Assignment 7.2

Scott Breitbach

7/14/2020

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

bc\_df <- read\_csv("completed/Week7/binary-classifier-data.csv")

## Parsed with column specification:  
## cols(  
## label = col\_double(),  
## x = col\_double(),  
## y = col\_double()  
## )

bc\_df$label <- factor(bc\_df$label)

head(bc\_df)

## # A tibble: 6 x 3  
## label x y  
## <fct> <dbl> <dbl>  
## 1 0 70.9 83.2  
## 2 0 75.0 87.9  
## 3 0 73.8 92.2  
## 4 0 66.4 81.1  
## 5 0 69.1 84.5  
## 6 0 72.2 86.4

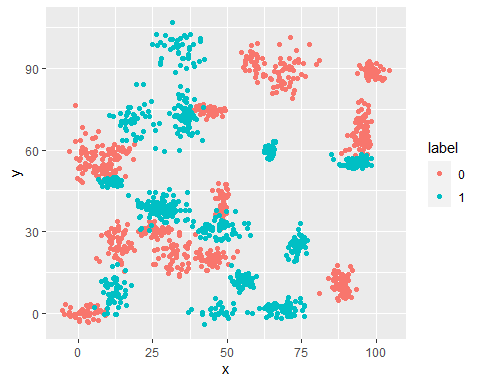
summary(bc\_df)

## label x y   
## 0:767 Min. : -5.20 Min. : -4.019   
## 1:731 1st Qu.: 19.77 1st Qu.: 21.207   
## Median : 41.76 Median : 44.632   
## Mean : 45.07 Mean : 45.011   
## 3rd Qu.: 66.39 3rd Qu.: 68.698   
## Max. :104.58 Max. :106.896

str(bc\_df)

## tibble [1,498 x 3] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ label: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ x : num [1:1498] 70.9 75 73.8 66.4 69.1 ...  
## $ y : num [1:1498] 83.2 87.9 92.2 81.1 84.5 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. label = col\_double(),  
## .. x = col\_double(),  
## .. y = col\_double()  
## .. )

ggplot(data = bc\_df, mapping = aes(x, y, color = label)) + geom\_point()



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

classifier\_glm <- glm(label ~ x + y, data = bc\_df, family = "binomial")  
  
res <- predict(classifier\_glm, bc\_df, type = "response")  
confmatrix <- table(Actual\_Value=bc\_df$label, Predicted\_Value = res > 0.5)  
confmatrix

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

(confmatrix[[1,1]] + confmatrix[[2,2]]) / sum(confmatrix)

## [1] 0.5834446

Applying a logistic regression to this data set yielded an accuracy of 58.3%. Looking at a plot of the data set, it would not appear that this would be the best tool for prediction.

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

bc\_df.2 <- bc\_df  
set.seed(42)  
ind <- sample(2, nrow(bc\_df.2), replace=TRUE, prob=c(0.67, 0.33))  
bc2.train <- bc\_df.2[ind==1, 2:3]  
bc2.test <- bc\_df.2[ind==2, 2:3]  
bc2.trainlabels <- bc\_df.2[ind==1,1]  
bc2.trainlabels <- bc2.trainlabels$label  
bc2.testlabels <- bc\_df.2[ind==2,1]  
bc2.testlabels <- bc2.testlabels$label

bc\_pred <- knn(train = bc2.train, test = bc2.test, cl = bc2.trainlabels, k=3)  
bc\_pred

## [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0  
## [38] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 1 0 0 0 0 0 1 0 0  
## [75] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0  
## [112] 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0  
## [149] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0  
## [186] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [223] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1  
## [260] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [297] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 0 0 1 1 1 1 1  
## [334] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [371] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1  
## [408] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [445] 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1  
## [482] 1 1 1  
## Levels: 0 1

CrossTable(x = bc2.testlabels, y = bc\_pred, prop.chisq=FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 484   
##   
##   
## | bc\_pred   
## bc2.testlabels | 0 | 1 | Row Total |   
## ---------------|-----------|-----------|-----------|  
## 0 | 241 | 11 | 252 |   
## | 0.956 | 0.044 | 0.521 |   
## | 0.968 | 0.047 | |   
## | 0.498 | 0.023 | |   
## ---------------|-----------|-----------|-----------|  
## 1 | 8 | 224 | 232 |   
## | 0.034 | 0.966 | 0.479 |   
## | 0.032 | 0.953 | |   
## | 0.017 | 0.463 | |   
## ---------------|-----------|-----------|-----------|  
## Column Total | 249 | 235 | 484 |   
## | 0.514 | 0.486 | |   
## ---------------|-----------|-----------|-----------|  
##   
##

sum(bc2.testlabels==bc\_pred)/length(bc\_pred)

## [1] 0.9607438

The k-nearest neighbors algorithm yielded an accuracy of 96.1%, much better than the logistic regression.

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

When the data is non-linear, as in this case where looking at the plot shows the data to be in small clusters, a logistic regression may not be the best approach. Since these data are clustered in groups, the nearest neighbors algorithm, which classifies points based on what other points are close to it, is a better option.