Exercise 14

Binay Prasanna Jena

October 24th 2020

# Fit a logistic regression model to the binary-classifier-data.csv dataset from the previous assignment.

Problem Statement : Fit a logistic regression model to the binary-classifier-data.csv dataset from the previous assignment.

1. What is the accuracy of the logistic regression classifier?
2. How does the accuracy of the logistic regression classifier compare to the nearest neighbors algorithm?
3. Why is the accuracy of the logistic regression classifier different from that of the nearest neighbors?

## Set the working directory to the root of your DSC 520 directory  
setwd("C:/Users/binay/Documents/GitHub/dsc520/")  
  
## Load the `caTools` library  
library(caTools)

## Warning: package 'caTools' was built under R version 4.0.3

#library(MASS)  
  
## Load the `data/binary-classifier-data.csv` to  
binary\_classifier\_df <- read.csv("data/binary-classifier-data.csv")  
head(binary\_classifier\_df)

## label x y  
## 1 0 70.88469 83.17702  
## 2 0 74.97176 87.92922  
## 3 0 73.78333 92.20325  
## 4 0 66.40747 81.10617  
## 5 0 69.07399 84.53739  
## 6 0 72.23616 86.38403

summary(binary\_classifier\_df)

## label x y   
## Min. :0.000 Min. : -5.20 Min. : -4.019   
## 1st Qu.:0.000 1st Qu.: 19.77 1st Qu.: 21.207   
## Median :0.000 Median : 41.76 Median : 44.632   
## Mean :0.488 Mean : 45.07 Mean : 45.011   
## 3rd Qu.:1.000 3rd Qu.: 66.39 3rd Qu.: 68.698   
## Max. :1.000 Max. :104.58 Max. :106.896

## Fit a logistic regression model to the binary-classifier-data.csv dataset from the previous assignment.

# Since label is number converting to factors, so that it becomes categorical  
binary\_classifier\_df$label <- as.factor(binary\_classifier\_df$label)  
  
# Splitting the dataset for the model into train and test datasets.  
myData <- sample.split(binary\_classifier\_df$label, SplitRatio=0.8)  
  
  
train <- subset(binary\_classifier\_df, myData==TRUE)  
test <- subset(binary\_classifier\_df, myData==FALSE)  
  
  
  
# This model includes all other parameters as dependent  
# Using train dataset to generate the model  
lrmodel.1 <- glm(label ~ . , family ='binomial' , data = train)  
summary(lrmodel.1)

##   
## Call:  
## glm(formula = label ~ ., family = "binomial", data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.3970 -1.1658 -0.9371 1.1599 1.4314   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.475602 0.131128 3.627 0.000287 \*\*\*  
## x -0.002575 0.002037 -1.264 0.206340   
## y -0.009110 0.002101 -4.337 1.45e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1661.5 on 1198 degrees of freedom  
## Residual deviance: 1637.5 on 1196 degrees of freedom  
## AIC: 1643.5  
##   
## Number of Fisher Scoring iterations: 4

## a. What is the accuracy of the logistic regression classifier?

# Using test dataset to see if the model is good  
result <- predict(lrmodel.1,test,type = "response")  
  
# result  
# validating - putting the actual value and counts of Predicted values in a matrix  
# Setting to T if result > 0.5  
confmatrix <- table(ActualValue=test$label, PredictedValue = result > 0.5)  
confmatrix

## PredictedValue  
## ActualValue FALSE TRUE  
## 0 83 70  
## 1 69 77

# accuracy - Cases where we predicted correctly by Total Predictions  
# from matrix, we see when Actual Value is T, confmatrix needs to pick 1,2  
# and when F it should pick 2,1  
(confmatrix[1,1]+confmatrix[2,2])/sum(confmatrix)

## [1] 0.5351171

So this model shows an accuracy ~55%.

# b. How does the accuracy of the logistic regression classifier compare to the nearest neighbors algorithm?

library(class)  
  
# Generating knn model with k=1  
knnmodel.1 <- knn(train[2:3],test[2:3],k=1,cl=train$label)  
summary(knnmodel.1)

## 0 1   
## 152 147

##create confusion matrix  
tab <- table(knnmodel.1,test$label)  
  
##this function divides the correct predictions by total number of predictions that tell us how accurate the model is.  
accuracy <- function(x){sum(diag(x)/(sum(rowSums(x)))) \* 100}  
accuracy(tab)

