Lecture 10: Boosting and Margins

Akshay Krishnamurthy
akshay@cs.umass.edu

October 5, 2017

1 Recap

Recall the AdaBoost algorithm. We are given a dataset $(x_1, y_1), \ldots, (x_n, y_n)$ of binary classification examples, and a weak learning algorithm, that finds hypothesis from $\mathcal{H}$. Starting with $D_1(i) = 1/n$ for all i, we iterate for each round $t = 1, \ldots, T$

1. Let h_t be the weak learner trained on D_t. Define
$$\epsilon_t = \sum_{i=1}^{n} D_t(i)\mathbf{1}\{h_t(x_i) \neq y_i\}.$$
2. Set $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$.
3. Update
$$D_{t+1}(i) = \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$
where Z_t normalizes the distribution.

After T iterations we output the weighted majority hypothesis $H(x) = \text{sign}(\sum_{t=1}^{T}\alpha_t h_t(x))$.

Last time we saw that after T iterations, AdaBoost has training error at most $\exp(-2\sum_{t=1}^{T}\gamma_t^2)$ where γ_t is the edge of hypothesis h_t, $\gamma_t = 1/2 - \epsilon_t$. We also connected AdaBoost to an ERM problem with the exponential loss, and proved a generalization bound based on VC dimension. In particular we saw that the excess training plus generalization error is

$$\exp(-2T\gamma^2) + O(\sqrt{\frac{dT\log(n/\delta)}{n}}),$$

if we always have edge at least γ. Thus we can choose T to optimize the bound. Since the training error decays exponentially but the generalization grows linearly, we expect that choosing T small should be the best.

It turns out that this bound does not explain the behavior of AdaBoost in practice. In fact it is somewhat common that AdaBoost gets zero training error, and then if you keep running the algorithm, the test error gets better and better. Is there a way to explain this behavior?

Actually you should not be super surprised by this if you did the Rademacher complexity problem on homework 2. There we saw that for a two-layer neural net, where the output layer weights are a distribution, the rademacher complexity does not grow with the number of hidden units. You can think of boosting in a similar way, it is taking a distribution over base learners/hidden units, so we should be able to get a Rademacher bound that is independent of the number of base learners used in the distribution, e.g., T. In a sense this is exactly what we're going to do, the main difference being that our loss function is not Lipschitz, so we cannot use Rademacher contraction right away. Well today we'll do it from first principles but on homework 3 you'll go through Rademacher complexity.

2 Margins and AdaBoost

Thus we seek another explanation for why boosting works and for this we will use the concept of **margins**.

Definition 1. *Let* $H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$ *and normalize the weights* $a_t = \frac{\alpha_t}{\sum_{i=1}^{T} \alpha_i}$ *and define* $f(x) = \sum_{i=1}^{T} a_i h_i(x)$. *Then the* ***margin*** *on point* (x, y) *is* $yf(x)$.

Margins also lead to a sufficient condition for weak learnability, which recall was important for ensuring low training error. Suppose that the training sample is such that there exists $g_1, \ldots, g_k \in \mathcal{H}$ and coefficients $a_1, \ldots, a_k$ forming a distribution such that for all $i \in [n]$

$$y_i \sum_{j=1}^{k} a_j g_j(x_i) \geq \theta > 0.$$

This means that $\langle a, g \rangle$ is perfect classifier, but also has some margin on all the points. This condition implies that weak learnability holds, since for any distribution D, we can take expectations

$$\sum_{j=1}^{k} a_j \mathbb{E} y_i g_k(x_i) \geq \theta > 0.$$

Then, since a_js are a distribution, it must be the case that one of the g_ks has $\mathbb{E} y g_k(x) > \theta$, which means that we have the weak learning condition with margin at least $\theta/2$, since

$$\mathbb{P}[y_i \neq g_k(x_i)] = \frac{1 - \mathbb{E} y_i g_k(x_i)}{2} \leq \frac{1 - \theta}{2}.$$

The calculation here uses that $y_i g_k(x_i) \in \{\pm 1\}$ so you have to translate from the expectation to the 0/1 loss.

The more important fact about margins is that they lead to a more predictive generalization bound.

Theorem 2. *Let* $\mathcal{D}$ *be a distribution over* $\mathcal{X} \times \{-1, +1\}$ *and let* S *be a sample of size* n. *Let* $\mathcal{H}$ *be a finite set of classifiers. Then for any* δ, *with probability at least* $1 - \delta$ *every weighted average function* $f \in conv(\mathcal{H})$ *satisfies*

$$\mathbb{P}_{\mathcal{D}}[yf(x) \leq 0] \leq \mathbb{P}_S[yf(x) \leq \theta] + O\left(\sqrt{\frac{\log(|\mathcal{H}|)}{n\theta^2} \log\left(\frac{n\theta^2}{\log|\mathcal{H}|}\right) + \frac{\log(1/\delta)}{n}}\right).$$

First let us briefly see how the proof will go. The idea is to use the slack provided by the margin assumption to discretize the set $\text{conv}(\mathcal{H})$. In particular, pick some m and define

$$\mathcal{A}_m = \{f : x \to \frac{1}{m} \sum_{i=1}^{m} h_i(x) \mid h_1, \ldots, h_m \in \mathcal{H}\}$$

Clearly this class is finite since $\mathcal{H}$ is finite as well. Ultimately we will only prove uniform convergence over this class.

