Homework number 2.

Theory question I: Consider the following variant of the Winnow algorithm, WINNOW2, that has two parameters $\theta \geq 1$ and $\beta \geq 1$. WINNOW2 learns weights $w_1, \ldots, w_n$ and classifies an example using $\text{sign}(\sum_{i=1}^n w_i x_i - \theta)$. (You can assume that $\theta >> 1$.) The algorithm’s response to mistakes:

<table>
<thead>
<tr>
<th>Name</th>
<th>Update Scheme</th>
<th>target</th>
<th>prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demotion</td>
<td>$\forall x_i = 1$ set $w_i = w_i/\beta$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Promotion</td>
<td>$\forall x_i = 1$ set $w_i = \beta \cdot w_i$</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Note that the variables with $x_i = 0$ do not change their weight. Also, the initial weights are 1, i.e., $w_i = 1$.

1. Let $u$ be the number of promotion steps, and let $v$ be the number of demotion steps. Show that $v \leq \frac{\beta}{\beta-1} \cdot n \theta + \beta \cdot u$.

2. Show that for all $i$, $w_i \leq \beta \cdot \theta$.

3. Assume there exists a hyperplane $(\mu_1, \ldots, \mu_n)$, $\mu_i \geq 0$ with a margin of $0 < \gamma \leq 1$, i.e., $\sum_{i=1}^n \mu_i x_i > 1$ and for any negative point $\sum_{i=1}^n \mu_i x_i < 1 - \gamma$.

If we run WINNOW2 with $\beta = 1 + \frac{\gamma}{2}$ and $\theta \geq 1$, then the number of mistakes is bounded by:

$$O \left( \frac{1}{\gamma^2} \cdot \frac{n}{\theta} + \frac{\log \theta}{\gamma^2} \cdot \sum_{i=1}^n \mu_i \right)$$

(You can show first that $\log \prod_{h=1}^n w_i^{\mu_i} = \sum_{i=1}^n \mu_i \log w_i \geq (u - (1 - \gamma)v) \log \beta$.)

Theory question II: Let the error of a hypothesis $h$ during the time interval $[\tau_1, \tau_2]$ be $\text{loss}(h, \tau_1, \tau_2) = \sum_{t=\tau_1}^{\tau_2} \ell_h^t$, where $\ell_h^t \in [0, 1]$ is the loss of $h$ at time $t$. Let the intervals regret of algorithm $A$ be $R = \max_{\tau_1 \leq \tau_2} \max_{h \in H} \{\text{loss}(A, \tau_1, \tau_2) - \text{loss}(h, \tau_1, \tau_2)\}$. The goal is to design an algorithm that will have a low regret for any interval.

Consider the following algorithm, which has $\beta \in (0, 1)$ as a parameter. For each hypothesis $h$ and time interval $[\tau_1, \tau_2]$ we maintain a weight $w_{h,\tau_1,\tau_2}$. At time $t$ we update the weights of $[\tau_1, \tau_2]$, such that $t \in [\tau_1, \tau_2]$, using the rule $w_{h,\tau_1,\tau_2}^{t+1} = w_{h,\tau_1,\tau_2}^t \beta^{\ell_h^t - \beta \ell_A^t}$, where $\ell_A^t$ is the loss of our online algorithm $A$ at time $t$, i.e., $\sum_h p_h^t \ell_h^t$. (Initially, $w_{h,\tau_1,\tau_2}^0 = 1$.)

At time $t$ we define $w_h^t = \sum_{\tau=\tau_1}^t \sum_{\tau_2=\tau}^{\tau_2} w_{h,\tau_1,\tau_2}^t$, $W^t = \sum_{h \in H} w_h^t$ and $p_h^t = w_h^t/W^t$. Our distribution over $H$ at time $t$ is $p_h^t$. 

1
1. Show that at any time $t$ we have $0 \leq \sum_{h,\tau_1,\tau_2} w_{h,\tau_1,\tau_2} \leq |H| \cdot |I|$, where $I$ is the set of all intervals $[\tau_1, \tau_2]$. (Note that for any $\beta \in [0, 1]$ and $x \in [0, 1]$ we have $\beta^x \leq 1 - (1 - \beta)x$ and $\beta^{-x} \leq 1 + (1 - \beta)x/\beta$. You can rewrite in the update rule $\beta^{(\ell^t_h - \beta \ell^t_A)} = \beta^{\ell^t_h - \beta \ell^t_A}$)

2. Show that the intervals regret $R$ is $O(\sqrt{T \log(|I| \cdot |H|)})$, where $T$ is the total number of time steps. (You need to set the parameter $\beta$.)

Theory question III:

1. Show in AdaBoost, that the error of $h_t$ on $D_{t+1}$ is exactly $1/2$ (assuming that $\epsilon_t \neq 0$).

2. Suppose that we decided to simplify Adaboost by setting $\alpha_t = \alpha$ for a fixed value $\alpha$ on each round. Assume that the weak learning assumption will always hold, so for every $t$, $1 \leq t \leq T$, we have $\epsilon_t \leq 1/2 - \gamma$. What is the best value of $\alpha$ to use (as a function of the sequence of $\epsilon_t$)? What is the guaranteed performance?

   Solving the optimization exactly is too difficult. Give an expression whose solution is the fixed optimal $\alpha$, show that there is a unique solution for $\alpha$, and show that if $\epsilon_t \leq 1/2 - \gamma$ then the fixed optimal $\alpha$ is in the range $\alpha \in [0, \frac{1}{2} \ln \frac{0.5 + \gamma}{0.5 - \gamma}]$.

Programming assignment:

Write a program to implement the AdaBoost algorithm. Run the program on the mnist data set at the home page of the course. (The data has handwritten digits of the number "4" and number "7").

File format: Each line is a different image. The images are $28 \times 28$ with each entry having a gray level. This implies that an image has 784 entries. There are 1000 training examples and 200 test examples. The training examples are in $X_{\text{train}}$ and their labels are in $Y_{\text{train}}$. The testing examples are in $X_{\text{test}}$ and their labels are in $Y_{\text{test}}$. Weak learner: A weak learner will correspond to a single pixel in the image and a threshold value, i.e., $x_{i,j} > \theta$. Select at each iteration the best weak learner available.

Your experiments and the graphs you will output should address the following questions:

1. What happens when the number of iterations increases (both in the training error and test error).

2. What happen when the size of the training set increases (you can use a random subset of the training set).

3. How does the distribution changes through different iterations (qualitatively).

The homework is due in two weeks.