COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire
Scribe: Neil Katuna

Lecture # 11
March 10, 2008

Last lecture, we focused on bounding the generalization error of AdaBoost, providing an upper bound in terms of the VC-dimension of the concept class of all linear threshold functions that express the final hypothesis in terms of its weak hypotheses. Using this bound, we saw that as the number of weak hypotheses grows large, our test error may increase correspondingly, causing *overfitting.*

This analysis said nothing of our confidence in our final hypothesis, though. We had expected that as the number of weak hypotheses increases, confidence in our final hypothesis would correspondingly increase, driving down the generalization error. Last time, to rigorously verify this notion, we introduced the *margin* of a labeled example under a hypothesis — intuitively, the weighted difference between the number of correctly labeled examples and the incorrectly labeled examples. We left off here, only sketching a proof of the following result — the focus of today's lecture.

1 Bounding the Margin

Recall that for a finite weak hypothesis space $\mathcal{H}$, we defined the convex hull of $\mathcal{H}$ to be

$$\mathrm{co}(\mathcal{H}) = \left\{ f : f(x) = \sum_{t=1}^{T} a_t h_t(x),\ T \geq 1,\ a_t \geq 0,\ \sum_{t=1}^{T} a_t = 1,\ h_1, \ldots, h_T \in \mathcal{H} \right\}.$$

For the rest of this lecture, $\mathcal{H}$ will denote such a finite weak hypothesis space. Also, it is important to note that the a_t define a distribution over $h_1, \ldots, h_t$. Lastly, note that we will never refer to the margin by name, but it will crop up frequently our discussion. Recall that $margin(x, y) = yf(x)$.

Theorem 1 *Let m be the size of the training set. For all hypotheses $f \in \mathrm{co}(\mathcal{H})$ and for all margin levels $\theta > 0$,*

$$\Pr_{\mathcal{D}}[\,yf(x) \leq 0\,] \leq \Pr_{S}[\,yf(x) \leq \theta\,] + O\left(\frac{1}{\sqrt{m}} \sqrt{\frac{\log m \log |\mathcal{H}|}{\theta^2} + \log 1/\delta} \right) \tag{1}$$

with probability $1 - \delta$.

Here, $\Pr_{\mathcal{D}}[\,\cdot\,]$ denotes the probability of an event given some (x, y) drawn from the true target distribution $\mathcal{D}$, and $\Pr_S[\,\cdot\,]$ denotes probability when (x, y) is chosen uniformly at random from the sample S. We will use this notation consistently throughout this article.

Note that Theorem 1 says nothing specifically about AdaBoost. Hence, it generalizes for any similar boosting algorithm, or any algorithm that combines hypotheses using a weighted majority vote.

Let us now give a sketch of a proof of our main result. Let

$$\mathcal{C}_N = \left\{ f : f(x) = \frac{1}{N} \sum_{j=1}^{N} h_j(x), \quad h_1, \ldots, h_N \in \mathcal{H} \right\}.$$

Clearly $\mathcal{C}_N \subset \mathrm{co}(\mathcal{H})$. We first claim that any $f \in \mathrm{co}(\mathcal{H})$ can be approximated by functions in $\mathcal{C}_N$. We will use this simple fact to exploit Chernoff bounds, specifically Hoeffding's inequality, to bound the probabilities we face.

Fix $f \in \mathrm{co}(\mathcal{H})$. By definition, $f(x) = \sum_{t=1}^{T} a_t h_t(x)$ for $a_t \geq 0$ and $\sum a_t = 1$. Now, for an appropriate N to be chosen later, define

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

where $g_j = h_t$ with probability a_t. That is, each g_j is chosen independently at random to be equal to one of the h_t's according to the distribution defined by the a_t's. Fixing x, we find that

$$\mathbb{E}_g[\, g_j(x)\,] = \sum_{t=1}^{t} \mathrm{Pr}_g[\, g_j = h_t\,] h_t(x) = \sum_{t=1}^{T} a_t h_t(x) = f(x).$$

where $\mathbb{E}_g$ and Pr_g refer to expectation and probability with respect to the random choice of g, respectively.

