EAS 506 Statistical Data Mining — Jaideep Reddy Kommera

Homework 5 — Class no. 28

Question 1:

Introduction: SPAM E-mail Database. A data frame with 4601 observations on the following 58 variables.

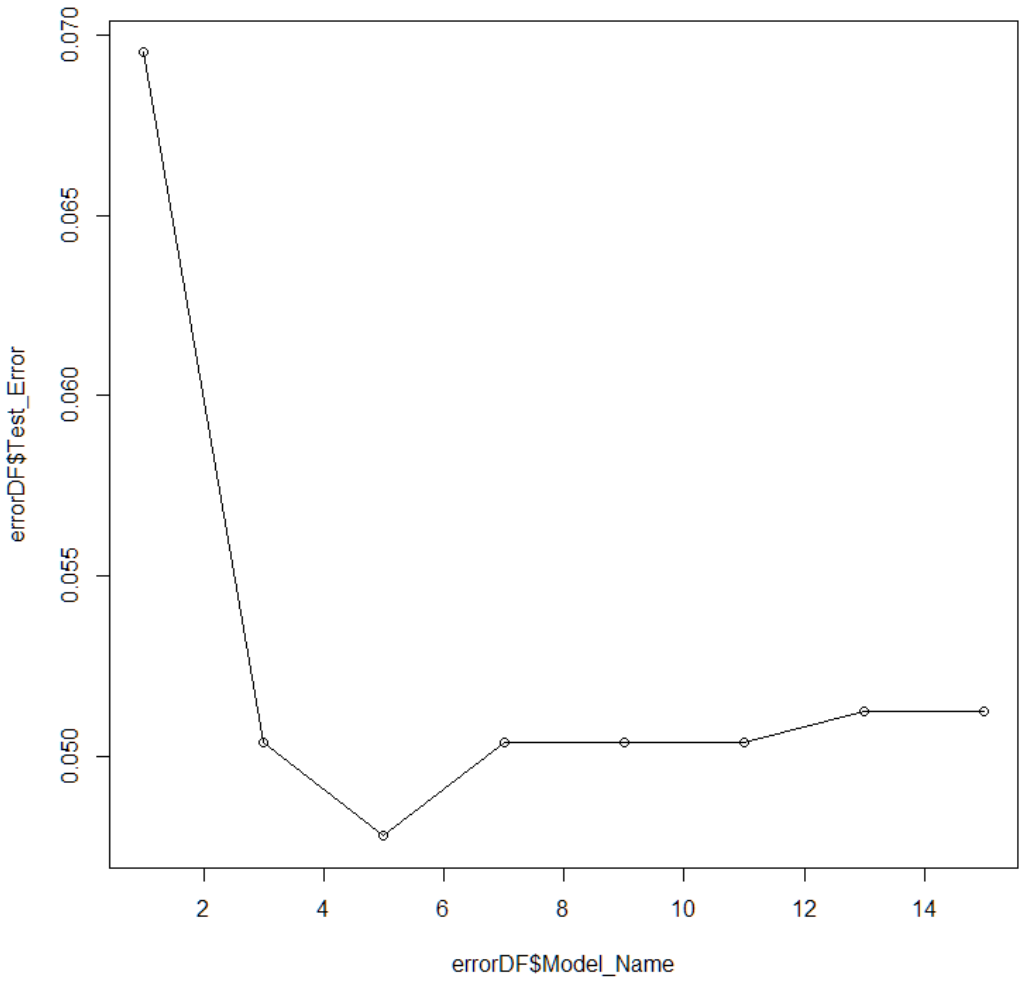
Pre-Processing:

