CS578 HW4

November 22, 2018

Total Points: 40
Due date: 11:59pm, December 5, 2018.

1 Boosting

(Points 20) In this problem you will have to use the AdaBoost algorithm to learn a function, mapping examples in $\mathbb{R}^2$ to a boolean value. The space of weak learner considered by the algorithm consists of hypotheses of the form: $x_i > A$ where A is an Integer, and $i = \{1, 2\}$. Run the AdaBoost algorithm for two rounds using the data appearing in the table below, at each round AdaBoost chooses the weak learner that minimizes the error (ϵ). Your answer should consist of :
(1) The weak hypothesis used at each round, and its error
(2) The distribution D_i over the examples for each round
(3) The final hypothesis after running two rounds.

index	x_1	x_2	y
1	1	10	-
2	4	4	-
3	8	7	+
4	5	6	-
5	3	16	-
6	7	7	+
7	10	14	+
8	4	2	-
9	4	10	+
10	8	8	-

2 PAC learning

2.1 Circles

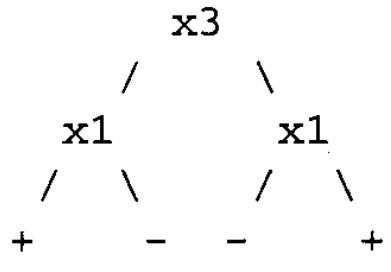
(Points 5) What is the VC-dimension of circles in the plane $\mathbb{R}^2$? I.e., examples are points in $\mathbb{R}^2, C = \{c_{p,r} : p \in \mathbb{R}^2, r \in \mathbb{R}\}$ and $c_{p,r} = 1$ if x is within distance r of p. Or, in words, a legal target function is specified by a circle, and labels any example positive if and only if it lies inside that circle.

2.2 Triangles

(Points 5) What is the VC-dimension of triangles in the plane $\mathbb{R}^2$? I.e., a legal target function is specified by a triangle, and labels any example positive if and only if it lies inside that triangle.

2.3 Trees

Consider the hypothesis class H_{rd2} of "regular, depth-2 decision trees" over n Boolean variables. A "regular, depth-2 decision tree" is a depth-2 decision tree (a tree with four leaves, all distance 2 from the root) in which the left and right child of the root are *required to contain the same variable.* For instance, the following tree is in H_{rd2}.



(a) **(Points 5)** As a function of n, how many syntactically distinct trees are there in H_{rd2}?

(b) **(Points 5)** Give an upper bound for the number of examples needed in the PAC model to learn $H_{rd2}2$ with error ϵ and confidence σ.

3 Report

Answer the above questions on AdaBoost and PAC learning precisely in a file named `report.pdf` and submit it following the instructions in Section 5.

4 Bonus

(20 Points) This task is not mandatory to do. You can do it as a compensation of the points you lost in the previous assignments.

4.1 Implementation of AdaBoost Algorithm:

- Remember the task of Homework 1, prediction of the white wine quality using Decision Tree and KNN. In this task you are required to solve the problem using AdaBoost algorithm where you will be using decision stumps as weak classifiers.
- You will be using the same dataset as Homework 1. You can use your previous implementation of decision tree and modify it to use as decision stumps.
- The decision stump has one level of branching, thus, max depth will be 1. You can remove recursion from your previous code of decision tree implementation or force the max depth to be 1.
- Now implement the AdaBoost algorithm using these decision stumps as weak classifiers.
- You can use same categorization of features as what you did in Homework 1 or can modify it.
- Cross validation is upto you.
- Compare the results of AdaBoost and your previous Decision Tree classifier's performance. Show a table describing the macro and micro F1 scores of both of your classifiers.

4.2 Report

Create a separate report for this additional part and name it `adaboost_report.pdf`. Clearly mention your name and login in the report. This report should contain following things. (Nothing else strictly)

- A table describing the performance of Decision Tree classifier and AdaBoost using Decision stumps.
- A short discussion on the performance.
- Running command for the `adaboost.py` script containing your code for AdaBoost.

4.3 Scripts

Submit a script named `adaboost.py` containing the implementation of AdaBoost. **There is no fixed input/output format. Clearly mention the running command for your code in the report.**

5 Submission Procedure

You are required to use LaTeX to type your solutions to questions, and report of your programming as well. https://www.overleaf.com/ is a website you can use freely as a Purdue student. Other formats of submission will **not** be accepted. A template named "homework_template.tex" is also provided for your convenience.

Your code will be tested on `data.cs.purdue.edu`, where you submit your homework as well. **Make sure that your program runs properly on `data.cs.purdue.edu`.** After logging into `data.cs.purdue.edu` (physically go to the lab or use ssh remotely, as you are all granted the accounts to CS data machines during this class), please follow these steps to submit your assignment:

1. Make a directory named *'yourname_yoursurname'* (all letters in lower case) and copy all of your files there. **Don't put the dataset file `winequality-white.csv`.**
2. While in the upper level directory (if the files are in /homes/dan/dan_goldwasser, go to/homes/dan), execute the following command:

   ```
   turnin -c cs578 -p HW4 *your_folder_name*
   ```

   (e.g. your instructor would use: `turnin -c cs578 -p HW4 dan_goldwasser` to submit his homework)

   Keep in mind that old submissions are overwritten with new ones whenever you execute this command.

3. You can verify the contents of your submission by executing the following command:

   ```
   turnin -v -c cs578 -p HW4
   ```

   Do **not** forget the -v flag here, as otherwise your submission would be replaced with an empty one.

Failure to follow the above instructions will incur the penalty when your homework is being graded.

5.1 Late Policies

As declared in the class.

5.2 Plagiarism Policies

Seriuosly, no cheating. If plagiarism was found, both you and the one whose homework you "referred to" receive 0 point on that homework. If you were found to have plagiarised more than once, we will report to the instructor and the department (and your home department if you are no