Discussion 07

2022.04.28

林维嘉 linwj@shanghaitech.edu.cn

Outline

Ensemble Learning

- Adaboost
- Perceptron

Adaboost

Given training set $(x_1, y_1), \dots, (x_m, y_m)$

$y_i \in Y = \{-1, +1\}$, true label of instance $x_i \in X$

For $t = 1,2, \dots, T$:

Construct distribution D_t on $\{x_1, x_2, \dots, x_m\}$

Find weak classifier:

$$h_t: X \to \{-1, +1\}$$

with small error ε_t on D_t

$$\varepsilon_t = P_{x_i \sim D_t}[h_t(x_i) \neq y_i]$$

Output final classifier: $H_{final}(x) = sign(\sum_{t=1} \alpha_t h_t(x))$

Adaboost

Construct D_t

$t = 1$, $D_t(i) = \frac{1}{m}$, uniform on $\{x_1, x_2, \dots, x_m\}$

$t \geq 2$, given D_t and h_t:

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} * \begin{cases} e^{-\alpha_t}, if\ y_i = h_t(x_i) \\ e^{\alpha_t}, if\ y_i \neq h_t\ (x_i) \end{cases}$$

$$= \frac{D_t(i)}{Z_t} e^{\left(-\alpha_t\, y_i h_t(x_i)\right)}$$

where Z_t is the normalization constant and $\alpha_t = \frac{1}{2} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) > 0$

Adaboost

Claim: D_{t+1} puts half of the weight on x_i where h_t was incorrect and half of the weight on x_i where h_t was correct.

Recall: $D_{t+1}(i) = \frac{D_t(i)}{Z_t} e^{(-\alpha_t\, y_i h_t(x_i))}$, $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$, $\varepsilon_t = \sum_{i:\, y_i \neq h_t(x_i)} D_t(i)$

$$\Pr[\, y_i \neq h_t\ (x_i)] = \sum_{i:\, y_i \neq h_t(x_i)} \frac{D_t(i)}{Z_t} e^{\alpha_t} = \frac{\varepsilon_t}{Z_t}\sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}} = \frac{\sqrt{\varepsilon_t(1-\varepsilon_t)}}{Z_t}$$

$$\Pr[\, y_i = h_t\ (x_i)] = \sum_{i:\, y_i = h_t(x_i)} \frac{D_t(i)}{Z_t} e^{-\alpha_t} = \frac{1-\varepsilon_t}{Z_t}\sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}} = \frac{\sqrt{\varepsilon_t(1-\varepsilon_t)}}{Z_t}$$

$$\Rightarrow \Pr[\, y_i \neq h_t\ (x_i)] = \Pr[\, y_i = h_t\ (x_i)]$$

Training Error Analysis

Theorem: $\varepsilon_t = \frac{1}{2} - \gamma_t$ (error of h_t over D_t)

$$err_S(H_{final}) \leq e^{-2\sum_t \gamma_t^2}$$

So, if $\forall t, \gamma_t \geqslant \gamma > 0$, then $err_S(H_{final}) \leq e^{-2\gamma^2 T}$

Adaboost is adaptive:

- Does not need to know γ or T as a priori
- Can exploit $\gamma_t \gg \gamma$

Proof

Let $f(x_i) = \sum_t \alpha_t h_t(x_i) \Rightarrow H_{final}(x_i) = sign(f(x_i))$

Step 1: unwrapping recurrence

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} e^{(-\alpha_t\, y_i h_t(x_i))}$$

$$= \frac{e^{(-\alpha_t y_i h_t(x_i))}}{Z_t} * \frac{e^{(-\alpha_{t-1} y_i h_{t-1}(x_i))}}{Z_{t-1}} * D_{t-1}(i)$$

Note that $D_1(i) = \frac{1}{m}$

$$D_{t+1}(i) = \frac{1}{m} \frac{e^{(-y_i(\alpha_t h_t(x_i) + \cdots + \alpha_1 h_1(x_i)))}}{\prod_t Z_t} = \frac{1}{m} \frac{e^{(-y_i f(x_i))}}{\prod_t Z_t}$$

