Assignment 5

Freya D Mello

December 2018

1 Question 1

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.

1.1 Setting the data

As the spam data was not spread correctly, the set was shuffled completely and then split into test ad train data with train having 55% of the data.

1.2 Inference

On performing Bagging over all the predictors at each split, test errors and OOB errors were computed. The minimum test error was 12 and minimum OOB error is 0.055 at the 5th m value.

2 Question 2

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.

2.1 Inference

Considering classification performance and interpretability, cross validation helped measure various test errors across different values of hidden neurons. The minimum error thus observed to be 0.06 at the hidden layer number 2. On using various m-try values in the additive models, we infer how neural networks has lesser errors on comparison. 1

3 Question 3

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?

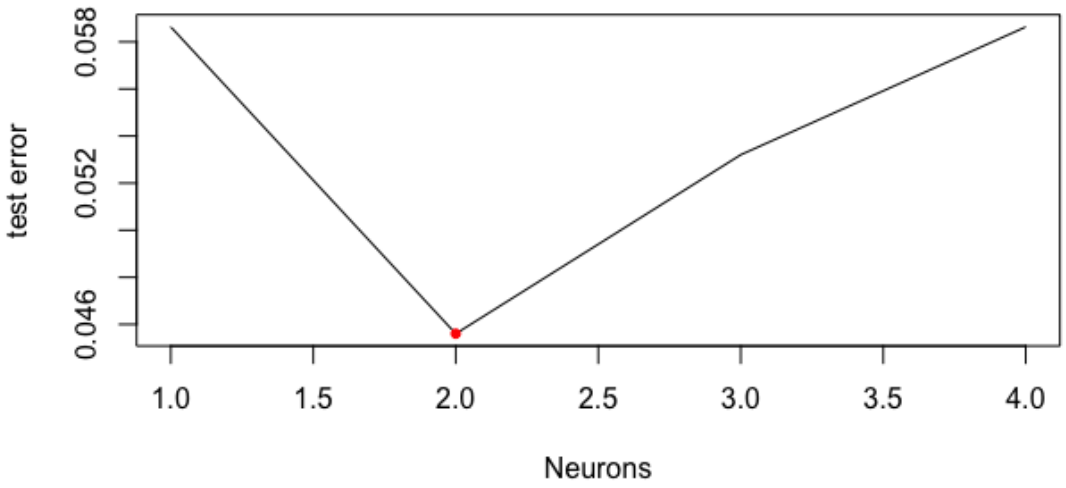

Figure 1: Test Error vs Number of Neurons

3.1 Inference

It shouldn't cause any major difference just by introducing one outlier since any data set has enough number of values, however when there are many outliers then neural network would give bad classification rates since neural network are bad at handling outliers due to its way of training the model. It's effects on the model are directly proportional to the difference between the actual values and the outlier thus introduced.2

4 Question 4

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.

4.1 Setting the data

The data was divided into test and train in a ratio 3:7.

4.2 (a)

The train and test errors were observed for all 3 types of SVM models - Linear Radial and Quad. Linear: Minimum training error was obtained at m=1 which had its train error value at 0.16. The test error stayed at its minima which is 0.16 for a value m = 1. Therefore the optimal cost for the test data is m=1 and for train it is 1. 3 4 5

4.3 (b)

Performance of radial kernel is better than linear and quadratic kernels as it has the least test error of 0.14 followed by quadratic with an error of 0.14 and then linear which had an error of 0.16. The train error though was 0.14 for both the non linear classifier(radial and quadratic), which happens

