---  
title: "WK8\_Exercise14\_logistic\_Binary\_Classifier"  
author: "SumbarajuAditya"  
date: "2/4/2021"  
output: word\_document  
bibliography: C:/BU/DSC520/assignment\_repo/dsc520/bibliography.bib  
---

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

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

library("caTools")  
bn\_classifier\_df <- read.csv("C:/BU/DSC520/assignment\_repo/dsc520/data/binary-classifier-data.csv")  
head(bn\_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(bn\_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

split<-sample.split(bn\_classifier\_df, SplitRatio=0.8)  
split

## [1] TRUE FALSE TRUE

train <- subset(bn\_classifier\_df, split="TRUE")  
test <- subset(bn\_classifier\_df, split="FALSE")  
bn\_classifier\_glm<-glm(label ~ x + y, data = train, family = "binomial")  
summary(bn\_classifier\_glm)

##   
## Call:  
## glm(formula = label ~ x + y, family = "binomial", data = train)  
##   
## 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

glm\_result <- predict(bn\_classifier\_glm, test, type="response")  
glm\_result <- predict(bn\_classifier\_glm, train, type="response")  
confmatrix\_glm <- table(Actual\_Value=train$label, Predicted\_Value= glm\_result >0.5)  
confmatrix\_glm

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

#Calculate accuracy based on confusion matrix  
(confmatrix\_glm[[1,1]] + confmatrix\_glm[[2,2]])/sum(confmatrix\_glm)

## [1] 0.5834446

**Answer** - Logistic regression indicates an accuracy of 58%.

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

library("class")  
#Identifying the value of K is decided as square root of number of observations   
sqrt(nrow(train))

## [1] 38.704

# the output indicates the value of K: K = 38 or K = 39. Let's try to calculate the accuracy for K = 38 and K = 39  
knn.38 <- knn(train=train, test=test, cl=train$label, k=38 )  
knn.39 <- knn(train=train, test=test, cl=train$label, k=39 )  
accuracy.38 <- 100 \* sum(test$label == knn.38)/nrow(test)  
accuracy.38

## [1] 97.46328

accuracy.39 <- 100 \* sum(test$label == knn.39)/nrow(test)  
accuracy.39

## [1] 97.39653

table(knn.38, test$label)

##   
## knn.38 0 1  
## 0 747 18  
## 1 20 713

table(knn.39, test$label)

##   
## knn.39 0 1  
## 0 746 18  
## 1 21 713

library("caret")

## Loading required package: lattice

## Loading required package: ggplot2

confusionMatrix(table(knn.39, test$label))

## Confusion Matrix and Statistics  
##   
##   
## knn.39 0 1  
## 0 746 18  
## 1 21 713  
##   
## Accuracy : 0.974   
## 95% CI : (0.9646, 0.9814)  
## No Information Rate : 0.512   
## P-Value [Acc > NIR] : <2e-16   
##   
## Kappa : 0.9479   
##   
## Mcnemar's Test P-Value : 0.7488   
##   
## Sensitivity : 0.9726   
## Specificity : 0.9754   
## Pos Pred Value : 0.9764   
## Neg Pred Value : 0.9714   
## Prevalence : 0.5120   
## Detection Rate : 0.4980   
## Detection Prevalence : 0.5100   
## Balanced Accuracy : 0.9740   
##   
## 'Positive' Class : 0   
##

**Answer** - KNN indicates an accuracy of 97% that is significantly higher than logistic regression.

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

**Answer:**

The KNN model is showing higher accuracy than logistic regression. LR can derive confidence level (about its prediction), whereas KNN can only output the labels. The reason behind this difference is because KNN is non parametric model and logistic regression is parametric model. (Varghese, 2019) Hence KNN tries to predict binary glm\_result by indicating outcome as 0 or 1. However logistic regression tries to find the probability of outcome so that the values lie between 0 and 1. But, KNN is comparatively slower than Logistic Regression.

references:

Varghese, D. (2019, May 10). Comparative study on classic machine learning algorithms. Retrieved February 05, 2021, from <https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222#>:~:text=KNN%20is%20a%20non%2Dparametric%20model%2C%20where%20LR%20is%20a,can%20only%20output%20the%20labels.