Our proof now reduces to finding the precise relations between

$$\mathrm{Pr}_{\mathcal{D}}[\, yf(x) \leq 0\,] \approx \mathrm{Pr}_{\mathcal{D}}[\, yg(x) \leq \theta/2\,] \approx \mathrm{Pr}_S[\, yg(x) \leq \theta/2\,] \approx \mathrm{Pr}_S[\, yf(x) \leq \theta\,].$$

Since f and g are approximately equal as already shown, we will use Chernoff bounds to clarify the first and third approximates. The second follows from the union bound. Let us now make this argument more formal with the use of some lemmas:

Lemma 1 *Fix an example x from $\mathcal{D}$. Then*

$$\mathrm{Pr}_g[\, |f(x) - g(x)| > \theta/2\,] \leq 2e^{-N\theta^2/8}.$$

Proof. Define random variables $Z_j = g_j(x) \in \{-1, 1\}$ so that $\mathbb{E}_g[\frac{1}{N}\sum Z_j] = f(x)$ as before. By Chernoff bounds,

$$\begin{aligned}
\mathrm{Pr}_g[\, |f(x) - g(x)| > \theta/2\,] &= \mathrm{Pr}_g\left[\left|\mathbb{E}_g\left[\frac{1}{N}\sum Z_j\right] - \frac{1}{N}\sum Z_j\right| > \theta/2\right] \\
&\leq 2e^{-2N(\theta/4)^2} \\
&= 2e^{-N\theta^2/8}.
\end{aligned}$$

Note that the Z_j must be rescaled to $\{0, 1\}$ to apply the Chernoff bounds as shown. For convenience, we denote $\beta_\theta := 2e^{-N\theta^2/8}$.

Lemma 2 *Say $(x, y) \sim \mathcal{P}$ for some distribution $\mathcal{P}$, where $y \in \{-1, 1\}$. Then,*

$$\mathrm{Pr}_{\mathcal{P},g}[\, |yf(x) - yg(x)| > \theta/2\,] \leq \beta_\theta.$$

Proof. With some computation, we find

$$\begin{aligned}
\mathrm{Pr}_{\mathcal{P},g}[\, |yf(x) - yg(x)| > \theta/2\,] &= \mathrm{Pr}_{\mathcal{P},g}[\, |f(x) - g(x)| > \theta/2\,] \\
&= \mathbb{E}_{\mathcal{P}}\left[\mathrm{Pr}_g[\, |f(x) - g(x)| > \theta/2\,]\right] \\
&\leq \mathbb{E}_{\mathcal{P}}[\, \beta_\theta\,] = \beta_\theta
\end{aligned}$$

as was to be shown.

Lemma 3 *Fix some $g \in \mathcal{C}_N$ and some $\theta > 0$. Then,*

$$\mathrm{Pr}_{S\sim\mathcal{D}^m}\left[\mathrm{Pr}_{\mathcal{D}}[\,yg(x) \leq \theta/2\,] > \mathrm{Pr}_S[\,yg(x) \leq \theta/2\,] + \epsilon\right] \leq e^{-2\epsilon^2 m}. \tag{2}$$

Here, $\mathrm{Pr}_{S\sim\mathcal{D}^m}[\,\cdot\,]$ denotes the probability of an event given a sample S of size m where each element of S is drawn independently at random from the distribution $\mathcal{D}$.

Proof. Define random variables

$$Z_i = \begin{cases} 1 & \text{if } y_i g(x_i) \leq \theta/2 \\ 0 & \text{otherwise} \end{cases}.$$

Hence, $\mathbb{E}[\,Z_i\,] = \mathrm{Pr}_{\mathcal{D}}[\,yg(x) \leq \theta/2\,]$ and $\frac{1}{m}\sum Z_i = \mathrm{Pr}_S[\,yg(x) \leq \theta/2\,]$. So, Equation 2 reduces to

$$\mathrm{Pr}_{S\sim\mathcal{D}^m}\left[\mathbb{E}[Z_i] > \frac{1}{m}\sum Z_i + \epsilon\right] \leq e^{-2\epsilon^2 m}$$

by Hoeffding's inequality.

Now we must show that the result in Lemma 3 holds for all $g \in \mathcal{C}_N$ and for all $\theta > 0$. We will accomplish this by taking the union bound.

