9.2 Introduction to Machine Learning

Arindam Samanta

February 09 2020

#### Loading the required libraries for our analysis

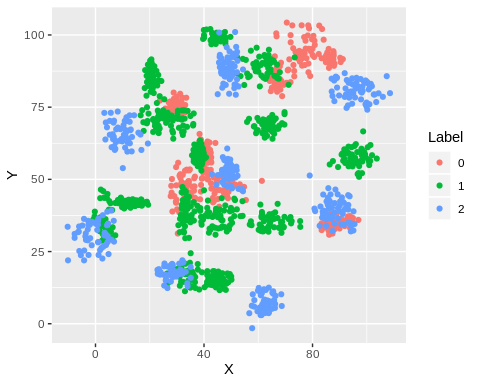
library(ggplot2)  
library(dplyr)

There are 1498 observations and 3 variables. Out of which the variable label is binary having 0 and 1 as the output. The other 2 variables are x and y.

#populating the binary classifier file to dataframe  
wd <- getwd()  
fname1 <- "binary-classifier-data.csv"  
path\_to\_file <- paste(wd,'/dataset/',fname1, sep = "")  
path\_to\_file  
  
my\_init\_bindf <- read.csv(path\_to\_file, header = TRUE)  
  
fname2 <- "trinary-classifier-data.csv"  
path\_to\_file <- paste(wd,'/dataset/',fname2, sep = "")  
path\_to\_file  
  
my\_init\_tridf <- read.csv(path\_to\_file,header = TRUE)  
#Details of the binary dataset  
  
summary(my\_init\_bindf)  
head(my\_init\_bindf)  
str(my\_init\_bindf)  
  
#Details of the trinary dataset  
  
summary(my\_init\_tridf)  
head(my\_init\_tridf)  
str(my\_init\_tridf)

### Plot the data for each data set using a scatter plot.

Plotting the data to visualize the relationship between the variables. Converting the label variable into factor as it has only two values. The plot shows that the values do not show any relationship.



### Fit a k nearest neighbors model for each dataset for k=3, k=5, k=10, k=15, k=20, and k=25. Compute the accuracy of the resulting models for each value of k.

Data splicing basically involves splitting the data set into training and testing data set.Here we are splitting the binary data set into train and test set.

## [1] 1048 3

## [1] 450 3

## [1] 1097 3

## [1] 471 3

The binary data has 2 labels and 0 and 1 and it is predicted using x and y predictors.Before building the KNN cluster we build the training and test sets.

The trinary data has 3 labels and 0 and 1 and 2 and it is predicted using x and y predictors.Before building the KNN cluster we build the training and test sets.

So now building a k NN clustering with k = (3,5,10,15,20,25) by removing the labels from the data frame.

### Plot the results in a graph where the x-axis is the different values of k and the y-axis is the accuracy of the model.

The below graphs shows that the accuracy is the highest for the binary cluster size of 5. So ideally there should be 5 clusters for the dataset. The trinary cluster has the maximum accuracy at k = 3.

