COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire — Lecture #12
Scribe: Tsung-Yen Yang — March 14, 2018

1 Margin for Boosting

Recall from the previous lecture that we proved that for the Adaboost algorithm, the training error decreases exponentially fast in the number of rounds of boosting. We also proved a bound on the generalization error in terms of the training error, number of rounds, number of examples and VC dimension of the weak hypothesis space. However, Occam's razor seems to contradict the behavior of the Adaboost algorithm. As we keep increasing the number of rounds T, we do not observe a worse generalization error even when training error is already 0 as shown in Fig. 1. So in this lecture, we aim to give an explanation about why Adaboost does not suffer from overfitting as we keep running the algorithm.

To understand what is happening, we need to consider the *confidence* in those predictions. Intuitively, as we keep running Adaboost, the predictions made by the combined classifier are getting more and more confident even if the training error remains the same. That increased confidence translates into better generalization performance as well.

To make this idea rigorous, we need to define *confidence*. Recall that our combined hypothesis makes predictions based on weighted majority vote. So the natural way to measure the confidence is by looking at *margin*, which is the difference between the weighted fraction of h_t's voting correctly and the fraction corresponding to those voting incorrectly. So we have:

$$\begin{aligned} H(x) &= sign(\sum_{t=1}^{T} \alpha_t h_t(x)) \\ &= sign(\sum_{t=1}^{T} a_t h_t(x)) \end{aligned}$$

where $a_t = \frac{\alpha_t}{\sum_{t'=1}^{T} \alpha_{t'}}$. In this way, we are normalizing the weights for each hypothesis, having $a_t \geq 0$, $\sum a_t = 1$. Then for an example x with correct label y, the margin is:

$$\begin{aligned} margin(x,y) &= \sum_{t:h_t(x)=y} a_t - \sum_{t:h_t(x)\neq y} a_t \\ &= \sum_t a_t \cdot \begin{cases} +1, & \text{if } h_t(x) = y \\ -1, & \text{else} \end{cases} \\ &= \sum_t a_t y h_t(x) \\ &= y \sum_t a_t h_t(x) \\ &= y f(x) \end{aligned}$$

where we define $f(x) = \sum_t a_t h_t(x)$. Unlike the normal margin we use in the real election, this margin has the sign. Its magnitude indicates the confidence as shown in Fig. 2.

In this lecture, we focus on showing that :

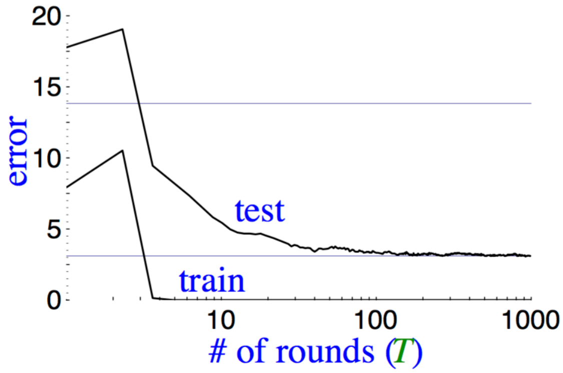

Figure 1: Error versus number of rounds of boosting

- The margins of examples tend to get larger as we keep running the boosting algorithm.
- Large margins on training examples results in better performance in generalization error.

1.1 Boosting Increases Margins of Training Examples

Next, we want to bound how many training examples have margins that are below the given value θ. We define $\widehat{Pr}_{\mathcal{S}}$ as empirical probability with respect to training set $\mathcal{S}$ such that $\mathcal{S} = \langle (x_1, y_1), \ldots, (x_m, y_m) \rangle$.

Theorem 1. *For $\theta \geq 0$, we have*

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

Furthermore, if for some $\gamma > 0$ and $\forall t$, $\epsilon_t \leq \frac{1}{2} - \gamma$, then

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

In particular, if $\theta < \gamma$, then this quantity goes to 0 as $T \to \infty$.

The second inequality comes from $\forall t : \epsilon_t \leq \frac{1}{2} - \gamma$ (weak learning assumption). Note that the last statement follows from the fact that the quantity inside the parentheses is strictly smaller than 1 under the weak learning condition.

Proof. The proof is similar to the proof of training error. □

Remark 1.1. *By setting $\theta = 0$ in the above theorem, we get the bound on training error proven in the previous lecture such that*

$$\widehat{err}(H) \leq \prod_{t=1}^{T} \left[2\sqrt{\epsilon_t(1-\epsilon_t)} \right]$$

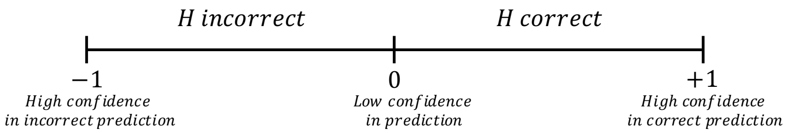
H incorrect
H correct
−1
High confidence in incorrect prediction
0
Low confidence in prediction
+1
High confidence in correct prediction

Figure 2: Diagram of margin

1.2 Large Margins on Training Set Reduce Generalization Error

Next, we define $\mathcal{H}$ as the weak hypothesis space and $d = VCdim(\mathcal{H})$. Also we find that $f(x)$ is a convex combination of $h_1, \dots, h_T$, so we define the convex hull of $\mathcal{H}$ to be

$$co(\mathcal{H}) = \Big\{ f : x \mapsto \sum_{t=1}^{T} a_t h_t(x) | a_1, \dots, a_T \geq 0, \sum_t a_t = 1, h_1, \dots, h_T \in \mathcal{H}, T \geq 1 \Big\}$$

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

$$err(h) \leq \widehat{err}(h) + \tilde{O}\Bigg(\sqrt{\frac{Td + \ln(1/\delta)}{m}}\Bigg)$$

We can rewrite this equivalently as

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

We don't want it to depend on T to capture the lack of overfitting, but instead on a parameter θ that we can relate to the margin.