Train-Test Split: 75% Train data

25% Test data

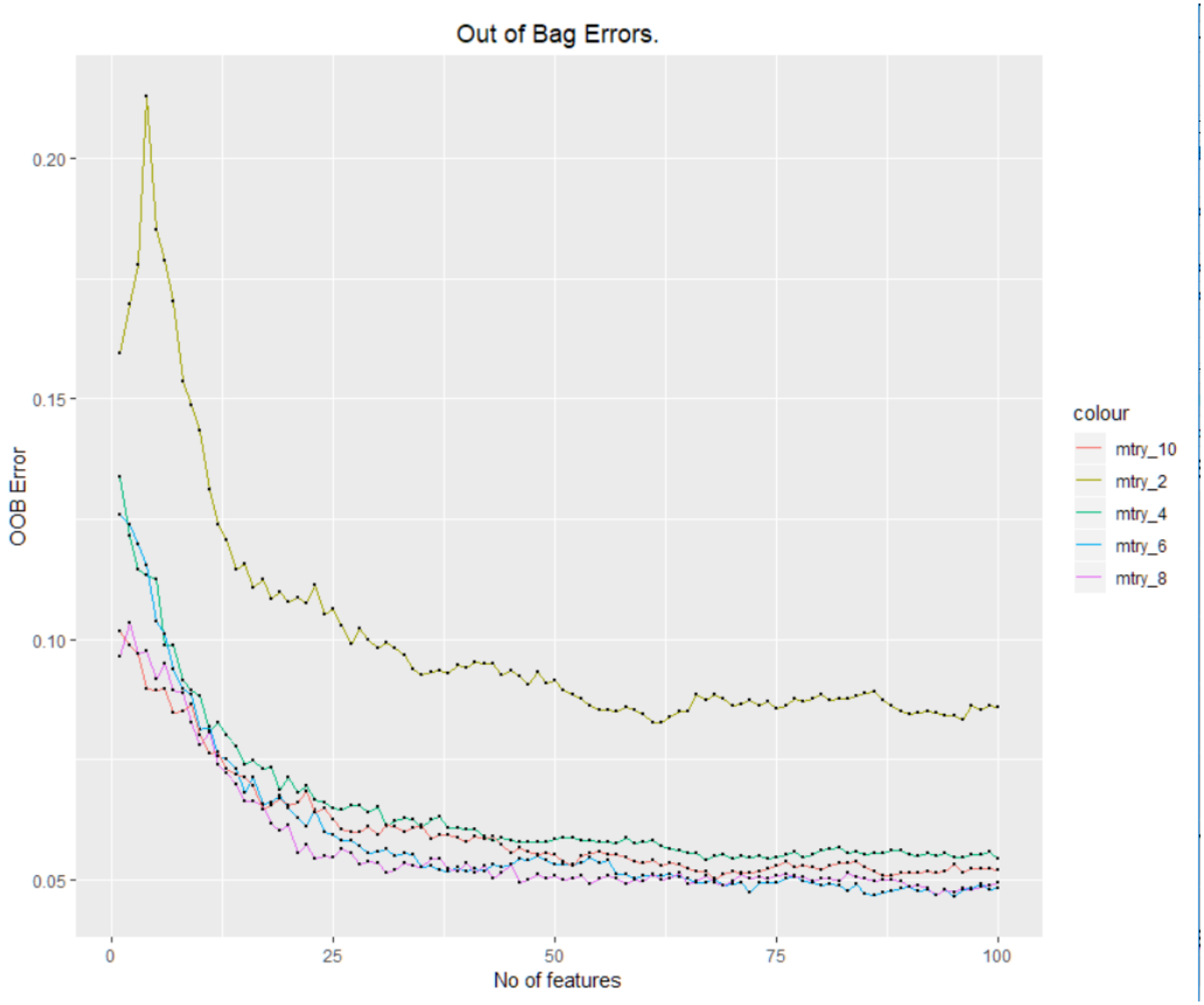
The data has no NA values

Plot for Test Error:



The test error plot for values greater than m values in the range 1 to 15 with step size of 2

The number of trees used for building Random Forest is 100



We can see that error decreases as the number of features increase.

Question 2:

Introduction: SPAM E-mail Database. A data frame with 4601 observations on the following 58 variables.

Pre-Processing:

