Machine Learning Theory (CSC 482A/581A) - Lecture 21

Nishant Mehta

1 AdaBoost and Margins

We thus far have seen that weak learnability implies strong learnability, and in the process we developed a PAC learning guarantee for AdaBoost. The risk bound we proved increases with the number of rounds T for which AdaBoost is run, suggesting that if we run AdaBoost for long enough it eventually will overfit; recall that our bound scaled with T because of a VC dimension-based argument we used for the class of linear threshold functions over T dimensions. On the other hand, in numerous examples it has been observed that when AdaBoost is run far beyond the point when it has first obtained zero training error, its test error continues to improve. *(I showed a figure in class that demonstrates this phenomenon)*.

We will now explore the margins explanation for why AdaBoost often does not overfit for seemingly excessively large T. The idea is that even after AdaBoost has obtained zero training error, as T increases, the weighted majority vote predictor becomes more and more *confident* in its predictions. The intuitive idea of the predictor being confident in making a prediction on an example x can be formalized as follows: consider the difference between the proportion of mass assigned to hypotheses predicting -1 on this example and the proportion of mass assigned to weak hypotheses predicting $+1$ on this example; then the greater this difference, the more confident the prediction. For any example, this difference will be called the margin. There are well-documented examples where the distribution of the AdaBoost's margins for the training examples continues to improve as T increases. *(I showed a figure with such an example in class)*.

The margins explanation for why AdaBoost tends not to overfit is based on two arguments. First, AdaBoost improves the margin distribution of the training examples. Second, voting methods (like AdaBoost) that predict with good margin admit risk bounds that do not grow with T but instead scale inversely with the margin.

2 AdaBoost and the empirical margin distribution

As before, let ε_t and γ_t be defined such that

$$\varepsilon_t = \Pr_{j \sim D_t} \left(h_t(X_j) \neq Y_j \right) = \frac{1}{2} - \gamma_t.$$

In conducting a margin-based analysis, it will be convenient to normalize the weights that appear in AdaBoost's weighted majority predictor. To this end, we define for each $j \in [n]$

$$a_j = \frac{\alpha_j}{\sum_{i=1}^n \alpha_i},$$

so that these new weights sum to 1.

Now, since the sign function is invariant to positive scaling, we can re-express AdaBoost's hypothesis in terms of the new weights as

$$\hat{f}(x) = \operatorname{sgn}\left(\hat{g}(x)\right) \tag{1}$$

for

$$\hat{g}(x_j) = \sum_{t=1}^{T} a_t h_t(x_j) = \frac{\sum_{t=1}^{T} \alpha_t h_t(x_j)}{\sum_{t=1}^{T} \alpha_t}$$

The next theorem provides an upper bound on the fraction of points in the training sample for which $\hat{g}$ does not achieve margin exceeding a prescribed threshold.

Theorem 1. *Let $\theta \geq 0$. If AdaBoost is run for T rounds on a training sample of size n, then*

$$\frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\left[Y_j \hat{g}(X_j) \leq \theta\right] \leq \prod_{t=1}^{T} \sqrt{(1+2\gamma_t)^{1+\theta}(1-2\gamma_t)^{1-\theta}}.$$

If we take $\theta = 0$, the above result recovers our bound on AdaBoost's empirical risk under zero-one loss, where, as an intermediate step, we had obtained the upper bound $\prod_{t=1}^{T}\sqrt{1-4\gamma_t^2}$. The proof of the above theorem is very similar to the proof for the special case of $\theta = 0$.

