Data 621: Assignment 2

Group 5: Christina Valore, Henry Vasquez, Chunhui Zhu, Chunmei Zhu, Yuen Chun Wong

Overview

In this homework assignment, you will work through various classification metrics. You will be asked to create functions in R to carry out the various calculations. You will also investigate some functions in packages that will let you obtain the equivalent results. Finally, you will create graphical output that also can be used to evaluate the output of classification models, such as binary logistic regression.

Q1.

Download the classification output data set (attached in Blackboard to the assignment).

Q2.

The data set has three key columns we will use:

- **class: the actual class for the observation**
- **scored.class: the predicted class for the observation (based on a threshold of 0.5)**
- **scored.probability: the predicted probability of success for the observation**

Use the table() function to get the raw confusion matrix for this scored dataset. Make sure you understand the output. In particular, do the rows represent the actual or predicted class? The columns?

Q3.

Write a function that takes the data set as a dataframe, with actual and predicted classifications identified, and returns the accuracy of the predictions.

$$Accuracy = TP + TN\ TP + FP + TN + FN$$

Q4.

Write a function that takes the data set as a dataframe, with actual and predicted classifications identified, and returns the classification error rate of the predictions.

$$Classification\ Error\ Rate = FP + FN\ TP + FP + TN + FN$$

Verify that you get an accuracy and an error rate that sums to one.

Q5.

Write a function that takes the data set as a dataframe, with actual and predicted classifications identified, and returns the precision of the predictions.

$$Precision = TP\ TP + FP$$

Q6.

Write a function that takes the data set as a dataframe, with actual and predicted classifications identified, and returns the sensitivity of the predictions. Sensitivity is also known as recall.

$$Sensitivity = TP\ TP + FN$$

Q7.

Write a function that takes the data set as a dataframe, with actual and predicted classifications identified, and returns the specificity of the predictions.

$$Specificity = TN\ TN + FP$$

Q8.

Write a function that takes the data set as a dataframe, with actual and predicted classifications identified, and returns the F1 score of the predictions.

$$F1\ Score = 2 \times Precision \times Sensitivity\ Precision + Sensitivity$$

Q9.

Before we move on, let's consider a question that was asked: What are the bounds on the F1 score? Show that the F1 score will always be between 0 and 1. (Hint: If $0 < a < 1$ and $0 < b < 1$ then $ab < a$.)

Using the hint above, we say a = precision and b = sensitivity and both numbers are between 0 and 1. So if a = .5 and b =.5 then a*b = .25 and so a*b < a or a*b < b. To prove that the F1 score is always between 0 and 1, we can run a simulation using the idea above with the p and s values also between 0 and 1.

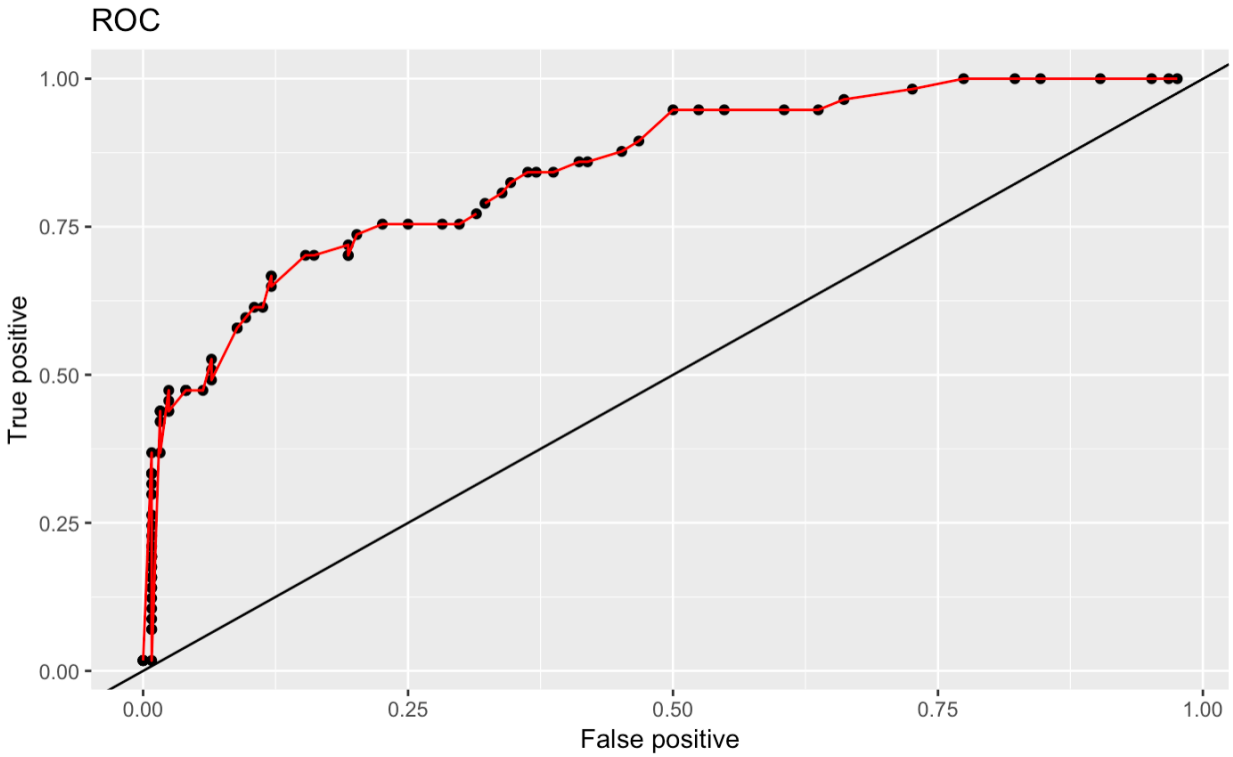
By generating random numbers and plugging them into the F1 score equation, we see that as we increase the amount of random numbers generated from 10 to 100 to 1000, we see that the max value gets close to 1 as the min value gets close 0 but never reaches either value. If we continue to generate more random numbers, we will see the max value continues to rise closer to 1, while the min closer to 0, however the max/min will never be equal to 1/0. Thus, the F1 value will always be between 0 and 1.

See the summary output below for random numbers generated at 10, 100 and 1000, taking note of min and max values approaching 0 and 1:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.04301 0.21419 0.47869 0.46052 0.67632 0.88619
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.004924 0.213325 0.346562 0.404481 0.578541 0.960901
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0005588 0.1889674 0.3670928 0.3981011 0.5990581 0.9950490
```

Q10.
Write a function that generates an ROC curve from a data set with a true classification column (class in our example) and a probability column (scored.probability in our example). Your function should return a list that includes the plot of the ROC curve and a vector that contains the calculated area under the curve (AUC). Note that I recommend using a sequence of thresholds ranging from 0 to 1 at 0.01 intervals.

We created a function that takes in the data and created the AUC segments using the scored probability and class variables. Then we took those segments and added them into two vectors and we added both of those vectors to a dataframe.



Q11.
Use your created R functions and the provided classification output data set to produce all of the classification metrics discussed above.

Q12.
Investigate the caret package. In particular, consider the functions confusionMatrix, sensitivity, and specificity. Apply the functions to the data set. How do the results compare with your own functions?

Q13.
Investigate the pROC package. Use it to generate an ROC curve for the data set. How do the results compare with your own functions?