COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire
Scribe: Jun Wei Ho

Lecture #12
March 13, 2014

1 Margin Theory for Boosting

Recall from the earlier lecture that we may write our hypothesis $H(x) = \text{sign}\left(\sum_{t=1}^{T} a_t h_t(x)\right)$, where $a_t = \alpha_t / \sum_s \alpha_s$ (so that $\sum_t a_t = 1$) and $h_1, \ldots, h_T$ are the weak hypotheses that we obtained over T iterations of AdaBoost.

Writing $f(x) = \sum_{t=1}^{T} a_t h_t(x)$, we define $\text{marg}_f(x, y) = yf(x)$ to be the *margin* of f for a training example (x, y). In the last lecture, we have seen that this quantity represents the weighted fraction of h_t's that voted correctly, minus the weighted fraction of h_t's that voted incorrectly, for the class y when given the data x.

A few remarks about the margin:

– $yf(x)$ takes values in the interval $[-1, 1]$

– $yf(x) > 0$ if and only if $H(x) = y$

– The magnitude $|yf(x)|$ represents the degree of 'confidence' for the classification $H(x)$. A number substantially far from zero implies high confidence, whereas a number close to zero implies low confidence.

It is therefore desirable for the margin $yf(x)$ to be 'large', since this represents a correct classification with high confidence. We will see that under the usual assumptions, AdaBoost is able to increase the margins on the training set and achieve a positive lower bound for these margins. In particular, this means that the training error will be zero, and we will see that larger margins help to achieve a smaller generalization error.

In this lecture, we aim to show that:

1. Boosting tends to increase the margins of training examples. Moreover, a bigger edge will result in larger margins after boosting.

2. Large margins on our training set leads to better performance on our test data (and this is independent of T, the number of rounds of boosting)

Notation

$\mathcal{S}$	Training set $\langle (x_1, y_1), \ldots, (x_m, y_m) \rangle$
$\mathcal{H}$	Weak hypothesis space
d	VCdim($\mathcal{H}$)
$co(\mathcal{H})$	Convex hull of $\mathcal{H}$, the set of functions given by $\left\{ f(x) = \sum_{t=1}^{T} a_t h_t(x) \ : \ a_1, \ldots, a_T \geq 0, \ \sum_t a_t = 1, \ h_1, \ldots, h_T \in \mathcal{H}, \ T \geq 1 \right\}$
$Pr_{\mathcal{D}}$	Probability with respect to the true distribution $\mathcal{D}$
$E_{\mathcal{D}}$	Expectation with respect to the true distribution $\mathcal{D}$
$\widehat{Pr}_{\mathcal{S}}$	Empirical probability with respect to $\mathcal{S}$
$\widehat{E}_{\mathcal{S}}$	Empirical expectation with respect to $\mathcal{S}$

1.1 Boosting Increases Margins of Training Examples

We will show that given sufficient rounds of boosting, we can guarantee that $y_i f(x_i) \geq \gamma \; \forall \, i$, where $\gamma > 0$ is the edge in our weak learning assumption. In particular, this means that $H(x)$ will classify each training example correctly, and do so with confidence at least γ. The main result we will use is the following.

Theorem 1. *For $\theta \in [-1, 1]$, we have*

$$\widehat{Pr}_{\mathcal{S}}[\, yf(x) \leq \theta \,] \leq \prod_{t=1}^{T} \left[2\sqrt{\epsilon_t^{1-\theta}(1-\epsilon_t)^{1+\theta}} \right] \tag{1}$$

Moreover, if $\epsilon_t \leq \frac{1}{2} - \gamma$ for $t = 1, \ldots, T$, then

$$\widehat{Pr}_{\mathcal{S}}[\, yf(x) \leq \theta \,] \leq \left[\sqrt{(1-2\gamma)^{1-\theta}(1+2\gamma)^{1+\theta}} \right]^{T} \tag{2}$$

Proof. Recall from the last lecture that

$$\frac{1}{m}\sum_{i=1}^{m} \exp\left(-y_i \sum_{t=1}^{T} \alpha_t h_t(x_i) \right) = \prod_{t=1}^{T} Z_t = \prod_{t=1}^{T} 2\sqrt{\epsilon_t(1-\epsilon_t)}$$

where we had set $\alpha_t = \frac{1}{2} \ln \frac{1-\epsilon_t}{\epsilon_t}$ to obtain the last equality.

