Rev 1.0 9/29

Homework 3 CSCI E-81

Due October 12 11:59pm EST

The third homework assignment is a partnered project focused on classification. We are covering a number of classification methods over the past several weeks and next week. This is your chance to explore and try out some of these methods, or any others that you find useful as part of a friendly competition.

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", "3", and "4" so there are four 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 4 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. Starting this weekend, your team will have the option of checking in one set of predictions on the test data to the HW3 dropbox. The file should be a text file with one line per prediction and no header. It should have exactly the same number of rows as the test data. Every evening, one of our teaching staff will run the dropbox and post the results to the HW3 page between 9pm and midnight. Our last run will be Oct 11 so you have time to complete your report, make plots, and clean up your code. If your output does not meet the format, we won't be able to generate any results for you.

The format of the predictions that will be used for the final evaluation as well should be a 5-column tab-delimited text file and we have a sample format included. The first column will be the predictions for each class where the class 1 vs. rest, where high values indicate class 1 and low values indicate classes 2-4. The next three columns are 2 vs. rest, 3 vs. rest, and 4 vs. rest. 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, 3, or 4) that you can predict using any method you wish. We will post 4 AUC scores for the first 4 columns. Note that each row of the file should match exactly that of the test data.

We plan to release the blind data set to you on Monday evening Oct 10 to have you make your predictions on that data set for submission. You will submit the 5-column format results as part of your submission.

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.

Rev 1.0 9/29

Requirements:

- Try several methods for classification and try to maximize your performance
- Exploratory points are available for those who go above and beyond—document how you achieved this if applicable
- Use any language or method

Write-up:

- Brief document highlighting what you tried and how it worked. We're looking for the story of how you tried methods, gained insights on the data, and leveraged that insight to making improvements. Show ROC curves comparing the at least the top approaches based on your training data.
- Your document should address the following questions:
    - 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 final prediction
- If you are utilizing public sources of code, be sure to cite them

Evaluation Criteria in decreasing order of importance:

- Overall performance of the classifiers
- Write-up showing how you navigated toward your solution which does not have to be long
- Readability of code so we can figure out what you did. Extensive commenting (if that means extra work) is not required.

What to submit:

- One partner only need submit:
    - Write-up with both partners' names
    - Code base
    - Final prediction results
    - Do not include the raw data
- Both partners submit:
    - Partner evaluation (use upcoming Excel sheet and save as CSV)