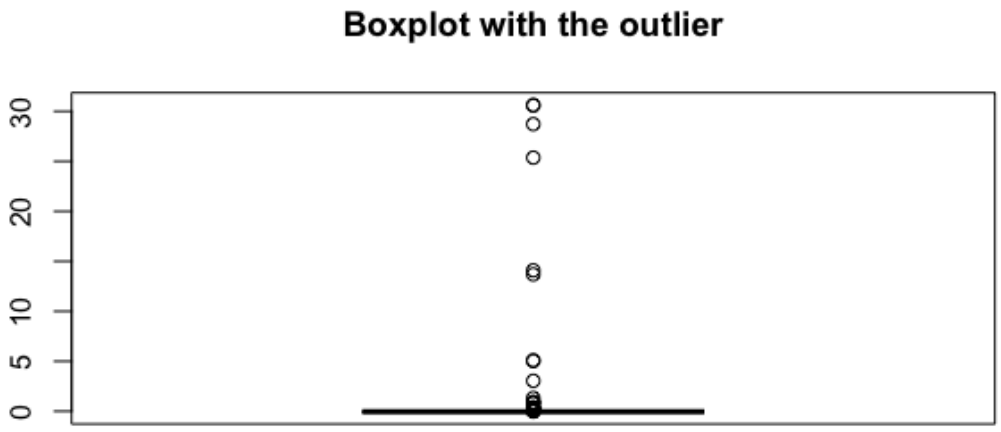

Figure 2: Boxplot with the outlier

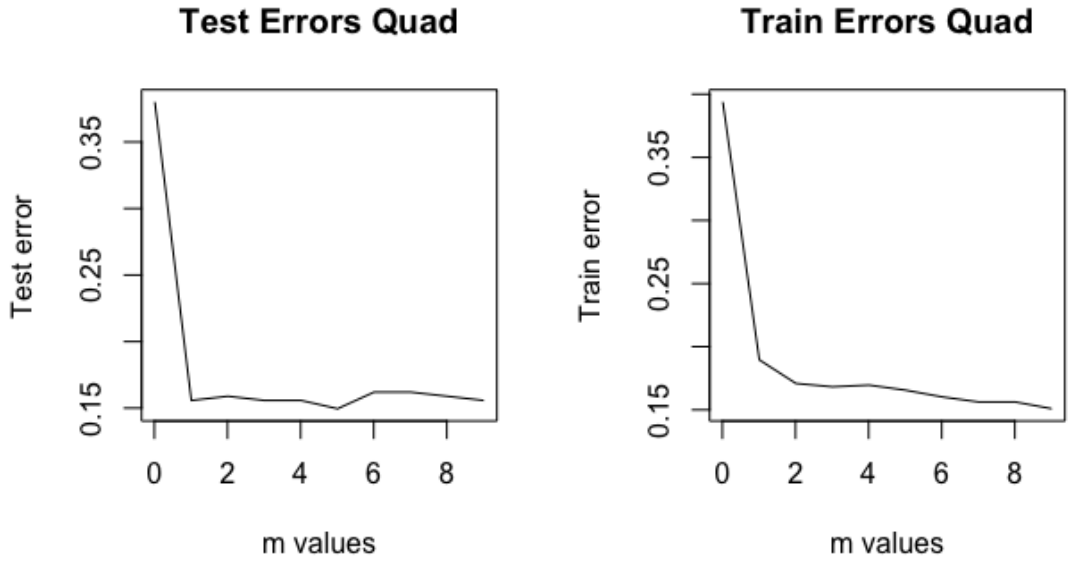

Figure 3: QUAD Test Train Errors vs m values

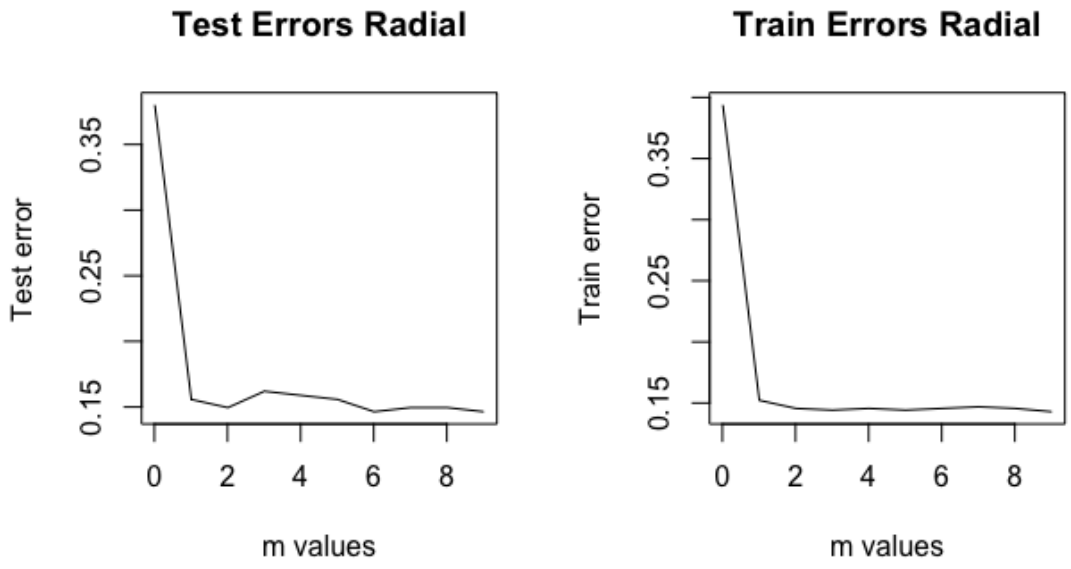

Figure 4: RADIAL Test Train Errors vs m values

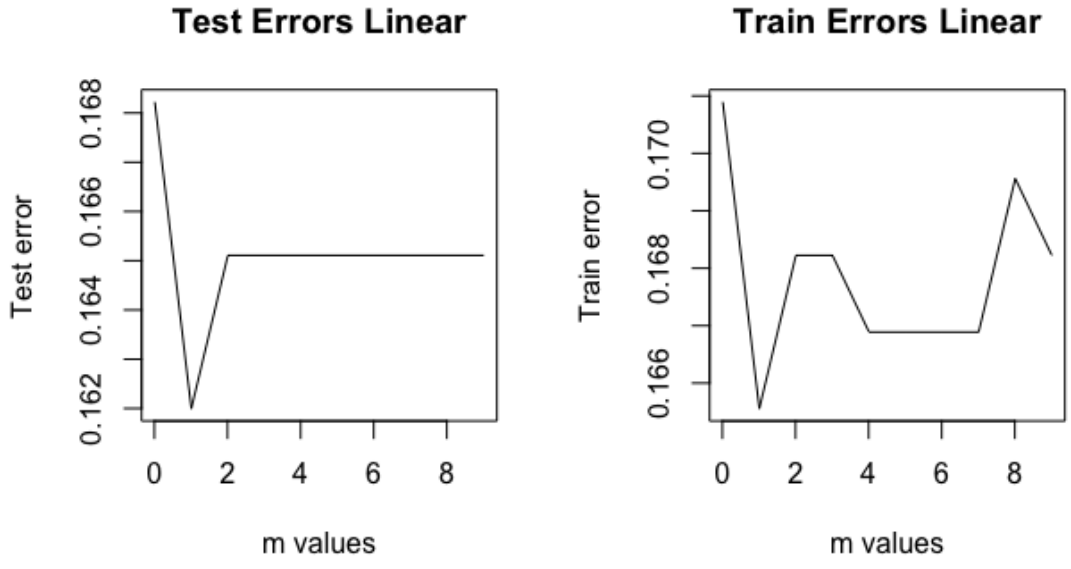

