MSRI Summer School on Mathematics of Machine Learning

Problem Set #3 Wednesday, July 31, 2019

Problem 3-1

a. Let $\mathcal{H}$ be a space of functions $h : X \to \{-1, +1\}$. As in class, let $\mathrm{co}(\mathcal{H})$ be the convex hull of $\mathcal{H}$. For any sample S, show that $\hat{\mathcal{R}}_S(\mathrm{co}(\mathcal{H})) = \hat{\mathcal{R}}_S(\mathcal{H})$.

b. Suppose the labels y belong to some space Y (while the instances x are in X, as usual). Let $\mathcal{F}$ be a predictor space of functions $f : X \to Y$. Let $L : Y \times Y \to [0, 1]$ be a loss function so that if an example x has predicted label $f(x)$ and true label y then the resulting loss is $L(f(x), y)$. For each f, we define the function $\ell_f : X \times Y \to [0, 1]$ by $\ell_f(x, y) = L(f(x), y)$, and let $\mathcal{L} = \{\ell_f : f \in \mathcal{F}\}$.

   Often, we are interested in bounding the true risk, $\mathrm{E}[\ell_f]$ (that is, the expected loss on the true distribution), uniformly in terms of the empirical risk, $\hat{\mathrm{E}}[\ell_f]$ (that is, the average loss on a random training set). As we have seen, the difference between these can be bounded uniformly for all $f \in \mathcal{F}$ in terms of the Rademacher complexity of $\mathcal{L}$. However, it is often more natural and convenient to state bounds in terms of the Rademacher complexity of $\mathcal{F}$.

   (i) Suppose, as in the usual classification setting, that $Y = \{-1, +1\}$ and $L(\hat{y}, y) = \mathbf{1}\{\hat{y} \neq y\}$. For any sample S, show that $\hat{\mathcal{R}}_S(\mathcal{L}) = \frac{1}{2}\hat{\mathcal{R}}_S(\mathcal{F})$. (Here and in the next part, S is a sequence of pairs $(x_1, y_1), \ldots, (x_m, y_m)$ from $X \times Y$; however, in computing $\hat{\mathcal{R}}_S(\mathcal{F})$, only the x_i's are relevant.)

   (ii) Suppose, as in a typical regression setting, that $Y = [0, 1]$ and $L(\hat{y}, y) = (\hat{y} - y)^2$. For any sample S, show that $\hat{\mathcal{R}}_S(\mathcal{L}) \leq 2\hat{\mathcal{R}}_S(\mathcal{F})$.

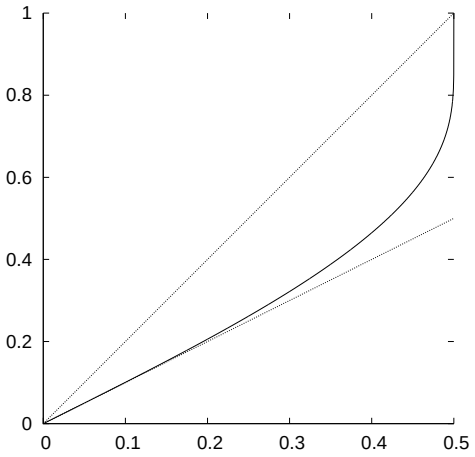

Figure 1: A plot of $\Upsilon(\gamma)$, as a function of γ. Also plotted are the linear lower and upper bounds, γ and 2γ.

Problem 3-2

a. For $i = 1, \ldots, n$, let $\mathcal{G}_i$ be a space of concepts ($\{0,1\}$-valued functions) defined on some domain X, and let $\mathcal{F}$ be a space of concepts defined on $\{0,1\}^n$. (That is, each $g_i \in \mathcal{G}_i$ maps X to $\{0,1\}$, and each $f \in \mathcal{F}$ maps $\{0,1\}^n$ to $\{0,1\}$.) Let $\mathcal{C}$ be the space of all concepts $c : X \to \{0,1\}$ of the form

$$c(x) = f(g_1(x), \ldots, g_n(x))$$

for some $f \in \mathcal{F}$, $g_1 \in \mathcal{G}_1, \ldots, g_n \in \mathcal{G}_n$.

Give a careful argument proving that

$$\Pi_{\mathcal{C}}(m) \leq \Pi_{\mathcal{F}}(m) \cdot \prod_{i=1}^{n} \Pi_{\mathcal{G}_i}(m).$$

b. AdaBoost outputs a combined hypothesis H of the form

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

where $\alpha_1, \ldots, \alpha_T \in \mathbb{R}$ and $h_1, \ldots, h_T$ are in some class $\mathcal{H}$, which we assume has VC-dimension $d \geq 1$ (and where we here regard T as fixed). Let $\mathcal{S}$ be the space of all functions H of this form. Use part (a) to first derive a bound on $\Pi_{\mathcal{S}}(m)$, and to then prove that, with probability at least $1 - \delta$, for every function $H \in \mathcal{S}$,

$$\text{err}(H) \leq \widehat{\text{err}}(H) + \tilde{O}\left(\sqrt{\frac{Td + \ln(1/\delta)}{m}}\right),$$

assuming $m \geq \max\{d, T\}$. (Here, the "soft-Oh" notation $\tilde{O}(\cdot)$ means that we are ignoring log terms in the bound, in the same way that big-Oh notation ignores constants.)

Hint for part (a): To get started, fix $g_1, \ldots, g_n$, and count how many behaviors can be realized on any set of m points by functions c of the form given in the problem (with f varying, but $g_1, \ldots, g_n$ fixed).

Problem 3-3

This problem gives a technique for relating edges and margins, 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. 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. 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. 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 $\Upsilon(\gamma)$.