Take some $f \in \text{conv}(\mathcal{H})$. This means we can write $f = \sum a_i h_i$ for some distribution a_i. The idea is that we can approximate f by a random function $\tilde{f}$ obtained by sampling h_i with probability a_i and repeating the experiment m times. Thus $\tilde{f} \in \mathcal{A}_m$. We'll use this $\tilde{f}$ to approximate f in the sense that we'll show

$$\mathbb{P}_{\mathcal{D}}[yf(x) \leq 0] \lessapprox \mathbb{P}_{\mathcal{D}}[y\tilde{f}(x) \leq \theta/2] \qquad \text{and} \qquad \mathbb{P}_S[y\tilde{f}(x) \leq \theta/2] \lessapprox \mathbb{P}_S[yf(x) \leq \theta]$$

Then we'll use a uniform-convergence type statement to analyze the margin distributions for the functions in $\mathcal{A}_m$, specifically,

$$\mathbb{P}_{\mathcal{D}}[y\tilde{f}(x) \leq \theta/2] \lessapprox \mathbb{P}_S[y\tilde{f}(x) \leq \theta/2]$$

Combining these, with the appropriate slack will prove the theorem.

Before diving in the proof, we can observe that in a sense we are using a covering-number style of proof. We have taken a discretization of the set $\text{conv}(\mathcal{H})$ and are approximating each function by an element from the discretization.

However there are several differences. First we are using a randomized covering element, which will be important in the proof. Second we are not working on the zero-one loss, but rather the margin distribution, which is more smooth and something that we can use discretization arguments to work with. In particular, we'll need some slack that we can tolerate by asking for a slightly larger margin.

There is a cleaner proof of the same result using Rademacher complexity that you'll do on the homework. You can almost apply Rademacher complexity as is, except that the 0/1 loss function is not Lipschitz, and this is why we have to introduce the margin. However, this proof is worth seeing, since in a sense we use the probabilistic method, which is a technique that all of you should be familiar with.

Proof. Let $f = \sum_j a_j h_j$ for some distribution a and recall how we defined the random variable $\tilde{f}$ by sampling m times from the distribution a to form a hypothesis based with uniform weights. We first prove

Lemma 3. *For fixed $f, x, \theta > 0$ and $n \geq 1$*

$$\mathbb{P}_{\tilde{f}}\left[|\tilde{f}(x) - f(x)| \geq \theta/2\right] \leq 2\exp(-n\theta^2/8) = \beta_{m,\theta}$$

Proof. Fixing x, the random variable $\tilde{h}_j(x)$ is in $\{-1, +1\}$ and when drawn according to a, it has mean $f(x)$. Since we are taking an average of m iid draws from this distribution, the statement follows by Hoeffding's inequality. □

Next we lift this result to any distribution over x, y

Lemma 4. *Let P be any distribution over (x, y) pairs. Then for $\theta > 0$, $m \geq 1$*

$$\mathbb{P}_{P,\tilde{f}}[|yf(x) - y\tilde{f}(x)| \geq \theta/2] \leq \beta_{m,\theta}.$$

Proof. The idea here is to marginalize and apply the Hoeffding bound from the previous lemma on the inner term.

$$\mathbb{P}_{P,\tilde{f}}[|yf(x) - y\tilde{f}(x)| \geq \theta/2] = \mathbb{E}_P \mathbb{P}_{\tilde{f}}[|yf(x) - y\tilde{f}(x)| \geq \theta/2] \leq \mathbb{E}_P \beta_{n,\theta} = \beta_{m,\theta}.$$ □

We can now establish the two inequalities relating f to $\tilde{f}$

$$\begin{aligned}
\mathbb{P}_{\mathcal{D}}[yf(x) \leq 0] &= \mathbb{P}_{\mathcal{D},\tilde{f}}[yf(x) \leq 0] \\
&\leq \mathbb{P}_{\mathcal{D},\tilde{f}}[y\tilde{f}(x) \leq \theta/2] + \mathbb{P}_{\mathcal{D},\tilde{f}}[yf(x) \leq 0, y\tilde{f}(x) \geq \theta/2] \\
&\leq \mathbb{P}_{\mathcal{D},\tilde{f}}[y\tilde{f}(x) \leq \theta/2] + \mathbb{P}_{\mathcal{D},\tilde{f}}[|yf(x) - y\tilde{f}(x) \geq \theta/2] \\
&\leq \mathbb{P}_{\mathcal{D},\tilde{f}}[y\tilde{f}(x) \leq \theta/2] + \beta_{m,\theta}.
\end{aligned}$$

Essentially the same calculation gives

$$\mathbb{P}_{S,\tilde{f}}[y\tilde{f}(x) \leq \theta/2] \leq \mathbb{P}_S[yf(x) \leq \theta] + \beta_{m,\theta}$$

So the last piece of the proof is a uniform convergence bound for the functions $\tilde{f} \in \mathcal{A}_m$. This is by now easy for us since we know that this set is finite in size. The counting is a bit tricky, but not anything particularly special.

