CUNY SPS DATA 621 - CTG5 - HW2

Betsy Rosalen, Gabrielle Bartomeo, Jeremy O'Brien, Lidiia Tronina, Rose Koh

March 13, 2019

Contents

Deliverables should use R functions and the other packages to generate the `classification metrics` for the provided data set.

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

2. 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?

	0	1
0	119	30
1	5	27

- rows = predicted, cols = actual
- 1 is positive, 0 is negative

3.-8.: Write a function that takes the data set as a dataframe, with actual and predicted classifications identified, and returns (3)Acurracy; (4)Error rate; (5)Precision; (6)Sensitivity(recall); (7)Specificity; (8)F1 score of the predictions. Verify that you get an accuracy and an error rate that sums to one.

```
## [1] 1
```

accuracy	error.rate	precision	sensitivity	specificity	f1
0.8066298	0.1933702	0.7986577	0.9596774	0.4736842	0.8717949

9. 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 $ab < a$)

10. 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.

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

12. 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?

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