Some explanations of adaboost

Lei Yang

December 11, 2014

Three topics [2]

1. Adaboost to Minimize Training Error
2. Direct Bounds on the Generalization Error
3. Margin Explanation for not overfitting

Adaboost review

1. Initialize $D_1(i) \equiv 1/m$
2. For $t = 1, \ldots, T$
   - 2.1 Get weak hypothesis h_t on $\mathcal{D}$: $\mathcal{X} \to \{-1, +1\}$ weighted by D_t
   - 2.2 Compute
   $$\epsilon_t \doteq Pr_{i \sim D_t}[h_t(x_i) \neq y_i]$$
   - 2.3 Find $\alpha_t = \frac{1}{2} \log\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$
   - 2.4 Set $D_{t+1} \leftarrow \frac{D_t \exp\{\alpha_t y_i h_t(x_i)\}}{Z_t}$
3. Output: $H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$

Adaboost to Minimize Training Error

We know the boosting algorithm came about to generate a strong learner from weak learner.

Above all, in practice we want to minimize training error by Adaboost first.

Here we assume the base learner can achieve a training error $\epsilon_t < 1/2 - \beta_t$ (better than guessing). We can prove that the training error of H will decrease exponentially.

Let $\gamma_t = 1/2 - \epsilon_t$ (then $\gamma_t > \beta_t > 0$), and let D_1 be an arbitrary initial distribution over the training set. Then the weighted training error of the combined classifier H with respect to D_1 is bounded as:

$$Pr_{i\sim D_1}\left[H(x_i) \neq y_i\right] \leq \prod_{t=1}^{T} \sqrt{1-4\gamma_t^2} \leq \exp\left(-2\sum_{t=1}^{T}\gamma_t^2\right)$$

Adaboost to Minimize Training Error

Let $F(x) \doteq \sum_{t=1}^{T} \alpha_t h_t(x)$. Since we have the recurrence of defining D_t,

$$
\begin{aligned}
D_{T+1}(i) &= D_1(i) \times \frac{e^{-y_i \alpha_1 h_1(x_i)}}{Z_1} \times \cdots \times \frac{e^{-y_i \alpha_T h_T(x_i)}}{Z_T} \\
&= \frac{D_1(i) \exp\left(-y_i \sum_{t=1}^{T} \alpha_t h_t(x_i)\right)}{\prod_{t=1}^{T} Z_t} \\
&= \frac{D_1(i) \exp\left(-y_i F(x_i)\right)}{\prod_{t=1}^{T} Z_t}
\end{aligned}
$$

Adaboost to Minimize Training Error

Since $H(x) = \text{sign}(F(x))$, $\mathbf{1}\{H(x) \neq y\} \leq e^{-yF(x)}$. Therefore, the (weighted) training error is:

$$
\begin{aligned}
Pr_{i\sim D_1}[H(x_i) \neq y_i] &= \sum_{i}^{m} D_1(i)\mathbf{1}\{H(x_i) \neq y_i\} \\
&\leq \sum_{i=1}^{m} D_1(i)\exp(-y_i F(x_i)) \\
&= \sum_{i=1}^{m} D_{T+1}(i)\prod_{t=1}^{T} Z_t \\
&= \prod_{t=1}^{T} Z_t
\end{aligned}
$$

Adaboost to Minimize Training Error

Finally, by our choice of α_t, we have that

$$
\begin{aligned}
Z_t &= \sum_{i=1}^{m} D_t(i) e^{-\alpha_t y_i h_t(x_i)} \\
&= \sum_{i: y_i = h_t(x_i)} D_t(i) e^{-\alpha_t} + \sum_{i: y_i \neq h_t(x_i)} D_t(i) e^{\alpha_t} \\
&= e^{-\alpha_t}(1 - \epsilon_t) + e^{\alpha_t} \epsilon_t \\
&= e^{-\alpha_t} \left(\frac{1}{2} + \gamma_t \right) + e^{\alpha_t} \left(\frac{1}{2} - \gamma_t \right) \\
&= \sqrt{1 - 4\gamma_t^2}
\end{aligned}
$$

Direct Bounds on the Generalization Error

Our ultimate goal is to find a classifier H with low generalization error

$$err(H) \doteq Pr_{(x,y)\sim D}[H(x) \neq y]$$

If the difference of $err(H) - \hat{err}(H)$ can be bounded, we may find a bound for $err(H)$ by our knowledge of $\hat{err}(H)$.

