Assignment 7.2: Fit a Logistic Regression Model to a Previous Dataset

Roni Kaakaty

7/18/2020

##Convert Label to factor

```
binary_df$label <- as.factor(binary_df$label)
```

##Logistic regression model

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

```
##
## 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
```

##Prediction and matrix creation

```
binary_pred <- predict(bin_log, binary_df, type = "response")
conf_matrix <- table(Actual_Value=binary_df$label, Predicted_Value= binary_pred >0.5)
conf_matrix
```

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

##Model accuracy

```r
bin_accuracy <- (429 + 445)/(429 + 445 + 286 + 338) *100
bin_accuracy
```

```
## [1] 58.34446
```

##Load library and normalize data

```r
library(class)
normalize_data <- function(x) {(x -min(x))/(max(x)-min(x))}
```

##Randomize dataset

```r
set.seed(9850)
order_binary<- runif(nrow(binary_df))
binary_df<-binary_df[order(order_binary),]
```

##Remove outcome from dataset

```r
binary_norm <- as.data.frame(lapply(binary_df[,c(2,3)], normalize_data))
str(binary_norm)
```

```
## 'data.frame':    1498 obs. of  2 variables:
##  $ x: num  0.157 0.849 0.12 0.199 0.417 ...
##  $ y: num  0.561 0.156 0.524 0.703 0.711 ...
```

##Create model to train and test

```r
binary_train <- binary_norm[1:1350, ]
binary_test <- binary_norm[1351:1498, ]
binary_train_target <-binary_df[1:1350, 1]
binary_test_target <-binary_df[1351:1498, 1]
```

##Find sq root of observations to find proper K value

```r
require(class)
sqrt(1498)
```

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

##Neighbor's Algorithm

```r
binary_knn <- knn(train = binary_train, test = binary_test, cl=binary_train_target, k=39)
```

##Confusion matrix to find accuracy

```r
table(binary_test_target, binary_knn)
```

```
##                   binary_knn
## binary_test_target  0  1
##                  0 78  1
##                  1  0 69
```

```r
knn_accuracy <-(78+69)/(78+69+1+0)*100
knn_accuracy
```

```
## [1] 99.32432
```

A.) The accuracy of the logistic regression classifier is 58.34%.

B.)The nearest neighbor's algorithm had an accuracy of 99.32% whereas the logistic regression classifier was at 58.34%.

C.)Knn produced better accuracy because it supports non-linear solutions whereas logistic regression only supports linear solutions. This provides better predictability for the actual value of the factor.