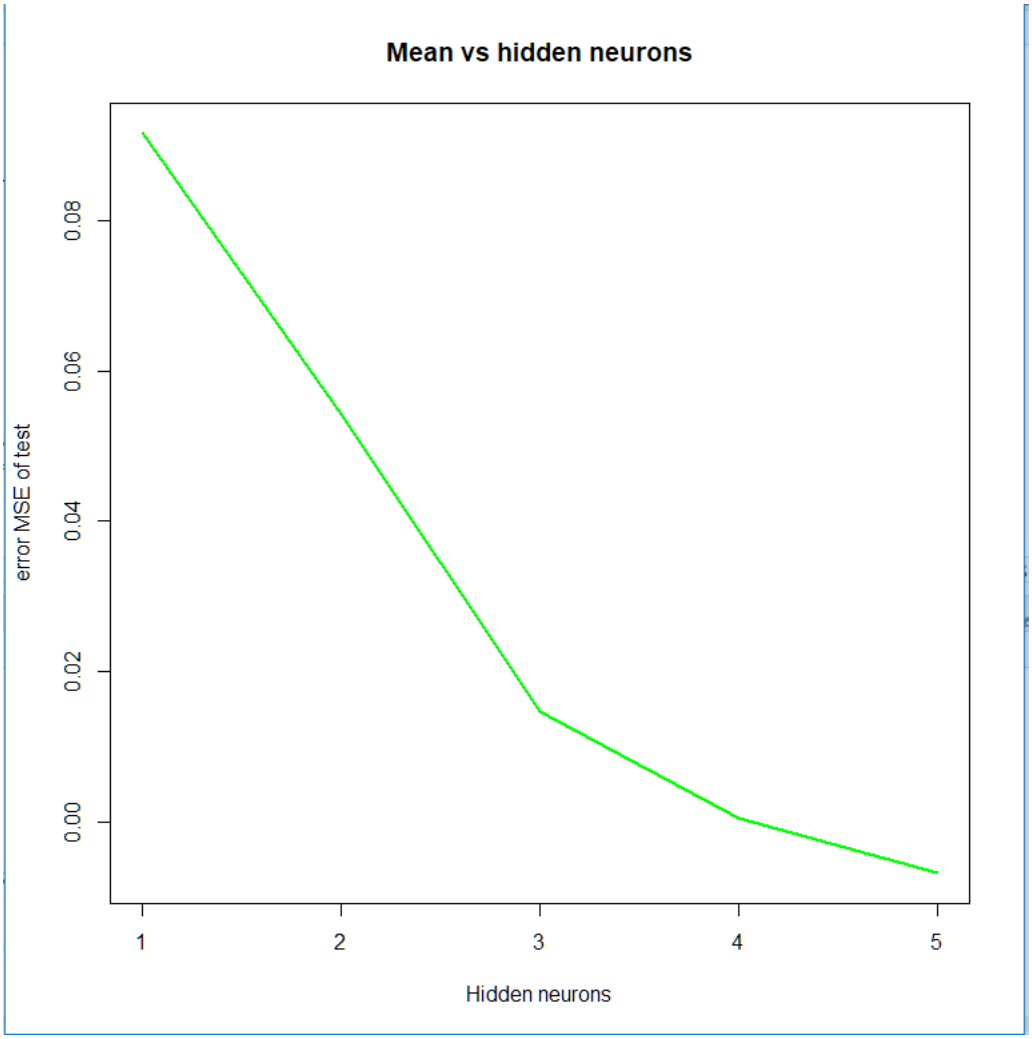
Train-Test Split: 75% Train data

25% Test data

The data has no NA values

Cross validation to find the best number of hidden neurons

Best model is chosen by taking minimum of mean error



No of hidden neurons that give us least error is 5

Error obtained from the model of neural network is 26.40

Error obtained from the additive model is 0.037

Question 3:

Introduction: SPAM E-mail Database. A data frame with 4601 observations on the following 58 variables.

Pre-Processing:

Train-Test Split: 75% Train data

25% Test data

The data has no NA values

Cross validation to find the best number of hidden neurons

Best model is chosen by taking minimum of mean error

The best number of hidden neurons was obtained from the previous question as 5

Data point (4,4) is set as an outlier and the value is gradually reduced.

Here are the results:

```
> outlier_vec
[1]  200.0  100.0   10.0    1.0    0.1 -100.0  -10.0   -1.0
> err_vec
[1] 0.06731813246 0.07491856678 0.06514657980 0.07383279045 0.07274701412 0.07600434311 0.06623235613 0.06080347448
```

