8.1 Assignment: Fit a Logistic Regression Model to the Thoracic Surgery Binary Dataset

Dan Siegel

July 22, 2018

# Assignment 8.2

## Fit the training data to a logistic regression model using the glm() function and use the test data to calculate the following classifier metrics.

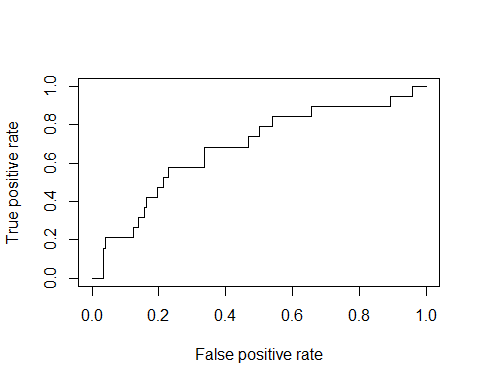
set.seed(123)  
n<- nrow(surg)  
shuffled <- surg[sample(n),]  
train <- shuffled[1:round(0.7 \* n),]  
test <- shuffled[(round(0.7 \* n) + 1):n,]  
testing\_survived <- glm(Risk1Yr ~ DGN + PRE4 + PRE5 + PRE6 + PRE7 + PRE8 +   
 PRE9 + PRE10 + PRE11 + PRE14 + PRE17 + PRE19 +   
 PRE25 + PRE30 + PRE32 + AGE, data=train)  
predicted\_data <- predict(testing\_survived, newdata = test)

## a. Calculate precision, recall, and F1 score for your model using the test dataset.

pred <- prediction(predicted\_data, test$Risk1Yr)  
RP.perf <- performance(pred, "prec", "rec")  
f1\_score <- performance(pred,"f")  
ROC.perf <- performance(pred, "tpr", "fpr")

## b. Plot the receiver operating characteristic (ROC) curve using the test dataset. Additionally, calculate the area under the curve (AUC) value.

plot (ROC.perf)



auc.tmp <- performance(pred,"auc");  
auc <- as.numeric(auc.tmp@y.values)  
auc

## [1] 0.6816221

## c. Consider the case where you fit a logistic regression model. When you calculate the classifier metrics, you get an accuracy of 96%, but an AUC of 53%. Is this a good predictive model? Explain why or not.

The AUC is the measure of performance of a binary classifier. The performance is averaged across all possible decision thresholds. Wikipedia has a great summarization of the issue at hand:

“One recent explanation of the problem with ROC AUC is that reducing the ROC Curve to a single number ignores the fact that it is about the tradeoffs between the different systems or performance points plotted and not the performance of an individual system, as well as ignoring the possibility of concavity repair, so that related alternative measures such as Informedness or DeltaP are recommended.”