COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire
Scribe: Linpeng Tang

Lecture #11
March 12, 2013

1 Complexity of Boosted Learners

In this section we aim to analyze the complexity of the output hypothesis of Boosting, in terms of its growth function. This will help us bound the generalization error of AdaBoost.

Note that the output hypothesis by AdaBoost after T iterations has the following form

$$H(x) = \text{sign}(\sum_{t=1}^{T} \alpha_t h_t(x)) \tag{1.1}$$

$$= g(h_1(x), \ldots, h_T(x)) \tag{1.2}$$

where $g(\mathbf{z}) = \text{sign}(\mathbf{w} \cdot \mathbf{z})$, $\mathbf{w} = \langle \alpha_1, \ldots, \alpha_T \rangle$. Let $\mathcal{F}$ to be the hypothesis space containing all H of such forms. And let $\mathcal{G}$ to be the set of linear threshold functions in $\mathbb{R}^T$ (without the offset term), then $\text{VCdim}(\mathcal{H}) = T$. And let $\mathcal{H}$ be the weak hypothesis space which we assume has $\text{VCdim}(\mathcal{H}) = d$.

By Homework 2, Problem 1 we can bound the growth function of $\mathcal{H}$ by

$$\Pi_{\mathcal{F}}(m) \leq \Pi_{\mathcal{G}}(m)[\Pi_{\mathcal{H}}(m)]^T \tag{1.3}$$

$$\leq \left(\frac{em}{T}\right)^T \left(\frac{em}{d}\right)^{dT} \qquad \text{(Sauer's Lemma)} \tag{1.4}$$

Given m examples, we have shown that with probability at least $1 - \delta$, for any $H \in \mathcal{F}$,

$$\text{err}(H) \leq \widehat{\text{err}}(H) + O\left(\sqrt{\frac{\ln \pi_{\mathcal{F}}(m) + \ln(1/\delta)}{m}}\right) \tag{1.5}$$

We introduce the "soft-Oh" notation, which hides the log factors just the same way that the "big-Oh" hides the constant factors. For example we would write $\frac{\ln m}{m} = \tilde{O}(1/m)$.

Then Equation (1.5) can be rewritten as

$$\text{err}(H) \leq \widehat{\text{err}}(H) + \tilde{O}\left(\sqrt{\frac{Td + \ln(1/\delta)}{m}}\right) \tag{1.6}$$

2 Margin Based Analysis

2.1 Definition of Margin

Observe Equation (1.6), as the number of iterations increase, the second term will increase to infinity and the formula predicts overfitting of Boosting.

However, in many experiments, we observe that the test error continues to decrease even after *the training error has reached zero*. How can we explain this phenomenon?

We need to realize that training error is only telling part of the story, and is an inadequate measure of the fit to the training set. As we continue to run AdaBoost, the predictions

of the combined classifier will become more "confident", and this increase in confidence translates into better performance. But how do we measure this confidence?

In politics, when two people compete for a position, we not only care about who has won the majority vote, but also how many more votes he/she has won over his/her competitor, i.e. "the margin of the victory". We also want to introduce this concept into learning.

Note that the output of AdaBoost is simply the weighted majority vote of weak hypotheses

$$H(x) = \text{sign}(\sum_{t=1}^{T} \alpha_t h_t(x)) \tag{2.1}$$

$$= \text{sign}(\sum_{t=1}^{T} a_t h_t(x)) \tag{2.2}$$

where $a_t \triangleq \alpha_t / (\sum_{i=1}^{T} \alpha_i)$, so $\langle a_1, \ldots, a_T \rangle$ is l_1 normalized (non-negative and add up to 1). Let

$$f(x) \triangleq \sum_{t=1}^{T} a_t h_t(x) \tag{2.3}$$

Then we define the margin on training example $\langle x, y \rangle$ to be

$$\text{margin}(x, y) = \sum_{t:y=h(t)} a_t - \sum_{t:y \neq h(t)} a_t \tag{2.4}$$

$$= \sum_{t} a_t y h_t(x) \tag{2.5}$$

$$= y f(x) \tag{2.6}$$

Note that in Equation (2.4), $\sum_{t:y=h(t)} a_t$ is simply the fraction of "correct" votes of weak learners, while $\sum_{t:y \neq h(t)} a_t$ is the fraction of "incorrect" votes of weak learners. So $\text{margin}(x, y)$ corresponds to the margin of victory in politics, assuming the winner is just the label of the training example.

