COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire
Scribe: Paula Gradu

Lecture #12
March 13, 2019

1 Boosting

1.1 Review

Last time we discussed the AdaBoost algorithm which uses a weak learning algorithm T times to get T weak hypotheses and combines them intelligently in order to achieve low training and generalization error. Given training set $S = \langle (x_1, y_1), ..., (x_m, y_m) \rangle$, we begin with a uniform distribution over examples and then update it as illustrated below:

Algorithm 1 AdaBoost

$D_1(i) = 1/m$
for $T = 1, ..., T$ **do**
  Run A on D_t and get weak hypothesis $h_t \in \mathcal{H}$
  $\alpha_t = \frac{1}{2} \ln \frac{1-\epsilon_t}{\epsilon_t} > 0$
  $D_{t+1}(i) = \frac{1}{Z_t} D_t(i) e^{-\alpha_t y_i h_t(x_i)}$ (where Z_t is a normalization factor)
end for
output $H(x) = \text{sign}(\sum_{i=1}^{T} \alpha_t h_t(x))$

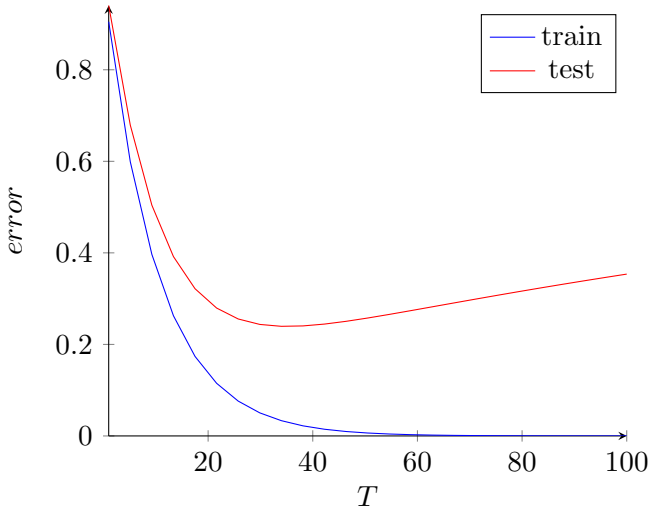
Last time we proved the following bounds on the training error and generalization error of the combined hypothesis H produced by AdaBoost:

Theorem 1.1. *$\widehat{err}(H) \leq \prod_{t=1}^{T} (2\sqrt{\epsilon_t(1-\epsilon_t)} \leq e^{-2\gamma^2 T}$, where the last inequality holds if $\gamma_t \geq \gamma$ for all t (i.e. if our weak learning assumption holds).*

Theorem 1.2. *With probability $1-\delta$, $err_D(H) \leq \widehat{err}(H) + \tilde{O}\left(\sqrt{\frac{Td + \ln 1/\delta}{m}}\right)$, where d is the VC-dimension of the weak hypothesis space $\mathcal{H}$.*

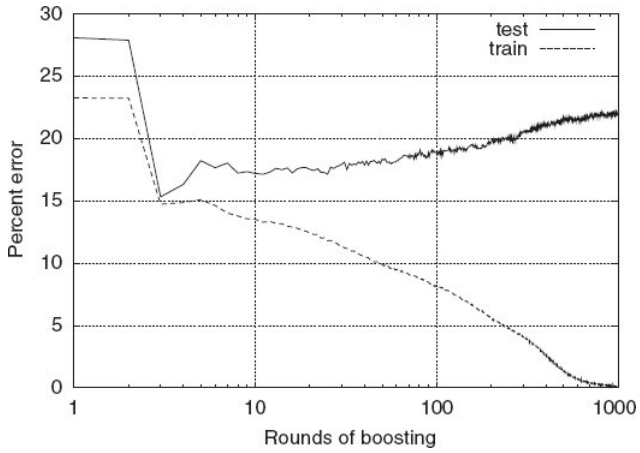
1.2 Theory vs. Reality

Based on the two theorems above, if we were to plot the performance of boosting against the number of rounds of boosting T, we expect that the training error would go down extremely fast (by Theorem 1.1), while the test error would initially go down but then eventually start going back up (by Theorem 1.2). In other words, we expect to see overfitting for large values of T. This expectation is summarized in the plot below:



So is this prediction consistent with practice? To answer this question, we will examine two different examples of AdaBoost's performance as a function of T on real data.