Direct Bounds on the Generalization Error

$\mathcal{H}$ is the base classifier space from which all of the h_t are selected, $\mathcal{C}_T$ is the space of combined classifiers that might potentially be generated by Adaboost if run for T rounds. We can write H as a compound function of σ and h_t:

$$H(x) = \sigma(h_1(x), \ldots, h_t(x))$$

where $\sigma : \mathbb{R}^T \to \{-1, +1\}$ is some linear threshold function of the form.

$$\sigma(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x})$$

for some $\mathbf{w} \in \mathbb{R}^T$.

Direct Bounds on the Generalization Error

Let Σ_T be the space of all such linear threshold functions, then $\mathcal{C}_T$ is simply the space of linear threshold functions defined over T hypothesis from $\mathcal{H}$. That is:

$$\mathcal{C}_T = \{x \to \sigma(h_1(x), \ldots, h_T(x)) : \sigma \in \Sigma_T, h_1, \ldots, h_T \in \mathcal{H}\}$$

Lemma: The space Σ_n of linear threshold functions over $\mathbb{R}^n$ has VC-dimension n.[2]
Finite Base Hypothesis Spaces: Assume H is finite. Let $m \geq T \geq 1$. For any set S of m points, the number of dichotomies realizable by $\mathcal{C}_T$ is bounded as follows:

$$|\Pi_{\mathcal{C}_T}(S)| \leq \Pi_{\mathcal{C}_T}(m) \leq \left(\frac{em}{T}\right)^T |\mathcal{H}|^T$$

Theorem: Suppose AdaBoost is run for T rounds on $m \geq T$ random examples, using base classifiers from a finite space H. Then, with probability at least $1 - \delta$ (over the choice of the random sample), the combined classifier H satisfies:

$$err(H) \leq \hat{err}(H) + \sqrt{\frac{32[T \log(em|\mathcal{H}|/T) + \log(8/\delta)]}{m}}$$

Direct Bounds on the Generalization Error

Plug in the bound for training error $\hat{err}(H)$.
With probability at least $1 - \delta$.

$$e^{-2\gamma^2 T} + O\left(\sqrt{\frac{T \log(m|\mathcal{H}|/T) + \log(1/\delta)}{m}}\right)$$

Figure taken from [1].

