CS534 — Implementation Assignment 3 — Due Nov 14th, 2015

General instruction.

1. The following languages are acceptable: Java, C/C++, Matlab, Python and R.
2. You can work in team of up to 3 people. Each team will only need to submit one copy of the source code and report.
3. Your source code and report will be submitted through the TEACH site `https://secure.engr.oregonstate.edu:8000/teach.php?type=want_auth`

   Please clearly indicate your team members' information.
4. Be sure to answer all the questions in your report. You will be graded based on your code as well as the report. So please write your report in clear and concise manner. Clearly label your figures, legends, and tables.
5. In your report, the results should always be accompanied by discussions of the results. Do the results follow your expectation? Any surprises? What kind of explanation can you provide?

Ensemble of Decision Stumps

In this assignment you will implement 1) the decision stump learning algorithm, which learns a decision tree with a single test; and 2) Bagging and AdaBoost using the decision stump algorithm as the base learner. You will test your implementation on the SPECT data sets, which is a two-class classification problem, with 22 binary features. You will train your classifies using the SPECT-train.csv file, and test on the SPECT-test.csv file. The first column of the data contains the class label, the remaining columns are the features.

You need to submit:

a. Source code of your implementations (Decision stump, bagging, Adaboost).

b. A report that contains the following components.

   1. For decision stump, please report the information gain of each binary feature. Present the final learned decision stump, and report its accuracy on the provided test data.
   2. For each ensemble method, please plot the training and testing errors as the size of the ensemble varies (please consider sizes 5, 10, 15, 20, 25 and 30). Note that for bagging, due to its stochastic nature, different random runs may lead to different results. To increase the robustness of the results, please show the average error rates over 10 random runs. Your plots should be clearly labeled with easy-to-read legends.
   3. A brief discussion of the results. Specifically, what trend do you observe in terms of the accuracy on the training and testing data respectively as we increase the number of stumps in the ensemble for bagging and for boosting? Comparing the performance of bagging and boosting, what do you observe? Are they consistent with your expectation? Please explain or comment on your expectation and the trend that you observe.

Grading rubric:

1) a correctly implemented decision stump and results (15 pts). Your implementation should correctly construct the decision stump for a given training set, report its accuracy (percentage of correctly classified examples) on the training set, and testing set.

2) a correctly implemented bagged decision stump (10 pts). Your implementation should correctly draw bootstrapped samples and apply your decision stump learning algorithm to the bootstrapped samples to construct an ensemble of stumps. It should also correctly carry out majority vote for prediction, and computes the ensemble's accuracy on the training set and testing set.

3) a correctly implemented boosted decision stump (20 pts). Your implementation should contain a version of the decision stump learning algorithm that takes instance weights into consideration, it should correctly carry out the boosting rounds and construct ensemble of boosted decision stumps. It should also correctly carry out weighted voting for prediction and computes the ensemble's accuracy on the training and testing data.

4) a well written report that provide a good discussion of the results and satisfactory answers to all the questions (15 pts)