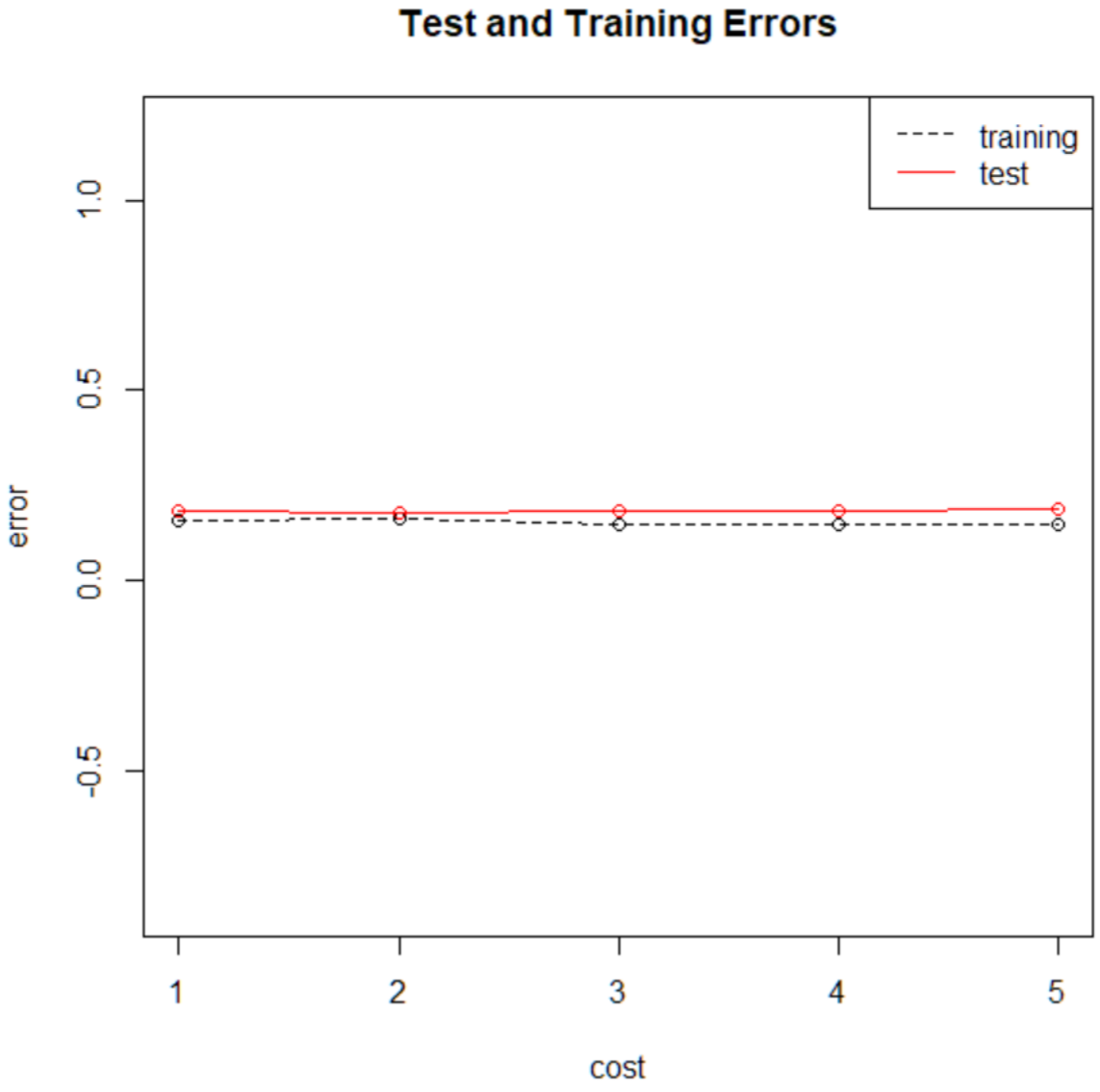
Question 4:

Introduction: 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. The data frame has 1070 observations on the following 18 variables.

Pre-Processing:

