ASSIGNMENT 8 Exercise 14: Fit a Logistic Regression Model to Previous Dataset

Abhijit Mandal

2020-10-26

## Assignment

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

## Question A:

**What is the accuracy of the logistic regression classifier?**

## Answer for A:

The accuracy came out to be 58.34%

setwd("~/Documents/GitHub/dsc520")  
binary\_df <- read.csv("data/binary-classifier-data.csv")  
  
binaryClassifier\_df <- read.csv("data/binary-classifier-data.csv")  
  
#Logistic regression model  
binaryClassifier\_glm <- glm(label ~ x + y, data=binaryClassifier\_df, family = binomial)  
  
summary(binaryClassifier\_glm)

##   
## Call:  
## glm(formula = label ~ x + y, family = binomial, data = binaryClassifier\_df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.3728 -1.1697 -0.9575 1.1646 1.3989   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.424809 0.117224 3.624 0.00029 \*\*\*  
## x -0.002571 0.001823 -1.411 0.15836   
## y -0.007956 0.001869 -4.257 2.07e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2075.8 on 1497 degrees of freedom  
## Residual deviance: 2052.1 on 1495 degrees of freedom  
## AIC: 2058.1  
##   
## Number of Fisher Scoring iterations: 4

res\_val <- predict(binaryClassifier\_glm, type="response")  
  
bcPredictionData <- table(Actual\_Value = binaryClassifier\_df$label, Predicted\_Value = res\_val > 0.5)  
  
bcPredictionData

## Predicted\_Value  
## Actual\_Value FALSE TRUE  
## 0 429 338  
## 1 286 445

dataModelAccuracy <- (bcPredictionData[[1,1]] + bcPredictionData[[2,2]]) / sum(bcPredictionData)  
  
dataModelAccuracy

## [1] 0.5834446

## Question B.

**How does the accuracy of the logistic regression classifier compare to the nearest neighbors algorithm?**

## Answer for B:

The accuracy of the logistic regression classifier as compared to the nearest neighbors algorithm is   
  
1. 71.33% with 100 nearest neighbors comparison  
2. 97.66% with 50 nearest neighbors comparison

library(class)  
# Splitting the binary classifier data in 80-20 ratio, 20 to train and 80 to test  
binaryClassifier\_split <- sample(1:nrow(binaryClassifier\_df), 0.8 \* nrow(binaryClassifier\_df))  
trainds <- binaryClassifier\_df[binaryClassifier\_split,]  
testds <- binaryClassifier\_df[-binaryClassifier\_split,]  
  
trained\_dataset <- binaryClassifier\_df[binaryClassifier\_split,1]  
test\_dataset <- binaryClassifier\_df[-binaryClassifier\_split,1]  
  
# Applying k nearest neighbour algorithm  
knnTestprediction <- knn(trainds,testds,cl=trained\_dataset,k=100)  
confusionMatrix <- table(test\_dataset,knnTestprediction)  
confusionMatrix

## knnTestprediction  
## test\_dataset 0 1  
## 0 125 31  
## 1 55 89

modelaccuracy <- (confusionMatrix[[1,1]] + confusionMatrix[[2,2]]) / sum(confusionMatrix)  
modelaccuracy

## [1] 0.7133333

# model accuracy with 100 neighbors

## Question C:

**Why is the accuracy of the logistic regression classifier different from that of the nearest neighbors?**

## Answer For C

KNN :- K-nearest neighbor works/predicts as per the surrounding datapoints (K). It is a deterministic algorithm, if you keep the value of K and run the algorithm n times, the results will be the same. KNN is lazy execution and can be applied to non-linear solutions, due to this it provides better accuracy than logistic regression  
  
Logistic Regression :- It works with algebraic calculation for best fit curve for the complete population. it is linear regression with some non-linear activation function at the end to covert the regular/continuous output into different classes.