Figure 5: LINEAR Test Train Errors vs m values

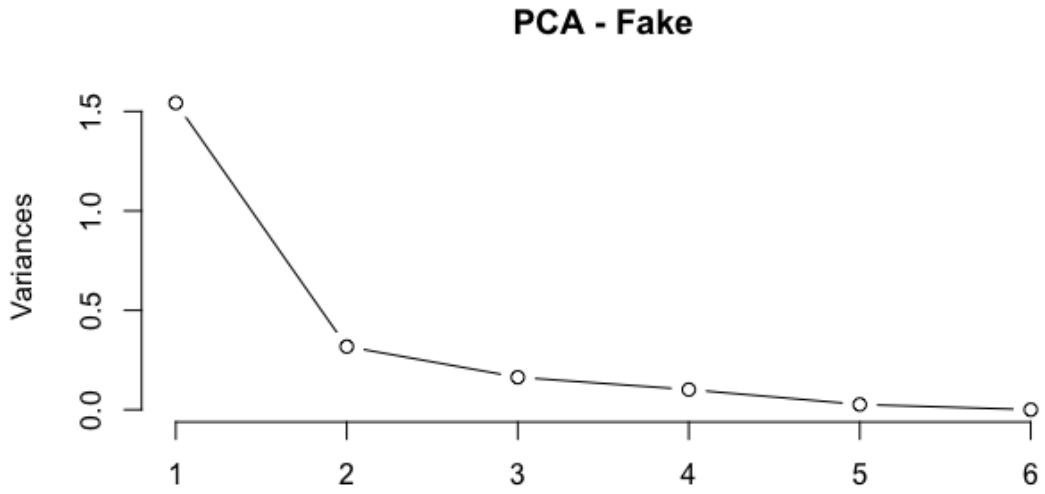

Figure 6: PCA - Fake

to be lower than support vector classifier. The optimal cost for radial kernel was 2.01 for test data and 3.01 and 5.01 for train as well. However, the quadratic kernel has an optimal cost of 9.01 for train and 5.01 for test.