We are going to show

- AdaBoost tends to increase the margin of the training examples as it continues to iterate.
- Large margin on training examples translates to low generalization error.

2.2 AdaBoost Increases Margin

Theorem 2.1. *For $\Theta > 0$, the fraction of training examples in S with margin less than Θ is bounded by*

$$\hat{\Pr}_S[y f(x) \leq \Theta] \leq \prod_{t=1}^{T} \left(2\sqrt{\epsilon_t^{1-\Theta} (1 - \epsilon_t)^{1+\Theta}} \right) \tag{2.7}$$

Then if weak learnability is satisfied, i.e. $\epsilon_t < \frac{1}{2} - \gamma$ for some $\gamma > 0$, then for any $\Theta < \gamma$,

$$\hat{\Pr}_S[y f(x) \leq \Theta] \leq \left(\sqrt{(1 - 2\gamma)^{1-\Theta} (1 + 2\gamma)^{1+\Theta}} \right)^T \to 0 \tag{2.8}$$

when $T \to \infty$.

Proof. The proof is very similar to that of the training error bound of AdaBoost.

$$\hat{\Pr}_s[yf(x) \le \Theta] = \frac{1}{m}\sum_{i=1}^{m} 1\{y_i f(x_i) \le \Theta\} \tag{2.9}$$

$$= \frac{1}{m}\sum_{i=1}^{m} 1\{y_i \sum_t a_t h_t(x_i) \le \Theta \sum_t a_t\} \tag{2.10}$$

$$= \frac{1}{m}\sum_{i=1}^{m} 1\{y_i \sum_t \alpha_t h_t(x_i) - \Theta \sum_t \alpha_t \le 0\} \tag{2.11}$$

$$\le \frac{1}{m}\sum_{i=1}^{m} \exp(-y_i \sum_t \alpha_t h_t(x_i) + \Theta \sum_t \alpha_t) \tag{2.12}$$

$$= \frac{1}{m}\exp(\Theta \sum_t \alpha_t) \sum_{i=1}^{m} \exp(-y_i \sum_t \alpha_t h_t(x_i)) \tag{2.13}$$

$$= \prod_t (Z_t e^{\Theta \alpha_t}) \tag{2.14}$$

$$= \prod_t 2\sqrt{\epsilon_t(1-\epsilon_t)} \left[\sqrt{\frac{1-\epsilon_t}{\epsilon_t}}\right]^{\Theta} \tag{2.15}$$

$$= \prod_t 2\sqrt{\epsilon_t^{1-\Theta}(1-\epsilon_t)^{1+\Theta}} \tag{2.16}$$

□

Equation (2.10) uses the definition of $f(\cdot)$ and the fact that $\sum_{t=1}^{T} a_t = 1$. Equation (2.11) multiplies both sides of $\le$ in Equation (2.10) by $\sum_{t=1}^{T} \alpha_t$ and does a simple transformation. Equation (2.12) uses the fact that $1\{-x\} \le e^{-x}$ for any x. Equation (2.14) follows from the derivation of AdaBoost that $\prod_t Z_t = \frac{1}{m}\sum_{i=1}^{m} \exp(-y \sum_t \alpha_t h_t(x_i))$. Equation (2.15) uses the values of Z_t and α_t calculated in AdaBoost.

2.3 Rademacher Complexity Analysis

We define the convex hull of $\mathcal{H}$ to be the set of all convex combinations of weak hypotheses in $\mathcal{H}$,

$$\mathrm{co}(\mathcal{H}) \triangleq \text{convex hull of } \mathcal{H} \tag{2.17}$$

$$= \left\{ f : x \mapsto \sum_{t=1}^{T} a_t h_t(x) \middle|\ a_t \ge 0, \sum_t a_t = 1, h_t \in \mathcal{H} \right\} \tag{2.18}$$

Although $\mathrm{co}(\mathcal{H})$ seems to be a more complex hypothesis space than $\mathcal{H}$, their Rademacher complexity turns out to be actually the same.

Lemma 2.2. $\mathcal{R}(\mathrm{co}(\mathcal{H})) = \mathcal{R}(\mathcal{H})$.

