Exercise 14 - Fit a logistic regression model

Kolekar, Shilpa

October 26th, 2020

## Set the working directory   
setwd("C:/Users/shilp/Documents/GitHub/dsc520/data")  
  
## Load the `data/binary-classifier-data.csv` to  
binary\_classifer\_df <- read.csv("binary-classifier-data.csv")  
summary(binary\_classifer\_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

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

newModel <- glm(label ~ x + y, data = binary\_classifer\_df, family = binomial)  
  
## install.packages("caTools")  
library("caTools")

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

split <- sample(1:nrow(binary\_classifer\_df), 0.8 \* nrow(binary\_classifer\_df))  
train <- binary\_classifer\_df[split,]  
test <- binary\_classifer\_df[-split,]  
  
# run the test data through model  
res <- predict(newModel, test, type= "response")  
res <- predict(newModel, train, type = "response")  
  
##validate the model confusion matrix  
confmatrix <- table(Actual\_Value = train$label, Predicted\_Value = res > 0.5)  
confmatrix

## Predicted\_Value  
## Actual\_Value FALSE TRUE  
## 0 337 256  
## 1 246 359

# Accuracy  
(confmatrix[[1,1]]+ confmatrix[[2,2]]) / sum(confmatrix)

## [1] 0.5809683

Accuracy if model is 58.10%.

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

##extract 1st column of train dataset because it will be used as 'cl' argument in knn function.  
target\_category <- binary\_classifer\_df[split,1]  
  
##extract 1st column if test dataset to measure the accuracy  
test\_category <- binary\_classifer\_df[-split,1]  
  
##load the package class  
## install.packages("class")  
library("class")

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

##run knn function, k=sqrt(1498)  
test\_pred <- knn(train,test,cl=target\_category,k=39)  
test\_pred

## [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1  
## [38] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0  
## [75] 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0  
## [112] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [149] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1  
## [186] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [223] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [260] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1  
## [297] 1 1 1 1  
## Levels: 0 1

##create confusion matrix  
table <- table(test\_category, test\_pred)  
table

## test\_pred  
## test\_category 0 1  
## 0 167 7  
## 1 2 124

# Accuracy  
(table[[1,1]]+ table[[2,2]]) / sum(table)

## [1] 0.97

Accuracy of the logistic regression classifier model is 58.10%, and the nearest neighbors algorithm model gave 97.00% of accuracy.

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

Nearest neighbors model has low biased and low variance data. To get the better accuracy rate on the model low biased (less error data in training data) and low variance (less error in the test data) is required. Based on the accuracy percentage, nearest neighbors model has low biased and low variance data.