Instructions

- This is a 40 point homework.
- Problem 4 is a programming assignment. You can use any programming language you like for this assignment.

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,...,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,...,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?

Solution

1. Since Geser selects a label uniformly at random, for a sample $x$, the probability that the selected label is the correct one is $\frac{1}{k}$, and the probability that it is incorrect is $1 - \frac{1}{k}$. Thus, Geser’s error is $1 - \frac{1}{k}$.

2. Zavulon’s guessing process is equivalent to the following probabilistic process: first draw a sample $x$ from the data distribution, and then draw a random variable $Y$, which is 1 w.p. $w_1$, 2 w.p. $w_2$, and so on. Let $E_i$ be the event that the true label of a selected sample is $i$; thus $\Pr(E_i) = w_i$. Thus Zavulon’s error is:

$$\sum_{i=1}^{k} \Pr(E_i) \Pr(Y \neq i|E_i) = \sum_{i=1}^{k} w_i(1 - w_i) = 1 - \sum_{i=1}^{k} w_i^2$$

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:

$$\Pr(X = i) = \frac{1}{4}, \ i \in \{1,2,3,4\}$$

$$\Pr(Y = 1|X = i) = 1, \ i \in \{1,4\}$$

$$\Pr(Y = 0|X = i) = 1, \ i \in \{2,3\}$$

$D_2$ is described as follows.

$$\Pr(X = i) = \frac{1}{4}, \ i \in \{1,2,3,4\}$$

$$\Pr(Y = 1|X = i) = \frac{i}{10}, \ i \in \{1,2,3,4\}$$
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 \( D_2 \).

Solution

1. According to the data distribution \( D_1 \), the true label \( Y = 1 \) for \( X = 1 \) and \( X = 4 \) whereas \( Y = 0 \) for \( X = 2 \) and \( X = 3 \). The classifier \( h(x) \) predicts the label 0 for \( x = 1 \) and label 1 otherwise. Therefore, \( h(x) \) makes mistakes for \( x \in \{1, 2, 3\} \). Hence, the true error of the classifier \( h(x) \) is given as follows.

\[
\text{True error} = \sum_{i=1}^{4} \Pr(X = i) \Pr(h(i) \neq Y | X = i)
\]

\[
= \Pr(X = 1) \Pr(Y \neq 0 | X = 1) + \Pr(X = 2) \Pr(Y \neq 1 | X = 2) + \Pr(X = 3) \Pr(Y \neq 1 | X = 3)
+ \Pr(X = 4) \Pr(Y \neq 1 | X = 4)
\]

\[
= \Pr(X = 1) \Pr(Y = 1 | X = 1) + \Pr(X = 2) \Pr(Y = 0 | X = 2) + \Pr(X = 3) \Pr(Y = 0 | X = 3)
+ \Pr(X = 4) \Pr(Y = 0 | X = 4)
\]

\[
= \frac{1}{4} \times 1 + \frac{1}{4} \times 1 + \frac{1}{4} \times 1 + \frac{1}{4} \times 0
\]

\[
= \frac{1}{2}
\]

2. Since we know that the random variable \( X \) can only take integer values from 1 to 4, we can easily come up with 5 classifiers that represent the entire concept class \( C \) such that one of the classifiers will classify all points as 1, the second classifier would classify \( x = 1 \) as 0 and all the rest as 1, the third would classify \( x = 1 \) and \( x = 2 \) as 0 and all the rest as 1, and so on. In order to do this, we vary \( t \) over the set \( \{0, 1, 2, 3, 4\} \) and observe the true errors of each of these 5 classifiers. We calculate the true error for each of these classifiers in a way similar to Problem 2.1.

For \( t = 0 \), \( h_0(x) = 1 \) if \( x > 0 \) and 0 otherwise.

\[
\text{True error } \epsilon_0 = \Pr(X = 1) \Pr(Y = 0 | X = 1) + \Pr(X = 2) \Pr(Y = 0 | X = 2) + \Pr(X = 3) \Pr(Y = 0 | X = 3)
+ \Pr(X = 4) \Pr(Y = 0 | X = 4)
\]

\[
= \frac{1}{4} \times 0 + \frac{1}{4} \times 1 + \frac{1}{4} \times 1 + \frac{1}{4} \times 0
\]

\[
= \frac{1}{2}
\]