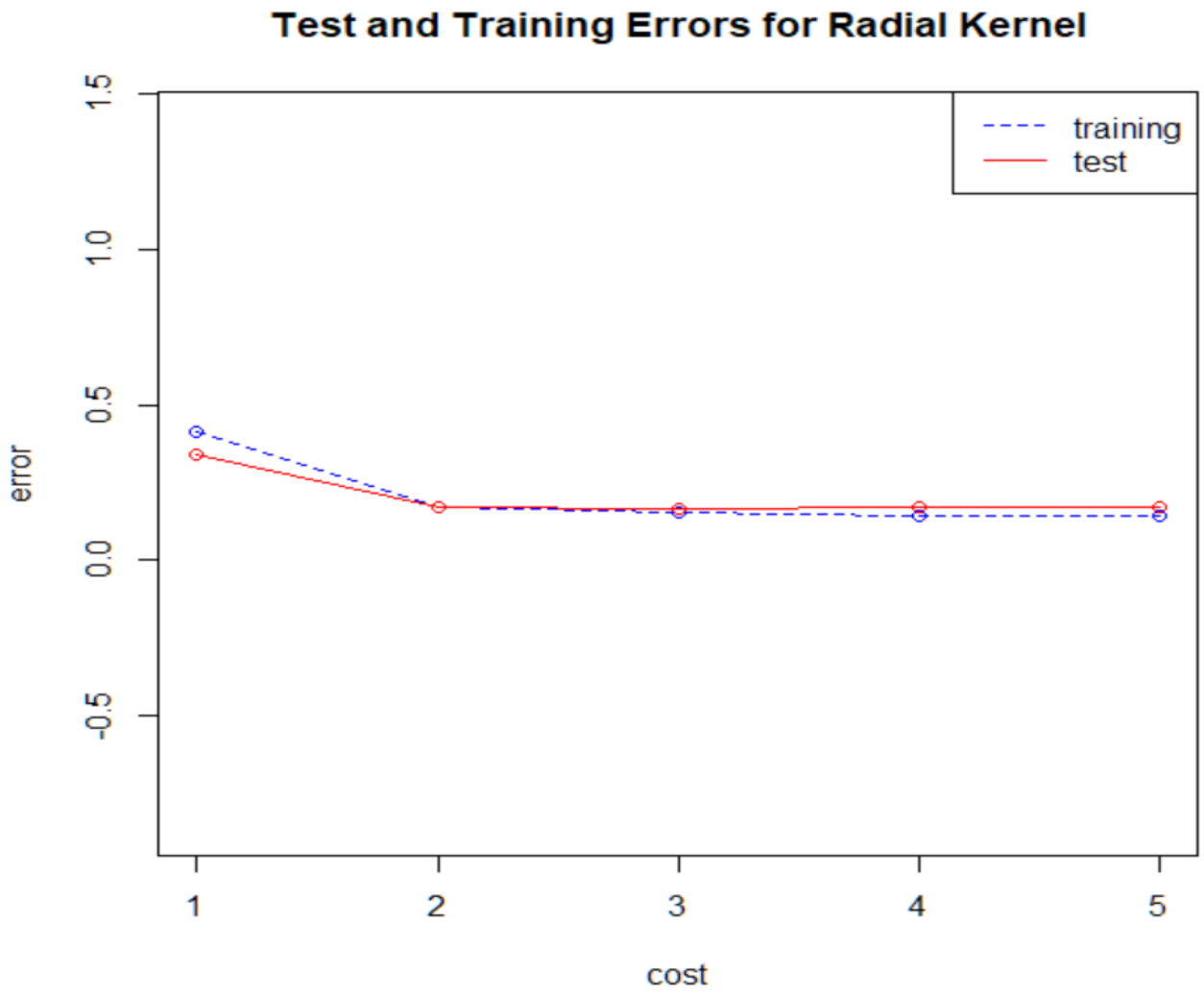
Train-Test Split: 65% Train data

35% Test data



SVM with cost parameters over the range 0.01 to 10.

