EAS506LEC000 : STATISTICAL DATA MINING

HW ASSIGNMENT – 5

SUBMITTED BY : DITHYA SRIDHARAN

CLASS NUMBER : 48

QUESTION 1

PROBLEM STATEMENT

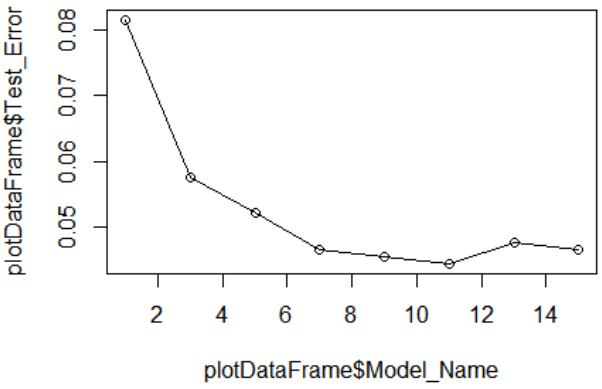
Fit a series of random-forest classifiers to the SPAM
data, to explore the sensitivity to m (the number of randomly selected inputs for each tree). Plot both the OOB error as well as the test error against a suitably chosen range of values for m.

1. DATA SUMMARY:

This is a multivariate data set containing 4601 observations with 58 attributes. The last column tells us if the email is spam or not.

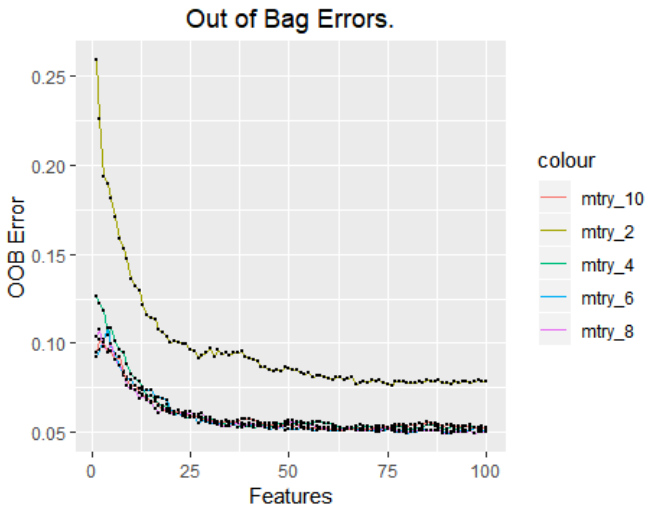
2. ESTIMATING OOB FOR DIFFERENT VALUES OF M:

Random Forest model was applied on the data set and the misclassification error was obtained. The OOB for different values of m were calculated. The m values were taken to be : 2,4,6,8,10,12,14. The OOB estimate came out to be 4.9% for each value. The errors were plotted and the following plot was obtained



3.PLOTTING OOB ESTIMATES

The OOB estimates and the test errors for different values of m = 2,4,6,8,10,12,14 were plotted as follows:



4.CONCLUSION

We can see that for as the value of m increases, the test error decreases. The test error is maximum when m=2 and is found to be less for higher values of m, especially when m =10

QUESTION 2

PROBLEM STATEMENT

Fit a neural network to the spam data of Section 9.1.2. The data is available through the package "ElemStatLearn". Use cross-validation or the hold out method to determine the number of neurons to use in the layer. Compare your results to those for the additive model given in the chapter. When making the comparison, consider both the classification performance and interpretability of the final model.

1.DATA SUMMARY

This is a multivariate data set containing 4601 observations with 58 attributes. The last column tells us if the email is spam or not.

2. CROSSVALIDATING

Cross validation was performed on the neural to find the number of hidden layers. The test and train error vectors were initialized and scaled to form neural network.

3.FITTING THE NEURAL NETWORK MODEL

The neural network model was fit and the test errors were calculated. They were obtained as follows : 0.0387622

The model was fit after crossvalidating with a default of number of neurons as 5. Later the model was fit and predicted.

The number of hidden layers was found using the which(min(test.error) == test.error) command. The number of hidden layers were found to be 11.

4. CONCLUSION

It can be seen that the test error of 0.0387622 which is much better than what was observed in the additive methods where the test errors were 0.058, 0.067 and so on for different values of m. Hence the neural networks makes for better classification performance since it has lesser mis-classification error. However the additive models were more interpretable. The interpretability of the neural networks is still questioned.

