicted 3	0	3	73

Table: Confusion Matrix QDA

	Actual 1	Actual 2	Actual 3
Predicted 1	29	3	0
Predicted 2	3	29	4
Predicted 3	0	4	72

Table: Accuracy Comparison

Method	Accuracy
LDA	11.11%
QDA	9.72%

QDA has slightly better accuracy as indicated in the table above. LDA, however, was slightly better at predicting classes 2 and 3. Unfortunately, I was unable to find the meaning of the classes so I am unable to definitively assess which model performs better at diagnosing diabetes.

2.3 Part 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?

Both LDA and QDA assign this individual to **Class 1**.

3 Exercise 3

3.1 Part 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.

Posterior Probability: of an observation (X), belongs to a class (G).

The log of the ratio of posterior probabilities are called log-odds, and it gives us the logistic regression curve classifying two classes.

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

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

$$\cdot\cdot$$

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

We can obtain the following using the above formulas

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

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

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

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

Hence, it can be shown $\sum_{k=1}^{K} Pr(G=k|X=x) = 1$

Under the assumptions of the Logistic Regression method, the sum of posterior probabilities of all classes should equal 1. This assumption holds true generalizing for values from 1 to K.

3.2 Part B

Using a little bit of algebra, show that the logistic function representation and the logit representation for the logistic regression model are equivalent.

For the Logistic function

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

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

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

Dividing the above equations we can show that the logistic function is equivalent to the following

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