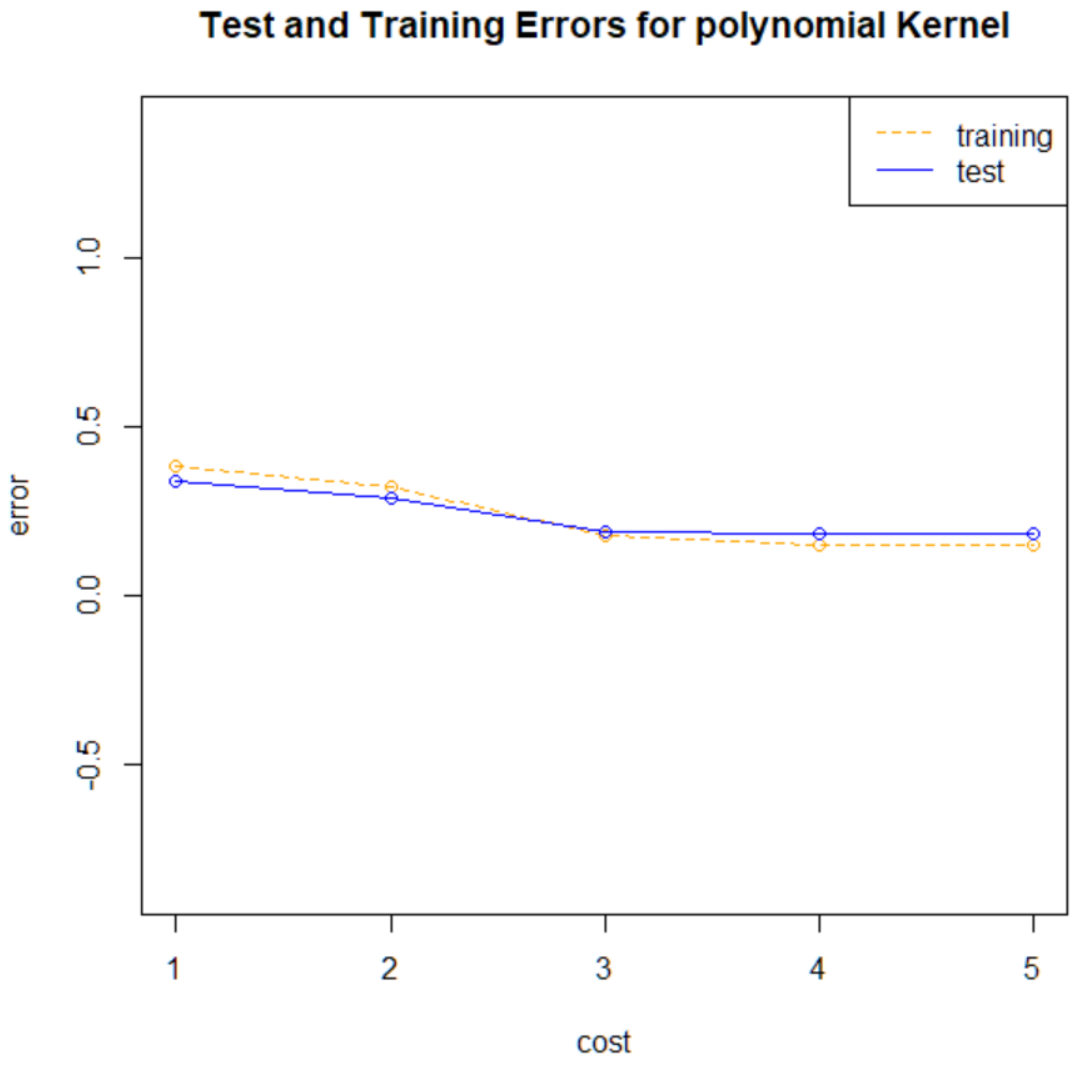
SVM with a Radial Kernel:



Errors:

```
> test_error_radial
[1] 0.3422460 0.1737968 0.1657754 0.1711230 0.1737968
> train_error_radial
[1] 0.4152299 0.1709770 0.1566092 0.1436782 0.1422414
```

SVM with Polynomial Kernel:



Errors:

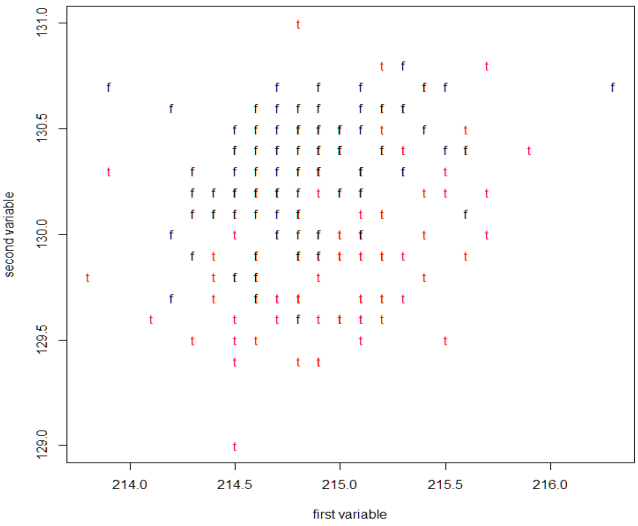
```
> test_error_poly
[1] 0.3368984 0.2914439 0.1898396 0.1818182 0.1818182
> train_error_poly
[1] 0.3864943 0.3204023 0.1767241 0.1494253 0.1494253
```

Question 5:

Introduction: Six measurements made on 100 genuine Swiss banknotes and 100 counterfeit ones.

