COS 511: Foundations of Machine Learning

Rob Schapire
Scribe: Kam Jim

Lecture #9
March 4, 2003

1 Previous Lecture

Last time we talked about *A*daBoost. The name comes from the word "adaptive", because $\epsilon_t \leq 1/2 - \gamma$, and we don't need to know γ ahead of time. Today we will talk about generalization error.

2 Overfitting

Look for *s*trong hypotheses for AdaBoost, and figure out the VC dimension, or better yet, the growth function. Suppose $h_1, ..., h_T \in \mathcal{H}$, where $|\mathcal{H}|$ is finite. Define $H(x)$ as

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

$H(x)$ is a strong hypothesis. Let G be the class of functions of the form $\sum_{t=1}^{T} \alpha_t h_t(x)$. Next lets determine the growth function $\Pi_G(m)$.

Fix $h_1, ..., h_T$. Given $S = x_1, ..., x_m$, how many dichotomies can we get of these m points? Take each x_i and pass it to the h's:

$$x_i \Rightarrow [h_1(x_i), ..., h_T(x_i)] = x_i'. \tag{2}$$

x_i' has dimension T. Pretend the x_i''s are the new sample points. Now we have a linear threshold function on inputs with dimension T (see Equation 1). From our homework, we know that the VC dimension is T, and therefore the growth function, with $h_1, \ldots, h_T$ fixed is at most

$$\left(\frac{em}{T}\right)^T$$

Since there are $|\mathcal{H}|^T$ choices for $h_1, \ldots, h_T$, the growth function for G is upper-bounded as follows:

$$\Pi_G(m) \leq |\mathcal{H}|^T \left(\frac{em}{T}\right)^T. \tag{3}$$

Using the fact that

$$err(H) \leq \widehat{err}(H) + O\left(\sqrt{\frac{\ln|\Pi_G(2m)| + \ln\frac{!}{\delta}}{m}}\right), \tag{4}$$

and plugging in for Π_G, we obtain:

$$err(H) \leq \widehat{err}(H) + O\left(\sqrt{\frac{T\ln|\mathcal{H}| + T\ln\frac{m}{T} + \ln\frac{1}{\delta}}{m}}\right). \tag{5}$$

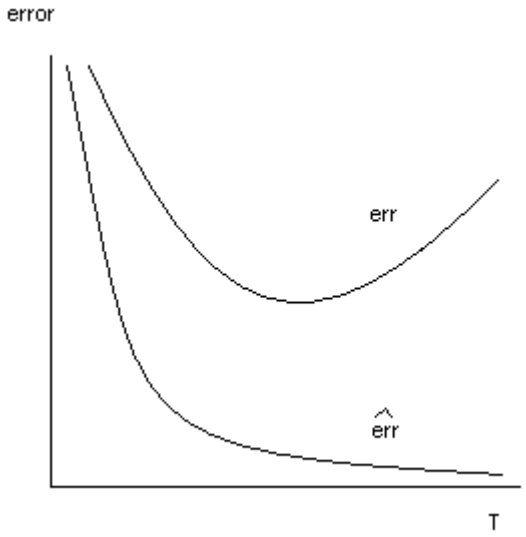

Figure 1: Expected generalization behavior due to overfitting.

This predicts overfitting. As T increases, $\widehat{err}(H)$ decreases, but the $O()$ term in Equation 5 increases and eventually overtakes the empirical error $\widehat{err}(H)$. The expected behavior looks like Figure 1.

However, this overfitting often does not happen with AdaBoost. For example, the performance on boosting the C4.5 decision tree learning algorithm is shown in Figure 2. C4.5 is a *weak* algorithm. The learned rules are more complicated with increasing values for T. In this example, the test error does not increase even after 1000 rounds, and in fact continues to drop even after the training error is zero. Hence, Occam's razor incorrectly predicts that simpler rules are better here.

3 Margins

Why don't we see overfitting more often? Claim: As you continue to boost, the predictions become more "confident." The *margin* is associated with confidence. The higher the margin, the better the generalization error.

