

Programming Assignment #2

The **coding component** and **report component** of this assignment should be **submitted through e-Learning** before **11:59 pm Central Daylight Time on Wednesday, March 27, 2019**. These deadlines are without exceptions unless permission was obtained from the instructor **in advance**.

You may collaborate with one other student, discuss the problems and work through solutions together.

However, your group **must write up your code and solutions on your own**, without copying others' work or letting another student copy your work. In your solution for each problem, you must write down the names of any person with whom you discussed it. This will **not** affect your grade.

Problem: Implement Bagging and AdaBoost based on the decision tree code that you developed in Programming Assignment 1. This code has to be modified to work with ensemble methods. Ensure that your code is modified and organized as follows:

- For this assignment, we will restrict ourselves to **binary trees**. This means that we will restrict ourselves to two splits instead of multiple possible splits at each node. More precisely, each node is restricted to be of the form $x_i = v_k?$, where it tests if the feature x_i takes the value v_k , which could make sure that is a binary tree.
- Implement three functions:** `bagging(x, y, max_depth, num_trees)`, `boosting(x, y, max_depth, num_stumps)` and `predict_example(x, h_ens)`, where `h_ens` is an ensemble of weighted hypotheses. The ensemble is represented as an array of pairs `[(alpha_i, h_i)]`, where each hypothesis and weight are represented by the pair: `(alpha_i, h_i)`. (You can add other self-defined functions if necessary.)

Data Sets: We will use the Mushroom Data Set¹ (attached on eLearning) from the UCI Repository for this assignment. There are 22 attributes in this data set, of which we have dropped the attribute (`stalk-root`) as it contains too many missing values. The data set has been converted from string to integer, with the unique feature values being assigned indices starting from 0 in alphabetical order. Also note that rather than perform the classical task of predicting whether the mushroom is poisonous or edible, our classification task is to predict (`bruises?`).

Experiments: Once you have debugged and tested your code, run the following experiments and write a brief report (**20 points**) answering the following questions:

- (**Bagging**, 30 points) Construct four models for each combination of maximum depth $d = 3, 5$ and bag size ($k = 10, 20$). Report the confusion matrix for these four settings.
- (**Boosting**, 30 points) Construct four models for each combination of maximum depth $d = 1, 2$ and bag size ($k = 20, 40$). Report the confusion matrix for these four settings.
- (**scikit-learn**, 20 points) Use `scikit-learn`'s bagging and AdaBoost learners and repeat the experiments as described in parts (a) and (b) above. Report the confusion matrices for these sets of settings. What can you say about the quality of your implementation's performance versus `scikit`'s performance?

Upload: Make sure all your code, including the (modified) decision tree functionality, is in a single file. Also ensure that your code can be executed by calling the main function. Upload your **code file** and **report** (it should be **PDF**) file through e-Learning before the deadline. **Only one submission for each group.**

¹<https://archive.ics.uci.edu/ml/datasets/mushroom>