Assignment 4

1 Description of the Assignment

1.1 Semi-supervised Learning

If the dataset, we have, contains some labeled and some unlabeled data, then it is better to use semi-supervised learning algorithm. Based on the given breast cancer dataset (dataset given for assignment 1), you have to run a semi-supervised learning algorithm using ID3 decision tree learner with information gain for evaluation criterion (definitely you can use your own code of assignment 1). The steps you should follow are given below:

1. Randomly divide the whole dataset into training data (80%) and test data (20%).
2. Divide the available training data into 2 sections. The first one is labeled data (D_l) and the second one is unlabeled data(D_u). For unlabeled data, do not consider the label/class/output feature of the data.
3. Build a decision tree T using D_l.
4. Based on T, give label to 10% (absolute 10%, not dependent on the current size of D_u) data of D_u and move these data from D_u to D_l.
5. Repeat the step 3 and 4 until D_u becomes empty.
6. Calculate the accuracy, precision, recall and F1-score using the test data.
7. Run the whole process 10 times and calculate the average accuracy, precision, recall and F1-score.

As the part of report writing,

- Compare and analyze the average accuracy, precision, recall and F1-score of semi-supervised learning with the ones obtained during the supervised learning.

1.2 Performance Evaluation

We want to evaluate the performance of the algorithm we have implemented using cross validation technique.

- **k-fold Cross Validation:** Divide the dataset into k subsets. Take i^{th} subset as the test data and the rest of the subsets as the training data. Do this for i=0,1,....,k. Calculate accuracy each time and take the average. Do it for k=5,10 and 20.
- **Leave-one-out Cross Validation:** If dataset contains m training examples, then take i^{th} labeled data as the test data and the rest (m-1) labeled data as the training data. Do this for i=0,1,....,m. Calculate accuracy each time and take the average.

As the part of report writing,

- Compare and analyze the accuracies obtained using cross validation for k=5, for k=10, for k=20, leave-one-out cross validation and without using any cross validation.

1.3 Ensemble Learning

A decision stump is a decision tree of depth one (i.e., it branches on only one attribute and then makes decision). Implement the discrete AdaBoosting algorithm using ID3 decision stump as the base learner. You should make your code as modular as possible. Namely, your main module of AdaBoosting should treat the base learner as a blackbox and communicate with it via a generic interface that inputs weighted examples and outputs a classifier, which then can classify any instances. For the decision stump, you can modify your ID3 implementation in assignment 1 or implement it from scratch. Use sampling with replacement strategy. Use information-gain as the evaluation criterion. Do not use pruning.

Run the experiments with the breast cancer dataset given for assignment 1 and answer the questions (for report writing).

- Compare and analyze the accuracies obtained by different learners: decision stump alone, boosting with 30 rounds, your ID3 implementation.
- Compare and analyze the accuracies obtained by boosting with different numbers of rounds: 5, 10, 20, 30.

2 50% Progress

For 50% progress, run the experiments and write the report of Semi-supervised learning and performance evaluation.

3 Instructions for Report Writing

1. Your final report will contain the following points:
   - Compare and analyze the average accuracy, precision, recall and F1-score of semi-supervised learning with the ones obtained during the supervised learning.
   - Compare and analyze the accuracies obtained using cross validation for k=5, for k=10, for k=20, leave-one-out cross validation and without using any cross validation.
   - Compare and analyze the accuracies obtained by different learners: decision stump alone, boosting with 30 rounds, your ID3 implementation.
   - Compare and analyze the accuracies obtained by boosting with different numbers of rounds: 5, 10, 20, 30.
2. Never copy the report. Just answer the questions precisely. Make it as simple as possible. Too much description is not needed!

4 Special Instructions

1. 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!
2. The report should be in .docx/.pdf (No hardcopy is required). Write precisely in your own language and keep it as simple as possible.
3. Upload the 50% progress submission in moodle within 6 P.M. of 12th November, 2016 (Friday).
4. Upload the final submission in moodle within 6 P.M. of 19th November, 2016 (Friday).
5. For python and matlab, you may not get supporting softwares in the lab. If you do program in these languages, bring your computer in the sessional.

6. You are allowed to show the assignment in your own laptop during the final submission. But in that case, ensure an internet connection as you have to instantly download your code from the moodle and show it.

5 Instructions for moodle upload

1. Upload the assignment within the specified time. Otherwise, we can't accept it :(
2. For both 50% progress and final submission, use the following rules:
   - If you write code in a single file, then rename it as <Student_id>_<code>. <extension>. For example, if your student id is 1105123 and you have done in java, then your file name should be "1105123_code.java".
   - If you write code in multiple files, then put all the necessary files in a folder and rename it as <Student_id>_<code>. For example, if your student id is 1105123 and you have done in java, then your folder name should be "1105123_code".
   - The report name should be <Student_id>_<report>.<extension>. For example, if your student id is 1105123 and it is in pdf format, then the report name should be "1105123_report.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, "1105123_Java". Then zip it and upload it. Done :)

6 About Version

A version is included with the assignment pdf name. This is because if any information is changed, then the version will be upgraded and the changes will be summarized in this section.