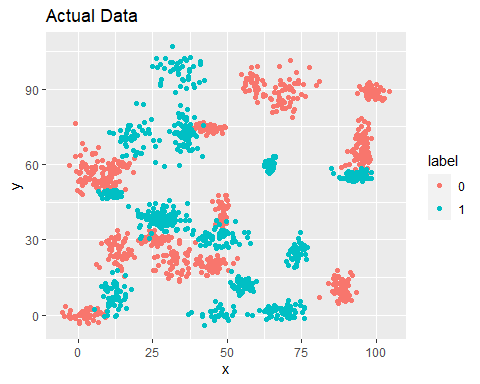
## [1] 96.32107

# Running for multiple K Values  
for(i in 1:20){  
 ##print(paste("Model with K=", i))  
 knnmodel.i <- knn(train[2:3],test[2:3],k=i,cl=train$label)  
 table.i <- table(knnmodel.i,test$label)  
 print(paste("Accuracy for Model with K=", i ," is ", accuracy(table.i)))  
}

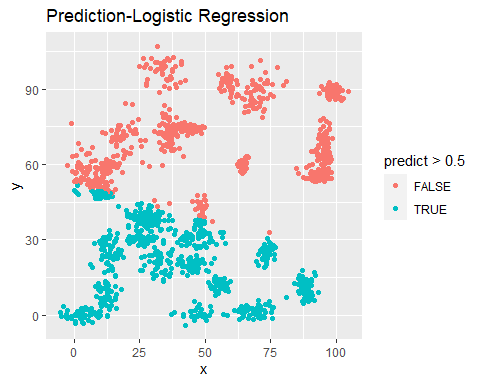
## [1] "Accuracy for Model with K= 1 is 96.3210702341137"  
## [1] "Accuracy for Model with K= 2 is 95.9866220735786"  
## [1] "Accuracy for Model with K= 3 is 96.6555183946488"  
## [1] "Accuracy for Model with K= 4 is 95.9866220735786"  
## [1] "Accuracy for Model with K= 5 is 95.9866220735786"  
## [1] "Accuracy for Model with K= 6 is 96.3210702341137"  
## [1] "Accuracy for Model with K= 7 is 97.3244147157191"  
## [1] "Accuracy for Model with K= 8 is 97.6588628762542"  
## [1] "Accuracy for Model with K= 9 is 97.3244147157191"  
## [1] "Accuracy for Model with K= 10 is 97.3244147157191"  
## [1] "Accuracy for Model with K= 11 is 97.6588628762542"  
## [1] "Accuracy for Model with K= 12 is 96.6555183946488"  
## [1] "Accuracy for Model with K= 13 is 97.3244147157191"  
## [1] "Accuracy for Model with K= 14 is 97.3244147157191"  
## [1] "Accuracy for Model with K= 15 is 97.6588628762542"  
## [1] "Accuracy for Model with K= 16 is 97.3244147157191"  
## [1] "Accuracy for Model with K= 17 is 97.6588628762542"  
## [1] "Accuracy for Model with K= 18 is 97.6588628762542"  
## [1] "Accuracy for Model with K= 19 is 97.3244147157191"  
## [1] "Accuracy for Model with K= 20 is 97.3244147157191"

# c. Why is the accuracy of the logistic regression classifier different from that of the nearest neighbors?

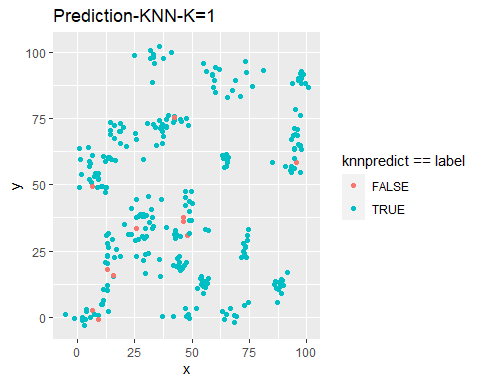
# creating df for plotting the comparison against actuals  
df <- binary\_classifier\_df  
df$predict <- predict(lrmodel.1, df,type = "response")  
  
# Adding details to test df for plotting  
test$knnpredict <- knn(train[2:3],test[2:3],k=1,cl=train$label)  
  
library(ggplot2)  
ggplot(data = binary\_classifier\_df, aes(y = y, x = x, color = label)) +  
 geom\_point() + ggtitle("Actual Data")



ggplot(data = df, aes(y = y, x = x, color = predict>0.5)) +  
 geom\_point() + ggtitle("Prediction-Logistic Regression")



ggplot(data = test, aes(y = y, x = x, color = knnpredict == label)) +  
 geom\_point() + ggtitle("Prediction-KNN-K=1")

 If we look at the Actual Data, its hard to divide the data with a line into two separate sections. So the Logistic Regression Model is not a good fit. If we look at “Prediction-KNN-K=1” plot, it seems to make sense as the values are plotted better.

# Reference

<https://towardsdatascience.com/k-nearest-neighbors-algorithm-with-examples-in-r-simply-explained-knn-1f2c88da405c>