Lemma 5. *Fix $\delta \in (0, 1)$. Then with probability at least $1 - \delta$ over the random training set, for all $m \geq 1$, for all $\tilde{f} \in \mathcal{A}_m$, and all $\theta > 0$*

$$\mathbb{P}_{\mathcal{D}}[y\tilde{f}(x) \leq \theta/2] \leq \mathbb{P}_S[y\tilde{f}(x) \leq \theta/2] + \sqrt{\frac{\log(m(m+1)^2|\mathcal{H}|^m/\delta)}{2n}}.$$

Proof. We can apply Hoeffding's inequality on each fixed $\tilde{f}, m, \theta$ since this is just asking for concentration of an empirical 0/1 loss to the population 0/1 loss. Specifically

$$\mathbb{P}_{\mathcal{D}}[y\tilde{f}(x) \leq \theta/2] = \mathbb{E}\mathbf{1}\{y\tilde{f}(x) \leq \theta/2\}$$

And the sample one is the sample average. Thus for each of these terms we get

$$\mathbb{P}_{S\sim\mathcal{D}^n}\left[\mathbb{E}\mathbf{1}\{y\tilde{f}(x)\le\theta/2\}-\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{y_i\tilde{f}(x_i)\le\theta/2\}\ge\epsilon\right]\le\exp(-2n\epsilon^2).$$

We just need to apply a union bound. The easy term is $\mathcal{A}_m$ which clearly has size $|\mathcal{H}|^m$. Then using integrality of the predictors $h(x)\in\{-1,+1\}$, we only have to consider $\theta\in\Theta_m=\{0,2/m,4/m,\ldots,2\}$ since for every x,y we have

$$\frac{1}{m}y\sum_{j=1}^{m}\tilde{h}_j(x)\le\frac{\theta}{2}\Rightarrow\frac{1}{m}y\sum_{j=1}^{m}\tilde{h}_j(x)\le\frac{1}{m}\left\lfloor\frac{m\theta}{2}\right\rfloor.$$

So only the integral values of $m\theta/2$ need to be considered. Note that this discretization is *not* dependent on the sample which is a subtle but important point. Thus we will take union bound over $m+1$ values of θ, and finally for m, we'll allocate failure probability $\frac{\delta}{m(m+1)}$. Taking the union bound, we get

$$\sum_{m\in\mathbb{N}}\sum_{\theta\in\Theta_m}\sum_{\tilde{f}\in\mathcal{A}_m}\exp(-2\epsilon_m^2n)\le\sum_{m\in\mathbb{N}}(m+1)|\mathcal{H}|^m\exp(-2\epsilon_m^2n)$$

Setting $\epsilon_m=\sqrt{\frac{\log(m(m+1)^2|\mathcal{H}|^m/\delta)}{2n}}$ the expression becomes

$$\sum_{m\in\mathbb{N}}\frac{\delta}{m(m+1)}\le\delta.$$

If you haven't seen it before, the series $1/m(m+1)$ can be written as a telescoping series $1/m-1/m+1$ which telescopes and evaluates to 1. □

We are now ready to finish the proof. Combining everything together, we have that with probability at least $1-\delta$, for any $m,\theta>0,f\in\mathcal{F}$

$$\begin{aligned}\mathbb{P}_{\mathcal{D}}[yf(x)\le0]&\le\mathbb{P}_S[yf(x)\le\theta/2]+2\beta_{m,\theta}+\sqrt{\frac{\log(m(m+1)^2|\mathcal{H}|^m/\delta)}{2n}}\\&=\mathbb{P}_S[yf(x)\le\theta/2]+4\exp(-m\theta^2/8)+\sqrt{\frac{\log(m(m+1)^2|\mathcal{H}|^m/\delta)}{2n}}\end{aligned}$$

We want to optimize this bound over m, but as usual it is a trascendental equation. At any rate if we set $m=\frac{4}{\theta^2}\log\left(\frac{4n\theta^2}{\log|\mathcal{H}|}\right)$ then we get the bound in the theorem statement. □

To recap the proof, we used a couple of new ideas. First we used the probabilistic method to discretize the function space $\mathcal{F}$. Part of this required using marginalization so that we could exploit the randomness in $\tilde{f}$ on every point in a distribution simultaneously. Then we used some clever counting to be uniform over the discretized space, which also had one continuous parameter.

Next time we'll study other "linear prediction" problems and revisit the notion of margin from a more geometric perspective.