James Yang

DATA 558

Spring 2022

Homework #5

1. .
2. Left side of data is just the mean value of kmeans$withinss = 1.219

Right side of data is the average distance between point and centroid squared = 1.219

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

This iteration created an equal value which indeed supports the idea in the textbook that it holds equivalent.

* 1. Producing the observations with different means so they can get spread out.

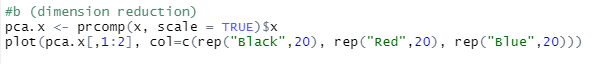
Table

Description automatically generated

* 1. Performing PCA, I get the following plot with the following code. Since it is dimension reduction, we only want the $x values from the output.

Chart, scatter chart

Description automatically generated



* 1. After doing K-means clustering = 3, we get all perfect labels.

A picture containing diagram

Description automatically generated

Table

Description automatically generated

* 1. After doing K = 2, we still get a perfect classification.

A picture containing graphical user interface

Description automatically generated

A picture containing table

Description automatically generated

* 1. After doing K = 4, it appears the 3rd column gets a split classification between 3 and 4.

A picture containing diagram

Description automatically generated

Table

Description automatically generated

* 1. After doing K = 3 with the PCA, it appears to perfectly classify.

A picture containing calendar

Description automatically generated

Table

Description automatically generated with low confidence

* 1. After doing the scale(x), it appears to also perfectly classify. The results appear to be the exact same as the ones found in part c except the classification flipped with column 2 and 3 finding 3 and 1 respectively. This could be due to randomization but standardizing the values because of their mean may alter their classification, but it does it on each set of values so they might just classify differently but still as a cluster.

A picture containing chart

Description automatically generated

Table

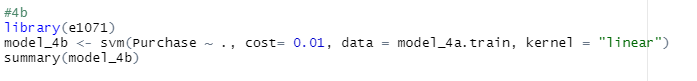
Description automatically generated

1. .
   1. The following code creates a training set of 800 observations in OJ.

Text

Description automatically generated

* 1. After calling svm, we get the following results.



Graphical user interface, application

Description automatically generated with medium confidence

It appears there were over 424 support vectors being plotted to classify this binary svm. 212 in one and 212 in the other. The reason it’s binary is because the number of classes being shown appears to be 2. It appears to be a linear kernel used with cost of 0.01.

* 1. Testing error appears to be 17.4%.

Training error appears to be 15.5%

With this code:

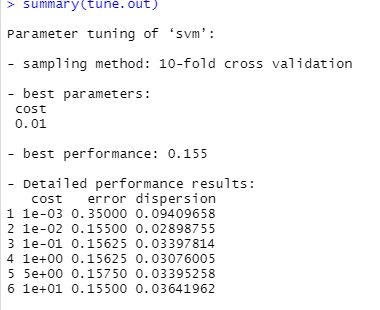
Text

Description automatically generated

* 1. When running the following code,

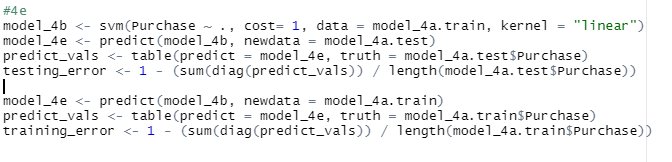


It produces the following:



Indicating that the most optimal cost is at 0.01 or 1 with an error of 15.5%.

* 1. When using the following code,



Training error: 14.875%

Testing error: 17.407%