COS 511: Theoretical Machine Learning

Homework #5 — Due:
Margins & SVM's — April 2, 2019

Problem 1

In class, we proved that the generalization error of a classifier of the kind produced by AdaBoost can be bounded by the number of "small-margin" training examples (with margin at most some threshold $\theta > 0$), plus an additional term that, in terms of the number of training examples m, goes to zero at the rate $\tilde{O}\left(1/\sqrt{m}\right)$.

In this problem, we will prove that a much better bound on the generalization error is possible when there are *no* small-margin training examples, in other words, when *all* of the training examples have margins larger than θ. In this case, we will see that the generalization error is only $\tilde{O}\left(1/m\right)$, a huge improvement over what we would get by plugging into the bound proved in class. Along the way, we will also explore a different technique for proving margin-based bounds.

The approach we take is based on the double-sample trick studied earlier in this course (in lectures 5 and 6). As usual, let $S = \langle (x_1, y_1), \ldots, (x_m, y_m) \rangle$ be the "real" training set of independent random examples from the target distribution $\mathcal{D}$. Let $S' = \langle (x'_1, y'_1), \ldots, (x'_m, y'_m) \rangle$ be a "ghost" sample (also i.i.d. from $\mathcal{D}$).

Let $\mathcal{H}$ be the weak hypothesis space, and let d be its VC-dimension, where we assume throughout that $m \geq d \geq 1$. Let $\mathrm{co}(\mathcal{H})$ be the convex hull of $\mathcal{H}$. Let $\theta > 0$ and let $\epsilon > 0$. Our aim will be to show that, with high probability, for every function $f \in \mathrm{co}(\mathcal{H})$, if all of the training examples have large margin (greater than θ) with respect to f, then the classifier corresponding to f (namely, $H(x) = \mathrm{sign}(f(x))$), must have low generalization error. More precisely, we wish to bound the probability (over the random selection of S) of the following "bad" event:

$$B \equiv \ \exists f \in \mathrm{co}(\mathcal{H}) : \hat{\Pr}_S \left[yf(x) \leq \theta \right] = 0 \ \wedge \ \Pr_{\mathcal{D}} \left[yf(x) \leq 0 \right] > \epsilon.$$

Here, as in class, $\hat{\Pr}_S [\cdot]$ means empirical probability with respect to the sample S, and $\Pr_{\mathcal{D}} [\cdot]$ means probability with respect to a random draw of an example $(x, y) \sim \mathcal{D}$. (Thus, $\hat{\Pr}_S \left[yf(x) \leq \theta \right]$ means the fraction of examples (x, y) in S with $yf(x) \leq \theta$; and $\Pr_{\mathcal{D}} \left[yf(x) \leq 0 \right]$ is the probability of selecting $(x, y) \sim \mathcal{D}$ with margin at most zero, which is the same as the generalization error of H.)

As was done in our earlier double-sample proof, we can effectively replace probability over $\mathcal{D}$ with empirical probability on the ghost sample. Thus, we can consider this alternative event:

$$B' \equiv \ \exists f \in \mathrm{co}(\mathcal{H}) : \hat{\Pr}_S \left[yf(x) \leq \theta \right] = 0 \ \wedge \ \hat{\Pr}_{S'} \left[yf(x) \leq 0 \right] > \frac{\epsilon}{2}.$$

You can take as given that $\Pr \left[B' | B \right] \geq 1/2$ (where probability is over the random choice of S and S'), assuming m is not too small, by the same argument used earlier. Therefore, as before, $\Pr \left[B \right] \leq 2 \Pr \left[B' \right]$.

a. [6] Let f be some fixed function in $\mathrm{co}(\mathcal{H})$; specifically, suppose

$$f(x) = \sum_{t=1}^{T} a_t h_t(x)$$

where $a_1, \ldots, a_T \in [0,1]$ with $\sum_{t=1}^{T} a_t = 1$ and $h_1, ..., h_T \in \mathcal{H}$. Notice how the a_t's form a probability distribution over the hypotheses $h_1, \ldots, h_T$, which means we can imagine using that distribution to select one of these T weak hypotheses at random. Suppose that we do that repeatedly, in other words, that we pick a sequence of weak hypotheses $g_1, \ldots, g_N$ independently at random from $\mathcal{H}$, where each g_j is selected (with replacement) according to the distribution given by the a_t's so that g_j is chosen to be h_t with probability a_t. Let g be their average:

$$g(x) = \frac{1}{N} \sum_{j=1}^{N} g_j(x).$$

For a fixed example (x, y), prove that

$$\Pr\left[|yf(x) - yg(x)| \geq \frac{\theta}{2}\right] \leq 2e^{-\theta^2 N/8},$$

so that this one example (x, y) is likely to have almost the same margin with respect to either f or g. (Here, probability is over the random choice of g.)

