COS 511: Theoretical Machine Learning

Homework #5 — Due:
Boosting & SVM's — April 1, 2014

Problem 1

[10] Suppose, in the usual boosting set-up, that the weak learning condition is guaranteed to hold so that $\epsilon_t \le \frac{1}{2} - \gamma$ for some $\gamma > 0$ which is not known before boosting begins. And suppose AdaBoost is run in the usual fashion, except that the algorithm is modified to halt and output the combined classifier H immediately following the first round on which it is consistent with all of the training examples (so that its training error is zero). Assume that the weak hypotheses are selected from a class of VC-dimension d. Prove that, with probability at least $1 - \delta$, the generalization error of the output combined classifier H is at most

$$\tilde{O}\left(\frac{(d/\gamma^2) + \ln(1/\delta)}{m}\right).$$

Give a bound in which all constants and log terms have been filled in explicitly.

Problem 2

Suppose AdaBoost is run for an unterminating number of rounds. In addition to our usual notation, we define for each $T \ge 1$:

$$F_T(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \quad \text{and} \quad s_T = \sum_{t=1}^{T} \alpha_t.$$

Recall that each $\alpha_t \ge 0$ (since $\epsilon_t \le \frac{1}{2}$). The *minimum margin* on round t, denoted θ_t, is the smallest margin of any training example; thus,

$$\theta_t = \min_i \frac{y_i F_t(x_i)}{s_t}.$$

Finally, we define the *smooth margin* on round t to be

$$g_t = \frac{-\ln\left(\frac{1}{m}\sum_{i=1}^{m} e^{-y_i F_t(x_i)}\right)}{s_t}.$$

a. [10] Prove that

$$\theta_t \le g_t \le \theta_t + \frac{\ln m}{s_t}.$$

Thus, if s_t gets large, then g_t gets very close to θ_t.

b. [10] Let us define the continuous function

$$\Upsilon(\gamma) = \frac{-\ln(1 - 4\gamma^2)}{\ln\left(\frac{1+2\gamma}{1-2\gamma}\right)}.$$

It is a fact (which you do not need to prove) that $\gamma \le \Upsilon(\gamma) \le 2\gamma$ for $0 \le \gamma \le \frac{1}{2}$.

Prove that g_T is a weighted average of the values $\Upsilon(\gamma_t)$, specifically,

$$g_T = \frac{\sum_{t=1}^{T} \alpha_t \Upsilon(\gamma_t)}{s_T}.$$

c. [10] Prove that if the edges γ_t converge (as $t \to \infty$) to some value γ, where $0 < \gamma < \frac{1}{2}$, then the minimum margins θ_t converge to $\Upsilon(\gamma)$.

Problem 3

a. [10] In class, we argued that if a function L satisfies the "minmax property"

$$\min_{\mathbf{w}} \max_{\boldsymbol{\alpha}} L(\mathbf{w}, \boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha}} \min_{\mathbf{w}} L(\mathbf{w}, \boldsymbol{\alpha}), \tag{1}$$

and if $(\mathbf{w}^*, \boldsymbol{\alpha}^*)$ are the desired solutions

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \max_{\boldsymbol{\alpha}} L(\mathbf{w}, \boldsymbol{\alpha}) \tag{2}$$

$$\boldsymbol{\alpha}^* = \arg\max_{\boldsymbol{\alpha}} \min_{\mathbf{w}} L(\mathbf{w}, \boldsymbol{\alpha}), \tag{3}$$

then $(\mathbf{w}^*, \boldsymbol{\alpha}^*)$ is a saddle point:

$$L(\mathbf{w}^*, \boldsymbol{\alpha}^*) = \max_{\boldsymbol{\alpha}} L(\mathbf{w}^*, \boldsymbol{\alpha}) = \min_{\mathbf{w}} L(\mathbf{w}, \boldsymbol{\alpha}^*). \tag{4}$$

(Here, it is understood that $\mathbf{w}$ and $\boldsymbol{\alpha}$ may belong to a restricted space (e.g., $\boldsymbol{\alpha} \geq 0$) which we omit for brevity.)

Prove the converse of what was shown in class. That is, prove that if $(\mathbf{w}^*, \boldsymbol{\alpha}^*)$ satisfies Eq. (4), then Eqs. (1), (2) and (3) are also satisfied. You should not assume anything special about L (such as convexity), but you can assume all of the relevant minima and maxima exist.

b. [10] Let $a_1, \ldots, a_n$ be nonnegative real numbers, not all equal to zero, and let $b_1, \ldots, b_n$ and c all be positive real numbers. Use the method of Lagrange multipliers to find the values of $x_1, \ldots, x_n$ which minimize

$$-\sum_{i=1}^{n} a_i \ln x_i$$

subject to the constraint that

$$\sum_{i=1}^{n} b_i x_i \leq c.$$

Show how this implies that relative entropy is nonnegative.

Problem 4 – Optional (Extra Credit)

[15] In class (as well as on Problem 1 of this homework), we showed how a weak learning algorithm that uses hypotheses from a space $\mathcal{H}$ of bounded VC-dimension can be converted into a strong learning algorithm. However, strictly speaking, the definition of weak learnability does *not* include such a restriction 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 every concept $c \in \mathcal{C}$ and for every distribution D on X, when given m_0 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 $1/2$,

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

Here, for simplicity, we have "hard-wired" the usual parameter δ to the constant $1/2$ so that A_0 takes a fixed number of examples and only needs to succeed with fixed probability $1/2$. Note that no restrictions are made on the form of hypothesis h used by A_0, nor on the cardinality or VC-dimension of the space from which it is chosen. For this problem, assume that A_0 is a deterministic algorithm.

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 $m = \mathrm{poly}(m_0, 1/\epsilon, 1/\delta, 1/\gamma)$ random examples x_i from D, each with its label $c(x_i)$, A outputs a hypothesis H such that, with probability at least $1 - \delta$,

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

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