RMarkdown Assignment # 15 - Week 09

Pushkar Chougule

Oct 29th 2020

## R Markdown

##Read the Binary csv dataset and use summary() on the dataframe.

binary\_csv <- read.csv('binary-classifier-data.csv', header = TRUE)  
  
binary\_csv$label <- as.factor(binary\_csv$label)  
  
str(binary\_csv)

## 'data.frame': 1498 obs. of 3 variables:  
## $ label: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ x : num 70.9 75 73.8 66.4 69.1 ...  
## $ y : num 83.2 87.9 92.2 81.1 84.5 ...

summary(binary\_csv)

## label x y   
## 0:767 Min. : -5.20 Min. : -4.019   
## 1:731 1st Qu.: 19.77 1st Qu.: 21.207   
## Median : 41.76 Median : 44.632   
## Mean : 45.07 Mean : 45.011   
## 3rd Qu.: 66.39 3rd Qu.: 68.698   
## Max. :104.58 Max. :106.896

trinary\_csv <- read.csv('trinary-classifier-data.csv', header = TRUE)  
  
trinary\_csv$label <- as.factor(trinary\_csv$label)  
  
str(trinary\_csv)

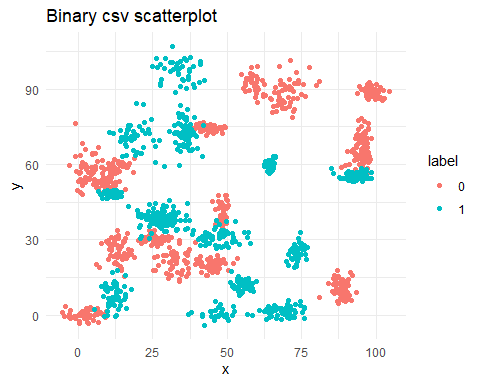
## 'data.frame': 1568 obs. of 3 variables:  
## $ label: Factor w/ 3 levels "0","1","2": 1 1 1 1 1 1 1 1 1 1 ...  
## $ x : num 30.1 31.3 34.1 32.6 34.7 ...  
## $ y : num 39.6 51.8 49.3 41.2 45.5 ...

summary(trinary\_csv)

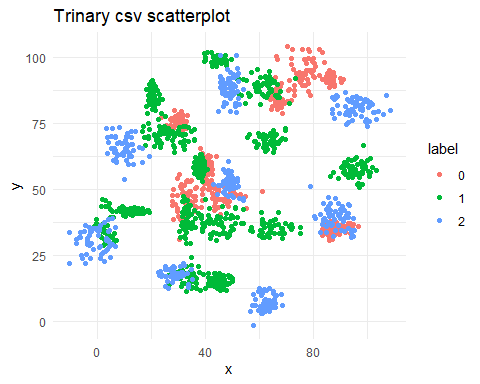
## label x y   
## 0:394 Min. :-10.26 Min. : -1.541   
## 1:722 1st Qu.: 31.15 1st Qu.: 35.906   
## 2:452 Median : 45.59 Median : 55.073   
## Mean : 48.86 Mean : 55.282   
## 3rd Qu.: 66.27 3rd Qu.: 77.403   
## Max. :108.56 Max. :104.293

**a. Plot scatter plots for both datasets**

library(ggplot2)  
theme\_set(theme\_minimal())  
  
ggplot(binary\_csv, aes(x = x, y = y, col=label)) + geom\_point() + ggtitle("Binary csv scatterplot")



ggplot(trinary\_csv, aes(x = x, y = y, col=label)) + geom\_point() + ggtitle("Trinary csv scatterplot")



##Split the respective data sets into train and test subsets. Also Eclidean distance calculated based on reference link.

library(caTools)  
library(philentropy)  
library(TSdist)  
  
set.seed(123456)  
  
split <- sample.split(binary\_csv, SplitRatio=.8)  
  
##############################################################  
  
binary\_csv\_matrix <- rbind(binary\_csv$x, binary\_csv$y)  
  
EuclideanDistance(binary\_csv$x,binary\_csv$y)

## [1] 1411.959

t1 <- stats::dist(binary\_csv\_matrix, method = "euclidean")  
  
cat("\n")

cat("Euclidean distance for Binary dataset\n")

## Euclidean distance for Binary dataset

t1

## 1  
## 2 1411.959

cat("\n")

bin\_train <- subset(binary\_csv, split == "TRUE")  
bin\_test <- subset(binary\_csv, split == "FALSE")  
  