Proof

Step 2: $err_S(H_{final}) \leq \prod_t Z_t$

$$err_S(H_{final}) = \frac{1}{m}\sum_i \begin{cases} 1, if\ y_i \neq H_{final}(x_i) \\ 0, if\ y_i = H_{final}(x_i) \end{cases}$$

$$= \frac{1}{m}\sum_i \begin{cases} 1, if\ y_i f(x_i) \leq 0 \\ 0, if\ y_i f(x_i) > 0 \end{cases}$$

$$\leq \frac{1}{m}\sum_i e^{(-y_i f(x_i))}$$

Note that $D_{t+1}(i) = \frac{1}{m}\frac{e^{(-y_i f(x_i))}}{\prod_t Z_t} \Rightarrow D_{t+1}(i)\prod_t Z_t = \frac{e^{(-y_i f(x_i))}}{m}$

$$err_S(H_{final}) = \prod_t Z_t \left(\sum_i D_{t+1}(i)\right) = \prod_t Z_t$$

Proof

Step 3: $Z_t = 2\sqrt{\varepsilon_t(1-\varepsilon_t)}$

$$Z_t = \sum_{i=1}^{m} D_t(i)e^{(-\alpha_t\, y_i h_t(x_i))}$$

$$= \sum_{i:y_i \neq h_t(x_i)} D_t(i)e^{\alpha_t} + \sum_{i:y_i = h_t(x_i)} D_t(i)e^{-\alpha_t}$$

$$= \varepsilon_t e^{\alpha_t} + (1-\varepsilon_t)e^{-\alpha_t}$$

Recall: $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$ $\leftarrow$ $\boxed{\text{How to get } \alpha_t\text{?}}$

Proof

$$\frac{\partial Z_t}{\partial \alpha_t} = \varepsilon_t e^{\alpha_t} - (1 - \varepsilon_t) e^{-\alpha_t} = 0$$

$$\Rightarrow \alpha_t + \ln(\varepsilon_t) = -\alpha_t + \ln(1 - \varepsilon_t)$$

$$\Rightarrow \alpha_t = \frac{1}{2} \ln\left(\frac{1 - \varepsilon_t}{\varepsilon_t}\right)$$

Plug α_t into Z_t:

$$Z_t = \varepsilon_t \sqrt{\frac{1 - \varepsilon_t}{\varepsilon_t}} + (1 - \varepsilon_t) \sqrt{\frac{\varepsilon_t}{1 - \varepsilon_t}} = 2\sqrt{\varepsilon_t (1 - \varepsilon_t)}$$

Proof

Step 4: $err_S\left(H_{final}\right) \leq e^{-2\sum_t \gamma_t^2}$

$$err_S\left(H_{final}\right) \leq \prod_t Z_t$$

$$= \prod_t 2\sqrt{\varepsilon_t(1-\varepsilon_t)}$$

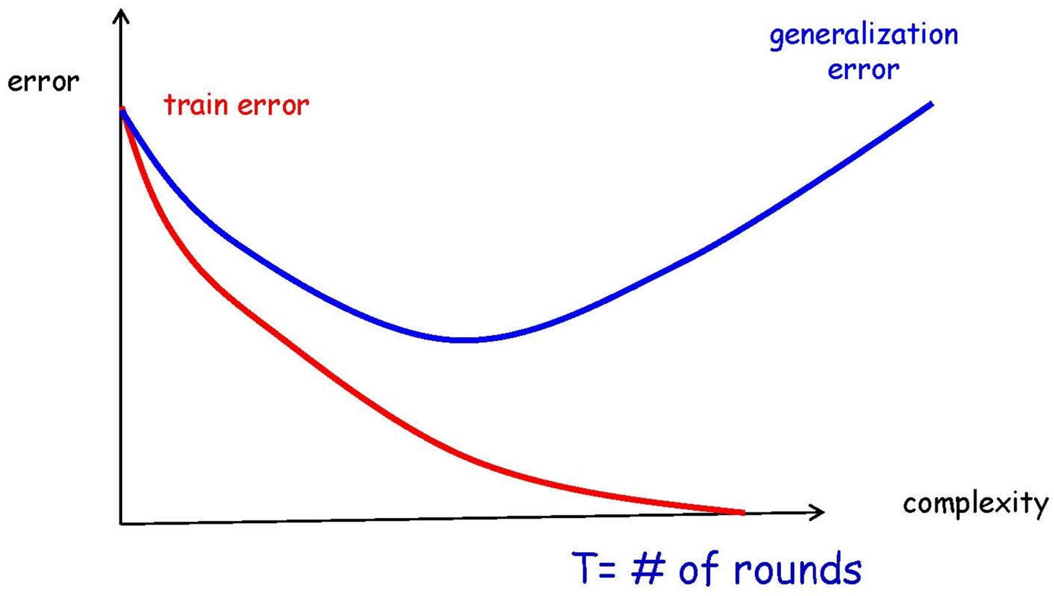
$$1 - x \leq e^{-x}$$
$$\sqrt{1-x} \leq e^{-\frac{x}{2}}$$

$$= \prod_t \sqrt{1 - 4\gamma_t^2}$$
$$\leq e^{-2\sum_t \gamma_t^2}$$

Training Error and Generalization Error

$$err_D(h) \leq err_S(h) + complexity(h)$$

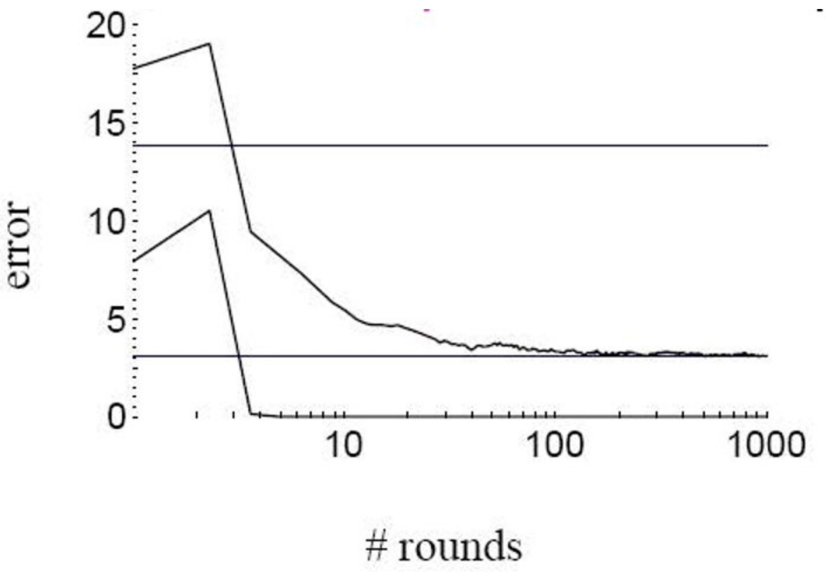
$$err_D(h) \leq err_S(h) + \tilde{O}\left(\sqrt{\frac{Td}{m}}\right)$$



Training Error and Generalization Error

$$err_D(h) \ \leq err_S(h) + complexity(h)$$

$$err_D(h) \ \leq err_S(h) + \tilde{O}(\sqrt{\frac{Td}{m}})$$



Training Error and Generalization Error

Key Idea:

- Training error does not tell the whole story.
- We need also to consider the classification confidence!!

