8.3 Assignment: Fit a Logistic Regression Model to Previous Dataset

Dan Siegel

July 22, 2018

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

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

confusionMatrix(table(data = as.numeric(binary\_pred>0.5), reference = dat\_test$label))

## Confusion Matrix and Statistics  
##   
## reference  
## data 0 1  
## 0 120 91  
## 1 113 125  
##   
## Accuracy : 0.5457   
## 95% CI : (0.4983, 0.5924)  
## No Information Rate : 0.5189   
## P-Value [Acc > NIR] : 0.1387   
##   
## Kappa : 0.0934   
## Mcnemar's Test P-Value : 0.1415   
##   
## Sensitivity : 0.5150   
## Specificity : 0.5787   
## Pos Pred Value : 0.5687   
## Neg Pred Value : 0.5252   
## Prevalence : 0.5189   
## Detection Rate : 0.2673   
## Detection Prevalence : 0.4699   
## Balanced Accuracy : 0.5469   
##   
## 'Positive' Class : 0   
##

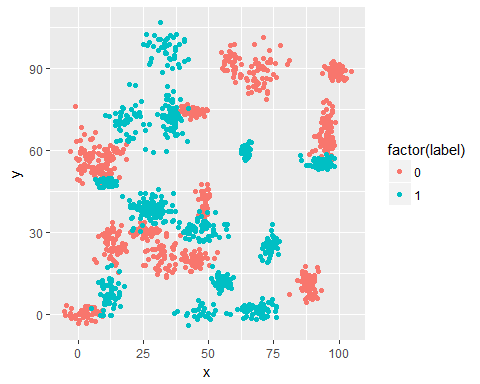
## # b. How does the accuracy of the logistic regression classifier compare to the nearest neighbors algorithm?

binary\_knn <- knn(dat\_train[,-1], dat\_test[,-1], factor(dat\_train$label), k=5, prob=TRUE)  
confusionMatrix(table(binary\_knn, dat\_test$label))

## Confusion Matrix and Statistics  
##   
##   
## binary\_knn 0 1  
## 0 228 10  
## 1 5 206  
##   
## Accuracy : 0.9666   
## 95% CI : (0.9455, 0.9812)  
## No Information Rate : 0.5189   
## P-Value [Acc > NIR] : <2e-16   
##   
## Kappa : 0.933   
## Mcnemar's Test P-Value : 0.3017   
##   
## Sensitivity : 0.9785   
## Specificity : 0.9537   
## Pos Pred Value : 0.9580   
## Neg Pred Value : 0.9763   
## Prevalence : 0.5189   
## Detection Rate : 0.5078   
## Detection Prevalence : 0.5301   
## Balanced Accuracy : 0.9661   
##   
## 'Positive' Class : 0   
##

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

1. Logistic regression predicts probabilities, which are a measure of the confidence of prediction. k-nearest neighbors predicts just the labels. 2.) Logistic regression assumes is that there is one smooth linear decision boundary. The data is clearly in clusters, and a little bit of EDA would help us determine which algorithm would be the best fit for this data.

 Looking at this data, there isn’t a single line we could put through and get the all the classifiers. Linear(or logistic) is simply not a good predictor of this data.