Using a similar argument as before,

$$\begin{aligned}
\widehat{Pr}_{\mathcal{S}}[\, yf(x) \leq \theta \,] &= \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{y_i f(x_i) \leq \theta\} \\
&= \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{y_i \sum_{t=1}^{T} \alpha_t h_t(x_i) \leq \theta \sum_{t=1}^{T} \alpha_t\} \\
&\leq \frac{1}{m}\sum_{i=1}^{m} \exp\left(-y_i \sum_{t=1}^{T} \alpha_t h_t(x_i) + \theta \sum_{t=1}^{T} \alpha_t \right) \\
&= \exp\left(\theta \sum_{t=1}^{T} \alpha_t \right) \frac{1}{m}\sum_{i=1}^{m} \exp\left(-y_i \sum_{t=1}^{T} \alpha_t h_t(x_i) \right) \\
&= \exp\left(\theta \sum_{t=1}^{T} \alpha_t \right) \prod_{t=1}^{T} Z_t \\
&= \prod_{t=1}^{T} e^{\theta \alpha_t} Z_t \\
&= \prod_{t=1}^{T} \left[2\sqrt{\epsilon_t^{1-\theta}(1-\epsilon_t)^{1+\theta}} \right]
\end{aligned}$$

where the inequality follows from $\mathbb{1}\{x \leq 0\} \leq e^{-x}$, and the final equality is achieved by setting $\alpha_t = \frac{1}{2} \ln \frac{1-\epsilon_t}{\epsilon_t}$.

The second result uses the fact that if $\epsilon_t \le \frac{1}{2} - \gamma$, then

$$\begin{aligned} e^{\theta\alpha_t} Z_t &= e^{\theta\alpha_t}\left(\epsilon_t e^{\alpha_t} + (1-\epsilon_t)e^{-\alpha_t}\right) \\ &\le e^{\theta\alpha_t}\left[\left(\frac{1}{2}-\gamma\right)e^{\alpha_t} + \left(\frac{1}{2}+\gamma\right)e^{-\alpha_t}\right] \\ &= \sqrt{(1-2\gamma)^{1-\theta}(1+2\gamma)^{1+\theta}} \end{aligned}$$

by setting $\alpha_t = \frac{1}{2}\ln\left(\frac{\frac{1}{2}+\gamma}{\frac{1}{2}-\gamma}\right)$. The reader should verify the inequality and work out the details. $\square$

Remark. *By setting $\theta = 0$ in the above result, we recover the bound on training error proven in the previous lecture. Moreover, it is possible to show that for any $0 < \theta \le \gamma$, the term $(1-2\gamma)^{1-\theta}(1+2\gamma)^{1+\theta} < 1$, hence as $T \to \infty$ the RHS of (2) goes to zero. As an easy consequence, we have the following:*

Corollary. *If the weak learning assumption holds, then given sufficiently large T, we have $y_i f(x_i) \ge \gamma \; \forall \; i$.*

1.2 Large Margins on Training Set Reduce Generalization Error

Previously, we have shown that with probability at least $1-\delta$,

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

We can rewrite this equivalently as

$$Pr_{\mathcal{D}}[\, yf(x) \le 0\,] \le \widehat{Pr}_{\mathcal{S}}[\, yf(x) \le 0\,] + \tilde{O}\left(\sqrt{\frac{Td + \ln(1/\delta)}{m}}\right)$$

We will now prove a variant of this result where the upper bound does not depend on T, but instead on a parameter θ that we can relate to the margin.

Theorem. *For $0 < \theta \le 1$, with probability at least $1-\delta$,*

$$Pr_{\mathcal{D}}[\, yf(x) \le 0\,] \le \widehat{Pr}_{\mathcal{S}}[\, yf(x) \le \theta\,] + \tilde{O}\left(\sqrt{\frac{d/\theta^2 + \ln(1/\delta)}{m}}\right).$$

Before we prove the theorem, we will first introduce two lemmas.