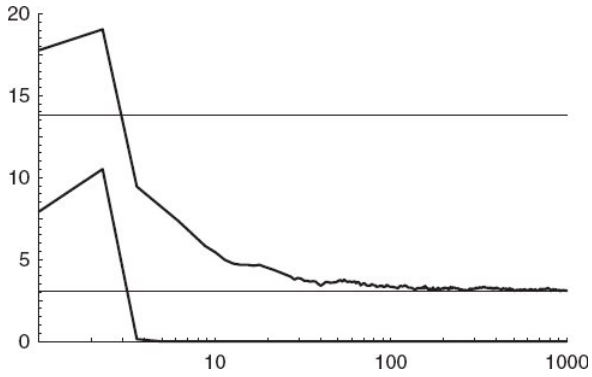
Example 1.3. *Performance of boosting with decision stumps on the heart-disease dataset:*



This example fits our expectations very well, achieving lowest test error for $T = 3$ and then degrading as we further increase T. The example above is not at all representative of an actual typical run of AdaBoost however.

A much more common outcome is illustrated in the example below:

Example 1.4. *Performance of boosting with decision trees on a dataset of handwritten letters of the alphabet:*



Clearly, the graph above does not seem to correspond to the theory we have developed so far. Even after the training error becomes 0 for $T = 5$, contrary to what Theorem 1.2 would suggest (if we were to assume it provides helpful information about AdaBoost's performance), the test error just keeps decreasing further. These results show that a very large T, such as $T = 1000$, despite producing a very large model (in this particular case $T = 1000$ creates a model of over $2,000,000$ nodes), can yield much lower generalization error than smaller values such as $T = 5$, even if the training error is zero in both cases. This goes against the intuition of Occam's Razor which suggests that, out of multiple consistent hypotheses, we should always pick the simplest one.

So what is the reason for this seeming paradox? First, clearly $\tilde{O}\left(\sqrt{\frac{Td+\ln 1/\delta}{m}}\right)$ is not a helpful bound. Secondly, training error only tells part of the story. It turns out that the generalization error depends not only on the fraction of correctly classified examples, but also on the confidence with which a particular class is assigned. So, in truth, although nothing seems to be happening if we only look at training error, the combined hypothesis gets more and more confident as we increase T which translates to increased performance on test data.

In order to properly understand and explain the observed behavior we need to introduce a mathematically rigorous interpretation of the idea of "confidence". In the following section, we will first do this and then reanalyze the performance of AdaBoost.

1.3 Confidence: From Intuition To Theory

We can think of the combined hypothesis as a weighted majority vote. Just like in elections, we will measure the confidence of a particular election outcome (i.e. of the class assigned to a particular example) by the following quantity:

$$\begin{aligned} \textit{margin} = &(\textit{weighted fraction of votes for correct class}) \\ &-(\textit{weighted fraction of votes for incorrect class}) \end{aligned}$$

Since $H(x) = \text{sign}(\sum_{i=1}^{T} \alpha_t h_t(x)) = \text{sign}(C \sum_{i=1}^{T} \alpha_t h_t(x)) \, \forall C > 0$, we can normalize the weights by taking $1/C = \sum_{t=1}^{T} \alpha_t > 0$ and considering $H(x) = \text{sign}(\sum_{i=1}^{T} a_t h_t(x))$ where $a_t = C\alpha_t \implies \sum_{t=1}^{T} a_t = 1$. From now on we will refer to $\sum_{i=1}^{T} a_t h_t(x)$ as $f(x)$.

