CSE472 (Machine Learning Sessional)
L-4, T-2, January 2018 Term

Assignment 1: Ensemble Learning

In ensemble learning, we combine decisions from multiple weak learners to solve a classification problem. It is expected that the combined decision will perform better than the individual models in the process of one model correcting the errors of the other. There are many ensemble methods such as stacking, bagging and boosting. In this assignment you will implement the **AdaBoost** algorithm. The necessary details are as follows.

1. As the weak/base learner use decision stump. A decision stump is a decision tree of depth one (i.e., it branches on only one attribute and then makes decision).
2. There are several implementations of AdaBoost algorithm. Follow the pseudocode given in the class.
3. You should make your code as modular as possible. Namely, your main module of Ada-Boosting should treat the base learner as a blackbox (in this case a decision stump) and communicate with it via a generic interface that inputs weighted examples and outputs a classifier, which then can classify any instances.
4. To incorporate effect of weighted dataset, create *training* data by sampling with replacement strategy. Use information-gain as the evaluation criterion.
5. For each stump check if the total weighted error is less than .5 to proceed to next step.
6. Use original data as *test* set and update weight vector for *test* set.
7. To train and test your model, use the file bank-additional-full.csv from the following link https://archive.ics.uci.edu/ml/datasets/Bank+Marketing. The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict whether the client will subscribe a term deposit (variable y) or not.
8. Analyze the expected performance of your model using k-fold cross validation for k=5, 10 and 20. Use F1 score as your evaluation metric.

Instructions for Report Writing

1. Your final report will contain the following points:
    a. Expected accuracies obtained using k-fold cross validation for k=5, 10 and 20.
    b. Compare the accuracies obtained by decision stump and boosting with 30 rounds.
    c. Compare the accuracies obtained by boosting with 5, 10 and 20 rounds.
2. Just answer the questions precisely. Make it as simple as possible.

Special Instructions

- Don't Copy anything! If you do copy from internet or from any other person or from any other source, you will be severely punished and it is obvious. More than that, we expect Fairness and honesty from you. Don't disappoint us!
- The report should be in *.pdf (No hardcopy is required). Write in your own language.
- You are encouraged to bring your computer in the sessional to avoid any hassle. But in that case, ensure an internet connection as you have to instantly download your code from the moodle and show it.

Submission Deadline

- Upload the codes in moodle within 9:00 A.M. of 21st April, 2018 (Saturday).

Instructions for moodle upload

Upload the assignment within the specified time. Otherwise, we can't accept it. For submission, use the following rules:

- If you write code in a single file, then rename it as<Studentid><code>.<extension>. For example, if your student id is 1305123 and you have done in java, then your file name should be "1305123code.java".
- If you write code in multiple files, then put all the necessary files in a folder and rename it as<Studentid><code>. For example, if your student id is 1205123 and you have done in java, then your folder name should be "1105123code".
- The report name should be <Studentid><report>.<extension>. For example, if your student id is 1205123 and it is in pdf format, then the report name should be "1205123report.pdf".
- Finally make a main folder, put the code (whether file or folder) and report in it, and rename the main folder as your<Student id><Programming language>. For example, "1205123Java". Then zip it and upload it.