QUESTION 3

1.PROBLEM STATEMENT

Take any classification data set and divide it up into a learning set and a test set. Change the value of one observation on one input variable in the learning set so that the value is now a univariate outlier. Fit separate single-hidden-layer neural networks to the original learning-set data and to the learning-set data with the outlier. Use cross-validation or the hold out method to determine the number of neurons to use in the layer. Comment on the effect of the outlier on the fit and on its effect on classifying the test set. Shrink the value of that outlier toward its original value and evaluate when the effect of the outlier on the fit vanishes. How far away must the outlier move from its original value that significant changes to the network coefficient estimates occur?

2.DATA SUMMARY

Spam is a multivariate data set containing 4601 observations with 58 attributes. The last column tells us if the email is spam or not.

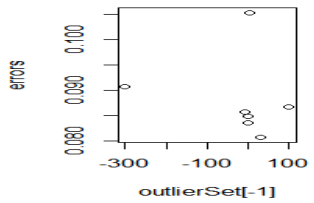
3.CHANGING THE OBSERVATION VALUES TO FORM UNIVARIATE OUTLIER:

The [1,4] observation in the spam data set was changed to form a univariate outlier to the following values : 300, 100, 30, 3, 0.1, -300, -10, -1.

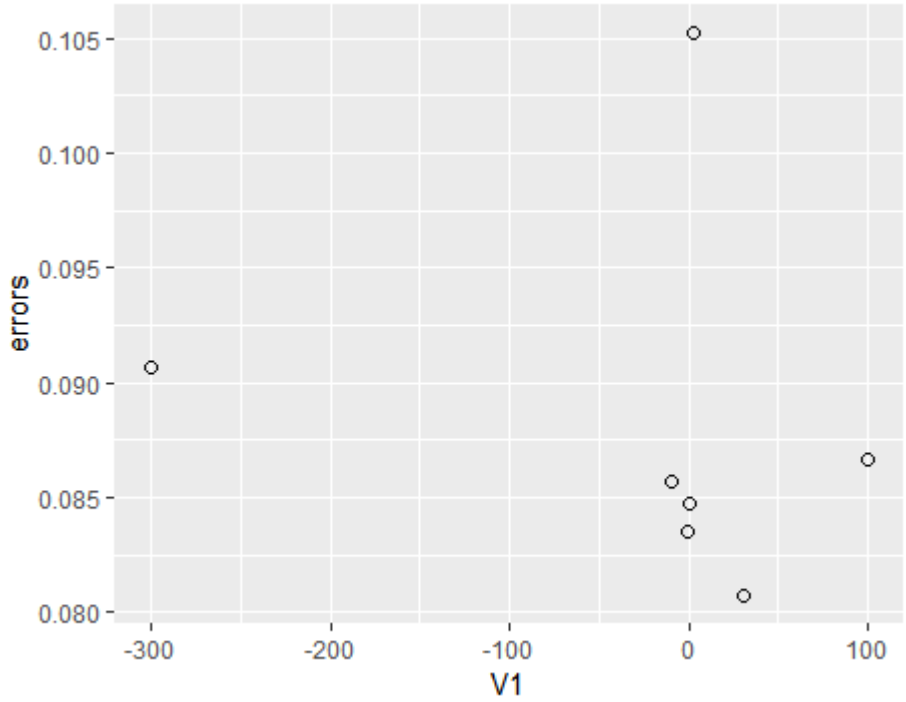
4. FITTING NEURAL MODEL :

A neural model was fit for each change in the observation [1,4] for the different values in the outlier data set. The test errors for the predicted values was calculated for each outlier set.

5.PLOTTING OUTLIER SET ERRORS :



We can see that when the observation values were changed to 30,10,0.1 they give less test error rates. While 0, -300 give massive test error rates.



6.CROSSVALIDATING AND ESTIMATING TEST ERROR

Neural networks have increasing complexity penalization so it is essential to perform cross validation for test errors. The data was scaled and model was fit after cross validation. It was found that the test error to be [0.06188925, 0.05863192, 0.05863192, 0.0576432, 0.0554673, 0.066798, 0.0384577, 0.0456666]. This shows the importance of cross validation.

7.CONCLUSION

Without outliers, the MSE value was recorded to be 0.048. Results shows that as outliers goes from (300, 100, 30, 3, 0.1, -300, -10, -1) the test errors were found to be 0.06188925, 0.05863192, 0.05863192, 0.0576432, 0.0554673, 0.066798, 0.0484577, 0.0456666. This indicates a decrease in modelling accuracy. The values closer to 30,3,0.1,-1,-10 give a value closer to original. Larger values are way off affecting the model accuracy.