Recall that for $\mathcal{S} = \langle z_1, \ldots, z_m \rangle$ and $\mathcal{F} = \{f : Z \to \mathbb{R}\}$, the empirical Rademacher complexity of $\mathcal{F}$ is given by

$$\widehat{R}_{\mathcal{S}}(\mathcal{F}) = E_\sigma\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i f(z_i)\right]$$

In the last lecture, we've seen that $\widehat{R}_{\mathcal{S}}(\mathcal{H}) = \tilde{O}\left(\sqrt{\frac{d}{m}}\right)$. The following lemma tells us how $\widehat{R}_{\mathcal{S}}(co(\mathcal{H}))$ relates to $\widehat{R}_{\mathcal{S}}(\mathcal{H})$.

Lemma 1. *The Rademacher complexity of $\mathcal{H}$ is equal to the Rademacher complexity of its convex hull. In other words, $\widehat{R}_{\mathcal{S}}(co(\mathcal{H})) = \widehat{R}_{\mathcal{S}}(\mathcal{H})$.*

Proof. Since $\mathcal{H} \subset co(\mathcal{H})$, it is clear that $\widehat{R}_{\mathcal{S}}(\mathcal{H}) \leq \widehat{R}_{\mathcal{S}}(co(\mathcal{H}))$. Moreover,

$$\begin{aligned}
\widehat{R}_{\mathcal{S}}(co(\mathcal{H})) &= E_\sigma\left[\sup_{f\in co(\mathcal{H})} \frac{1}{m}\sum_{i=1}^{m}\sigma_i \sum_t a_t h_t(x_i)\right] \\
&= E_\sigma\left[\sup_{f\in co(\mathcal{H})} \frac{1}{m}\sum_t a_t \sum_{i=1}^{m}\sigma_i h_t(x_i)\right] \\
&\leq E_\sigma\left[\sup_{f\in co(\mathcal{H})} \frac{1}{m}\sum_t a_t \sup_{h\in\mathcal{H}} \sum_{i=1}^{m}\sigma_i h(x_i)\right] \\
&= E_\sigma\left[\sup_{f\in co(\mathcal{H})} \frac{1}{m} \sup_{h\in\mathcal{H}} \sum_{i=1}^{m}\sigma_i h(x_i)\right] \\
&= E_\sigma\left[\frac{1}{m} \sup_{h\in\mathcal{H}} \sum_{i=1}^{m}\sigma_i h(x_i)\right] \\
&= \widehat{R}_{\mathcal{S}}(\mathcal{H})
\end{aligned}$$

To obtain the fourth line we had used the fact that $\sum_t a_t = 1$, and for the fifth line we note that the expression in $\sup_f(..)$ does not depend on f, so we could omit the $\sup_f$ function. We therefore conclude that $\widehat{R}_{\mathcal{S}}(co(\mathcal{H})) = \widehat{R}_{\mathcal{S}}(\mathcal{H})$. □

Next, for any function $\phi : \mathbb{R} \to \mathbb{R}$, and $f : Z \to \mathbb{R}$, we define the composition $\phi \circ f : Z \to \mathbb{R}$ by $\phi \circ f(z) = \phi(f(z))$. We also define the space of composite functions $\phi \circ \mathcal{F} = \{\phi \circ f : f \in \mathcal{F}\}$.

Lemma 2. *Suppose ϕ is Lipschitz-continuous, that is, $\exists\ L_\phi > 0$ such that $\forall\ u, v \in \mathbb{R}$, $|\phi(u) - \phi(v)| \leq L_\phi |u - v|$. Then $\widehat{R}_{\mathcal{S}}(\phi \circ \mathcal{F}) \leq L_\phi \widehat{R}_{\mathcal{S}}(\mathcal{F})$.*

Proof. See Mohri et al. □

Equipped with the two lemmas, we are now ready to prove the main theorem. We will state the result once more:

Theorem 2. *For $0 < \theta \leq 1$, with probability at least $1 - \delta$,*

$$Pr_{\mathcal{D}}[\,yf(x) \leq 0\,] \leq \widehat{Pr}_{\mathcal{S}}[\,yf(x) \leq \theta\,] + \tilde{O}\left(\sqrt{\frac{d/\theta^2 + \ln(1/\delta)}{m}}\right).$$