b. [12] Let $\mathcal{A}_N$ be the space of all functions that, like the function g above, are the average of N (not necessarily distinct) weak hypotheses:

$$\mathcal{A}_N = \left\{ x \mapsto \frac{1}{N} \sum_{j=1}^{N} g_j(x) : g_1, \ldots, g_N \in \mathcal{H} \right\}.$$

Consider the following event:

$$B'' \equiv \quad \exists g \in \mathcal{A}_N : \hat{\Pr}_S\left[yg(x) \leq \frac{\theta}{2}\right] = 0 \ \wedge \ \hat{\Pr}_{S'}\left[yg(x) \leq \frac{\theta}{2}\right] > \frac{\epsilon}{2}.$$

Prove that $\Pr[B'] \leq \Pr[B'']$ for a suitable choice of N with $N = O\left(\frac{\ln m}{\theta^2}\right)$. Give an expression for N with explicit constants. (Here, probability is only over the random choice of S and S'.)

c. [12] Prove that

$$\Pr[B''] \leq \left(\frac{2me}{d}\right)^{dN} \cdot 2^{-m\epsilon/2}.$$

Combining the steps above, this immediately shows that $\Pr[B]$ is at most twice the bound given in part (c). As usual, we can set the resulting bound equal to δ and solve for ϵ to conclude that $\Pr[B] \leq \delta$ if we choose

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

This means that, with probability at least $1 - \delta$, if the classifier produced by AdaBoost yields margins exceeding θ on all of the training examples (as in Problem 2(c), for suitable θ, under the weak learning assumption), then its generalization error is at most the bound given above for ϵ.

Problem 2

In this problem, we will give an alternative technique relating edges and margins, and specifically showing that, when the weak learning assumption holds, all examples will eventually have "large" margins (at least some positive value).

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

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

Recall that each $\alpha_t \geq 0$ (since $\epsilon_t \leq \frac{1}{2}$). 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

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

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

b. [10] For $0 \leq \gamma \leq \frac{1}{2}$, let us define the continuous function

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

(where, by continuity, $\Upsilon(0) = 0$ and $\Upsilon(\frac{1}{2}) = 1$). A plot of this function is shown in Figure 1. It is a fact (which you do not need to prove) that $\gamma \leq \Upsilon(\gamma) \leq 2\gamma$, and also that $\Upsilon(\gamma)$ is (strictly) increasing.

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. [6] Suppose that, for some $\gamma > 0$, and for all t, $\gamma_t \geq \gamma$. Prove that, for all t,

$$\theta_t \geq \Upsilon(\gamma) - \frac{C}{t}$$

where $C > 0$ is a number that may depend on m and γ, but should not depend on t. Give an explicit expression for C. This shows that the minimum margin θ_t (and therefore the margins of all the training examples) must in the limit be at least $\Upsilon(\gamma)$.

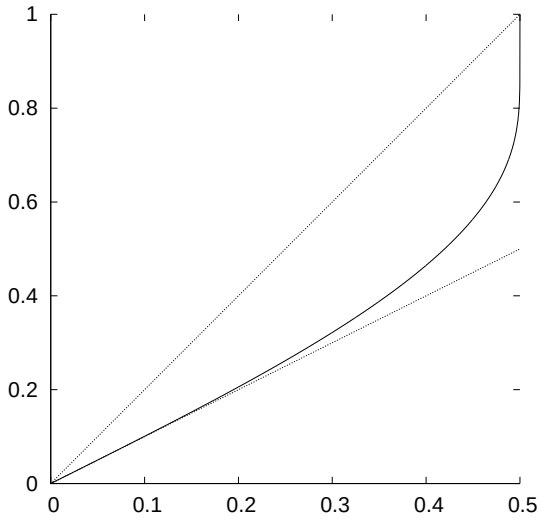

Figure 1: A plot of $\Upsilon(\gamma)$, as a function of γ. Also plotted are the linear lower and upper bounds, γ and 2γ.

Problem 3

[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.

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

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

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