7.2 Assignment

Anna Harvey

7/19/2020

## 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?

#Split the data into training and testing sets  
split <- sample.split(binary, SplitRatio = 0.8)  
train <- subset(binary, split == TRUE)  
test <- subset(binary, split == FALSE)  
  
#Create the GLM with the training set  
mod <- glm(label ~ x + y, data = binary)  
  
#Run the data through the model  
res <- predict(mod, test, type = "response")  
res <- predict(mod, train, type = "response")  
  
#Validate model with confusion matrix  
confmatrix <- table(Actual\_value = train$label, Predicted\_value = res > 0.5)  
confmatrix

## Predicted\_value  
## Actual\_value FALSE TRUE  
## 0 285 227  
## 1 191 296

#Calculate accuracy  
accresult <- signif((((confmatrix[[1,1]] + confmatrix[[2,2]]) / sum(confmatrix))\*100), digits=5)  
print(paste("Accuracy = ", accresult, "%"))

## [1] "Accuracy = 58.158 %"

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

#Generate random number that is 90% of the total number of rows in dataset  
ran <- sample(1:nrow(binary), 0.9 \* nrow(binary))  
  
#Normalization function  
nor <- function(x) {(x-min(x)/(max(x)-min(x)))}  
  
#Run normalization on the 3 columns of dataset because they are the predictors  
binary\_norm <- as.data.frame(lapply(binary[,c(1,2,3)], nor))  
summary(binary\_norm)

## label x y   
## Min. :0.000 Min. : -5.153 Min. : -3.983   
## 1st Qu.:0.000 1st Qu.: 19.816 1st Qu.: 21.244   
## Median :0.000 Median : 41.807 Median : 44.668   
## Mean :0.488 Mean : 45.121 Mean : 45.047   
## 3rd Qu.:1.000 3rd Qu.: 66.438 3rd Qu.: 68.734   
## Max. :1.000 Max. :104.623 Max. :106.932

#Extract training set  
binary\_train <- binary\_norm[ran,]  
  
#Extract testing set  
binary\_test <- binary\_norm[-ran,]  
  
#Extract 1st column of train dataset because it will be used as 'cl' argument in knn function  
binary\_target\_category <- binary[ran, 1]  
  
#Extract 1st column of test dataset to measure the accuracy  
binary\_test\_category <- binary[-ran, 1]  
library(class)  
  
#Run knn function  
pr <- knn(binary\_train, binary\_test, cl=binary\_target\_category, k = 39)  
  
#Create confusion matrix  
tab <- table(pr, binary\_test\_category)  
  
#Divides correct predictions by total number of predictions to tell us how accurate the model is  
accuracy <- function(x) (sum(diag(x))/(sum(rowSums(x)))\*100)  
print(paste("Accuracy = ", accuracy(tab), "%"))

## [1] "Accuracy = 96.6666666666667 %"

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

The KNN algorithm calculates accuracy by checking the 39 nearest neighbors of the predicted value (k=39) to determine the category. The logistic regression classifier calculates the overall accuracy of the model.