Project 3: Classification Algorithms

Demo time: Start from 10:30am, December 3
Code&report submission due: 10:30am December 3

Two datasets (*project3_dataset1*, *project3_dataset2*) can be found on Piazza. Please check the README file first for a short description of the two datasets. This is a team project. Each team consists of at most three members.

Complete the following tasks:

- Implement three classification algorithms by yourself: **Nearest Neighbor, Decision Tree,** and **Naïve Bayes**.
- Implement **Random Forests** based on your own implementation of Decision Tree.
- Adopt 10-fold **Cross Validation** to evaluate the performance of all methods on the provided two datasets in terms of **Accuracy, Precision, Recall**, and **F-1 measure**.
- We will send you an invite for a Kaggle competition. For that dataset, we hold out the class labels for testing data. Apply various tricks on top of any classification algorithm discussed in class (including nearest neighbor, decision tree, Naïve Bayes, SVM, bagging, AdaBoost, random forests) and tune parameters using training data. You can call packages for these algorithms but need to implement any improvement on top of these algorithms. Submit your classification result for the testing data. Your efforts towards improving these algorithms will be evaluated. Those who are among the top on the leaderboard after the deadline will receive bonus points.

Your final submission should be a zip file named as *project3.zip*. In the zip file, you need to include a folder *Code* and a folder *Report*:

- Code: Implementation of four methods and your implementation for the Kaggle competition. **The four methods must be implemented by yourself**. The implementation for the Kaggle competition can use learning packages of the algorithms that were mentioned, but the improvement should be implemented by yourself. Together with your code submission, a README file should be included to explain how to execute your code.
- Report: For the four methods: Describe the flow of all the implemented methods, and describe the choice you make (such as parameter setting, pre-processing, etc.). Compare their performance, and state their pros and cons based on your findings. For the competition: Explain why a certain base algorithm is chosen, state clearly the parameters you choose for this algorithm, and discuss the improvement you have made towards improving its performance.
- Log in any CSE department server and submit your zip file as follows:
  >> **submit_cse601 project3.zip**

The details about Demo will be released **two days before the demo date through Piazza**. Please note:

- New datasets will be given to check your implemented classification methods and performance measures. The data format will be consistent with the README file that we already provided.
- During the demo, you will be asked to adopt specific setting and run your code.
- We will ask you to explain the basic idea of the method you use for the competition.