5 Question 5

Access the Swiss Bank Notes 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.

5.1 Inference

PCA was performed to infer that the attributes Top and Bottom were orthogonal to each other and hence have a correlation value of 0 from the 1st and 2nd PC perspective. On observing the genuine and fake notes it can be seen that these two attributes are negatively correlated wrt 1st and 2nd PCs. If we consider fake notes, left and right and length attributes don't have much variation, the ones low on loadings, for the first two Principle components as it was compared to genuine notes where it was infered to contribute equally and correlated. For fake the right left and length attributes are positively correlated and for genuine, the attributes are negatively correlated.
7 6 8

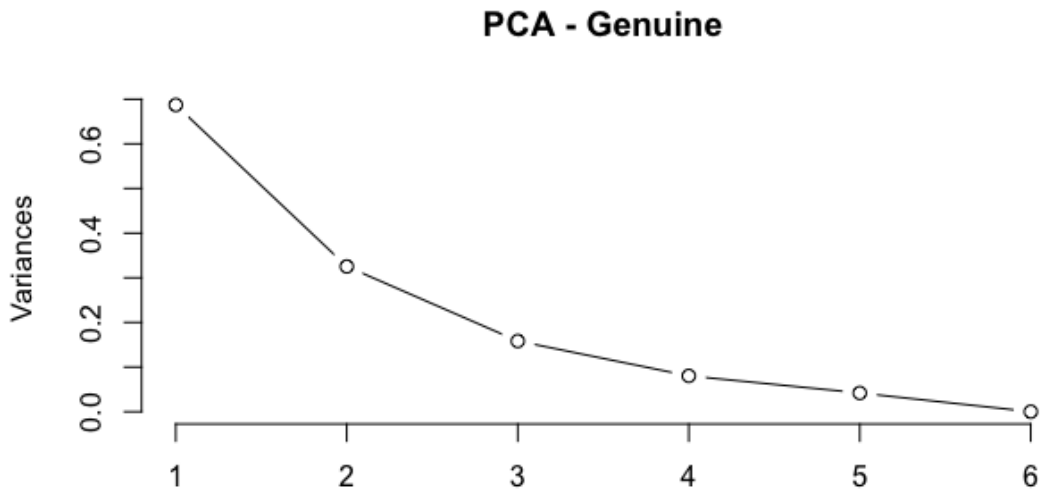

Figure 7: PCA - Genuine

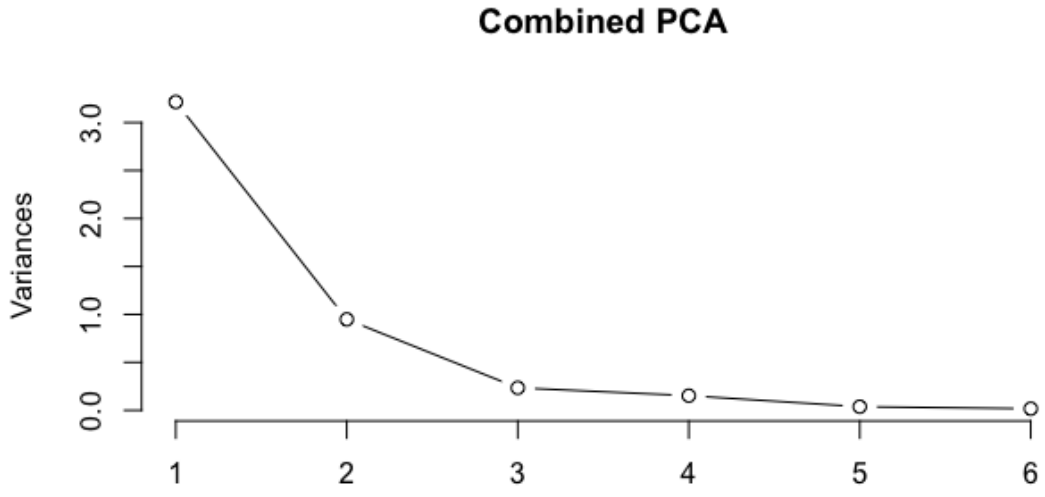

Figure 8: PCA - Combined