The first 100 notes are genuine and the next 100 notes are counterfeit.

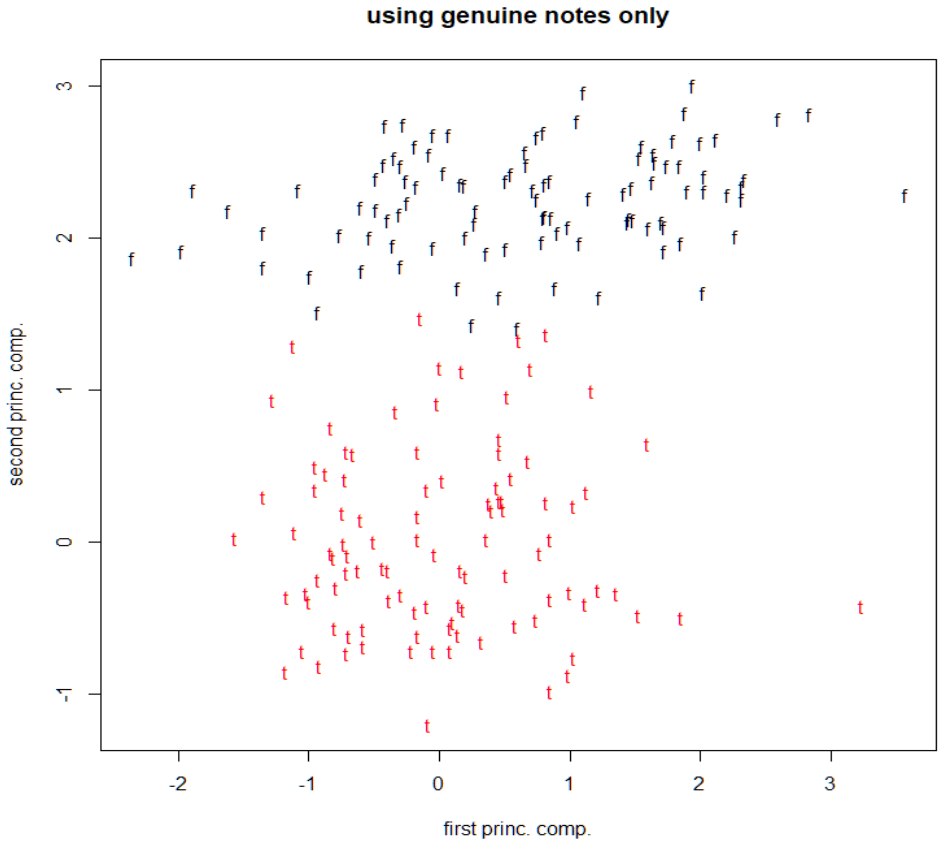
Without PCA:



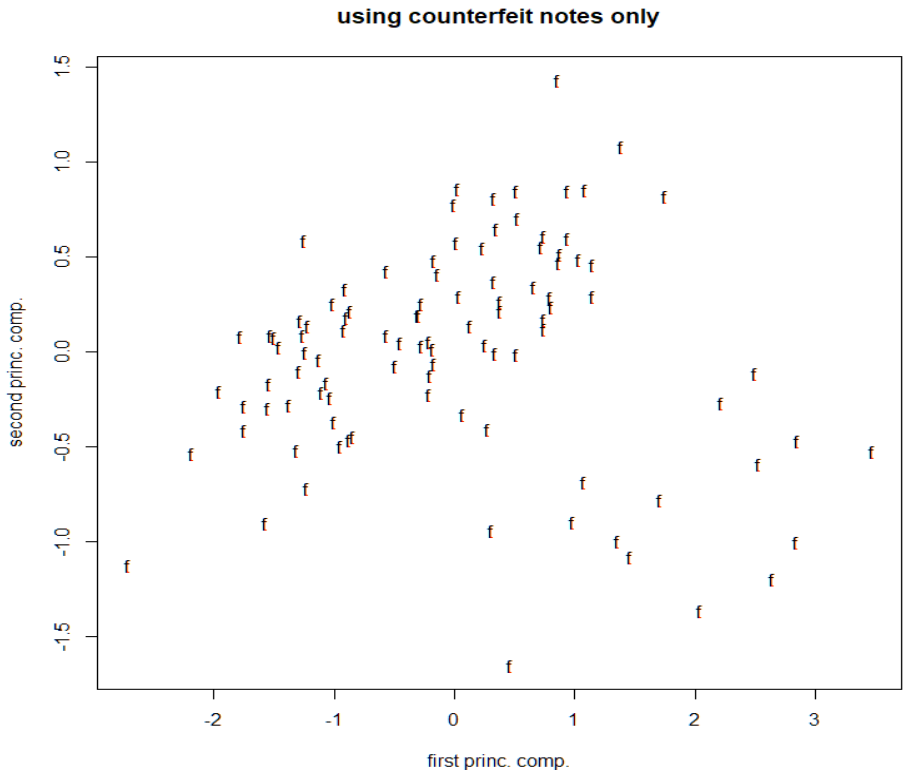
We see that the data cannot be separated at all using the first two variables.