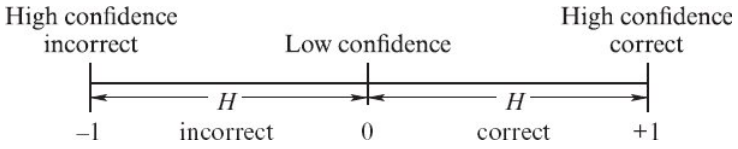
Using the notations above, we define the margin (with respect to f) to be:

Definition 1.5. $margin(x, y) = \sum_{t:h_t(x)=y} a_t - \sum_{t:h_t(x)\neq y} a_t$.

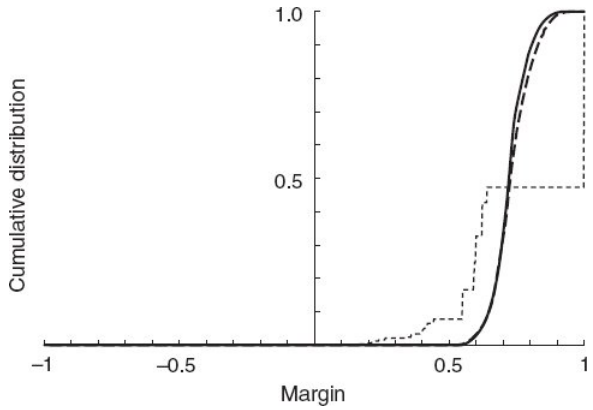
Observation 1.6. $margin(x, y) = yf(x)$.

Proof. $\sum_{t:h_t(x)=y} a_t - \sum_{t:h_t(x)\neq y} a_t = \sum_t a_t \cdot \begin{cases} +1 & \text{if } h_t(x) = y \\ -1 & \text{if } h_t(x) \neq y \end{cases} = y \sum_{i=1}^{T} a_t h_t(x) = yf(x).$ □

Since $y \in \{\pm 1\}$ and f is a convex combination of $h_t(x) \in \{\pm 1\}$, we have that $f(x) \in [-1, 1]$ and $margin(x, y) \in [-1, 1]$. Moreover, the combined hypothesis H is correct if and only if $yf(x) > 0$. Furthermore, the closer $yf(x)$ is to zero, the more evenly divided the "votes" are and conversely for $yf(x)$ close to 1 or -1. Therefore, the magnitude of $yf(x)$ represents the level of confidence of the prediction.



Since the motivation for introducing this quantity was to hopefully better capture what is actually happening in AdaBoost as T increases, let's look at T's influence on the margins of the examples in S. The graph below shows the cumulative distribution of margins of the training instances after 5, 100, and 1000 iterations, indicated by short-dashed, long-dashed, and solid curves, respectively.



So, as T increases, the lower margin is being pushed to the right (i.e. H is becoming increasingly confident). Clearly, based on the performance in example 1.4, it would be reasonable to guess that higher margins lead to better generalization error. Our goal in the next section is to explore the theory that goes along with the empirical evidence above. More specifically, we will (1) show this phenomenon is not a coincidence (i.e. that AdaBoost tends to produce higher margins as T gets large), and (2) prove that large margins lead to better generalization error. We will do so by deriving a bound that is independent of T which will show that what matters in driving down generalization error is not the number of rounds of boosting, but the margins obtained.

1.4 Number of Rounds of Boosting vs. Margin Sizes on Training Data

We first observe that the training error on S can be rewritten as $\widehat{Pr}_S[yf(x) \leq 0]$. What we are actually interested in is $\widehat{Pr}_S[yf(x) \leq \theta]$ where $\theta > 0$ is the margin we care about. By proceeding in the same manner as when proving Theorem 1.1, we get that:

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

If weak learning holds (so that $\gamma_t \geq \gamma$ for all t), then this is at most

$$[(1-2\gamma)^{1-\theta}(1+2\gamma)^{1+\theta}]^T$$

It is also possible to show that for $\theta < \gamma$:

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

So, under the weak learning hypothesis assumption, as $T \to \infty$, we have that $y_i f(x_i) \geq \gamma$, $\forall i$. That is, $\forall \theta < \gamma$, every example will eventually have margin at least γ which is what we hoped for. Based on this result, we can also note that a larger edge leads to larger margins.

