COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire — Lecture #10
Scribe: Haipeng Zheng — March 5, 2008

1 Review of AdaBoost Algorithm

Here is the AdaBoost Algorithm:

```
input: $(x_1, y_1), \dots, (x_m, y_m)$, where $x_i \in \mathcal{X}$, and $y_i \in \{-1, +1\}$;
initialization: $D_1(i) = 1/m$;
for $t$ from 1 to $T$ do
    run $A$ on $D_t$ and get $h_t : \mathcal{X} \to \{-1, +1\}, h_t \in \mathcal{H}$;
    $D_{t+1}(i) = \frac{D_t(i)}{Z_t} e^{-\alpha_t y_i h_t(x_i)}$;
    where $\alpha_t = \frac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right), \epsilon_t = \mathrm{err}_{D_t}(h_t)$, and $Z_t$ is normalization factor.
end
output: final/combined hypothesis: $H(x) = \mathrm{sign}(\sum_{t=1}^{T} \alpha_t h_t(x))$.
```

2 Upper-bound of Generalization Error of AdaBoost

Last time, we have shown that the training error goes down very quickly with respect to the number of rounds of boosting T[1]. However, what is the performance of the *generalization error* of AdaBoost? How do we estimate the upper-bound of the generalization error?

2.1 An Upper-bound Based on VC-dimension

The combined hypothesis has a "weighted-vote" form. Define

$$\mathscr{G} = \{\text{all functions of form sign}\left(\textstyle\sum_{t=1}^{T} \alpha_t h_t(x)\right)\}.$$

If H is consistent and $|\mathscr{G}|$ is finite, then with probability $1-\delta$,

$$\mathrm{err}(H) \le O\left(\frac{\ln|\mathscr{G}| + \ln(1/\delta)}{m}\right).$$

However, since the number of choices of α_t is uncountable, $|\mathscr{G}|$ is infinite, hence the above bound is futile. Therefore, we need to use VC-dimension to bound the generalization error. Using the theorem in the previous two lectures and Homework 3, the bound is of the form

$$\mathrm{err}(H) \le \widehat{\mathrm{err}}(H) + O\left(\sqrt{\frac{\ln \Pi_{\mathscr{G}}(2m) + \ln(1/\delta)}{m}}\right). \tag{1}$$

On the right hand side, the first term $\widehat{\mathrm{err}}(H)$ addresses the error caused by inconsistent H, and the second term $O\left(\sqrt{\left(\ln \Pi_{\mathscr{G}}(2m) + \ln(1/\delta)\right)/m}\right)$ is the hard part to prove. Now the key question is to figure out the growth function $\Pi_{\mathscr{G}}(2m)$ for the combined hypothesis.

[1] $\widehat{\mathrm{err}}(H) \le \exp\left\{-2\sum_{t=1}^{T} \gamma_t^2\right\}$

To make our life easier, assume $|\mathcal{H}| < \infty$. In order to bound $\Pi_{\mathscr{G}}(2m)$, we further assume that $x_1, \ldots, x_m$ and $h_1, \ldots, h_T$ are fixed. Now $\alpha_1, \ldots, \alpha_T$ are the only variables. Also notice $\sum_t \alpha_t h_t(x)$ is a linear function, and $\operatorname{sign}(\cdot)$ is a threshold function. By Problem 2 in Homework 2, we know that linear threshold functions in $\mathbb{R}^n$ have VC-dimension n.

Define $\mathbf{z}_i = \langle h_1(\mathbf{x}_i), \ldots, h_T(\mathbf{x}_i)\rangle$, then

$$H(\mathbf{x}_i) = \operatorname{sign}(\boldsymbol{\alpha} \cdot \mathbf{z}_i),$$

where $\boldsymbol{\alpha} = \langle \alpha_1, \ldots, \alpha_T\rangle$. Since we have m different inputs, and $\operatorname{sign}(\boldsymbol{\alpha}\cdot\mathbf{z}_i)$ has VC-dimension T, by Sauer's Lemma, the number of dichotomies, given $h_1, \ldots, h_T$ fixed, is upper-bounded by

$$\#\text{ of dichotomies} \le \sum_{i=0}^{T} \left(\begin{array}{c} m \\ i \end{array}\right) \le \left(\frac{em}{T}\right)^T.$$

For each choice of $h_1, \ldots, h_T$, we have no more than $\left(\frac{em}{T}\right)^T$ dichotomies, and we have a total of at most $|\mathcal{H}|^T$ choices of $h_1, \ldots, h_T$. Therefore,

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

Plugging (2) into (1), we get

$$\operatorname{err}(H) \le \widehat{\operatorname{err}(H)} + O\left(\sqrt{\frac{T \ln|H| + T\ln(m/T) + \ln(1/\delta)}{m}}\right), \forall H \in \mathscr{G}$$

with probability at least $1 - \delta$.