For \( t = 1 \), \( h_1(x) = 1 \) if \( x > 1 \) and 0 otherwise.

\[
\text{True error } \epsilon_1 = \Pr(X = 1) \Pr(Y = 1 | X = 1) + \Pr(X = 2) \Pr(Y = 0 | X = 2) + \Pr(X = 3) \Pr(Y = 0 | X = 3)
+ \Pr(X = 4) \Pr(Y = 0 | X = 4)
\]

\[
= \frac{1}{4} \times 1 + \frac{1}{4} \times 1 + \frac{1}{4} \times 1 + \frac{1}{4} \times 0
\]

\[
= \frac{3}{4}
\]
For \( t = 2 \), \( h_2(x) = 1 \) if \( x > 2 \) and 0 otherwise.

True error \( \epsilon_2 = \Pr(X = 1) \Pr(Y = 1|X = 1) + \Pr(X = 2) \Pr(Y = 1|X = 2) + \Pr(X = 3) \Pr(Y = 0|X = 3) \\
+ \Pr(X = 4) \Pr(Y = 0|X = 4) \\
= \frac{1}{4} \times 1 + \frac{1}{4} \times 0 + \frac{1}{4} \times 1 + \frac{1}{4} \times 0 \\
= \frac{1}{2} \\
For \( t = 3 \), \( h_3(x) = 1 \) if \( x > 3 \) and 0 otherwise.

True error \( \epsilon_3 = \Pr(X = 1) \Pr(Y = 1|X = 1) + \Pr(X = 2) \Pr(Y = 1|X = 2) + \Pr(X = 3) \Pr(Y = 1|X = 3) \\
+ \Pr(X = 4) \Pr(Y = 0|X = 4) \\
= \frac{1}{4} \times 1 + \frac{1}{4} \times 0 + \frac{1}{4} \times 0 + \frac{1}{4} \times 0 \\
= \frac{1}{4} \\
For \( t = 4 \), \( h_4(x) = 1 \) if \( x > 4 \) and 0 otherwise.

True error \( \epsilon_4 = \Pr(X = 1) \Pr(Y = 1|X = 1) + \Pr(X = 2) \Pr(Y = 1|X = 2) + \Pr(X = 3) \Pr(Y = 1|X = 3) \\
+ \Pr(X = 4) \Pr(Y = 1|X = 4) \\
= \frac{1}{4} \times 1 + \frac{1}{4} \times 0 + \frac{1}{4} \times 0 + \frac{1}{4} \times 1 \\
= \frac{1}{2} \\
Since \( t = 3 \) achieves the least true error, \( h_3(x) \) is the classifier which minimizes the true error. The true error of this classifiers is \( \frac{1}{4} \). Also, the bias of the concept class \( C \) is \( \frac{1}{4} \) since this is the least error achieved by any classifier in \( C \) when the data distribution is \( D_1 \). Thus, we have a non-zero bias for the concept class \( C \) and the true distribution \( D_1 \).

3. (a) Similar to Problem 2.1, the true error of the classifier \( h(x) \) when the data distribution is \( D_2 \) is given as follows.

True error \( \sum_{i=1}^{4} \Pr(X = i) \Pr(h(i) \neq Y|X = i) \\
= \Pr(X = 1) \Pr(Y \neq 0|X = 1) + \Pr(X = 2) \Pr(Y \neq 1|X = 2) + \Pr(X = 3) \Pr(Y \neq 1|X = 3) \\
+ \Pr(X = 4) \Pr(Y \neq 1|X = 4) \\
= \frac{1}{4} \times \frac{1}{10} + \frac{1}{4} \left( 1 - \frac{2}{10} \right) + \frac{1}{4} \left( 1 - \frac{3}{10} \right) + \frac{1}{4} \left( 1 - \frac{4}{10} \right) \\
= \frac{11}{20} \\
(b) Similar to Problem 3.2, we vary \( t \) over the set \( \{0, 1, 2, 3, 4\} \) and observe the true errors for these 5 classifiers.
For $t = 0$, $h_0(x) = 1$ if $x > 0$ and 0 otherwise.