1.5 Large Margins Lead to Low Generalization Error

Let $d = \text{VC-dim}(\mathcal{H})$. We will start with what we had from Theorem 1.2:

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

We can rewrite this as:

$$Pr_D[yf(x) \leq 0] \leq \widehat{Pr}_S[yf(x) \leq 0] + \tilde{O}\left(\sqrt{\frac{Td + \ln 1/\delta}{m}} \right)$$

What we want is to not have overfitting which means we would like to get rid of the T in the numerator: so we hope for $\tilde{O}\left(\sqrt{\frac{d+\ln 1/\delta}{m}} \right)$. To get this, we need to give something up.

Based on the previous section, we will attempt to bound generalization error in terms of $\widehat{Pr}_S[yf(x) \leq \theta]$ instead of $\widehat{Pr}_S[yf(x) \leq 0]$ (i.e. in terms of the fraction of examples with small margin instead of in terms of training error). But when θ is very close to zero, we are back to simply measuring the training error, so we expect there to be a degradation in the bound. This is reflected in the second term of our bound, where Td is in fact replaced by d/θ^2.

To state our bound, we introduce the convex hull, the set of all convex combinations of hypotheses in the space $\mathcal{H}$.

Definition 1.7. $co(\mathcal{H}) = \{f : X \mapsto \sum_{t=1}^{T} a_t h_t(x) \,|\, a_t \geq 0, \sum_{t=1}^{T} a_t = 1, h_t \in \mathcal{H}\, T \geq 1\}$.

Observation 1.8. *All possible outputs of boosting are included in* $co(\mathcal{H})$.

Using this, we will prove the following theorem:

Theorem 1.9. *With probability* $\geq 1-\delta$, $\forall f \in co(\mathcal{H})$ *and* $\forall\, \theta > 0$, *we have that:*

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

Although the bound holds for all $\theta > 0$, we will only prove it for a fixed value of θ. To do so, we will use the theory of Radamacher complexity which we have discussed previously.

Lemma 1.10. $\widehat{R}_S(\mathcal{H}) \leq \tilde{O}(\sqrt{d/m})$.

Proof. See lecture 10. □

Lemma 1.11. $\widehat{R}_S(\mathcal{H}) = \widehat{R}_S(co(\mathcal{H}))$.

Proof.

$$\mathcal{H} \subseteq co(\mathcal{H}) \implies \widehat{R}_S(co(\mathcal{H})) \geq \widehat{R}_S(\mathcal{H}) \tag{1}$$

On the other hand,

$$\begin{aligned}
\widehat{R}_S(co(\mathcal{H})) &= E_\sigma\left[\frac{1}{m}\sup_{f\in co(\mathcal{H})}\sum_i \sigma_i f(x_i)\right] \\
&= E_\sigma\left[\frac{1}{m}\sup_{f\in co(\mathcal{H})}\sum_i \sigma_i \sum_t a_t h_t(x_i)\right] \\
&= E_\sigma\left[\frac{1}{m}\sup_{f\in co(\mathcal{H})}\sum_t a_t \sum_i \sigma_i h_t(x_i)\right] \\
&\leq E_\sigma\left[\frac{1}{m}\sup_{f\in co(\mathcal{H})}\sum_t a_t \sup_{h\in\mathcal{H}}\sum_i \sigma_i h(x_i)\right] \\
&= E_\sigma\left[\frac{1}{m}\sup_{f\in co(\mathcal{H})}\sup_{h\in\mathcal{H}}\sum_i \sigma_i h(x_i)\right] \\
&= E_\sigma\left[\frac{1}{m}\sup_{h\in\mathcal{H}}\sum_i \sigma_i h(x_i)\right] \\
&= \widehat{R}_S(\mathcal{H})
\end{aligned} \tag{2}$$

(1) and (2) $\implies \widehat{R}_S(\mathcal{H}) = \widehat{R}_S(co(\mathcal{H}))$.

□

So, although it may feel like extending the space to $co(\mathcal{H})$ should increase the Radamacher complexity, doing so actually maintains the complexity of the original space.

