COS 511: Theoretical Machine Learning

Homework #4 — Due: March 13, 2008
Boosting

Special late policy: Because this homework is due during midterms week and right before break, there will be a special late policy in which the Friday, Saturday and Sunday following the deadline count as a single late "day." To be clear, this table shows how many "days" late your assignment will be counted if turned in on the following dates:

Calendar date	**Number of late days charged**
Thursday, March 13	0
Sunday, March 16	1
Monday, March 17	2
Tuesday, March 18	3
Wednesday, March 19	4
Thursday, March 20	5
Friday, March 21	*Not accepted for credit*

For this homework only, if you are out of the Princeton area over break, you may mail or email your assignment to me. If mailed, your homework is considered submitted on the post mark date, and should be sent to this address: Robert Schapire, Princeton University, Department of Computer Science, 35 Olden Street, Princeton, NJ 08540. (It would be wise to send me email at the same time you mail your assignment so that I can look out for it; also, save a photocopy of your work.)

Problem 1

Consider a variant of AdaBoost in which we set $\alpha_t = \alpha$ on every round of boosting, where $\alpha > 0$ is a fixed parameter that is set ahead of time. Call this algorithm Boost(α).

Assume that on every round t of boosting, it is known ahead of time that ϵ_t will be at most $1/2 - \gamma$, for some number $\gamma > 0$. Suppose that we set

$$\alpha = \tfrac{1}{2} \ln \left(\frac{1 + 2\gamma}{1 - 2\gamma} \right).$$

a. [10] Show how to modify the training-error analysis of AdaBoost to derive an upper bound on the training error of the final hypothesis H produced by Boost(α) after T rounds. Your bound should be in terms of γ and T only (and should not depend on α, ϵ_t, etc.).

b. [5] Use your result in part (a) to show that the final hypothesis H will be consistent with m training examples (i.e., have zero training error) after T rounds if

$$T > \frac{\ln m}{2\gamma^2}.$$

c. [10] Assume that the weak learning algorithm generates hypotheses h_t which belong to a finite class $\mathcal{H}$. Use general-purpose error bounds proved in this class to show that if we choose T as in part (b), then with probability $1 - \delta$, the generalization error of H is at most

$$\frac{T \ln |\mathcal{H}| + \ln(1/\delta)}{m}.$$

Problem 2

Let $X = \{-1, +1\}^n$. For $\mathbf{x} \in X$, let $x(i)$ denote the ith component of $\mathbf{x}$ so that $\mathbf{x} = \langle x(1), x(2), \ldots, x(n) \rangle$.

Let $\mathcal{M}_k$ be the set of concepts $c : X \to \{-1, +1\}$ for which there exist $i_1, \ldots, i_k \in \{1, \ldots, n\}$ (not necessarily distinct) such that

$$c(\mathbf{x}) = \text{sign}\left(\sum_{j=1}^{k} x(i_j)\right) \tag{1}$$

for all $\mathbf{x} \in X$. (As usual, we define sign(0) = 0. Note that this implies that, if $c \in \mathcal{M}_k$, then $\sum_j x(i_j)$ cannot be equal to 0 for any $\mathbf{x}$ since $c(\mathbf{x}) \in \{-1, +1\}$.)

Roughly speaking, in this problem, you will show that a concept c is γ-weakly learnable by the features $x(1), \ldots, x(n)$ for $\gamma = \Omega(1/\text{poly}(n))$ if and only if c is in $\mathcal{M}_k$ for $k = \text{poly}(n)$. (The "if" direction is shown in part (c); the "only if" direction is shown in part (d).)

a. [5] For any concept $c : X \to \{-1, +1\}$ and distribution D on X, show that

$$\text{E}_{\mathbf{x} \sim D}\left[c(\mathbf{x})x(i)\right] = 1 - 2\text{Pr}_{\mathbf{x} \sim D}\left[c(\mathbf{x}) \neq x(i)\right].$$

b. [5] Let c be as in Eq. (1). Argue that

$$\sum_{j=1}^{k} \text{E}_{\mathbf{x} \sim D}\left[c(\mathbf{x})x(i_j)\right] \geq 1$$

for every distribution D on X.

c. [10] Let $c \in \mathcal{M}_k$. Use parts (a) and (b) to show that for every distribution D on X, there exists an index $i \in \{1, \ldots, n\}$ such that

$$\text{Pr}_{\mathbf{x} \sim D}\left[x(i) \neq c(\mathbf{x})\right] \leq \frac{1}{2} - \frac{1}{2k}.$$

d. [10] Consider a weak learning algorithm A that works as follows: Given a training set $\langle (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m) \rangle$, and a distribution D_t over the examples, A outputs the hypothesis $h_t(\mathbf{x}) = x(i^*)$ where i^* has minimum error with respect to the examples and distribution. That is,

$$i^* = \arg\min_{1 \leq i \leq n} \sum_{j : x_j(i) \neq y_j} D_t(j).$$

Let $c : X \to \{-1, +1\}$ be *any* concept, and let $\gamma > 0$. Suppose that, for every distribution D on X, there exists an index i such that

$$\text{Pr}_{\mathbf{x} \sim D}\left[x(i) \neq c(\mathbf{x})\right] \leq \tfrac{1}{2} - \gamma.$$

Use your analysis of Boost(α) and the weak learner described above to show that $c \in \mathcal{M}_k$ if

$$k > \frac{n \ln 2}{2\gamma^2}.$$

Problem 3 – Extra Credit

[15] In class, we showed how a weak learning algorithm that uses hypotheses from a space $\mathcal{H}$ of bounded cardinality can be converted into a strong learning algorithm. This result can be generalized to weak hypothesis spaces of bounded VC-dimension. However, strictly speaking, the definition of weak learnability does *not* include such restrictions on the weak hypothesis space. The purpose of this problem is to show that weak and strong learnability are equivalent, even without these restrictions.

Let $\mathcal{C}$ be a concept class on domain X. Let A_0 be a weak learning algorithm and let $\gamma > 0$ be a (known) constant such that, for $\delta > 0$, for every concept $c \in \mathcal{C}$ and for every distribution D on X, when given $m_0 = \text{poly}(1/\delta)$ random examples x_i from D, each with its label $c(x_i)$, A_0 outputs a hypothesis h such that, with probability at least $1 - \delta$,

$$\Pr_{x \in D}[h(x) \neq c(x)] \leq \tfrac{1}{2} - \gamma.$$

Note that no restrictions are made on the form of h, or on the cardinality or VC-dimension of the space from which it is chosen.

Show that A_0 can be converted into a strong learning algorithm using boosting. That is, construct an algorithm A such that, for $\epsilon > 0$, $\delta > 0$, for every concept $c \in \mathcal{C}$ and for every distribution D on X, when given $m = \text{poly}(m_0, 1/\epsilon, 1/\delta, 1/\gamma)$ random examples x_i from D, each with its label $c(x_i)$, A outputs a hypothesis H such that, with probability at least $1 - \delta$,

$$\Pr_{x \in D}[H(x) \neq c(x)] \leq \epsilon.$$

Be sure to show that the number of examples needed by this algorithm is polynomial in m_0, $1/\epsilon$, $1/\delta$ and $1/\gamma$.