Exercise 14: Fit a Logistic Regression Model to Previous Dataset

Kolekar, Ajit

October 23rd, 2020

# Assignment description

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

binary\_df <- read.csv("C:/Users/Shilp/Documents/GitHub/dsc520/data/binary-classifier-data.csv")  
  
  
# Split data to use 80% of data to train the model and 20% of data to test the model  
data\_split <- sample(1:nrow(binary\_df), 0.8 \* nrow(binary\_df))  
train <- binary\_df[data\_split,]  
test <- binary\_df[-data\_split,]  
  
nrow(binary\_df)

## [1] 1498

nrow(train)

## [1] 1198

nrow(test)

## [1] 300

## Question a

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

binary\_glm <- glm(label ~ x + y, data=binary\_df, family = binomial)  
summary(binary\_glm)

##   
## Call:  
## glm(formula = label ~ x + y, family = binomial, data = binary\_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 <- predict(binary\_glm, type="response")  
  
confmatrix <- table(Actual\_Value = binary\_df$label, Predicted\_Value = res > 0.5)  
confmatrix

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

accuracy <- (confmatrix[[1,1]] + confmatrix[[2,2]]) / sum(confmatrix)  
accuracy

## [1] 0.5834446

The accuracy of the model using logistic regression classifier is 58.34%.

## Question 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\_df[data\_split,1]  
  
# extract 1st column if test dataset to measure the accuracy  
test\_category <- binary\_df[-data\_split,1]  
  
# load the package class  
library(class)  
  
# run knn function, using k = sqrt(nrow(binary\_df))  
test\_pred <- knn(train,test,cl=target\_category,k=39)  
  
# create confusion matrix  
confmatrix <- table(test\_category,test\_pred)  
confmatrix

## test\_pred  
## test\_category 0 1  
## 0 135 3  
## 1 2 160

accuracy <- (confmatrix[[1,1]] + confmatrix[[2,2]]) / sum(confmatrix)  
accuracy

## [1] 0.9833333

The accuracy of the model using neighbor’s algorithm is 98.33%

## Question c

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

Logistic regression is not a regression model, it is a classification model. It works similar to linear regression and calculates linear output. It cannot be applied to non-linear classification problems. Colinearity and outliers tamper the accuracy of Logistic Regression model. K nearest neighbors (KNN) supports non-linear solutions. Based on the accuracy of both models, it appears that the reason why KNN provided better accuracy than Logistic Regression is because the data set classification could be non-linear.