In the above expression, we dislike T, the number of weak hypotheses in H. Intuitively, T increases the complexity of the combined hypothesis. The bound indicates the trend of generalization error shown in Figure 1. It decreases at first, yet finally increases as T increases. This is exactly the kind of overfitting behavior we expect since intuitively, the complexity of the combined hypothesis is equal to its size, which is proportional to the number of rounds T.

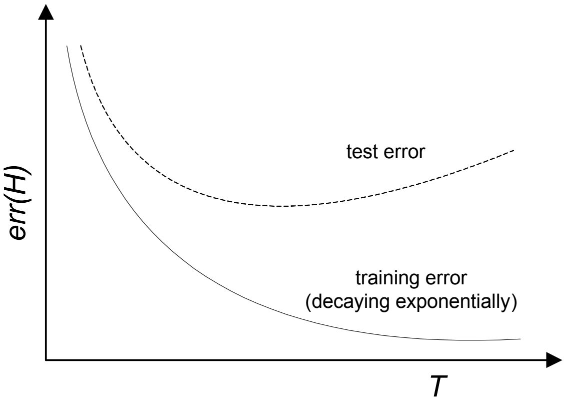

Figure 1: The upper-bound of training error and test error of AdaBoost.

But what actually happens to AdaBoost for real data?

number of rounds	5	100	1000
training error	0.0	0.0	0.0
test error	8.4	3.3	3.1
% margins ≤ 0.5	7.7	0.0	0.0
minimum margin	0.14	0.52	0.55

Table 1: The training error, test error and margins of a typical run of AdaBoost Algorithm with C4.5 Decision Tree as weak learner.

2.2 A Counter-Intuitive Example

Running AdaBoost on the "letter" data set (16,000 training examples and 4,000 test examples), with C4.5 decision tree as the weak learner, we get results shown in Figure 2 and Table 1. The training error decreases very fast, and it hits zero after only 5 rounds. Meanwhile, the test error continually decreases as the number of rounds of boosting increases up to 1000.

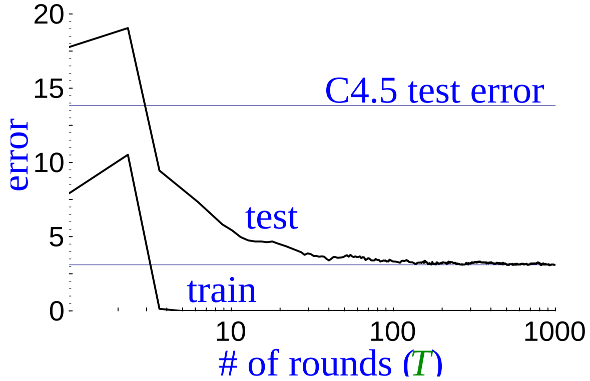

Figure 2: The curves of training error and test error with respect to the number of rounds of boosting on real data.

The result contradicts our previous argument, since the test error continues to drop even after training error reaches zero, instead of increasing as T gets larger. In this case, it seems that Occam's razor *wrongly* predicts that "simpler" rules are better. The result is, after all, good news, because we don't see overfitting as T increases, but it is also bad news because this good performance is not explained by our previous theory.

This paradoxical phenomenon is due to the fact that the training error $\widehat{\mathrm{err}}(H)$ does not tell the whole story. Not only does one need to know whether hypothesis H is "right or wrong", but also the "confidence" of the hypothesis. Actually, even when training error hits 0, the confidence continues to increase, and high confidence reduces the generalization error. This intuition is supported by results shown in Figure 3 and Table 1.

But how to measure confidence quantitatively?

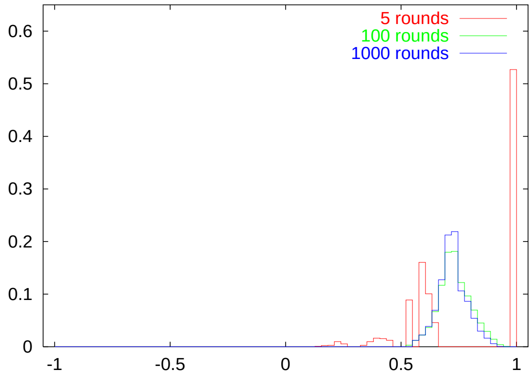

Figure 3: The probability density of letter margin.

