COS 511: Foundations of Machine Learning

Rob Schapire
Scribe: Alina Ene

Lecture #10
March 9, 2006

1 Last Lecture

Last lecture we looked at the **AdaBoost** algorithm. This lecture we will analyze the generalization error of **AdaBoost**:

Pseudocode for **AdaBoost:**

```
given ($x_1$, $y_1$),...,($x_m$, $y_m$) $y_i \in \{-1, +1\}$
    for t = 1,..,T
        construct $D_t$
        train $h_t$ on $D_t$
            $\epsilon_t = err_{D_t}(h_t) \leq \frac{1}{2} - \gamma$
    output
        $H(x) = sign\left(\sum_t \alpha_t h_t(x)\right)$
```

2 Analyzing the generalization error of AdaBoost

The main thing we are interested in is the form of $H(x)$:

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

$H(x)$ is a linear threshold function.
Suppose h_1,..,$h_T \in \mathcal{H}$ where $\mathcal{H}$ is finite. Let $\mathcal{G} = \{$ linear threshold functions over $\mathcal{H}$ $\}$
We want to determine the growth function $\Pi_{\mathcal{G}}(m)$.
Let's fix h_1,...,h_T. Given S = $\{x_1$,..,$x_m\}$, how many dichotomies are there on these m points? We know that the VC-dimension of a linear combination with T variables is T (see problem 2 on Homework 2), therefore the number of dichotomies over m points is at most

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

There are $|\mathcal{H}|^T$ choices for $h_1, .., h_T$, therefore we can bound the growth function as follows:

$$\Pi_{\mathcal{G}}(m) \leq |\mathcal{H}|^T \cdot \left(\frac{em}{T}\right)^T \tag{2}$$

By taking the logarithm, we get:

$$\ln \Pi_{\mathcal{G}}(2m) \leq T\left(\ln\left(\frac{em}{T}\right) + \ln|\mathcal{H}|\right) \tag{3}$$

We know that with probability at least $1 - \delta$:

$$err(H) \leq \widehat{err}(H) + O\left(\sqrt{\frac{\ln(\Pi_{\mathcal{G}}(2m)) + \ln(\frac{1}{\delta})}{m}}\right) \tag{4}$$

Putting together equation (3) and equation (4):

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

3 Overfitting

The equation that we derived in the previous section predicts overfitting. As T increases, $\widehat{err}(H)$ decreases, but the $O(\cdot)$ term increases and we expect the true error to increase as well. The expected behaviour is depicted in **Figure 1**.
However, overfitting often does not happen with AdaBoost. Let's take a look at an actual typical run. **Figure 2** depicts the error curves for boosting C4.5 on the letter dataset. C4.5 is a weak learner. As T increases, the learned rules get increasingly complicated. However, even after 1000 rounds, the test error does not increase, and actually continues to drop even after the training error is zero. Therefore Occam's razor wrongly predicts that simpler rules are better.
Explanation:
Consider the confidence of each prediction, instead of whether the prediction is right or wrong. The increasing confidence on the training set translates to better performance on the test set. How can we measure confidence? We can consider each weak hypothesis as a voter.The classifier $H(x)$ is a weighted majority vote. Let's define the *margin* to be the difference between the weighted number of hypotheses voting for the right label and the weighted number of hypotheses voting for the wrong label.
Let's rewrite $H(x)$ as follows:

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

where $a_t \geq 0$, $\sum_t a_t = 1$
We have normalized the α_t's without changing the predictions.
Let $f(x) = \sum_{t=1}^{T} a_t h_t(x)$. Let's note that $f(x)$ is a linear combination of weak hypotheses.
We define the margin as follows:

$$\begin{aligned} margin(x,y) &= y\cdot f(x) && (7)\\ &= y\cdot \sum_{t=1}^{T} a_t h_t(x) && (8)\\ &= \sum_{t=1}^{T} a_t y h_t(x) && (9)\\ &= \sum_{t:h_t(x)=y} a_t - \sum_{t:h_t(x)\neq y} a_t && (10) \end{aligned}$$

The margin is positive if and only if H is correct. Also, $|yf(x)| = |f(x)|$ represents the confidence in 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 on the letter dataset. We can see that the margins are pushed to the right by boosting, and this is correlated with the drop of the test error.

4 The margin distribution

First, let's introduce some notation:
$\mathcal{H}$ = the space of the weak hypotheses (we assume that $\mathcal{H}$ is finite)
$co(\mathcal{H})$ = the convex hull of $\mathcal{H}$
$co(\mathcal{H}) = \{$ f of the form $f(x) = \sum_{t=1}^{T} a_t h_t(x)$, where $a_t \geq 0$, $\sum a_t = 1, h_t \in \mathcal{H}$ $\}$
$Pr_{\mathcal{D}}[\cdot]$ = the probability with respect to the true distribution $\mathcal{D}$
$Pr_{\mathcal{D}}[H(x) \neq y]$ = generalization error
$Pr_S[\cdot]$ = the probability with respect to the uniform distribution on the training set S
$Pr_S[H(x) \neq y]$ = the training error