Lemma 4

$$\mathrm{Pr}_{S\sim\mathcal{D}^m}\left[\forall g \in \mathcal{C}_n, \forall\theta > 0 : \mathrm{Pr}_{\mathcal{D}}[\,yg(x) \leq \theta/2\,] \leq \mathrm{Pr}_S[\,yg(x) \leq \theta/2\,] + \epsilon\right] \geq 1 - \delta \tag{3}$$

if

$$\epsilon = \sqrt{\frac{\log\left(\left(\frac{N}{2}+1\right)|\mathcal{H}|^N/\delta\right)}{2m}}.$$

Proof. First, observe that for any $y \in \{-1, 1\}$ and any $g \in \mathcal{C}_N$, where $g(x) = \frac{1}{N}\sum g_j(x)$,

$$\begin{aligned} yg(x) \leq \frac{\theta}{2} \quad &\Longleftrightarrow \quad \frac{y}{N}\sum g_j(x) \leq \frac{\theta}{2} \\ &\Longleftrightarrow \quad y\sum g_j(x) \leq \frac{N\theta}{2} \\ &\Longleftrightarrow \quad y\sum g_j(x) \leq \left\lfloor \frac{N\theta}{2} \right\rfloor \\ &\Longleftrightarrow \quad yg(x) \leq \frac{\widetilde{\theta}}{2} \end{aligned}$$

where $\widetilde{\theta} = 2\lfloor N\theta/2 \rfloor/N$. Note that these steps hold because $y\sum g_j(x)$ is always an integer. Furthermore, notice that $\lfloor N\theta/2 \rfloor$ takes values in $\{0, \ldots, N/2\}$. Thus, when bounding the probability in Equation 3, we only need consider $(N/2 + 1)$ values of θ.

Therefore, by the union bound,

$$\begin{aligned} &\mathrm{Pr}_{S\sim\mathcal{D}^m}\left[\ \exists g \in \mathcal{C}_n, \exists\theta > 0 : \mathrm{Pr}_{\mathcal{D}}[\,yg(x) \leq \theta/2\,] > \mathrm{Pr}_S[\,yg(x) \leq \theta/2\,] + \epsilon\ \right] \\ &= \quad \mathrm{Pr}_{S\sim\mathcal{D}^m}\left[\ \exists g \in \mathcal{C}_n, \exists\theta > 0 : \mathrm{Pr}_{\mathcal{D}}[\,yg(x) \leq \widetilde{\theta}/2\,] > \mathrm{Pr}_S[\,yg(x) \leq \widetilde{\theta}/2\,] + \epsilon\ \right] \\ &\leq \quad |\mathcal{C}_N|\left(\frac{N}{2}+1\right)e^{-2\epsilon^2 m} \\ &\leq \quad |\mathcal{H}|^N\left(\frac{N}{2}+1\right)e^{-2\epsilon^2 m}. \end{aligned} \tag{4}$$

Note that the inequality follows from Lemma 3 and the fact that we only care about those values of θ which affect $\widetilde{\theta}$. Setting Equation 4 above equal to δ and solving gives the result.

We are now in a position to tie our four lemmas together to complete the proof of our main result.

Proof of Theorem 1. For all hypotheses $f \in \mathrm{co}(\mathcal{H})$ and for all margin levels $\theta > 0$, observe that

$$\begin{aligned}
&\Pr_{\mathcal{D}}[\, yf(x) \leq 0\,] \\
&\qquad = \Pr_{\mathcal{D},g}[\, yf(x) \leq 0 \wedge yg(x) \leq \theta/2\,] + \Pr_{\mathcal{D},g}[\, yf(x) \leq 0 \wedge yg(x) > \theta/2\,] \\
&\qquad \leq \Pr_{\mathcal{D},g}[\, yg(x) \leq \theta/2\,] + \Pr_{\mathcal{D},g}[\, |yf(x) - yg(x)| > \theta/2\,] \\
&\qquad \leq \mathbb{E}_g[\, \Pr_{\mathcal{D}}[\, yg(x) \leq \theta/2 \,|\, g\,]\,] + \beta_\theta
\end{aligned}$$

by Lemma 2. However, by Lemma 4, with probability greater than or equal to $1 - \delta$,

$$\mathbb{E}_g[\, \Pr_{\mathcal{D}}[\, yg(x) \leq \theta/2 \,|\, g\,]\,] + \beta_\theta \leq \mathbb{E}_g[\, \Pr_{\mathcal{S}}[\, yg(x) \leq \theta/2 \,|\, g\,] + \epsilon\,] + \beta_\theta.$$

Using Lemma 2 once more,

$$\begin{aligned}
&\mathbb{E}_g[\, \Pr_{\mathcal{S}}[\, yg(x) \leq \theta/2 \,|\, g\,] + \epsilon\,] + \beta_\theta \\
&\qquad = \Pr_{S,g}[\, yg(x) \leq \theta/2\,] + \epsilon + \beta_\theta \\
&\qquad = \Pr_{S,g}[\, yg(x) \leq \theta/2 \wedge yf(x) \leq \theta\,] + \Pr_{S,g}[\, yg(x) \leq \theta/2 \wedge yf(x) > \theta\,] + \epsilon + \beta_\theta \\
&\qquad \leq \Pr_{S,g}[\, yf(x) \leq \theta\,] + \Pr_{S,g}[\, |yf(x) - yg(x)| > \theta/2\,] + \epsilon + \beta_\theta \\
&\qquad \leq \Pr_{S}[\, yf(x) \leq \theta\,] + \beta_\theta + \epsilon + \beta_\theta.
\end{aligned}$$