3 Generalization Error Based on Margin

3.1 Concept of Margin

We introduce the concept of *margin* to measure the confidence of a hypothesis quantitatively. Mathematically, the margin (of a specific example) is defined as

$$\begin{aligned}\text{margin}(x,y) &= yf(x)\\ &= y\sum_t a_t h_t(x)\\ &= \sum_t a_t y h_t(x)\\ &= \sum_{t:h_t(x)=y} a_t - \sum_{t:h_t(x)\neq y} a_t\end{aligned}$$

where y is the correct label of instance x, and a_t is a normalized version of α_t such that $\alpha_t \geq 0$ and $\sum_t a_t = 1$. The expression $\sum_{t:h_t(x)=y} a_t$ stands for the weighted fraction of correct votes, and $\sum_{t:h_t(x)\neq y} a_t$ stands for the weighted fraction of incorrect votes. Margin is a number between -1 and 1 as shown in Figure 4.

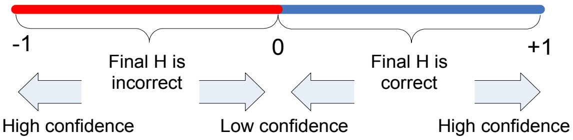

Figure 4: The definition of margin.

3.2 Margin and Generalization Error of AdaBoost

Empirically, AdaBoost pushes examples of low margin to the right end of the segment in Figure 4. In order to analyze what is going on, we do a two-part analysis:

1. Show that AdaBoost increases the margins, i.e. the margin distribution of the training examples is pushed to the right end of $[-1, +1]$.
2. Large margin in training indicates lower generalization error, independent of the number of rounds of boosting.

Before we dive into the detailed analysis, we need to clarify some definitions:

- $\mathscr{D}$: target distribution on $\mathcal{X} \times \{-1, +1\}$;
- S: a sample;
- $\Pr_{\mathscr{D}}[\cdot]$: probability over $(x, y) \sim \mathscr{D}$;
- $\Pr_S[\cdot]$: empirical probability, in others words, the fraction of the training set for which the event holds;

For instance, $\Pr_{\mathscr{D}}[h(x) \neq y] = \mathrm{err}_{\mathscr{D}}(h)$ is the usual generalization error, and $\Pr_S[h(x) \neq y] = \widehat{\mathrm{err}}(h)$ is the usual training error.

In the first part, we show that AdaBoost increases the margins.

Theorem

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

The proof is actually only a slight generalization of the training error theorem we proved last time, so we don't show it here. It's worth pointing out that when $\theta = 0$, the theorem reduces to the bound proven in last lecture. But what does the theorem indicate?

Assume the weak learner can do better than random guessing, say, $\epsilon_t \leq 1/2 - \gamma$ with $\gamma > 0$. Then according to the theorem,

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

As long as $\theta < \gamma$, the term $(1-2\gamma)^{1-\theta}(1+2\gamma)^{1+\theta}$ is strictly less than 1. Hence, $\forall \theta < \gamma$,

$$\Pr_S[yf(x) \leq \theta] \to 0 \text{ as } T \to \infty.$$

This indicates that

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

i.e. the number of points with margin less than γ tends to 0. And the better the weak hypothesis (larger γ), the larger the margin we are going to end up with.

In the second part of the analysis, we show that the generalization error is independent of the number of rounds of boosting, and larger margin indicates lower generalization error. First define

- $\mathcal{H}$ = weak hypothesis space, and $|\mathcal{H}| < \infty$;
- Convex hull of $\mathcal{H}$, $\mathrm{co}(\mathcal{H}) = \{$all functions of the form $f(x) = \sum_{t=1}^{T} a_t h_t(x),$ where $a_t \geq 0, \sum_{t=1}^{T} a_t = 1, T \geq 1, h_1, \ldots, h_T \in \mathcal{H}\}$

Theorem With probability $1-\delta$, $\forall f \in \mathrm{co}(\mathcal{H})$, $\forall \theta > 0$,

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

Surprisingly, the generalization error is independent of the number of rounds of boosting T. It only depends on the size of the weak hypothesis space and the margin distance. There is no penalty for large T at all. Also, from the second term on the right hand side, we know that larger margin θ indicates lower generalization error.

Proof (Sketch) The basic idea of the proof is: do a "survey" on $h_t(x)$ by picking a random subset. As long as the margin θ is large, the sampled survey can predict the final result correctly.

Define $f(x) = \sum_{t=1}^T a_t h_t(x)$, then we can approximate $f(x)$ by

$$g(x) \in \mathcal{C}_N = \{f \text{ of the form } f(x) = \tfrac{1}{N}\textstyle\sum_{t=1}^N h_t(x)\},$$

which comes from a subspace of the space that $f(x)$ is in. First, select $h_t(x)$ according to distribution defined by the a_t to be g_j, i.e. $g_j = h_t$ with probability a_t. Average all g_j and get $g(x) = \frac{1}{N}\sum_{j=1}^N g_j(x)$. Then,

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

We then will use Chernoff bounds to show that $g(x)$ and $f(x)$ are close enough to each other. Then we will apply a uniform convergence theorem to the small class $\mathcal{C}_N$ to show the upper-bound of generalization error.