COS 511: Theoretical Machine Learning

Homework #6 — Due: April 17, 2008

Winnow and Widrow-Hoff

Problem 1

In class, we discussed a version of the winnow algorithm that makes few mistakes when examples $\mathbf{x}, y$ are such that $y(\mathbf{u} \cdot \mathbf{x}) > 0$ for some unknown vector $\mathbf{u}$. Effectively, the inner product $\mathbf{u} \cdot \mathbf{x}$ is being compared to the threshold 0 to determine $\mathbf{x}$'s classification. In this problem, we will consider the case in which some threshold other than 0 is to be used. Thus, we now suppose that examples are such that

$$y(\mathbf{u} \cdot \mathbf{x} - b) > 0$$

for some known threshold $b \in \mathbb{R}$, and some unknown vector $\mathbf{u}$.

To be more precise, as in class, assume $\mathbf{x}_t \in [-1, +1]^N$ and $y_t \in \{-1, +1\}$. Assume further that there exists $\delta > 0$, $\mathbf{u} \in [0, 1]^N$ with $||\mathbf{u}||_1 = 1$ such that

$$y_t(\mathbf{u} \cdot \mathbf{x}_t - b) \geq \delta$$

where $b \in \mathbb{R}$ is known. To learn, we use the following variant of winnow: Initially, $w_{1,i} = 1/N$ (as usual). On each round t, if $y_t(\mathbf{w}_t \cdot \mathbf{x}_t - b) > 0$ (no mistake), then we do nothing (i.e., $\mathbf{w}_{t+1} = \mathbf{w}_t$). Otherwise, we update $\mathbf{w}_t$ as follows:

$$\begin{aligned} \text{if } y_t = +1 \text{ then} \quad w_{t+1,i} &= \frac{w_{t,i} \exp\left(\overline{\eta} x_{t,i}\right)}{Z_t} \\ \text{if } y_t = -1 \text{ then} \quad w_{t+1,i} &= \frac{w_{t,i} \exp\left(-\underline{\eta} x_{t,i}\right)}{Z_t} \end{aligned}$$

where Z_t is a normalization constant, and where $\overline{\eta} > 0$ and $\underline{\eta} > 0$ are parameters of the algorithm.

Let $\overline{m}$ and $\underline{m}$ be the number of mistakes made by this algorithm on rounds on which $y_t = +1$ and $y_t = -1$ respectively. Thus, $\overline{m} + \underline{m}$ is the total number of mistakes.

a. [12] Use a potential argument as in class to prove that

$$\overline{m}\,\overline{C} + \underline{m}\,\underline{C} \leq \ln N$$

where

$$\begin{aligned} \overline{C} &= \overline{\eta}(\delta + b) - \ln\left[\frac{e^{\overline{\eta}} + e^{-\overline{\eta}}}{2} + \frac{e^{\overline{\eta}} - e^{-\overline{\eta}}}{2} b\right] \\ \underline{C} &= \underline{\eta}(\delta - b) - \ln\left[\frac{e^{\underline{\eta}} + e^{-\underline{\eta}}}{2} - \frac{e^{\underline{\eta}} - e^{-\underline{\eta}}}{2} b\right] \end{aligned}$$

b. [8] Show how to choose $\overline{\eta}$ and $\underline{\eta}$ as functions of δ and b to prove that

$$\overline{m}\,\mathrm{RE}\left(\frac{1+b+\delta}{2} \| \frac{1+b}{2}\right) + \underline{m}\,\mathrm{RE}\left(\frac{1+b-\delta}{2} \| \frac{1+b}{2}\right) \leq \ln N.$$

c. [5] Suppose $\mathbf{x}_t \in \{-1, +1\}^N$ and that there exists a set of indices $S \subseteq \{1, \ldots, N\}$ such that $y_t = +1$ if and only if $x_{t,i} = +1$ for at least one of the indices $i \in S$. In other words, y_t is a disjunction of the variables indexed by S. Assume $k = |S|$ is known. Show how the winnow algorithm and analysis *given in class* can be applied to this case and that the number of mistakes is at most $O(k^2 \ln N)$.

d. [5] Now show how the version of winnow developed in parts (a) and (b) can be applied to this problem to obtain a mistake bound of $O(k \ln N)$. (For this problem, you may freely approximate $\ln(1 + \epsilon)$ by ϵ when $|\epsilon|$ is small.)

Problem 2

In class, we proved that the loss of the Widrow-Hoff (WH) algorithm is at most

$$\min_{\mathbf{u} \in \mathbb{R}^n} \left(pL_{\mathbf{u}} + q||\mathbf{u}||_2^2 \right) \tag{1}$$

for constants $p = 1/(1-\eta)$ and $q = 1/\eta$. In this problem, we will show that these constants are the best possible, in other words, that no algorithm can achieve a bound that is strictly better.

Let A be *any* deterministic, on-line learning algorithm (not necessarily WH or even a weight-update algorithm), and assume that the cumulative loss of A,

$$L_A = \sum_{t=1}^{T} (\hat{y}_t - y_t)^2$$

is at most the bound given in Eq. (1). As usual,

$$L_{\mathbf{u}} = \sum_{t=1}^{T} (\mathbf{u} \cdot \mathbf{x}_t - y_t)^2.$$

Consider training A on the following examples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_T, y_T)$: each $\mathbf{x}_t$ is a unit vector with a 1 in the t-th coordinate, and 0's in all other coordinates. (Thus, $\mathbf{x}_t \in \mathbb{R}^n$ where $n \geq T$.) The y_t's are all in $\{-1, +1\}$ and can be chosen adversarially.

a. [8] Show how an adversary can choose the y_t's to ensure that $L_A \geq T$.

b. [12] Show that, regardless of how the y_t's are chosen in (a), the upper bound on L_A in Eq. (1) is equal to:

$$\frac{pq}{p+q} T.$$

c. [5] Combine parts (a) and (b) to show that

$$\frac{1}{p} + \frac{1}{q} \leq 1.$$

Show how this implies that the bounds for WH are the best possible, i.e., that it cannot be the case that $p < 1/(1-\eta)$ and simultaneously $q < 1/\eta$ for any $\eta \in (0, 1)$.