We think of θ as an arbitrary cut-off, with margins considered *large* or *small* based on whether they are above or below θ. We are then replacing the training error, which is the same as the fraction of training examples with margin at most zero, with the fraction of training examples with margin at most θ. But it is then inevitable that we must somehow *pay a penalty* for choosing θ too close to zero, so that $1/\theta^2$ ends up replacing T in the bound.

Theorem 2. *For $\theta > 0$ and $\forall f \in co(\mathcal{H})$, with probability at least $1 - \delta$,*

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

To prove this theorem, there are three lemmas need to be proved first.

Lemma 3. *Suppose that $\mathcal{S} = \langle x_1, \dots, x_m \rangle$. Then the empirical Rademacher complexity of $\mathcal{H}$ is given by*

$$\mathcal{R}_{\mathcal{S}}(\mathcal{H}) \leq \sqrt{\frac{2d \ln(\frac{em}{d})}{m}} = \tilde{O}\Bigg(\sqrt{\frac{d}{m}}\Bigg)$$

Proof. See an earlier class (lecture #10). □

Lemma 4. *The Rademacher complexity of $\mathcal{H}$ is equal to the Rademacher complexity of its convex hull. In other words, $\mathcal{R}_{\mathcal{S}}(co(\mathcal{H})) = \mathcal{R}_{\mathcal{S}}(\mathcal{H})$*

Proof. $\mathcal{R}_{\mathcal{S}}(co(\mathcal{H})) \geq \mathcal{R}_{\mathcal{S}}(\mathcal{H})$ since $\mathcal{H} \subseteq co(\mathcal{H})$. Moreover,

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

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

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 $\phi \circ \mathcal{F} = \{\phi \circ f : f \in \mathcal{F}\}$. Now we want to find the Rademacher complexity of this space $\phi \circ \mathcal{F}$.

Lemma 5. *Suppose ϕ is L_ϕ-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 $\mathcal{R}_{\mathcal{S}}(\phi \circ \mathcal{F}) \leq L_\phi \mathcal{R}_{\mathcal{S}}(\mathcal{F})$.*

Proof. Please refer to Mohri et al. □

We are now ready to prove the main theorem.

Proof. Define $\text{marg}_f(x, y) = yf(x)$, which is the margin function associated with f that maps a labeled example (x, y) to its margin under f. And define the space of all such functions

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

Then

$$\begin{aligned}
\mathcal{R}_{\mathcal{S}}(\mathcal{M}) &= E_\sigma\Big[\sup_{f\in co(\mathcal{H})} \frac{1}{m}\sum_i (y_i\sigma_i) f(x_i)\Big] \\
&= \mathcal{R}_{\mathcal{S}}(co(\mathcal{H})) \\
&= \mathcal{R}_{\mathcal{S}}(\mathcal{H})
\end{aligned}$$

Note that in the first line, since y_i is $+1$ or -1, we can treat $y_i\sigma_i$ as a new Rademacher random variable. And the last line is from Lemma 4.

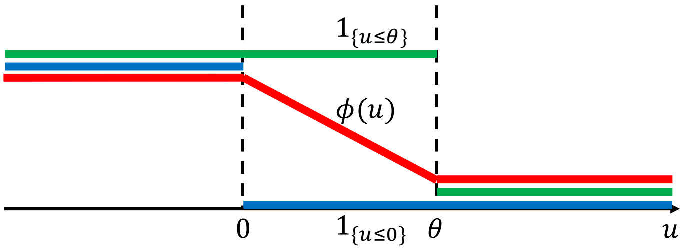

Figure 3: Diagram of $\phi(u)$

We want to use the general result we earlier proved to prove our theorem, namely, that with high probability,

$$\forall g \in \mathcal{F}, E[g] \leq \widehat{E}_{\mathcal{S}}[g] + 2\widehat{R}_{\mathcal{S}}(\mathcal{F}) + O().$$

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

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

Fig. 3 shows $\phi(u)$. Then we have:

$$Pr_{\mathcal{D}}[yf(x) \leq 0] = E_{\mathcal{D}}[\mathbb{1}\{yf(x) \leq 0\}] \leq E_{\mathcal{D}}[\phi(yf(x))]$$
$$\widehat{Pr}_{\mathcal{S}}[yf(x) \leq \theta] = \widehat{E}_{\mathcal{S}}[\mathbb{1}\{yf(x) \leq \theta\}] \geq \widehat{E}_{\mathcal{S}}[\phi(yf(x))]$$

Moreover, ϕ is Lipschitz-continuous with $L_\phi = \frac{1}{\theta}$. Therefore, Lemma 3 and Lemma 5 give us

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

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

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

as desired. □