With PCA

a) For 100 genuine notes: PCA using only the data from genuine notes. Computing the whole data in these new coordinates and plotting all the 200 notes. Using only the first two Principle Components.

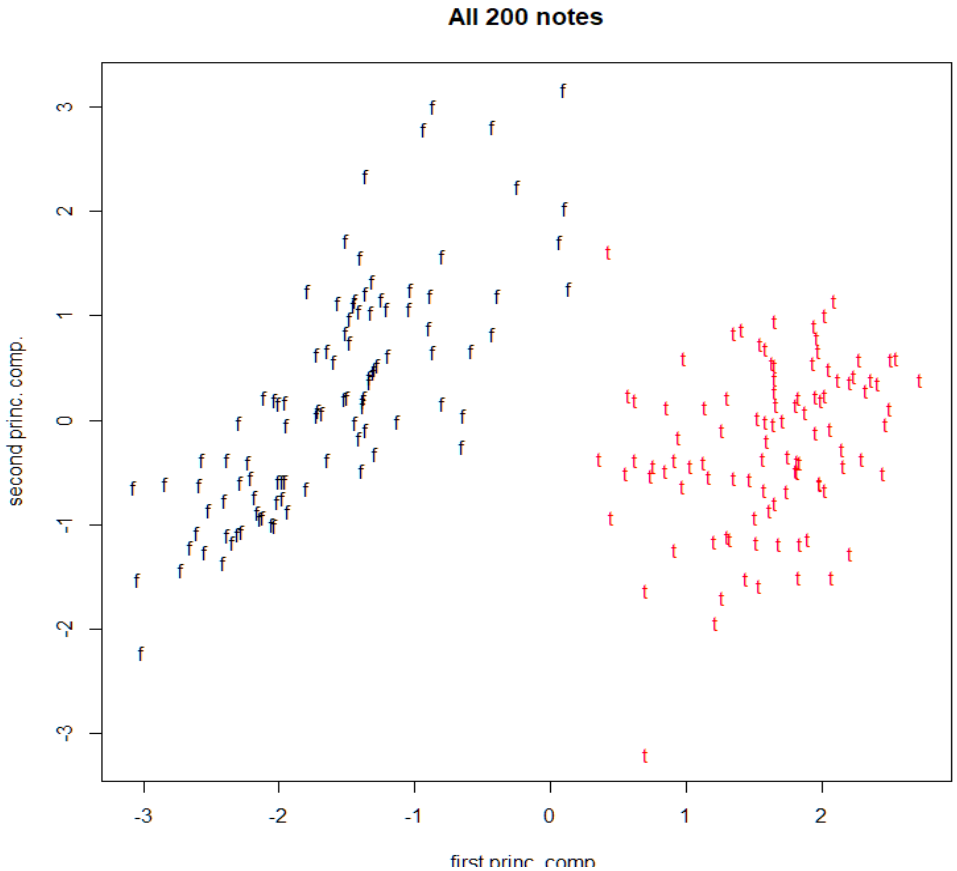


b) <u>For 100 counterfeit notes:</u> PCA using only the data from counterfeit notes. Computing the whole data in these new coordinates and plotting all the 200 notes



Totally cannot differentiate between genuine and counterfeit notes.

c) PCA using all 200 notes



A decision boundary can be drawn to clearly distinguish between genuine notes and counterfeit notes using just the first two principle components.