Proof of Theorem 1. First, for any $z \in \mathbb{R}$, observe that $\mathbf{1}\left[z \geq 0\right] \leq e^z$. Therefore, for any $j \in [n]$,

$$\begin{aligned}
\mathbf{1}\left[Y_j \hat{g}(X_j) \leq \theta\right] &= \mathbf{1}\left[\theta - Y_j \hat{g}(X_j) \geq 0\right] \\
&= \mathbf{1}\left[\theta - Y_j \sum_{t=1}^{T} a_t h_t(X_j) \geq 0\right] \\
&= \mathbf{1}\left[\theta \sum_{t=1}^{T} \alpha_t - Y_j \sum_{t=1}^{T} \alpha_t h_t(X_j) \geq 0\right] \\
&\leq \exp\left(\theta \sum_{t=1}^{T} \alpha_t - Y_j \sum_{t=1}^{T} \alpha_t h_t(X_j)\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\left[Y_j \hat{g}(X_j) \leq \theta\right] \\
&\leq \frac{1}{n}\sum_{j=1}^{n} \exp\left(\theta \sum_{t=1}^{T} \alpha_t - Y_j \sum_{t=1}^{T} \alpha_t h_t(X_j)\right) \\
&= \exp\left(\theta \sum_{t=1}^{T} \alpha_t\right) \cdot \frac{1}{n}\sum_{j=1}^{n} \exp\left(-Y_j \sum_{t=1}^{T} \alpha_t h_t(X_j)\right) \\
&= \exp\left(\theta \sum_{t=1}^{T} \alpha_t\right) \cdot \prod_{t=1}^{T} Z_t \\
&= \exp\left(\theta \sum_{t=1}^{T} \alpha_t\right) \cdot \prod_{t=1}^{T} \sqrt{(1-4\gamma_t^2)},
\end{aligned}$$

where we proved the last two equalities when analyzing the rate at which AdaBoost's training error (under zero-one loss) converges to zero. Plugging in the value of α_t used by AdaBoost,

$$\alpha_t = \frac{1}{2}\log\frac{1-\varepsilon_t}{\varepsilon_t} = \frac{1}{2}\log\frac{1+2\gamma_t}{1-2\gamma_t},$$

yields

$$\begin{aligned}
&\prod_{t=1}^{T}\left(\frac{1+2\gamma_t}{1-2\gamma_t}\right)^{\theta/2}\sqrt{1-4\gamma_t^2} \\
&= \prod_{t=1}^{T}\sqrt{\left(\frac{1+2\gamma_t}{1-2\gamma_t}\right)^{\theta}(1+2\gamma_t)(1-2\gamma_t)} \\
&= \prod_{t=1}^{T}\sqrt{(1+2\gamma_t)^{1+\theta}\cdot(1-2\gamma_t)^{1-\theta}}.
\end{aligned}$$

□

To get a feel for the upper bound in Theorem 1, let's consider the pessimistic special case where $\gamma_t = \gamma$ for all $t \in [T]$, so that the weak learning assumption is exactly satisfied with no slack. Then the upper bound reduces to

$$\left((1+2\gamma)^{1+\theta}(1-2\gamma)^{1-\theta}\right)^{T/2}.$$

In the special case of $\theta = 0$, we see that the above quantity decays to zero exponentially in T as long as $\gamma > 0$. How large can θ be before exponential decay fails to occur? To answer this, we need only find, for fixed $\gamma > 0$, the largest θ such that

$$(1+2\gamma)^{1+\theta}(1-2\gamma)^{1-\theta} < 1.$$

Isolating θ yields the upper bound

$$\theta < \frac{-\log(1-4\gamma^2)}{\log\frac{1+2\gamma}{1-2\gamma}}.$$

This complicated inequality is perhaps a less than satisfying answer for how large θ can be; however, in the range $0 \le \gamma \le \frac{1}{2}$, it holds that

$$\gamma \le \frac{-\log(1-4\gamma^2)}{\log\frac{1+2\gamma}{1-2\gamma}} \le 2\gamma.$$

(I demonstrated this via a plot in class)

Thus, as long as $\theta < \gamma$, the empirical θ-margin error decays to zero exponentially quickly in T, and moreover, this upper bound is tight up to a factor of 2.