Let's rewrite Equation 1 slightly:

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} a_t h_t(x)\right) \quad (6)$$

$$a_t \geq 0 \quad (7)$$

$$\sum_{t=1}^{T} a_t = 1 \quad (8)$$

In other words, we have normalized the α_t's for convenience – this doesn't change the

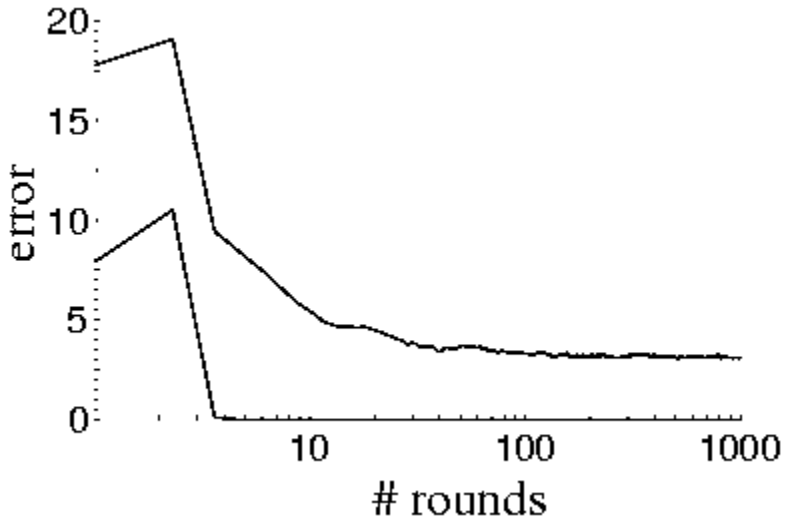

Figure 2: Error curves for boosting C4.5 on the letter dataset. Taken from Schapire, "The Boosting Approach to Machine Learning - An Overview", 2001.

predictions. Define the margin as follows

$$\begin{aligned} margin &= yf(x), \quad \text{where } y \in \{+1, -1\} \text{ is the correct label} && (9) \\ &= y\sum_{t=1}^{T} a_t h_t(x) = \sum_{t=1}^{T} a_t y h_t(x) && (10) \\ &= \sum_{t:y=h_t(x)} a_t - \sum_{t:y\neq h_t(x)} a_t && (11) \end{aligned}$$

The first term in Equation 11 corresponds to the weak hypotheses with correct predictions, and the second term in Equation 11 corresponds to weak hypotheses with incorrect predictions. The margin is positive if and only if $y = H(x)$. In other words, $yf(x) > 0$ iff $y = H(x)$. Also, call $|yf(x)| = |f(x)|$ the "strength" or confidence of the vote.

Figure 3 shows the cumulative distribution of the margins of the training examples after 5, 100, and 1000 rounds of boosting on the C4.5 algorithm applied to the letter dataset. Here we see that boosting increases the margins, and continues to increase the margins even after the training error reaches zero (after 10 rounds).

4 Connection Between Margin and Generalization Error

We will use the following notation. Let $\mathcal{H}$ be the space of weak hypotheses. Assume $|\mathcal{H}| < \infty$. Define $co(\mathcal{H})$ to be the convex hull of $\mathcal{H}$. In other words,

$$co(\mathcal{H}) = \left\{ f \text{ of the form } f(x) = \sum_{t=1}^{T} a_t h_t(x), \text{ where } a_t \geq 0, \sum a_t = 1, h_t \in \mathcal{H} \right\} \quad (12)$$

$\mathcal{D}$ is the distribution on $\mathcal{X} \times \{-1, 1\}$. $\mathcal{S}$ is a sample of size m. $\Pr_{\mathcal{D}}[\cdot]$ is the probability when (x, y) is chosen from $\mathcal{D}$: for example, $\Pr_{\mathcal{D}}[y \neq H(x)] = \text{error}$. $\Pr_{\mathcal{S}}[\cdot]$ is the probability when

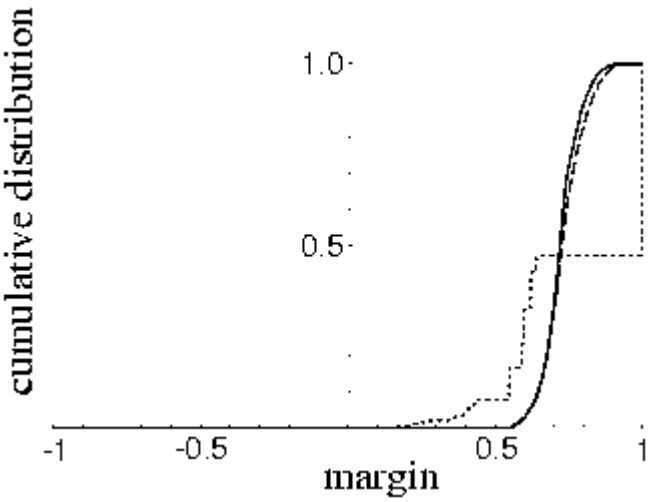

Figure 3: Cumulative distribution of margins for boosting C4.5 on the letter dataset. Taken from Schapire, "The Boosting Approach to Machine Learning - An Overview", 2001.

