```
---
title: "ASSIGNMENT 7.2_BinaryClassifier"
author: "Bhushan Suryawanshi"
date: '2020-07-14'

---
```

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")
```

```
## Warning: package 'caTools' was built under R version 4.0.2
```

```
classifier_df <- read.csv("binary-classifier-data.csv")
head(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(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(classifier_df, SplitRatio=0.8)
split
```

```
## [1]  TRUE FALSE  TRUE
```

```
train <- subset(classifier_df, split="TRUE")
test <- subset(classifier_df, split="FALSE")

logistic_model<-glm(label ~  x + y, data = train, family = "binomial")
summary(logistic_model)
```

```
##
## 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
```

```r
result <- predict(logistic_model, test, type="response")


result <- predict(logistic_model, train, type="response")



confusion_matrix <- table(Actual_Value=train$label, Predicted_Value= result >0.5)
confusion_matrix
```

```
##             Predicted_Value
## Actual_Value FALSE TRUE
##            0   429  338
##            1   286  445
```

```r
#Accuracy calculation based on confusion matrix

(confusion_matrix[[1,1]] + confusion_matrix[[2,2]])/sum(confusion_matrix)
```

```
## [1] 0.5834446
```

Answer - Logistic regression is showing accuracy of 58%.

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

```r
#KNN implementation
library("class")
```

```
## Warning: package 'class' was built under R version 4.0.2
```

```
#The value of K is decided as square root of number of observations
sqrt(nrow(train))
```

```
## [1] 38.704
```

```
# Based on above value we get K = 38 or K = 39 (if we round to nearest integer)

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] 97.32977
```

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

accuracy.39
```

```
## [1] 97.39653
```

```
table(knn.38, test$label)
```

```
##
## knn.38   0   1
##      0 745  18
##      1  22 713
```

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

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

```
library("caret")
```

```
## Warning: package 'caret' was built under R version 4.0.2
```

```
## 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 is showing accuracy of 97% much 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. The reason behind this difference is because KNN is non parametric model and logistic regression is parametric model. Hence KNN tries to predict binary result by indicating outcome as 0 or 1. However LR tries to find the probability of outcome so that the values lie between 0 and 1.