QUESTION 4

PROBLEM STATEMENT

This problem involves the OJ data set in the
ISLR package. We are interested in the prediction of "Purchase". Divide the data into test and training.
(A) Fit a support vector classifier with varying cost parameters over the range [0.01, 10]. Plot the training and test error across this spectrum of cost parameters, and determine the optimal cost.
(B) Repeat the exercise in (A) for a support vector machine with a radial kernel. (Use the default parameter for gamma). Repeat the exercise again for a support vector machine with a polynomial kernel of degree=2. Reflect on the performance of the SVM with different kernels, and the support vector classifier, i.e., SVM with a linear kernel.

1. DATA SUMMARY:

The data contains 1070 purchases where the customer either purchased Citrus Hill or Minute Maid Orange Juice. A number of characteristics of the customer and product are recorded.

2.SUPPORT VECTOR CLASSIFIER [0.01,10]

SVM classifier was fit for cost from 0.01 to 10. The summary was obtained for each cost to s ummarise statistics and the results obtained. The training and the test error came out to be 0.1 60046729 and 0.1682242991 respectively.

3.SVM WITH RADIAL KERNAL

For a default value of gamma, support vector machine with radial kernel was fitted on the dat a. Truth matrix for training and test were as follows : [489 312] and [164 105] respectively.

4.SVM WITH POLYNOMIAL KERNAL

For a default value of gamma degree is set to 0, support vector machine with polynomial was fitted on the data. Truth matrix for training and test were as follows : [489 312] and [164 105].

5.CONCLUSION

We can see that radial fit is not good enough. We get better results for linear approach.

QUESTION 5

PROBLEM STATEMENT

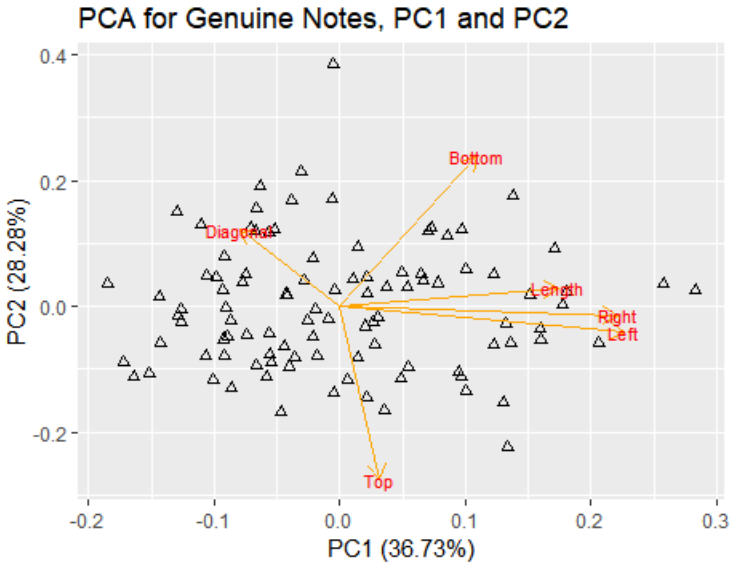
Access the SwissBankNotes data (posted with assignment). The data consists of six variables measured on 200 old Swiss 1,000 -franc bank notes. The first 100 are genuine and the second 100 are counterfeit. The six variables are length of the bank note, height of the bank note, measured on the left, height of the bank note measured on the right, distance of the inner frame to the lower border, distance of inner frame to upper border, and length of the diagonal. Carry out a PCA of the 100 genuine bank notes, of the 100 counterfeit bank notes, and all of the 200 bank notes combined. Do you notice any differences in the results? Show all work in the selection of Principal Components, including diagnostic plots

1. DATA SUMMARY:

The data set contains six measurements made on 100 genuine and 100 counterfeit old-Swiss 1000-franc bank notes.