Sometimes we want to consider passing all of the functions in some space through some fixed function $\phi : \mathbb{R} \to \mathbb{R}$. The natural question which arises is: what is the Rademacher complexity of the new space of transformed functions?

More concretely, given a space of functions $\mathcal{F}$, we are interested in the Radamacher complexity of the space $\phi \circ \mathcal{F} = \{\phi \circ f : f \in \mathcal{F}\}$, where $\phi \circ f$ denotes the composition of ϕ with f, so $(\phi \circ f)(z) = \phi(f(z))$.

Usually, a general formula can't be deduced without some assumptions on ϕ. An often useful assumption is Lipschitz continuity which provides a bound on the slope of the function:

Definition 1.12. *f is said to be Lipschitz continuous if $|\phi(u) - \phi(v)| \leq L_\phi |u - v|$ $\forall u, v$, where L_ϕ is called the Lipschitz constant.*

For such functions, we have the following important result:

Lemma 1.13. *(Talagrand's Concentration Lemma) For ϕ a Lipschitz continuous function,*

$$\widehat{R}_S(\phi \circ \mathcal{F}) \leq L_\phi \widehat{R}_S(\mathcal{F}).$$

Proof. See Mohri et al. □

For any function f, we will be interested in the associated margin function, $marg_f(x, y) = yf(x)$, which computes the margin of any example (x, y) with respect to f. We'll look at all the functions of the form $\mathcal{M} = \{marg_f : f \in co(\mathcal{H})\}$.

In a manner similar to a related proof from lecture #10, we can also obtain the following result:

Lemma 1.14. $\widehat{R}_S(\mathcal{M}) = \widehat{R}_S(co(\mathcal{H}))$.

Now we are finally able to prove the main theorem (Theorem 1.9):

Proof. We want to use the previous result that with probability $\geq 1 - \delta$, $\forall g \in \mathcal{F}$:

$$E[g] \leq \widehat{E}_S[g] + 2\widehat{R}_S(\mathcal{F}) + O(...)$$

To do so, we will look at:

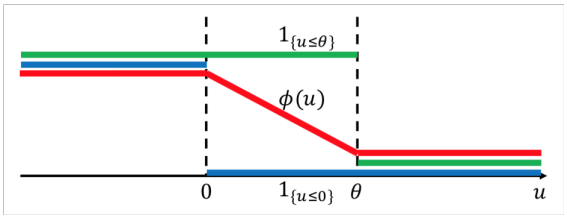
$$Pr_D[yf(x) \leq 0] = E_D[1\{yf(x) \leq 0\}]$$

$$\widehat{Pr}_S[yf(x) \leq \theta] = E_S[1\{yf(x) \leq \theta\}]$$

To resolve the mismatch between 0 and θ, we define $\phi : \mathbb{R} \to [0, 1]$ to be:

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

That is, we stretch the function between 0 and θ as illustrated below:



Therefore, we have that:

$$1\{yf(x) \leq 0\}(x) \leq \phi(x) \leq 1\{yf(x) \leq \theta\}(x) \implies$$

$$E_D[1\{yf(x) \leq 0\}] \leq E_D[\phi(yf(x))]$$

$$\widehat{E}_S[1\{yf(x) \leq \theta\}] \geq \widehat{E}_S[\phi(yf(x))]$$

Since ϕ is Lipschitz continuous with Lipschitz constant $1/\theta$, we can use Talagrand's Concentration Lemma to get that:

$$\widehat{R}_s(\phi \circ \mathcal{M}) \leq L_\phi R_S(\mathcal{M}) = (1/\theta)\, \widehat{R}_S(\mathcal{H}) \text{ by lemmas 1.11 and 1.14}$$

$$\leq (1/\theta)\, \tilde{O}(\sqrt{d/m}) \text{ by lemma 1.10}$$

Since $\forall f \in \mathcal{F}$:

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

We have that, with probability $\geq 1 - \delta$:

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

□

So we have succeeded in proving that we can indeed use boosting without overfitting and in bounding its generalization error in terms of the margin θ. Now the theory corresponds to the empirical evidence suggesting that larger margins lead to better test results.