True error $\epsilon_0 = \Pr(X = 1) \Pr(Y = 0|X = 1) + \Pr(X = 2) \Pr(Y = 0|X = 2) + \Pr(X = 3) \Pr(Y = 0|X = 3)
+ \Pr(X = 4) \Pr(Y = 0|X = 4)$
$$= \frac{1}{4} \left( 1 - \frac{1}{10} \right) + \frac{1}{4} \left( 1 - \frac{2}{10} \right) + \frac{1}{4} \left( 1 - \frac{3}{10} \right) + \frac{1}{4} \left( 1 - \frac{4}{10} \right)$$
$$= \frac{3}{4}$$

For $t = 1$, $h_1(x) = 1$ if $x > 1$ and 0 otherwise.

True error $\epsilon_1 = \Pr(X = 1) \Pr(Y = 1|X = 1) + \Pr(X = 2) \Pr(Y = 0|X = 2) + \Pr(X = 3) \Pr(Y = 0|X = 3)
+ \Pr(X = 4) \Pr(Y = 0|X = 4)$
$$= \frac{1}{4} \times \frac{1}{10} + \frac{1}{4} \times \frac{1}{10} + \frac{1}{4} \left( 1 - \frac{3}{10} \right) + \frac{1}{4} \left( 1 - \frac{4}{10} \right)$$
$$= \frac{11}{20}$$

For $t = 2$, $h_2(x) = 1$ if $x > 2$ and 0 otherwise.

True error $\epsilon_2 = \Pr(X = 1) \Pr(Y = 1|X = 1) + \Pr(X = 2) \Pr(Y = 1|X = 2) + \Pr(X = 3) \Pr(Y = 0|X = 3)
+ \Pr(X = 4) \Pr(Y = 0|X = 4)$
$$= \frac{1}{4} \times \frac{1}{10} + \frac{1}{4} \times \frac{2}{10} + \frac{1}{4} \times \frac{3}{10} + \frac{1}{4} \left( 1 - \frac{4}{10} \right)$$
$$= \frac{2}{5}$$

For $t = 3$, $h_3(x) = 1$ if $x > 3$ and 0 otherwise.

True error $\epsilon_3 = \Pr(X = 1) \Pr(Y = 1|X = 1) + \Pr(X = 2) \Pr(Y = 1|X = 2) + \Pr(X = 3) \Pr(Y = 1|X = 3)
+ \Pr(X = 4) \Pr(Y = 0|X = 4)$
$$= \frac{1}{4} \times \frac{1}{10} + \frac{1}{4} \times \frac{2}{10} + \frac{1}{4} \times \frac{3}{10} + \frac{1}{4} \times \frac{4}{10}$$
$$= \frac{3}{10}$$

For $t = 4$, $h_4(x) = 1$ if $x > 4$ and 0 otherwise.

True error $\epsilon_4 = \Pr(X = 1) \Pr(Y = 1|X = 1) + \Pr(X = 2) \Pr(Y = 1|X = 2) + \Pr(X = 3) \Pr(Y = 1|X = 3)
+ \Pr(X = 4) \Pr(Y = 1|X = 4)$
$$= \frac{1}{4} \times \frac{1}{10} + \frac{1}{4} \times \frac{2}{10} + \frac{1}{4} \times \frac{3}{10} + \frac{1}{4} \times \frac{4}{10}$$
$$= \frac{1}{4}$$

Since $t = 4$ achieves the least true error, $h_4(x)$ is the classifier which minimizes the true error. The true error of this classifier is $\frac{1}{4}$. Also, the bias of the concept class $C$ is $\frac{1}{4}$ since this is the least error achieved by any classifier in $C$ when the data distribution is $D_2$. Thus, we have a non-zero bias for the concept class $C$ and the true 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 linguistics 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:

\[
h_{i,+}(x) = 1, \quad \text{if word } i \text{ occurs in email } x
\]

\[
= -1, \quad \text{otherwise}
\]

Similarly, the classifier \(h_{i,-}\) is the rule:

\[
h_{i,-}(x) = 1, \quad \text{if word } i \text{ does not occur in email } x
\]

\[
= -1, \quad \text{otherwise}
\]

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

<table>
<thead>
<tr>
<th>Round</th>
<th>Training</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.064</td>
<td>0.039</td>
</tr>
<tr>
<td>7</td>
<td>0.029</td>
<td>0.031</td>
</tr>
<tr>
<td>10</td>
<td>0.016</td>
<td>0.039</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>0.023</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0.023</td>
</tr>
</tbody>
</table>