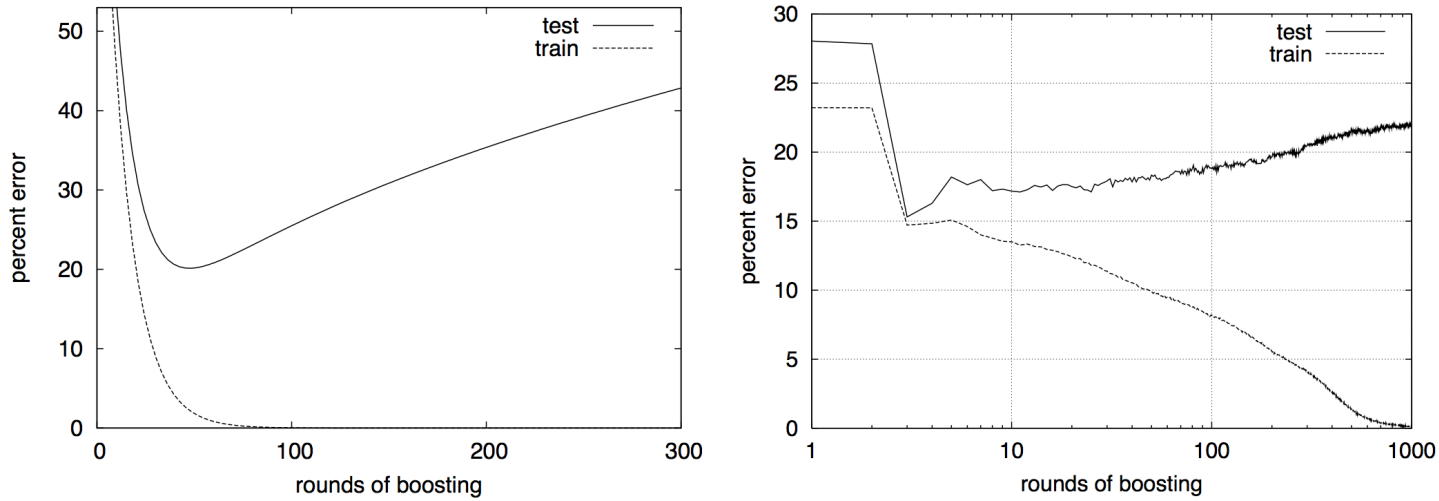

Fig. 2 *Left:* A plot of the theoretical training and test percent errors for AdaBoost, as predicted by the arguments of Section 2. *Right:* The training and test percent error rates obtained using boosting on the Cleveland heart-disease benchmark dataset. (Reprinted from [30] with permission of

Direct Bounds on the Generalization Error

Assume H has finite VC-dimension $d \geq 1$. Let $m \geq \max\{d, T\}$. For any set S of m points, the number of dichotomies realizable by $\mathcal{C}_T$ is bounded as follows:

$$|\Pi_{\mathcal{C}_T}(S)| \leq \Pi_{\mathcal{C}_T}(m) \leq \left(\frac{em}{T}\right)^T \left(\frac{em}{d}\right)^{dT}$$

so with probability $1 - \delta$:

$$err(H) \leq \hat{err}(H) + \sqrt{\frac{32[T(\log(em/T) + d\log(em/d)) + \log(8/\delta)]}{m}}$$

Margin as a Measure of Confidence

Recall $H(x) = \text{sign}(F(x))$, $F(x) \doteq \sum_{t=1}^{T} \alpha_t h_t(x)$. It will be convenient to normalize the nonnegative weights α_t on the base classifiers. Let $a_t \doteq \dfrac{\alpha_t}{\sum_{t'=1}^{T} \alpha_{t'}}$ and let $f(x) \doteq \sum_{t=1}^{T} a_t h_t(x) = \dfrac{F(x)}{\sum_{t=1}^{T} \alpha_t}$. Define the margin simply to be $yf(x)$.

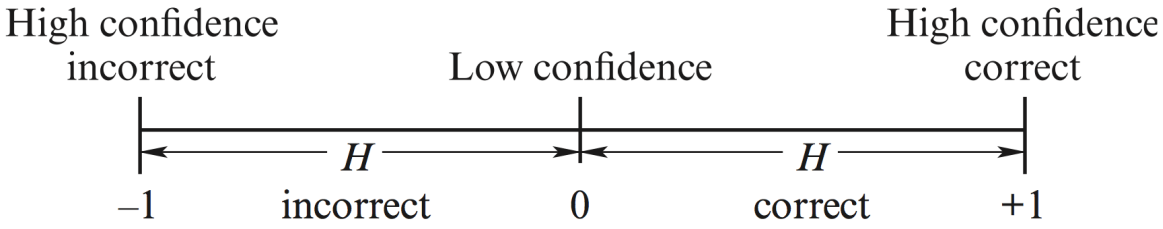

Figure 5.1
An example's margin is in the range $[-1, +1]$ with a positive sign if and only if the combined classifier H is correct. Its magnitude measures the confidence in the combined classifier's prediction.