  
##############################################################  
  
trinary\_csv\_matrix <- rbind(trinary\_csv$x, trinary\_csv$y)  
  
EuclideanDistance(trinary\_csv$x,trinary\_csv$y)

## [1] 1357.734

t2 <- stats::dist(trinary\_csv\_matrix, method = "euclidean")  
  
cat("\n")

cat("Euclidean distance for Trinary dataset\n")

## Euclidean distance for Trinary dataset

t2

## 1  
## 2 1357.734

cat("\n")

trin\_train <- subset(trinary\_csv, split == "TRUE")  
trin\_test <- subset(trinary\_csv, split == "FALSE")

**Answer a.**

Euclidean distance calculated for Binary csv dataset based on couple different ways is 1411.959.

Euclidean distance calculated for Trinary csv dataset based on couple different ways is 1357.734.

**b. The k nearest neighbors algorithm. Accuracy results in a graph**

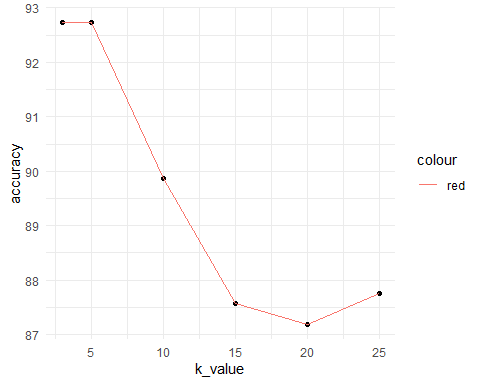
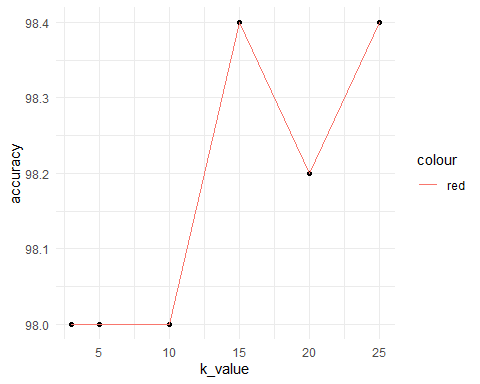
Calculated the accuracy levels for both the datasets based on the k values provided in the exercise. Stored the values of k in a list. Stored the corresponding combinations of k values and accuracy levels in a dataframe, later to be used for plotting the results.

## [1] 0 0 0 0 0 0  
## Levels: 0 1

## k\_value accuracy  
## 1 3 98.0  
## 2 5 98.0  
## 3 10 98.0  
## 4 15 98.4  
## 5 20 98.2  
## 6 25 98.4

## k\_value accuracy  
## 1 3 92.73423  
## 2 5 92.73423  
## 3 10 89.86616  
## 4 15 87.57170  
## 5 20 87.18929  
## 6 25 87.76291

**Plots of accuracy levels for respective k values**



**Answer b.**

As we notice above, K nearest neighbors algorithm produces models with accuracy levels closer to 98% for binary data csv, for given set of values of k. And this accuracy level is fairly varies in the small range between 97.8% to 98.4% for the suggested K values. Accuracy levels are highest for k=15 and k=25 and lower for k=20.

For trinary data csv, K nearest neighbors algorithm produces models with accuracy levels varying between the range 87.6% to 92.7% for the suggested K values. Accuracy levels are highest for k=3 and k=5 and goes on decreasing later on up to k=20. Then for k=25, it goes up a bit.

**c. Linear classifiers and decision plots**

**Answer c.**

Looking at the scatter plot of Binary csv, the x vs. y scatter plots of labels 0 and 1 data points aren’t distributed in a way as to to be able to easily classify them with linear classifier (imaginary line shown for question c). i.e. since the data points for these two labels are together at many places, we won’t be able to use linear classifiers to easily differentiate between them

Similarly, for the scatter plot of Trinary csv, the x vs. y scatter plots of labels 0, 1 and 2 data points aren’t distributed in a way as to to be able to easily classify them with linear classifier (imaginary line shown for question c). i.e. since the data points for these three labels are together at many places, we won’t be able to use linear classifiers to easily differentiate between them

## References

<https://cran.r-project.org/web/packages/philentropy/vignettes/Distances.html>

<https://www.analyticsvidhya.com/blog/2015/08/learning-concept-knn-algorithms-programming/#>:~:text=Unhesitatingly%2C%20using%20kNN%20Algorithm.,points%20into%20well%20defined%20groups

<https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/>