To summarize, we have found that

$$\Pr_{\mathcal{D}}[\, yf(x) \leq 0\,] \leq \Pr_{S}[\, yf(x) \leq \theta\,] + 4e^{-N\theta^2/8} + \sqrt{\frac{\log\left(\left(\frac{N}{2}+1\right)|\mathcal{H}|^N/\delta\right)}{2m}}$$

with probability greater than or equal to $1 - \delta$. Setting

$$N = \left\lceil \frac{4}{\theta^2} \log\left(\frac{m}{\log|\mathcal{H}|}\right)\right\rceil$$

and slogging through the algebra gives the desired result.

Now that we have analyzed boosting in enough depth, let us set it aside for now. Boosting, while very intuitively satisfying, was not explicitly designed to maximize margins. It is time that we turn to a new way of calculating consistent hypotheses called *support vector machines.*

2 Support Vector Machines

Given a collection of labeled examples $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)$ where $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \{-1, 1\}$, we wish to find a hypothesis consistent with all m examples. One could think of the coordinates of $\mathbf{x}_i$ as specific attributes, which, taken together, form a vector which lies in $\mathbb{R}^n$. Assuming for now that the data is linearly separable, we can take our hypotheses

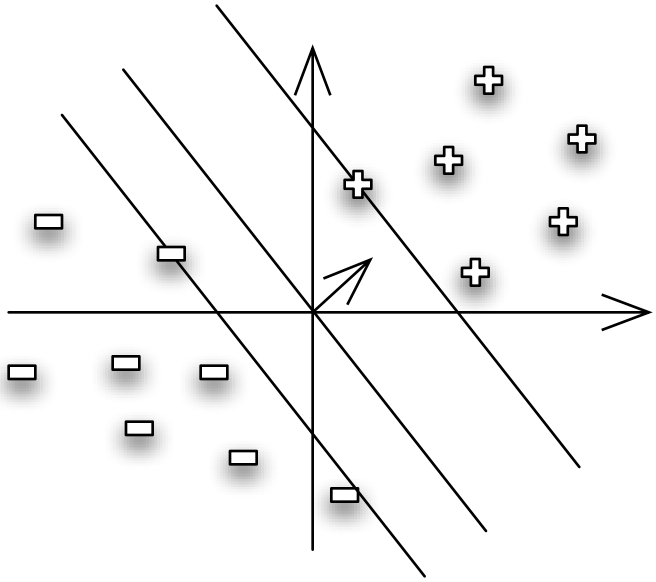

Figure 1: Labeled examples in the plane with a separating hyperplane (a line) passing through the origin. We wish to maximize the distance δ between the line through the origin and the other two parallel lines.

to be hyperplanes which separate positively labeled examples from negatively labeled ones. See Figure 1.

We define our hyperplane by a perpendicular vector $\mathbf{v} \in \mathbb{R}^n$ and always assume that it passes through the origin. The quantity $\mathbf{v} \cdot \mathbf{x}_i$ therefore represents how far $\mathbf{x}_i$ is from the separating hyperplane. Specifically,

$$\mathbf{v} \cdot \mathbf{x}_i = \begin{cases} > 0 & \text{if } \mathbf{x}_i \text{ is above the hyperplane} \\ = 0 & \text{if } \mathbf{x}_i \text{ is on the hyperplane} \\ < 0 & \text{if } \mathbf{x}_i \text{ is below the hyperplane} \end{cases}$$

Hence, we take our prediction rule to be $sign(\mathbf{v} \cdot \mathbf{x}_i)$ and we similarly define the margin as $margin(\mathbf{x}_i, y_i) = y_i(\mathbf{v} \cdot \mathbf{x}_i)$. The points for which the margin is exactly δ are called *support vectors*. Consequently, the best choice of $\mathbf{v}$ maximizes $\delta > 0$ such that $\|\mathbf{v}\| = 1$ and $y_i(\mathbf{v} \cdot \mathbf{x}_i) \geq \delta$ for all i. This criterion will be the goal of our further study.