CSE 151: Introduction to Machine Learning Winter 2017

Problem Set 6

Instructor: Kamalika Chaudhuri **Due on:** Mar 17, 2017

Instructions

- This is a 40 point homework.
- Problem 3 is a programming assignment. For this problem, you are free to use any programming language you wish.
- To submit your code, please send email your code to cse151homeworks@gmail.com. **Please submit your solution to Problem 3 with the homework, and email only the code to this address.**

Problem 1: 6 points

In class, we saw that if we have two labels, then the error of a classifier which guesses completely randomly is 0.5. In this problem, we look at what happens when there are $k > 2$ labels.

1. Random guesser Geser knows that there are k labels, and for each example, selects a label out of $\{1, \dots, k\}$ uniformly at random. What is the error of Geser ?
2. Now suppose we have a more sophisticated random guesser Zavulon who knows that w_1 fraction of the data distribution has label 1, w_2 fraction has label 2, and so on. For each example, Zavulon also selects a label out of $\{1, \dots, k\}$ at random, but he selects label 1 with probability w_1, label 2 with probability w_2 and so on. What is the error of Zavulon?

Problem 2: 14 points

Consider the following two data distributions D_1 and D_2 over labeled examples. There is a single feature, denoted by X which takes values in the set $\{1, 2, 3, 4\}$ and a binary label $Y \in \{0, 1\}$. D_1 is described as follows:

$$\begin{aligned}\Pr(X = i) &= \frac{1}{4}, \;\; i \in \{1, 2, 3, 4\} \\ \Pr(Y = 1|X = i) &= 1, \;\; i \in \{1, 4\} \\ \Pr(Y = 0|X = i) &= 1, \;\; i \in \{2, 3\}\end{aligned}$$

D_2 is described as follows.

$$\begin{aligned}\Pr(X = i) &= \frac{1}{4}, \;\; i \in \{1, 2, 3, 4\} \\ \Pr(Y = 1|X = i) &= \frac{i}{10}, \;\; i \in \{1, 2, 3, 4\}\end{aligned}$$

1. Consider the following classifier h: $h(x) = 1$ if $x > 1.5$ and 0 otherwise. What is the true error of h when the true data distribution is D_1?
2. Suppose our concept class C is the set of all classifiers of the form: $h_t(x) = 1$ if $x > t$ and 0 otherwise. Write down a classifier in this concept class that minimizes the true error when the data distribution is D_1. What is the true error of this classifier? Do we have a non-zero bias when the concept class is C and the data distribution is D_1?
3. Repeat parts (1) and (2) for the data distribution D_2.

Problem 3: 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 `hw6train.txt`, `hw6test.txt` and `hw6dictionary.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 `hw6train.txt` and `hw6test.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 `hw6dictionary.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:

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

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

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

   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.]