Homework 5 CSCI E-81

Due November 23 11:59pm EST

The final homework assignment is a partnered project focused on classification. We have covered a number of classification methods over the past several weeks. This is your chance to explore and try out some of these methods, or any others that you find useful.

We have a data set of a reasonable size that has 3 parts:

a) Training data* and Training labels*
b) Test data* and Test labels
c) Blind data and Blind labels

We are providing you only with the data with asterisks but will be assessing you on the performance on all three sets. The labels for the data are "1", "2", and "3" so there are three different classes shown in the labels of each of the a) b) and c) data sets. Since we have mostly focused on two-class classification problems, we will be assessing these as 3 separate classifications of each label against the other two. Note that the training data has missing values denoted NA. The test data and blind data do not have any NA so you are welcome to handle the NA any way that you feel is appropriate.

Your task is to develop your best algorithms on the training data. You can use it any way you like but should cross-validate and optimize your classifiers for this data set. Every evening or so, you have the option of checking in a set of predictions on the test data and checking them into the dropbox. Starting this weekend, one of our teaching staff will take all the predictions, run them, and post the AUC scores on canvas. We haven't tried this yet, but we suggest posting your predictions under a coded filename.

The format of the predictions that will be used for the final evaluation as well should be a 4-column tab-delimited text file. The first 3 columns will be the predictions for each class where the class 1 vs. 2 and 3 would have high values for class 1 and low values for classes 2 and 3. The next two columns are 2 vs. 1 and 3, and 3 vs. 1 and 2. The final column (which will not be used until the end of the project will contain for your final class predicted class labels (i.e. 1, 2, or 3) that you can predict using any method you wish. We will post 3 AUC scores for the first 3 columns. Note that each row of the file should match exactly that of the test data.

We haven't decided the approach for the blind data. We will either run your training algorithms and make our own predictions, or release the blind data shortly before the deadline for you to test. However, you should be able to output the 4 columns described earlier.

Some of the classification algorithms that we have discussed include a variable importance type of feature. It is likely that one of the algorithms that you will try has that feature, although it may not be the final algorithm that you use. We would like you to show us the top features for one of your later classification algorithms. If none of your algorithms have a relatively obvious way of extracting that information, you can skip this but otherwise we would like to see which features are among the best.

Requirements:

- Try at least a few methods for classification and try to maximize your performance
- Bonus points are available for those who go above and beyond
- Use any language or method

Write-up:

- Document what you tried and how it worked. We're looking for your approach to the problem and evidence of what approaches you tried out.
- Show ROC curves comparing the at least the top approaches based on your training data.
- How did you handle the NAs?
- How did you produce the final class for each data point?
- What are the most important features and how many seem important?
- Document how we can utilize your code/method to re-train and re-do your top prediction
- If you are utilizing public sources of code, be sure to cite them

Evaluation Criteria in decreasing order of importance:

- Classification methods you tried
- Overall performance of the classifiers
- Quality of the write-up
- Readability of code so we can figure out what you did. Extensive commenting (if that means extra work) is not required.