Homework #4

The **coding component** of this assignment should be **submitted through e-Learning** before **12:59 pm Central Daylight Time on Monday, April 15, 2019**. The **report component** of this assignment is due **due at the start of class** on **Monday, April 15, 2019**. These deadlines are without exceptions unless permission was obtained from the instructor **in advance**.
You may collaborate with other students, discuss the problems and work through solutions together. However, you **must write up your code and solutions on your own**, without copying another student's 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 Homework 3. This code has to be modified to work with ensemble methods. Ensure that your code is modified and organized as follows:

a. For this assignment, we will continue to restrict ourselves to binary trees. However, the functions to compute entropy, mutual information, id3 and error must be modified in order to take weighted examples. Specifically, make sure your code contains the modified the function headers:

- `def entropy(y, w=None):`
- `def mutual_information(x, y, w=None):`
- `def id3(x, y, attribute_value_pairs=None, max_depth=5, depth=0, w=None):`
- `def compute_error(y_true, y_pred, w=None):`

b. Implement three new 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)`.

Data Sets: We will use the Mushroom Data Set[1] 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 answering the following questions:

a. (**Bagging**, 20 points) Construct four models for each combination of maximum depth $d = 3, 5$ and bag size $(k = 5, 10)$. Report the confusion matrix for these four settings.

b. (**Boosting**, 20 points) Construct four models for each combination of maximum depth $d = 1, 2$ and ensemble size $(k = 5, 10)$. Report the confusion matrix for these four settings.

c. (`scikit-learn`, 40 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 it can be executed by calling the main function. Upload your file through e-Learning before the deadline.

[1] https://archive.ics.uci.edu/ml/datasets/mushroom