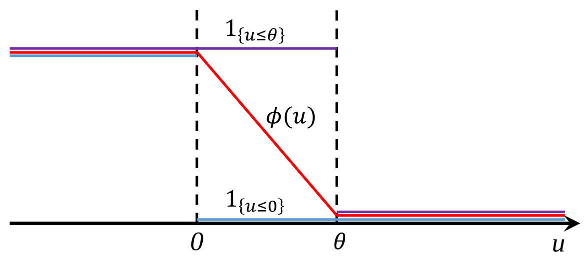
Proof. Write $\text{marg}_f(x, y) = yf(x)$. Define $\mathcal{M} = \{\text{marg}_f : f \in co(\mathcal{H})\}$. Then

$$\begin{aligned}
\widehat{R}_{\mathcal{S}}(\mathcal{M}) &= E_\sigma\left[\sup_{f\in co(\mathcal{H})} \frac{1}{m}\sum_{i=1}^{m}(\sigma_i y_i) f(x_i)\right] \\
&= \widehat{R}_{\mathcal{S}}(co(\mathcal{H})) \\
&= \widehat{R}_{\mathcal{S}}(\mathcal{H}) \qquad \text{(by Lemma 1)}
\end{aligned}$$

Next, we define the function $\phi : \mathbb{R} \to [0, 1]$ by

$$\phi(u) = \begin{cases} 1 & \text{if } u \leq 0 \\ 1 - u/\theta & \text{if } 0 < u \leq \theta \\ 0 & \text{if } u > \theta \end{cases}$$

A plot of $\phi(u)$ is shown in the diagram below:



Note that for all $u \in \mathbb{R}$, we have

$$\mathbb{1}\{u \leq 0\} \leq \phi(u) \leq \mathbb{1}\{u \leq \theta\}$$

Moreover, ϕ is clearly Lipschitz-continuous with $L_\phi = \frac{1}{\theta}$. Therefore, Lemma 2 gives us

$$\widehat{R}_{\mathcal{S}}(\phi \circ \mathcal{M}) \;\leq\; \frac{1}{\theta}\widehat{R}_{\mathcal{S}}(\mathcal{M}) \;=\; \frac{1}{\theta}\widehat{R}_{\mathcal{S}}(\mathcal{H}) \;\leq\; \tilde{O}\left(\sqrt{\frac{d/\theta^2}{m}}\right)$$

Using the result from a previous lecture[1] and the results above, we have

$$\begin{aligned} Pr_{\mathcal{D}}[\,yf(x) \leq 0\,] &= E_{\mathcal{D}}[\mathbb{1}\{yf(x) \leq 0\}] \\ &\leq E_{\mathcal{D}}[\phi \circ (yf)(x)] \\ &\leq \widehat{E}_{\mathcal{S}}[\phi \circ (yf)(x)] + 2\widehat{R}_{\mathcal{S}}(\phi \circ \mathcal{M}) + O\left(\sqrt{\frac{\ln(1/\delta)}{m}}\right) \\ &\leq \widehat{E}_{\mathcal{S}}[\mathbb{1}\{yf(x) \leq \theta\}] + \tilde{O}\left(\sqrt{\frac{d/\theta^2}{m}}\right) + O\left(\sqrt{\frac{\ln(1/\delta)}{m}}\right) \\ &= \widehat{Pr}_{\mathcal{S}}[\,yf(x) \leq \theta\,] + \tilde{O}\left(\sqrt{\frac{d/\theta^2 + \ln(1/\delta)}{m}}\right) \end{aligned}$$

as desired. $\square$

Remark. *The larger the value of θ we use, the smaller the $\tilde{O}(...)$ term on the RHS. With larger margins on the training set, we are able to choose larger values of θ while keeping the $\widehat{Pr}_{\mathcal{S}}[\,yf(x) \leq \theta\,]$ term zero (or close to zero), and this will give us a sharper upper bound on the generalization error. This suggests that by increasing the margin on the training set, we may expect to see a smaller generalization error.*

[1] In an earlier lecture, we had proved that with probability at least $1 - \delta$, $\forall\, f \in \mathcal{F}$,

$$E_{\mathcal{D}}[f] \leq \widehat{E}_{\mathcal{S}}[f] + 2\widehat{R}_{\mathcal{S}}(\mathcal{F}) + O\left(\sqrt{\frac{\ln(1/\delta)}{m}}\right)$$