Theorem 4.1.

$$Pr_S[y\cdot f(x) \leq \theta] \leq \prod_{t=1}^{T} \left(2\sqrt{{\epsilon_t}^{1-\theta}(1-\epsilon_t)^{1+\theta}} \right)$$

If $\epsilon_t \leq \frac{1}{2} - \gamma$:

$$Pr_S[y\cdot f(x) \leq \theta] \leq \left(\sqrt{(1-2\gamma)^{1-\theta}(1+2\gamma)^{1+\theta}} \right)^T$$

If $\theta < \gamma$:

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

Therefore:

$$\lim_{T\to\infty} \left(\sqrt{(1-2\gamma)^{1-\theta}(1+2\gamma)^{1+\theta}} \right)^T = 0$$

and

$$Pr_S[y\cdot f(x) \leq \theta] \to 0$$

Putting it all together, the theorem tells us that if $\epsilon_t \leq \frac{1}{2} - \gamma$, then

$$\lim_{T\to\infty} \min_i y_i f(x_i) \geq \gamma$$

Theorem 4.2. *With probability at least* $1-\delta$, $\forall f \in co(\mathcal{H}), \forall \theta > 0$*:*

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

Let's note that the $O(\cdot)$ term is independent of T.

Proof:
Let:
$f(x) = \sum a_t h_t(x)$
$g_j =$ randomly chosen h_t
$Pr[g_j \equiv h_t] = a_t$
$g(x) = \frac{1}{N}\sum\limits_{j=1}^{N} g_j(x)$
$\mathcal{C}_N = \{g$ of the form $g(x) = \frac{1}{N}\sum\limits_{j=1}^{N} g_j(x)$, where $g_j = h_t$ for some t $\}$
That is, we construct $g(x)$ by randomly choosing N elements from the set of the h_t's, where each h_t is chosen with probability a_t. We claim that $g(x)$ will be a good approximation of $f(x)$ when the margin is large.
Before continuing with the proof, let's note that if we fix x, then:

$$\begin{aligned} E_g[g_j(x)] &= \sum_{t=1}^{T} a_t h_t(x) && (11) \\ &= f(x) && (12) \end{aligned}$$

We will now give the outline of the proof, and we will prove the theorem next lecture.
We want to show the following equivalences:

$$Pr_{\mathcal{D}}[yf(x) \leq 0] \approx Pr_{\mathcal{D}}[yg(x) \leq \frac{\theta}{2}] \approx Pr_S[yg(x) \leq \frac{\theta}{2}] \approx Pr_S[yf(x) \leq \theta] \qquad (13)$$

We will obtain the first approximation using Chernoff bounds, the second approximation using uniform convergence, and the third approximation using Chernoff bounds.

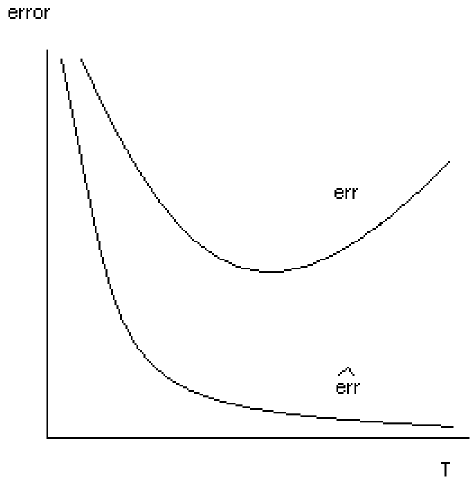

Figure 1: Expected generalization error due to overfitting.

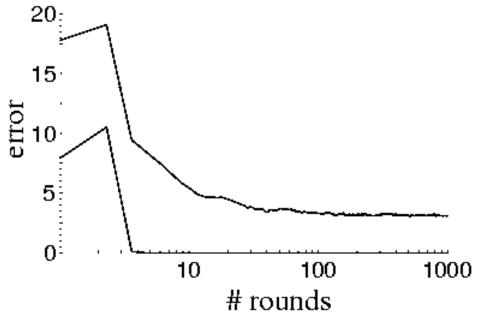

Figure 2: Error curves for boosting C4.5 on the letter dataset. (Robert E. Schapire, "The boosting approach to machine learning: An overview")

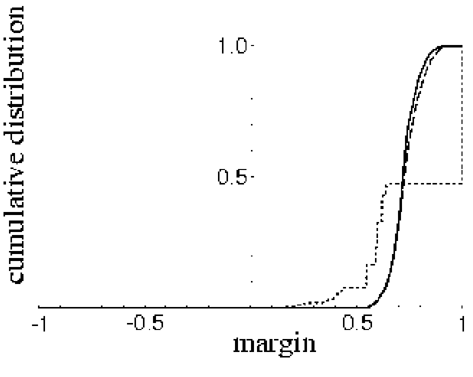

Figure 3: Cumulative distribution of margins for boosting C4.5 on the letter dataset. (Robert E. Schapire, "The boosting approach to machine learning: An overview")