

# AdaBoost

Daniel Hsu (COMS 4771)

## The algorithm

The input training data is  $\{(x_i, y_i)\}_{i=1}^n$  from  $\mathcal{X} \times \{-1, +1\}$ .

- Initialize  $D_1(i) := 1/n$  for each  $i = 1, \dots, n$ .
- For  $t = 1, \dots, T$ , do:
  - Give  $D_t$ -weighted examples to Weak Learner; get back  $h_t: \mathcal{X} \rightarrow \{-1, +1\}$ .
  - Compute weight on  $h_t$  and update weights on examples:

$$s_t := \sum_{i=1}^n D_t(i) \cdot y_i h_t(x_i)$$
$$\alpha_t := \frac{1}{2} \ln \frac{1 + s_t}{1 - s_t}$$

where  $1, \dots, n$

$Z_t := \sum_{i=1}^n D_t(i) \cdot e^{\alpha_t h_t(x_i)}$

is the normalizer that makes  $D_{t+1}$  a probability distribution

- Final hypothesis is  $\hat{h}$  defined by  $\hat{h}(x) := \text{sign}\left(\sum_{t=1}^T \alpha_t \cdot h_t(x)\right)$  for  $x \in \mathcal{X}$ .

## Training error rate bound

Let  $\hat{\ell}$  be the function defined by

$$\hat{\ell}(x) := \sum_{t=1}^T \alpha_t \cdot h_t(x) \quad \text{for } x \in \mathcal{X}$$

so  $\hat{h}(x) = \text{sign}(\hat{\ell}(x))$ . The training error rate of  $\hat{h}$  can be bounded above by the average exponential loss of  $\hat{\ell}$ :

$$\frac{1}{n} \sum_{i=1}^n \mathbf{1}\{\hat{h}(x_i) \neq y_i\} \leq \frac{1}{n} \sum_{i=1}^n \exp(-y_i \hat{\ell}(x_i)).$$

This holds because

$$\hat{h}(x_i) \neq y_i \Leftrightarrow -y_i \hat{\ell}(x_i) \geq 0 \Leftrightarrow \exp(-y_i \hat{\ell}(x_i)) \geq 1.$$

Furthermore, the average exponential loss of  $\hat{\ell}$  equals the product of the normalizers from all rounds:

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n \exp(-y_i \hat{\ell}(x_i)) &= \sum_{i=1}^n D_1(i) \cdot \exp\left(-\sum_{t=1}^T \alpha_t \cdot y_i h_t(x_i)\right) \\
&= Z_1 \sum_{i=1}^n \frac{D_1(i) \cdot \exp(-\alpha_1 \cdot y_i h_1(x_i))}{Z_1} \cdot \exp\left(-\sum_{t=2}^T \alpha_t \cdot y_i h_t(x_i)\right) \\
&= Z_1 \sum_{i=1}^n D_2(i) \cdot \exp\left(-\sum_{t=2}^T \alpha_t \cdot y_i h_t(x_i)\right) \\
&= Z_1 Z_2 \sum_{i=1}^n \frac{D_2(i) \cdot \exp(-\alpha_2 \cdot y_i h_2(x_i))}{Z_2} \cdot \exp\left(-\sum_{t=3}^T \alpha_t \cdot y_i h_t(x_i)\right) \\
&= Z_1 Z_2 Z_3 \sum_{i=1}^n D_3(i) \cdot \exp\left(-\sum_{t=3}^T \alpha_t \cdot y_i h_t(x_i)\right) \\
&= \dots \\
&= \prod_{t=1}^T Z_t.
\end{aligned}$$

Since each  $y_i h_t(x_i) \in \{-1, +1\}$ , the normalizer  $Z_t$  can be written as

$$Z_t = \sum_{i=1}^n D_t(i) \cdot \exp(-\alpha_t \cdot y_i h_t(x_i))$$

$$= \sum_{i=1}^n D_t(i) \cdot \frac{1 + y_i h_t(x_i)}{2} \frac{1 - s}{1 + s} \frac{1 - y h(x)}{1 + s}$$

$$= \frac{1}{\sqrt{(1+s_t)(1-s_t)}} = \frac{1}{\sqrt{1-s_t^2}}$$

$$= \sqrt{1-s_t^2}.$$

So, we conclude the following bound on the training error rate of  $\hat{h}$ :

$$\frac{1}{n} \sum_{i=1}^n \mathbf{1}\{\hat{h}(x_i) \neq y_i\} \leq \frac{1}{n} \sum_{i=1}^n \exp(-y_i \hat{\ell}(x_i)) = \prod_{t=1}^T Z_t = \prod_{t=1}^T \sqrt{1-s_t^2} \leq \exp\left(-\frac{1}{2} \sum_{t=1}^T s_t^2\right)$$

where the last step uses the fact that  $1+x \leq e^x$  for any real number  $x$ .

(The bound is usually written in terms of  $\gamma_t := s_t/2$ , i.e., as  $\exp(-2 \sum_{t=1}^T \gamma_t^2)$ .)

## Margins on training examples

Let  $\hat{g}$  be the function defined by

$$\hat{g}(x) := \frac{\sum_{t=1}^T \alpha_t \cdot h_t(x)}{\sum_{t=1}^T |\alpha_t|} \quad \text{for } x \in \mathcal{X}$$

so  $y_i \hat{g}(x_i)$  is the margin achieved on example  $(x_i, y_i)$ . We may assume without loss of generality that  $\alpha_t \geq 0$  for each  $t = 1, \dots, T$  (by replacing  $h_t$  with  $-h_t$  as needed.) Fix a value  $\theta \in (0, 1)$ , and consider the fraction of training examples on which  $\hat{g}$  achieves a margin at most  $\theta$ :

$$\frac{1}{n} \sum_{i=1}^n \mathbf{1}\{y_i \hat{g}(x_i) \leq \theta\}.$$

This quantity can be bounded above using the arguments from before:

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n \mathbf{1}\{y_i \hat{g}(x_i) \leq \theta\} &= \frac{1}{n} \sum_{i=1}^n \mathbf{1}\left\{y_i \hat{\ell}(x_i) \leq \theta \sum_{t=1}^T \alpha_t\right\} \\
&\leq \exp\left(\theta \sum_{t=1}^T \alpha_t\right) \cdot \frac{1}{n} \sum_{i=1}^n \exp(-y_i \hat{\ell}(x_i)) \\
&= \exp\left(\theta \sum_{t=1}^T \alpha_t\right) \cdot \prod_{t=1}^T \sqrt{1 - s_t^2} \\
&= \prod_{t=1}^T \sqrt{(1 + s_t)^{1+\theta} (1 - s_t)^{1-\theta}}.
\end{aligned}$$

Suppose that for some  $\gamma > 0$ ,  $s_t \geq 2\gamma$  for all  $t = 1, \dots, T$ . If  $\theta < \gamma$ , then using calculus, it can be shown that each term in the product is less than 1:

$$\sqrt{(1 + s_t)^{1+\theta} (1 - s_t)^{1-\theta}} < 1.$$

Hence, the bound decreases to zero exponentially fast with  $T$ .

# Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro