CSCI 5525 Machine Learning Fall 2019

Lecture 15: AdaBoost

Oct 24th 2019

Lecturer: Steven Wu *Scribe: Steven Wu*

AdaBoost

Let us revisit *AdaBoost*. The algorithm takes as input a training dataset $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \{\pm 1\}$, and returns an ensemble predictor by performing:

- Initialize D_1 as the uniform distribution over the examples.
- For $t = 1, \ldots, T$:
  - Train weak classifier ("rule of thumb") h_t on D_t by minimizing the weighted risk
  $$\sum_i D_t(i) \mathbf{1}\left[h(x_i) \neq y_i\right]$$
  - Let $\epsilon_t = \sum_i D_t(i) \mathbf{1}\left[h_t(x_i) \neq y_i\right]$ be the weighted error and choose parameter α_t
  $$\alpha_t = \frac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$$
  We assume that h_t is at least as accurate as constant classifiers that always predict 1 or -1 on all examples, and so $\epsilon_t < 1/2$ and $\alpha_t > 0$.
  - Compute new distribution D_{t+1}: for each example i, the weight is
  $$D_{t+1}(i) \propto D_t(i) \exp\left(-\alpha_t y_i h_t(x_i)\right)$$
  ($\propto$ means "proportional to", which hides the normalization step.)
- Output final classifier: $\hat{f} \colon x \to \text{sign}\left(\sum_t \alpha_t h_t(x)\right)$

Under some "weak learning" assumption, one can show that the training error goes to zero exponentially fast.

Theorem 0.1. *Suppose the* weak learning assumption *holds for all t: each h_t is better than random guessing: for some $\gamma > 0$,*

$$\epsilon_t \leq 1/2 - \gamma$$

Then the training error

$$\hat{\mathcal{R}}_{01}(\hat{f}) \leq \exp\left(-2\gamma^2 T\right).$$

Ideas behind AdaBoost

The predictor class AdaBoost selects from is

$$\left\{ x \to \text{sign}(f(x)) \mid f(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \text{ for some } \alpha_1, \ldots, \alpha_T \geq 0 \text{ and } h_1, \ldots, h_T \in \mathcal{H}, T \geq 1 \right\}$$

where $\mathcal{H}$ is the weak predictor class (e.g., decision stumps). Recall that one of the surrogate losses we introduced but haven't used yet is the *exponential loss*: $\ell(y, \hat{y}) = \exp(-y_i f(x_i))$. AdaBoost is essentially trying to greedily minimize the empirical exponential loss $\frac{1}{n}\sum_i \exp(-y_i f(x_i))$. In particular, let $f_t = \sum_{\tau=1}^{t} \alpha_\tau h_\tau$. At round t, the algorithm would like to greedily improve on f_{t-1} by finding h_t and α_t to minimize

$$\begin{aligned}\sum_{i=1}^{n} \exp\left(-y_i f_t(x_i)\right) &= \sum_{i=1}^{n} \exp\left(-y_i f_{t-1}(x_i)\right) \exp\left(-y_i \alpha_t h_t(x_i)\right) \\ &\propto \sum_{i=1}^{n} D_t(i) \exp\left(-y_i \alpha_t h_t(x_i)\right)\end{aligned}$$

The last line follows by the definition of D_t:

$$D_t(i) \propto D_{t-1}(i) \exp\left(-y_i \alpha_{t-1} h_{t-1}(x_i)\right) \propto \cdots \propto \exp\left(-y_i f_{t-1}(x_i)\right)$$

Now the problem becomes finding h_t and α_t to minimize the following quantity:

$$\begin{aligned}Z_t =& \sum_{i=1}^{n} D_t(i) \exp\left(-y_i \alpha_t h_t(x_i)\right) \\ =& \sum_{i: y_i \neq h_t(x_i)} D_t(i) \exp\left(\alpha_t\right) + \sum_{i: y_i = h_t(x_i)} D_t(i) \exp\left(-\alpha_t\right) \\ =& \epsilon_t \exp\left(\alpha_t\right) + (1 - \epsilon_t) \exp\left(-\alpha_t\right) && (\epsilon_t \text{ is the error of } h_t) \\ =& \epsilon_t (\exp(\alpha_t) - \exp(-\alpha_t)) + \exp(-\alpha_t)\end{aligned}$$

Since we only consider $\alpha_t \geq 0$, then $(\exp(\alpha_t) - \exp(-\alpha_t)) \geq 0$, so we will pick h_t to minimize the weighted error ϵ_t over the distribution D_t. Now when h_t (and ϵ_t) are fixed, we will further optimize α_t. It can be shown that the optimal choice of α_t is

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1 - \epsilon_t}{\epsilon_t}\right).$$

With this choice, $Z_t = \sqrt{1 - 4\gamma_t^2}$, with $\gamma_t = 1/2 - \epsilon_t$.

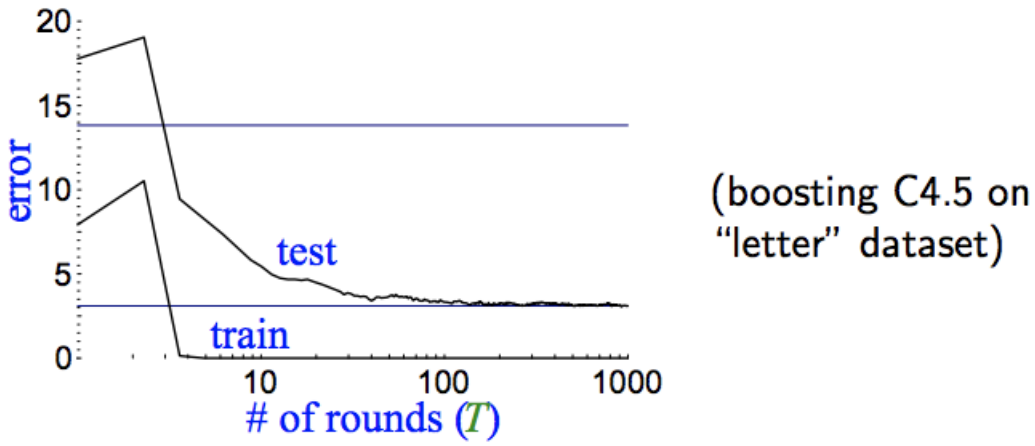

Figure 1: AdaBoost's resistance to overfitting. Test error doesn't increase with T. Image taken from Rob Schapire's slides.

Resistance to Overfitting

From Theorem 0.1, we know that AdaBoost is extremely good at driving the training error down to zero, as long as the weak learning assumption holds. Suppose that we run AdaBoost for T rounds, the final predictor consists of T weak predictors h_t's from the class $\mathcal{H}$. How good is the true error of the predictor? Standard tools from VC theory would suggest that the generalization error bound tend to increase with the complexity of the ensemble class, which in this case is roughly $T \operatorname{VCD}(\mathcal{H})$, and so:

$$\mathcal{R}_{01}(\hat{f}) \leq \hat{\mathcal{R}}_{01}(\hat{f}) + \tilde{O}\left(\sqrt{\frac{T \operatorname{VCD}(\mathcal{H})}{n}}\right)$$

According to this bound, the generalization bound should go up as we increase T. However, empirically it is observed that larger ensemble tend to have better test error, and so the VC theory is not quite adequate to explain this.

To explain this, we will consider a margin-based analysis. To define margin, we will consider the following *convex hull* $\operatorname{co}(\mathcal{H})$ as the set of all mappings that are given by a weighted average of classifiers from $\mathcal{H}$:

$$\operatorname{co}(\mathcal{H}) = \left\{ f \colon f(x) = \sum_{t=1}^{T} a_t h_t(x) \;\middle|\; a_1, \ldots, a_T \geq 0, \sum_{t=1}^{T} a_t = 1 \text{ and } h_1, \ldots, h_T \in \mathcal{H}, T \geq 1 \right\}$$

Note that this is just normalizing the weight α_t by $a_t = \frac{\alpha_t}{\sum_{t'=1}^{T} \alpha_{t'}}$, and so the sign of the weighted sum remain the same. For each f in $\operatorname{co}(\mathcal{H})$, the margin of f on each example (x_i, y_i) is $y_i f(x_i)$, which is a value in $[-1, 1]$.

Here is a nice slides show that demonstrates how the margin distribution changes during AdaBoost training: link.

Theorem 0.2 (Margin-based generalization bound)**.** *Let P be a distribution over $\mathcal{X} \times \{\pm 1\}$, and let $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be i.i.d. draws from P. Suppose that $\mathcal{H}$ is a finite class. Then for any $\delta > 0$, with probability $1 - \delta$ over the random draws from the examples, for all $f \in co(\mathcal{H})$*

$$\mathbf{Pr}_{(x,y)\sim P}[yf(x) \leq 0] \leq \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[y_i f(x_i) \leq \theta] + O\left(\sqrt{\frac{\log|\mathcal{H}|}{n\theta^2}\log\left(\frac{n\theta^2}{\log|\mathcal{H}|}\right) + \frac{\log(1/\delta)}{n}}\right)$$

for all $\theta > \sqrt{\ln(|\mathcal{H}|)/(4n)}$

We can also obtain similar results for function class $\mathcal{H}$ with bounded VC dimension.