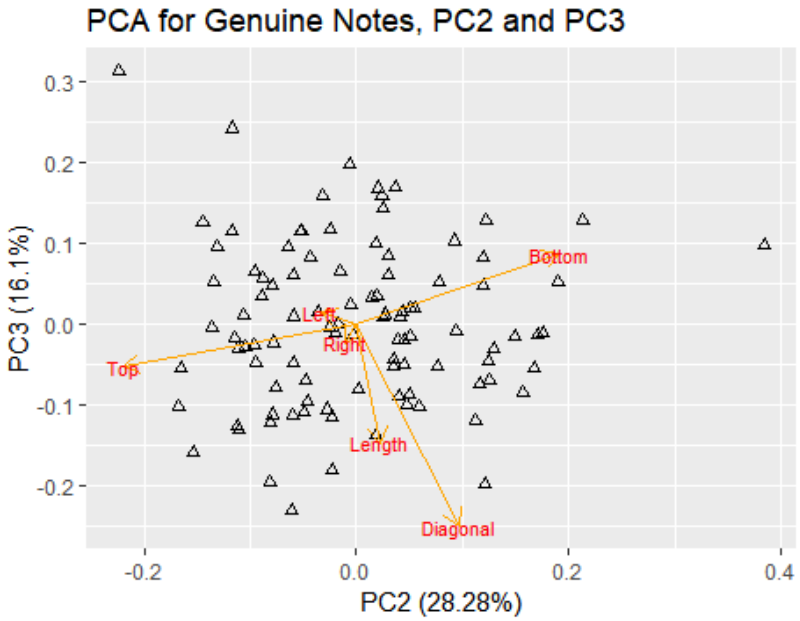
2.PCA FOR GENUINE NOTES

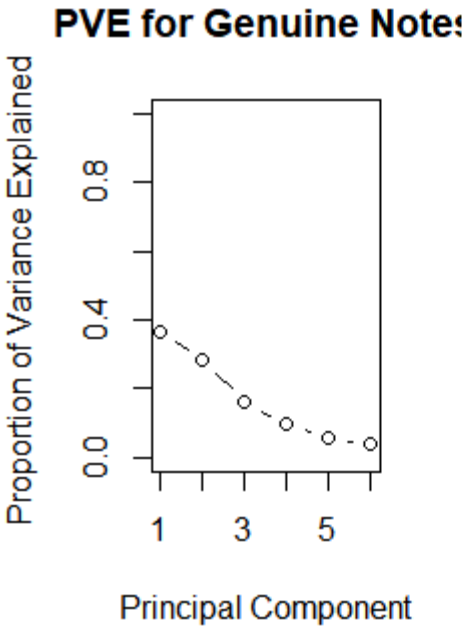
We will first perform PCA on all the Swiss banknote data. We will be using the prcomp function in R to build the PCA model.

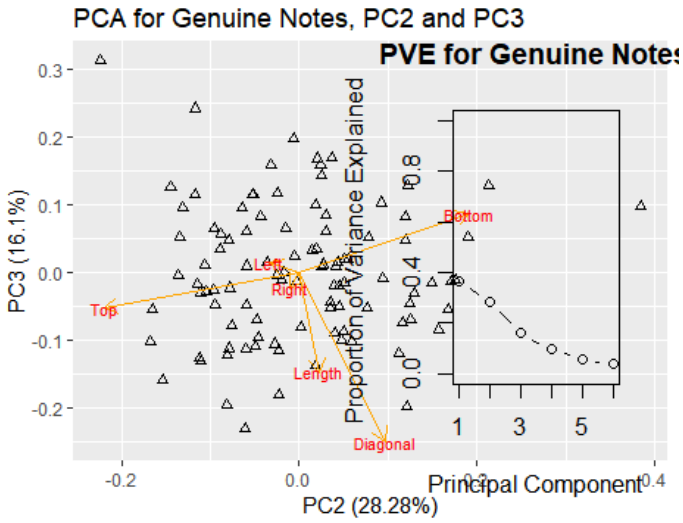


The first two pc's provide a very good approximation of the 6-dimensional data: Therefore the visualization of the sample on the cartesian plane of the first two pc's is a reliable picture of the original 6-dimensional configuration.

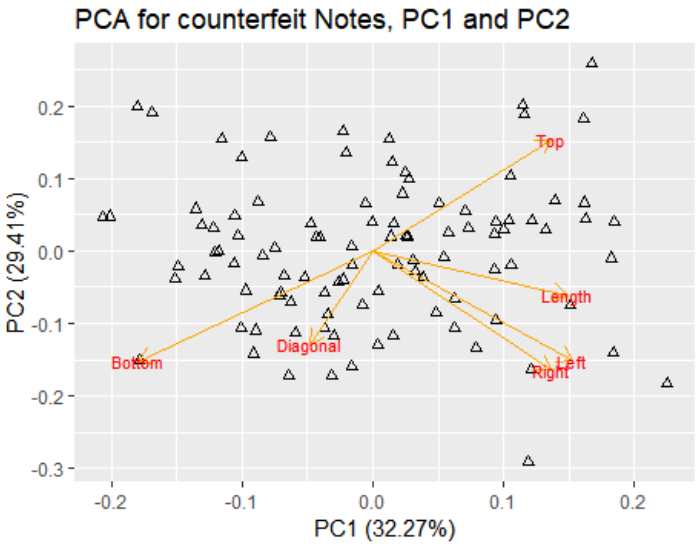
The criterion for PCA is a high variance in the principal components. The question involves "how much the PCs explain the variation within the whole data."