(x, y) is chosen from $\mathcal{S}$: for example, $\Pr_{\mathcal{S}}[y \neq H(x)]$ = training error. We want to prove the following theorem:

Theorem 4.1. *With probability at least* $1 - \delta$, $\forall f \in co(\mathcal{H})$, $\forall \theta > 0$,

$$\Pr_{\mathcal{D}}[yf(x) \leq 0] \leq \Pr_{\mathcal{S}}[yf(x) \leq \theta] + O\left(\frac{1}{\sqrt{m}}\sqrt{\frac{\ln m \ln |\mathcal{H}|}{\theta^2} + \ln \frac{1}{\delta}}\right). \tag{13}$$

Note that the $O(\cdot)$ term is independent of the number of rounds of boosting.

We want to approximate $co(\mathcal{H})$ with a much smaller class. Let $\mathcal{C}_N$ be the set of unweighted averages over N elements from $\mathcal{H}$:

$$\mathcal{C}_N = \left\{ g \text{ of the form } g(x) = \frac{1}{N}\sum_{j=1}^{N} h_j(x) \right\} \tag{14}$$

Given

$$f(x) = \sum a_t h_t(x),$$

we will construct a $g \in \mathcal{C}_N$ that approximates f. We will construct g randomly, and we will write $\Pr_g[\cdot]$ to denote probabilities over the random choice of g. Notice that $|\mathcal{C}_N| \leq |\mathcal{H}|^N$, because we have $|\mathcal{H}|$ choices of h_j, and there are N of them. Also, we expect $N \ll T$. We construct each g by randomly choosing N elements from the set of h_t's, where each h_t is chosen with probability a_t. We can rewrite this as

$$g(x) = \frac{1}{N}\sum_{j=1}^{N} g_j(x), \tag{15}$$

where $g_j = h_t$ with probability a_t.

Before proceeding with the proof, notice that if we fix x, then

$$E_g[g_j(x)] = \sum_{t=1}^{T} a_t h_t(x) = f(x) \tag{16}$$

$$E_g[g(x)] = f(x) \tag{17}$$

Hence our intuition for the proof is the following: since $g(x) = \frac{1}{N}\sum g_j(x)$, then using Chernoff bounds $|g(x) - f(x)| \leq \epsilon$; i.e. $g(x)$ and $f(x)$ should be close. The outline of the proof is to show the following:

$$\Pr_{\mathcal{D}}[yf(x) \leq 0] \approx \Pr_{\mathcal{D}}[yg(x) \leq \theta/2] \tag{18}$$

$$\approx \Pr_{\mathcal{S}}[yg(x) \leq \theta/2] \tag{19}$$

$$\approx \Pr_{\mathcal{S}}[yf(x) \leq \theta] \tag{20}$$

Proof The generalization error can be rewritten as $\Pr_{\mathcal{D}}[yf(x) \leq 0] = \Pr_{\mathcal{D},g}[yf(x) \leq 0]$, because g doesn't appear on the left-hand side equation, so it's OK to randomize on it. Separate this probability into 2 terms:

$$\Pr_{\mathcal{D},g}[yf(x) \leq 0] = \Pr_{\mathcal{D},g}\left[yf(x) \leq 0 \wedge yg(x) \leq \theta/2\right] + \Pr_{\mathcal{D},g}\left[yf(x) \leq 0 \wedge yg(x) > \theta/2\right] \tag{21}$$

The second term can be rewritten as

$$\Pr_{\mathcal{D},g}\left[yf(x) \leq 0 \wedge yg(x) > \theta/2\right] = E_{\mathcal{D}}\left[\Pr_g\left[yf(x) \leq 0 \wedge yg(x) > \theta/2\right]\right] \tag{22}$$

