Homework 2: Nov 28,2010

Lecturer: Yishay Mansour

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 $sign(\sum_{i=1}^{n} w_i x_i - \theta)$. (You can assume that $\theta >> 1$.) The algorithm's response to mistakes:

Name	Update Scheme	target	prediction
Demotion	$\forall x_i = 1 \;\; set \;\; w_i = w_i/\beta$	0	1
Promotion	$\forall x_i = 1 \;\; set \;\; w_i = \beta \cdot w_i$	1	0

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

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., or any positive point $\sum_{i=1}^{n} \mu_i x_i > 1$ 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 $\sum_{i=1}^{n} \mu_i \log w_i \geq (\mathbf{u} - (1-\gamma)\mathbf{v}) \log\beta$.)

Theory question II: Let the error of a hypothesis h during the time interval $[\tau_1, \tau_2]$ be $loss(h, \tau_1, \tau_2) = \sum_{t=\tau_1}^{\tau_2} \ell_h^t$, where $\ell_h^t \in [0, 1]$ is the loss of h at time t. Let the intervals regret of algorithm A be $R = \max_{\tau_1 \leq \tau_2} \max_{h \in H}\{loss(A, \tau_1, \tau_2) - loss(h, \tau_1, \tau_2)\}$. 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,τ_1,τ_2}^t. 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 ℓ_A^t is the loss of our online algorithm A at time t. (Initially, $w_{h,\tau_1,\tau_2}^0 = 1$.)

At time t we define $w_h^t = \sum_{\tau_1=1}^{t} \sum_{\tau_2=t}^{T} w_{h,\tau_1,\tau_2}^t$, $W^t = \sum_{h \in H} w_h^t$ and $p_h^t = w_h^t / W^t$. Our distribution over H at time t is p^t.

1. Show that at any time t we have $0 \leq \sum_{h,\tau_1,\tau_2} w^t_{h,\tau_1,\tau_2} \leq |H| \cdot |I|$, 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 $\beta^{-x} \leq 1 + (1-\beta)x/\beta$.)

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 β.)

Programming assignment:

You will need to write the programming assignments in the `R` (You can download it for free, see http://www.r-project.org/) or in MATLAB.

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 28×28 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_train` and their labels are in `Y_train`. The testing examples are in `X_test` and their labels are in `Y_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).

READE FIRST THE SUBMISSION GUIDELINES AT THE COURSE HOMEPAGE !

The homework is due in two weeks