Proof.

$$\hat{\mathcal{R}}_S(\mathcal{H}) \le \hat{\mathcal{R}}_S(\text{co}(\mathcal{H})) \tag{2.19}$$

$$= E_\sigma[\frac{1}{m}\sup_{f\in\text{co}(\mathcal{H})}\sum_i \sigma_i f(x_i)] \qquad \text{(by definition of co}(\mathcal{H})) \tag{2.20}$$

$$\le E_\sigma[\frac{1}{m}\sup_{h\in\mathcal{H}}\sum_i \sigma_i h(x_i)] \tag{2.21}$$

$$= \hat{\mathcal{R}}_S(\mathcal{H}) \tag{2.22}$$

For Equation (2.21), since any $f \in \text{co}(\mathcal{H})$ is a weighted average of some functions in $\mathcal{H}$, there exists some $h \in \mathcal{H}$ s.t.

$$\sum_i \sigma_i f(x_i) \le \sum_i \sigma_i h(x_i).$$

So for any σ, $\sup_{f\in\text{co}(\mathcal{H})}\sum_i \sigma_i f(x_i) \le \sup_{h\in\mathcal{H}}\sum_i \sigma_i h(x_i)$, and the inequality follows. □

Let function $\phi : \mathbb{R} \to \mathbb{R}$, define the function composition operator as

$$\phi \circ f : z \mapsto \phi(f(z)) \tag{2.23}$$

And similarly we define function composition on function class

$$\phi \circ \mathcal{F} = \{\phi \circ f : f \in \mathcal{F}\} \tag{2.24}$$

What is the relationship between $\mathcal{R}(\mathcal{F})$ and $\mathcal{R}(\phi \circ \mathcal{F})$? Generally this is a difficult question, but when ϕ is Lipschitz continuous, we can prove some results. Especially, we have the following lemma. The proof is presented in the textbook and is omitted here.

Lemma 2.3. *If ϕ is Lipschitz continuous, i.e. $\forall u, v$,*

$$|\phi(u) - \phi(v)| \le L_\phi |u - v| \tag{2.25}$$

for some constant L_ϕ, then

$$\hat{\mathcal{R}}_S(\phi \circ \mathcal{F}) \le L_\phi \hat{\mathcal{R}}_S(\mathcal{F}). \tag{2.26}$$

Define

$$\mathcal{M} = \{(x, y) \mapsto yf(x) : f \in \text{co}(\mathcal{H})\} \tag{2.27}$$

We can show the Rademacher complexity of $\mathcal{M}$ is actually the same as $\mathcal{H}$.

$$\hat{\mathcal{R}}_S(\mathcal{M}) = E_\sigma[\frac{1}{m}\sup_{f\in\text{co}(\mathcal{H})}\sum_i \sigma_i y_i f(x_i)] \tag{2.28}$$

$$= E_\sigma[\frac{1}{m}\sup_{f\in\text{co}(\mathcal{H})}\sum_i \sigma_i f(x_i)] \tag{2.29}$$

$$= \hat{\mathcal{R}}_S(\text{co}(\mathcal{H})) \tag{2.30}$$

$$= \hat{\mathcal{R}}_S(\mathcal{H}) \qquad \text{(by Lemma 2.3)} \tag{2.31}$$

Equation (2.29) holds because we can do a one-to-one mapping from $\{-1, 1\}^m$ to itself by

$$\langle \sigma_1, \ldots, \sigma_m \rangle \mapsto \langle y_1\sigma_1, \ldots, y_m\sigma_m \rangle \tag{2.32}$$

so the two expectations are the same.

For any function class $\mathcal{F}$, by the main theorem about Rademacher complexity we have proved previously, with probability at least $1-\delta$ (over the choice of S), $\forall f \in \mathcal{F}$,

$$E_D[f] \leq \hat{E}_S[f] + 2\hat{\mathcal{R}}_S(\mathcal{F}) + O\left(\sqrt{\frac{\ln(1/\delta)}{m}}\right) \tag{2.33}$$

We want to prove similar results with margins:

$$\Pr_D[yf(x) \leq 0] \leq \hat{\Pr}_S[yf(x) \leq \Theta] + \hat{O}\left(\sqrt{\frac{d/\Theta^2 + \ln 1/\delta}{m}}\right) \tag{2.34}$$

To show this, first note that

$$\hat{\Pr}_S[yf(x) \leq \Theta] = \hat{E}_S[1\{yf(x) \leq \Theta\}] \tag{2.35}$$

$$\Pr_D[yf(x) \leq 0] = E_D[1\{yf(x) \leq 0\}] \tag{2.36}$$

The two expectations in Equation (2.35) and Equation (2.36) are over two different functions. These seems to be a problem at first sight, but actually turns out to be the key for connecting $\Pr_D[yf(x) \leq 0]$, $\hat{\Pr}_S[yf(x) \leq \Theta]$ and $\hat{\mathcal{R}}_S(\mathcal{F})$. We will complete the proof next time.