Next, we will see risk bounds for AdaBoost which motivate taking θ to be as large as possible, in perfect analogy to the margin bounds we developed for support vector machines.

3 Margin bounds for ensemble methods

When we studied margin bounds for SVM's, we saw the following Rademacher complexity-based bound on the risk of a classifier under zero-one loss:

Theorem 2. *Let $\mathcal{F}$ be a class of real-valued functions. Then for any $\theta > 0$, with probability at least $1 - \delta$, for all $f \in \mathcal{F}$,*

$$\mathsf{E}\left[\mathbf{1}\left[Yf(x) \leq 0\right]\right] \leq \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\left[Y_j f(X_j) < \theta\right] + \frac{2}{\theta}\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{\log\frac{1}{\delta}}{2n}}.$$

Now, using (1), Adaboost's real-valued hypothesis (prior to taking the sign) $\hat{g}(x)$ can be expressed as a convex combination of weak hypotheses $h_1, \ldots, h_T$. Let us suppose that each h_t belongs to a set of hypotheses $\mathcal{H}$ with $\mathrm{VCdim}(\mathcal{H}) = V < \infty$, a reasonable state of affairs since these are weak hypotheses. Then the function $\hat{g}$ belongs to the class $\mathcal{F} = \mathrm{conv}(\mathcal{H})$. Next, recall that for any class $\mathcal{H}$,

$$\mathcal{R}_n(\mathrm{conv}(\mathcal{H})) = \mathcal{R}_n(\mathcal{H}). \tag{2}$$

Lastly, since $\mathcal{H}$ is a set of classifiers, we can upper bound $\mathcal{R}_n(\mathcal{H})$ using the VC dimension as[1]

$$\mathcal{R}_n(\mathcal{H}) \leq \sqrt{\frac{2V\log\frac{en}{V}}{n}}. \tag{3}$$

Applying Theorem 3 with $\mathcal{R}_n(\mathcal{F}) = \mathcal{R}_n(\mathrm{conv}(\mathcal{H}))$ bounded from (2) and (3), we have the following margin bound for classifiers that predict according to the convex hull of a set of classifiers of finite VC dimension.

Theorem 3. *Let $\mathcal{H}$ be a set of classifiers with $\mathrm{VCdim}(\mathcal{H}) = V$ and let $\mathcal{F} = \mathrm{conv}(\mathcal{H})$. Then, for any $\theta > 0$, with probability at least $1 - \delta$, for all $f \in \mathcal{F}$,*

$$\mathsf{E}\left[\mathbf{1}\left[Yf(x) \leq 0\right]\right] \leq \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\left[Y_j f(X_j) < \theta\right] + \frac{2}{\theta}\sqrt{\frac{2V\log\frac{en}{V}}{n}} + \sqrt{\frac{\log\frac{1}{\delta}}{2n}}.$$

The above theorem applies to *any* ensemble predictor that predicts according to a weighted majority vote of classifiers belonging to a VC-class $\mathcal{H}$. Therefore, it applies in particular to AdaBoost. Moreover, if $\theta < \gamma$, the empirical risk with respect to the θ-margin error decays exponentially quickly as then number of rounds for which AdaBoost runs increases, and thus for sufficiently large T and taking $\theta = \frac{\gamma}{2}$, the bound in the above theorem becomes

$$\mathsf{E}\left[\mathbf{1}\left[Yf(x) \leq 0\right]\right] \leq \frac{4}{\gamma}\sqrt{\frac{2V\log\frac{en}{V}}{n}} + \sqrt{\frac{\log\frac{1}{\delta}}{2n}}.$$

Technically, we should not be entirely satisfied with the above result: in the realizable case, we should hope to dispense with the square root to obtain a faster convergence rate with respect to n. This can be accomplished by a more refined analysis.

[1] We proved this bound in lecture 14 (see the top of page 6 in the lecture notes for Lectures 13–15).