C S 487/519 Applied Machine Learning
Ensemble approaches

1 Objective

In this *individual* homework, you are required to understand and utilize ensemble approaches.

2 Requirements

2.1 Tasks

(1) (20 points) AdaBoost algorithm. Given a dataset shown in the first three columns of the table below. Assume that it is at the boosting round i, the weights used in this round are shown in column 4 and the predicted class labels are shown in column 5. Please show the detailed steps to calculate the updated weight that will be used in the boosting round $(i+1)$. You can calculate these values manually using calculators, or by writing a simple program. In the answer of this question, you need to show the steps 2(c), 2(d), 2(e), and 2(f) clearly. Put the detailed calculation and final results to `report.pdf` file.

index	x	y	weight	$\hat{y}$	updated weight
1	1.0	1	0.072	1	
2	2.0	1	0.072	1	
3	3.0	1	0.072	1	
4	4.0	-1	0.072	-1	
5	5.0	-1	0.072	-1	
6	6.0	-1	0.072	-1	
7	7.0	1	0.167	1	
8	8.0	1	0.167	-1	
9	9.0	1	0.167	-1	
10	10.0	-1	0.072	-1	

(2) (35 points) Write classification code by utilizing several ensemble learning approaches: (1) Random forest, (2) Bagging, and (3) AdaBoost.

(3) (20 points) Each classifier needs to be tested using two datasets: (1) the digits dataset offered by scikit-learn library, and (ii) the Mammographic Mass Data Set in the UCI repository. If there are missing values in the dataset, you may want to properly process them.

(4) (20 points) Properly analyze the performance of your ensemble approach. You may want to compare the ensemble approaches with the base classifiers (by reusing results from your previous project). You may also want to test which parameter(s) affect the performance more. Put the analysis in a report.pdf file.

(5) (5 points) Write a readme file `readme.txt` with detailed instructions to run your program.

2.2 Other requirements

- Your Python code should be written for **Python version 3.5.2 or higher**.
- Please write proper **comments** in your code to help the instructor and teaching assistants to understand it.
- Please properly organize your Python code (e.g., create proper classes, modules).
- You can put your code to Jupyter Notebook or a `.py` file.

3 Submission instructions

Put all your files (Python code, readme file, report, etc.) to a zip file named `hw.zip` and upload it to Canvas.

4 Grading criteria

(1) **ZERO point will be given if your code does not work. Please do not submit code that you did not test and make sure it works.**

(2) The score allocation has been put beside the questions.

(3) FIVE points will be deducted if files are not submitted in the required format.

(4) If the total points are more than 100. Your grades will be scaled to the range of [0,100].

(5) Please make sure that you test your code thoroughly by considering all possible test cases. For this homework, your code will NOT be tested using more datasets. Thus, it does not need to be flexible to accept other datasets as input. However, you may not hardcode the datasets in your Python code.