COM 307 Machine Learning/Data Mining Project 3: Ionosphere free electrons (due by October 27 before class start)

In this project you will be using real-life data to see how different variables (the same kind of measurement taken from different locations) associate with structure in the earth's ionosphere. If you don't know what the ionosphere is and don't care, that's fine. If you don't and do care, feel free to look into it! If you do know, that's fine too. This dataset will have a binary "good" or "bad" to indicate ionosphere structure ("good") or lack of structure ("bad"). This is what we will know in our training dataset but would like to predict in any future test data. The data also contain 17 high-frequency antenna pulse data, which is continuous. Each pulse has 2 data points, for a total of 34 pieces of data. This is known in the training data and would also be known in any test data. For more information, please see http://archive.ics.uci.edu/ml/datasets/Ionosphere. To do this, you will be implementing a random forest as discussed in lecture 13. In order to do this, you will need to access the data from the website listed above.

For the actual implementation, you will want to select *m* random variables from amongst the 34 (you can select a value for *m*), figure out which does the best job separating "good" from "bad," and at what value (i.e. variable < 1.048 yields "good"-enriched node otherwise "bad"-enriched node). You will then recurse to build a decision tree based on those randomly selected variables. We have not yet talked about bagging or boosting, so you are free to determine how you merge the trees that you make into a random forest. The simplest (and worst) would be to just take the first random tree. A better approach would be to "average" the trees, to find the "majority votes best" tree. Full credit will be given to properly styled and documented code with strong write-up that uses only the first random tree. Additional "flexibility" in grading will be given for more sophisticated code/methodolgy.

[Submit]

Submit as a zipped file the following files:

1) Your code (in whatever language you used)

2) A not too official write-up of what you found during this exercise and what you did when you reached decision points. What did you select for *m*, the number of variables at random to select from among the 34 (*p*) total variables?

3) Feel free to submit additional results if you code generates them, hand-drawn graphics, or anything else that you feel helps you convey your strategy or conclusions.