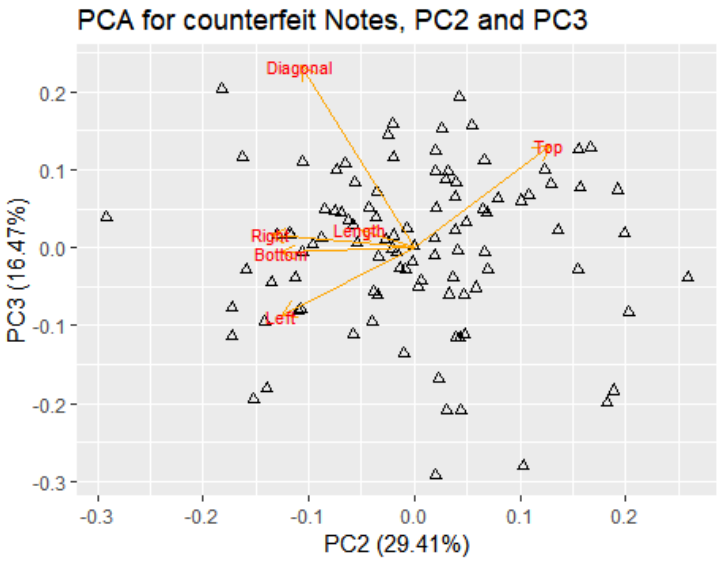


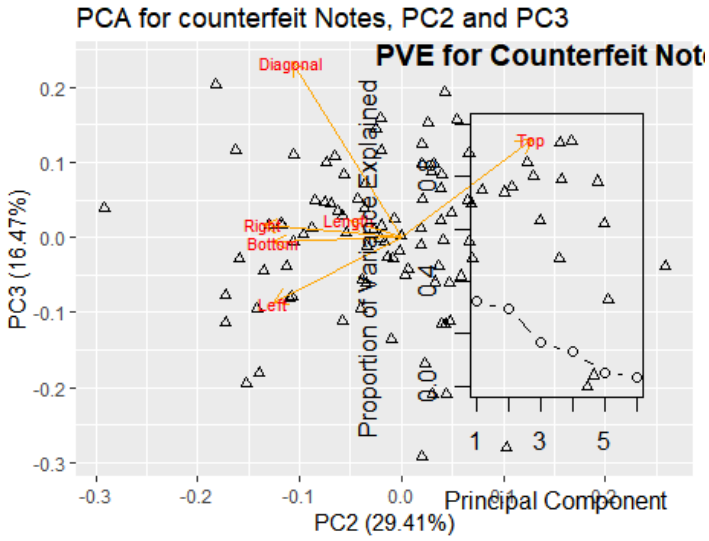


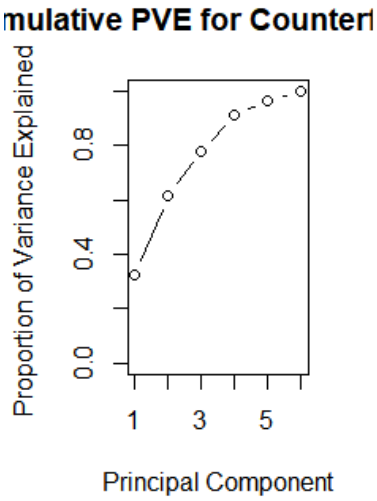


4.PCA FOR COUNTERFEIT NOTES


PCA for counterfeit Notes, PC1 and PC2
Top
Length
Left
Right
Diagonal
Bottom
PC2 (29.41%)
PC1 (32.27%)
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.2
-0.1
0.0
0.1
0.2


PCA for counterfeit Notes, PC2 and PC3
Diagonal
Top
Right
Length
Bottom
Left
PC3 (16.47%)
PC2 (29.41%)
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.3
-0.2
-0.1
0.0
0.1
0.2





5.CONCLUSION

- An important task in multidimensional data analysis is reduction in complexity, complexity reduction can be achieved either by reduction of variables or by reduction of objects. Here we consider variables to be numerical in nature.
- A first, and obvious, method for simplification of features is to select a subset of features able to retain the desired information. A second, more general, method is to look {transformations} of the observed features able to retain the desired information.

- Principal component analysis (PCA) belongs to this second family of methods and is a typical step when trying to understand the structure of multidimensional numerical data.
- If we want to preserve the main information given by the observed features, it is clear that complexity reduction is only possible when there is some \textit{redundancy} in the data
- PCA is not scale invariant.