Homework 2

Group 1

Contents

Prepared for:

Dr. Nathan Bastian
City University of New York, School of Professional Studies - Data 621

Prepared by:

Group 1

Senthil Dhanapal
Yadu Chittampalli
Christophe Hunt

1 Data Source

We download the data from our public GitHub repository which was originally provided through Blackboard.

2 Data Explained

We will be using the following columns from the data source:

- 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

Below is the raw confusion matrix for our scored dataset

	Predicted Failure	Predicted Success
Actual Failure	119	5
Actual Success	30	27

In particular, do the rows represent the actual or predicted class? The columns? - Yadu

3 Accuracy of Predictions

We developed the below function that takes the data set as a dataframe, with actual and predicted classifications identified, and returns the accuracy of the predictions. - Yadu

```
accuracy <- function(actual, prediction){
  require(scales)
  truefalse <- as.data.frame(table(actual == prediction)) %>%
               spread('Var1', 'Freq')
  accuracy <- unlist(truefalse[2]/nrow(classification))
  paste0("The prediction accuracy is ", percent(accuracy))
}
```

```
accuracy(actual = scores$class, prediction = scores$scored.class)
```

[1] "The prediction accuracy is 80.7%"

4. 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.Verify that you get an accuracy and an error rate that sums to one. - Yadu
5. Write a function that takes the data set as a dataframe, with actual and predicted classifications identified, and returns the precision of the predictions. - Senthil
6. 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. - Senthil
7. Write a function that takes the data set as a dataframe, with actual and predicted classifications identified, and returns the specificity of the predictions. - Christophe
8. 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. - Christophe
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 < ???? < 1 and 0 < ???? < 1 then ????????- Christophe
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. - Senthil
11. Use your created R functions and the provided classification output data set to produce all of the classification metrics discussed above. - Christophe
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? - Yadu
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? - Senthil