$\Pr_g[yf(x) \leq 0]$ is 0 if $yf(x) > 0$, otherwise we know $yf(x) \leq 0$ (hence we will obtain an upper bound for the second term by assuming that $yf(x) \leq 0$). Also $\Pr_g[yg(x) > \theta/2] \leq e^{-N\theta^2/8}$ by applying Hoeffding's Inequality to the expectation of $g(x)$ (see Equation 17). Hence we can bound the second term as follows

$$\Pr_{\mathcal{D},g}\left[yf(x) \leq 0 \wedge yg(x) > \theta/2\right] \leq E_{\mathcal{D}}\left[e^{-N\theta^2/8}\right] = e^{-N\theta^2/8} \tag{23}$$

To get an upper bound on the first term, we begin by using the fact that $\Pr[A \wedge B] \leq \Pr[B]$ and rewrite the first term using expectations:

$$\Pr_{\mathcal{D},g}\left[yf(x) \leq 0 \wedge yg(x) \leq \theta/2\right] \leq \Pr_{\mathcal{D},g}[yg(x) \leq \theta/2] \tag{24}$$

$$= E_g\left[\Pr_{\mathcal{D}}\left[yg(x) \leq \theta/2\right]\right] \tag{25}$$

We want to show that $\Pr_{\mathcal{D}}[yg(x) \leq \theta/2] \approx \Pr_{\mathcal{S}}[yg(x) \leq \theta/s]$. Use the fact that

$$\Pr_{\mathcal{S}}[yg(x) \leq \theta/2] = \frac{1}{m}\sum_{i=1}^{m}[[y_i g(x_i) \leq \theta/2]] \tag{26}$$

(recall that $[[\cdot]]$ is +1 if the argument is true, 0 otherwise). Observe that the expectation of $[[y_i g(x_i) \leq \theta/2]]$ is $\Pr_{\mathcal{D}}[yg(x) \leq \theta/2]$. Hence we can use Chernoff bounds by fixing g and θ:

$$\Pr_{sample}\left[\Pr_{\mathcal{D}}\left[yg(x) \leq \theta/2\right] > \Pr_{\mathcal{S}}\left[yg(x) \leq \theta/2\right] + \epsilon_\theta\right] \leq e^{-2\epsilon_\theta^2 m} \tag{27}$$

Here $\Pr_{sample}[\cdot]$ is the probability taken over a random choice of $\mathcal{S}$, whereas $\Pr_{\mathcal{S}}[\cdot]$ is the probability with an already chosen sample S. Using the Union Bound for a *fixed* θ,

$$\Pr_{sample}\left[\exists g \in \mathcal{C}_N : \Pr_{\mathcal{D}}\left[yg(x) \leq \theta/2\right] > \Pr_{\mathcal{S}}\left[yg(x) \leq \theta/2\right] + \epsilon_\theta\right] \leq |\mathcal{H}|^N e^{-2\epsilon_\theta^2 m}. \tag{28}$$

To account for θ, we note that it's not necessary to consider an infinite number of θ's. $yg(x)$ is always a multiple of $1/N$, and is between 0 and 1. Hence we only need to consider values of θ of the form $2i/N$ for $i = 0, ..., N$. (I.e. there are $(N + 1)$ values for θ.) Hence, by the union bound,

$$\Pr_{sample}\left[\exists\theta > 0, \exists g \in \mathcal{C}_N : \Pr_{\mathcal{D}}\left[yg(x) \leq \theta/2\right] > \Pr_{\mathcal{S}}\left[yg(x) \leq \theta/2\right] + \epsilon_\theta\right] \leq \delta$$

if we choose ϵ_θ so that

$$|\mathcal{H}|^N e^{-2\epsilon_\theta^2 m} = \delta/(N + 1). \tag{29}$$

Now we have, with probability at least $1 - \delta$, $\forall g \in \mathcal{C}_N$, $\forall\theta > 0$, $\Pr_{\mathcal{D}}[yg(x) \leq \theta/2] \leq \Pr_g[yg(x) \leq \theta/2] + \epsilon_\theta$. Solving for ϵ_θ by using Equation 29 gives:

$$\epsilon_\theta = \sqrt{\frac{1}{2m} \ln\left(\frac{(N + 1)|\mathcal{H}|^N}{\delta}\right)} \tag{30}$$

To be continued next lecture.