8.2 Logistic Regression vs Nearest Neighbour

Arindam Samanta

February 02 2020

### Source for the Data

#### Loading the required libraries for our analysis

library(ggplot2)  
library(dplyr)

There are 1498 observations and 3 variables. Out of which the variable label is binary having 0 and 1 as the output. The other 2 variables are x and y.

#populating the housing\_data dataframe  
wd <- getwd()  
fname <- "binary-classifier-data.csv"  
path\_to\_file <- paste(wd,'/dataset/',fname, sep = "")  
path\_to\_file  
  
my\_init\_df <- read.csv(path\_to\_file, header = TRUE)  
  
summary(my\_init\_df)  
  
head(my\_init\_df)

Plotting the data to visualize the relationship between the variables. Converting the label variable into factor as it has only two values. The plot shows that the values do not show any relationship.



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

Data splicing basically involves splitting the data set into training and testing data set.

## [1] 1048 3

## [1] 450 3

Fit a logistic regression model with the binary-classifier-data.csv dataset from the previous assignment. Fitting a logictic regression model using the training data set.

##   
## Call:  
## glm(formula = label ~ x + y, family = binomial, data = train.binary\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.3292 -1.1476 -0.9918 1.1761 1.3996   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.327218 0.139800 2.341 0.019252 \*   
## x -0.001078 0.002185 -0.493 0.621848   
## y -0.007499 0.002213 -3.389 0.000702 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1452.0 on 1047 degrees of freedom  
## Residual deviance: 1439.1 on 1045 degrees of freedom  
## AIC: 1445.1  
##   
## Number of Fisher Scoring iterations: 4

In order to get the accuracy of the logistic regression we need to build the confusion matrix. To compute the confusion matrix, we need to have a set of predictions so that they can be compared to the actual targets.

##   
## FALSE TRUE  
## 0 132 96  
## 1 110 112

In the above prediction model prediction type = “response”, indicates to compute the response probability. In order to compute the confusion matrix, predict>0.5 means it returns 1 if the predicted probabilities are above 0.5, else 0.

Now each row in a confusion matrix represents an actual target, while each column represents a predicted target. The model accuracy can be calculated by summing the TP+TN over the total observations.

## [1] 0.5422222

The model shows an accuracy of around 54%. So not a good classifier.

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

The binary data has 2 labels and 0 and 1 and it is predicted using x and y predictors. So now building a k NN clustering with k = 2 by removing the labels from the data frame.

## pred  
## test\_label 0 1  
## 0 224 4  
## 1 5 217

## [1] 0.98

So the accuracy of the K Nearest neighbor model is 98% while that of the logistic regression was 54%.

As it is a big issue with k-Nearest Neighbors is the choice of a suitable k. How many neighbors should you use to decide on the label of a new observation? But the below method proves that our choice of K=5 is appropriate.



## [1] 5

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

KNN is a non parametric model while Logistic Regression is a parametric model. So in KNN there are no parameters to be defined. Because of the way this data was distributed there was no linear relationship between the predictors and the outcome. Hence the outcome of the logistic regression classifier was much less accurate than the K nearest neighbor.

1. Why is the accuracy of the logistic regression classifier different from that of the nearest neighbors? When we plotted the data it showed that there were no linear or non liner relationship between the X and Y predictor variables rather they showed some type of clustering features. So the K Nearest neighbor was able to predict the outcomes more accurately than the logistic model.