CSE 151: Introduction to Machine Learning Winter 2020

Programming Assignment 5

Instructor: Kamalika Chaudhuri **Due on:**

Problem 1: Programming Assignment: 20 points

In this assignment, we will look at the task of spam classification using boosting. Our raw data is a set of emails, which were collected from a liguistics mailing list; the emails are labeled as spam or not spam. For your benefit, we have already preprocessed the emails to remove stop-words, punctuation, and to do some preliminary preprocessing that lemmatises the words (for example, that maps words such as *include*, *includes* and *included* to the same word), and converted them to vectors of features.

Download files `pa5train.txt`, `pa5test.txt` and `pa5dictionary.txt` from the class website. The first two files contain your training and test datasets respectively. The third file is a dictionary and contains a list of words. Each line in the files `pa5train.txt` and `pa5test.txt` correspond to an email followed a label which can be 1 or -1. An email is represented by a feature vector of length 4003; a label 1 indicates that the email is a spam message, and a label -1 indicates that it is not spam. Coordinate i of the feature vector corresponding to an email is 1 when word i in `pa5dictionary.txt` is present in the email and 0 otherwise.

1. Write down the training and test errors of the classifiers obtained after $t = 3, 7, 10, 15, 20$ rounds of boosting. Use the following weak learning procedure. Each weak learner corresponds to a classifier $h_{i,+}$ or $h_{i,-}$, where i is a word in the dictionary and the classifier $h_{i,+}$ is the rule:

$$\begin{aligned} h_{i,+}(x) = &\quad 1, \quad \text{if word i occurs in email x} \\ = &\quad -1, \quad \text{otherwise} \end{aligned}$$

   Similarly, the classifier $h_{i,-}$ is the rule:

$$\begin{aligned} h_{i,-}(x) = &\quad 1, \quad \text{if word i does not occur in email x} \\ = &\quad -1, \quad \text{otherwise} \end{aligned}$$

   The set of weak learners C is the collection of such classifiers for all i, and your weak learning procedure should select the weak learner which has the *highest accuracy* in C with respect to the current weighted set of examples.

2. Based on the dictionary file, write down the words corresponding to the weak learners chosen in the first 10 rounds of boosting.

[Hint: If your code is correct, you should get a training error of 0.051 and a test error of 0.039 after 4 rounds of boosting.]

Solution

1.

Round	Training	Test
3	0.064	0.039
7	0.029	0.031
10	0.016	0.039
15	0	0.023
20	0	0.023

2.

Round	Word
1	remove
2	language
3	free
4	university
5	money
6	linguistic
7	click
8	fax
9	want
10	de