Week 8: Assignment 14

Megan Holford

2020-10-24

Exercise 14: Fit a logistic regression model to the binary-classifier-data.csv dataset

```r
binary_df <- read.csv("data/binary-classifier-data.csv")
head(binary_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
```

```r
summary(binary_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
```

```r
# Split data to use 80% of data to train the model and 20% of data to test the model
library(caTools)
binary_split <- sample.split(binary_df$label, SplitRatio=0.8)

train <- subset(binary_df, binary_split==TRUE)
test <- subset(binary_df, binary_split==FALSE)
```

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

```r
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
```

```r
binary_predict <- predict(binary_glm, type="response")

binary_cm <- table(Actual_Value = binary_df$label, Predicted_Value = binary_predict > 0.5)

binary_cm
```

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

```r
binary_accuracy <- (binary_cm[[1,1]] + binary_cm[[2,2]]) / sum(binary_cm)

binary_accuracy
```

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

The accuracy of the logistic regression model is approximately 58%.

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

```r
library(class)

# Generating knn model with k=1
```

```r
binary_knn <- knn(train[2:3],test[2:3],k=1,cl=train$label)

summary(binary_knn)
```

```
##   0   1
## 155 144
```

```r
##create confusion matrix

binaryknn_cm <- table(binary_knn,test$label)



##this function divides the correct predictions by total number of predictions that tell us how accurat

accuracy <- function(x){sum(diag(x)/(sum(rowSums(x)))) * 100}

accuracy(binaryknn_cm)
```

```
## [1] 95.98662
```

```r
# Running for multiple K Values

for(i in 1:20){

  ##print(paste("Model with K=", i))

  binary_knn2 <- knn(train[2:3],test[2:3],k=i,cl=train$label)

  binaryknn_cm2 <- table(binary_knn2,test$label)

  print(paste("Accuracy of K=", i, accuracy(binaryknn_cm2)))
}
```

```
## [1] "Accuracy of K= 1 95.9866220735786"
## [1] "Accuracy of K= 2 95.9866220735786"
## [1] "Accuracy of K= 3 96.3210702341137"
## [1] "Accuracy of K= 4 96.989966555184"
## [1] "Accuracy of K= 5 96.989966555184"
## [1] "Accuracy of K= 6 97.3244147157191"
## [1] "Accuracy of K= 7 97.3244147157191"
## [1] "Accuracy of K= 8 96.989966555184"
## [1] "Accuracy of K= 9 96.989966555184"
## [1] "Accuracy of K= 10 96.989966555184"
## [1] "Accuracy of K= 11 96.989966555184"
## [1] "Accuracy of K= 12 96.989966555184"
## [1] "Accuracy of K= 13 96.989966555184"
## [1] "Accuracy of K= 14 96.989966555184"
## [1] "Accuracy of K= 15 96.3210702341137"
## [1] "Accuracy of K= 16 96.3210702341137"
## [1] "Accuracy of K= 17 96.6555183946488"
## [1] "Accuracy of K= 18 96.3210702341137"
## [1] "Accuracy of K= 19 96.6555183946488"
## [1] "Accuracy of K= 20 96.3210702341137"
```

The accuracy of the KNN model is approximately 96 to 97%

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

Logistic regression is similar to linear regression and calculates a linear output and is not as accurate when working with non-linear problems. The KNN is good for non-linear problems. Because the accuracy is beeter with the KNN model, the data classification seems to be non-linear and therefore linear models would not be as accurate.