Margin Explanation for not overfitting

Finite Base Hypothesis Spaces: With probability $1-\delta$

$$Pr_{\mathcal{D}}[yf(x) \leq 0] \leq Pr_S[yf(x) \leq \theta] + O\left(\sqrt{\frac{\log|\mathcal{H}|}{m\theta^2} \cdot \log\left(\frac{m\theta^2}{\log|\mathcal{H}|}\right) + \frac{\log(1/\delta)}{m}}\right)$$

for all $\theta > \sqrt{\frac{log(\mathcal{H})}{4m}}$. Infinite Base Hypothesis Spaces:

$$Pr_{\mathcal{D}}[yf(x) \leq 0] \leq Pr_S[yf(x) \leq \theta] + O\left(\sqrt{\frac{d\log(m/d)}{m\theta^2} \cdot \log\left(\frac{m\theta^2}{d}\right) + \frac{\log(1/\delta)}{m}}\right)$$

for all $\theta > \sqrt{8d\log(em/d)/m}$.

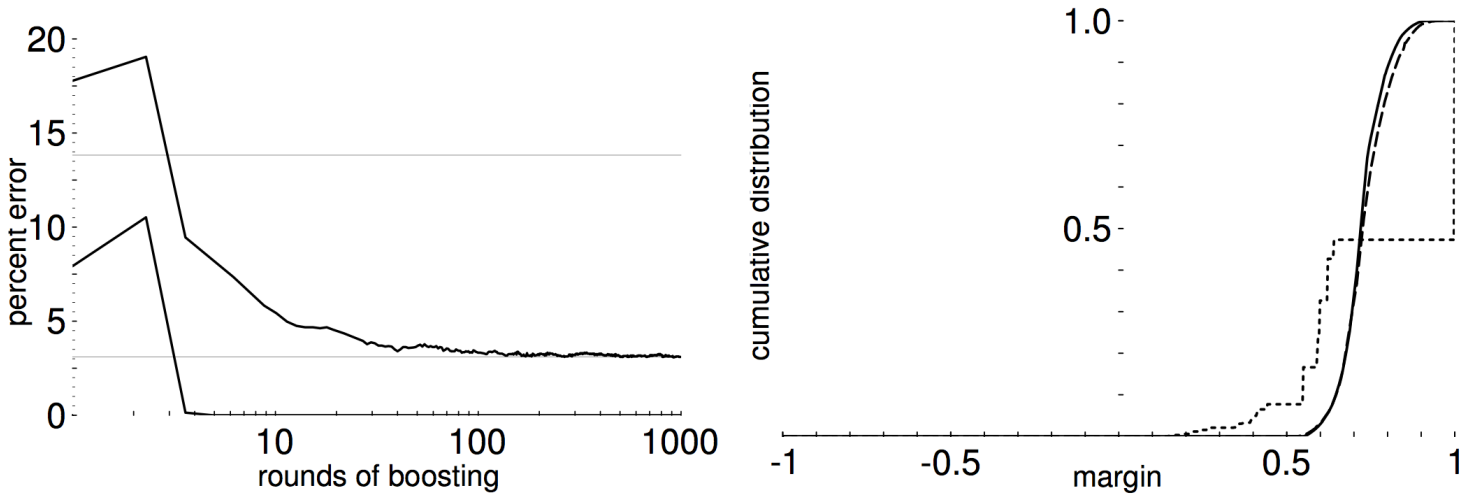

Fig. 3 *Left:* The training and test percent error rates obtained using boosting on an OCR dataset with C4.5 as the base learner. The top and bottom curves are test and training error, respectively. The top horizontal line shows the test error rate using just C4.5. The bottom line shows the final test error rate of AdaBoost after 1000 rounds. *Right:* The margin distribution graph for this same case showing the cumulative distribution of margins of the training instances after 5, 100 and 1000 iterations, indicated by short-dashed, long-dashed (mostly hidden) and solid curves, respectively. (Both figures are reprinted from [31] with permission of the Institute of Mathematical Statistics.)

References

[1] Robert E Schapire. Explaining adaboost. In *Empirical Inference*, pages 37–52. Springer, 2013.

[2] Robert E. Schapire and Yoav Freund. *Boosting: Foundations and Algorithms*. The MIT Press, 2012.