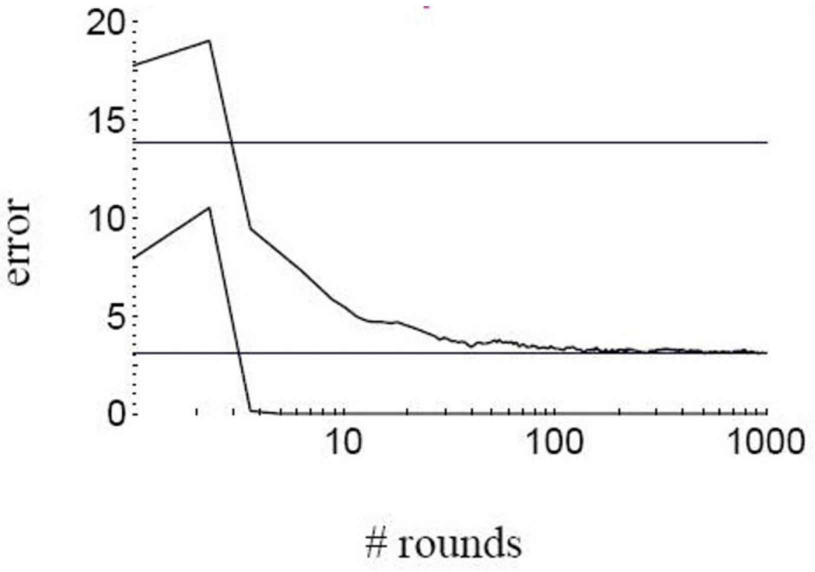
Solution: **Margin**

Definition: margin of H_f on example (x, y) to be $yf(x)$

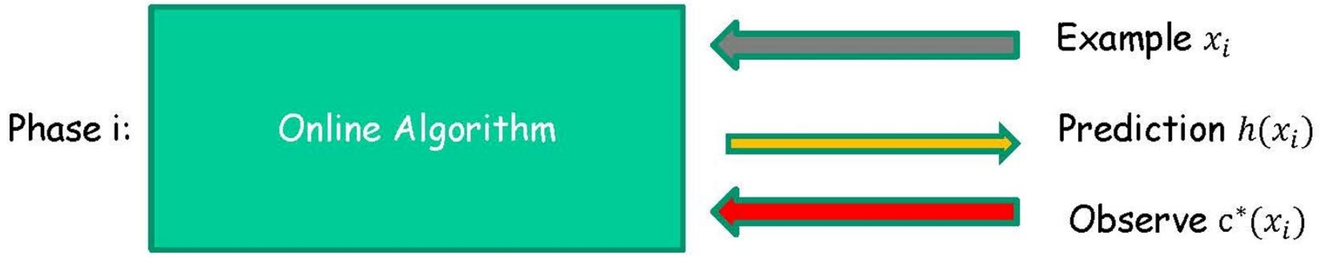
$$yf(x) = y\sum_t \alpha_t h_t(x) = \sum_t y\alpha_t h_t(x) = \sum_{t:y=h_t(x)} \alpha_t - \sum_{t:y\neq h_t(x)} \alpha_t$$

Training Error and Generalization Error

$$Pr_D[yf(x) \leq 0] \leq Pr_s[yf(x) \leq \theta] + \tilde{O}(\sqrt{\frac{d}{m\theta^2}})$$



Online Learning



Mistake bound model:

- Analysis wise, make no distributional assumptions.
- Goal: Minimize the number of mistakes.

Perceptron

Set $t = 1$, start with the all zero vector w_1.

Given example x, predict positive iff $w_t \bullet x \geq 0$

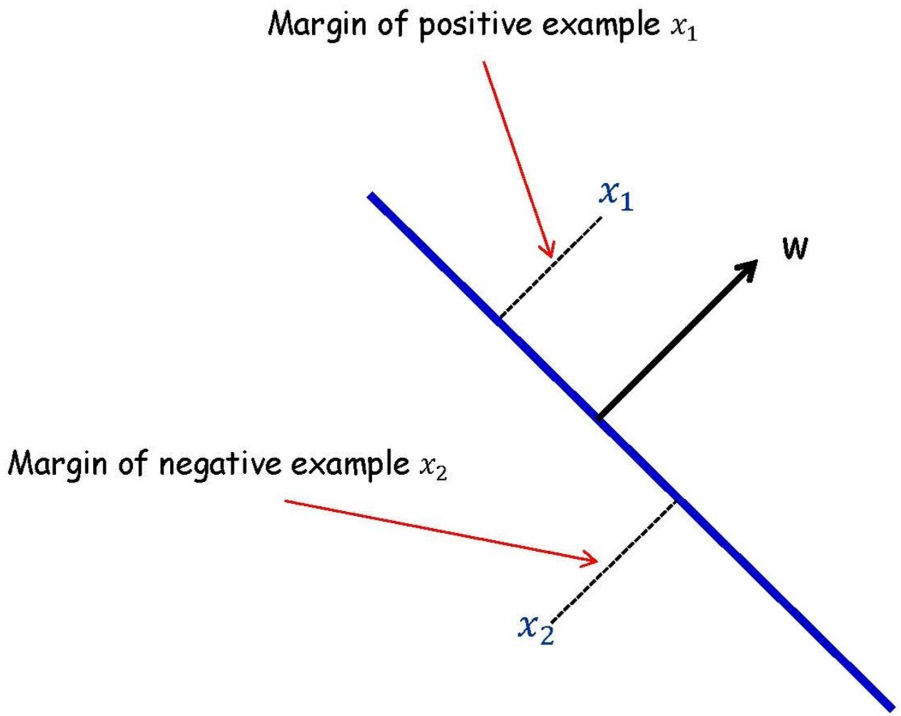
On a mistake, update as follows:

Mistake on positive, then update $w_{t+1} \leftarrow w_t + x$

Mistake on negative, then update $w_{t+1} \leftarrow w_t - x$

Geometric Margin

Definition: The margin of example x w.r.t. a linear separator w is the distance from x to the plane $w \bullet x = 0$(or the negative if on wrong side).



$$\begin{cases} \gamma_1 \frac{w}{||w||} = x_1 - x_0 \\ w^T x_0 = 0 \end{cases}$$

$$\Rightarrow \gamma_1 = \frac{w^T x_1}{||w||}$$

Similarly, $\gamma_2 = -\frac{w^T x_2}{||w||}$

$$\gamma_i = y_i \frac{w^T x_i}{||w||}$$

Geometric Margin

Definition: The margin γ_w of a set of examples S w.r.t. a linear separator w is the smallest margin over points $x \in S$.

$$\gamma_w = \min_{x_i \in S} \gamma_i = \min_{x_i \in S} y_i \frac{w^T x_i}{||w||}$$

Definition: The margin γ of a set of example S is the maximum γ_w over all linear separators w.

$$\gamma = \max_{w \in R^d} \gamma_w = \max_{w \in R^d} \min_{x_i \in S} y_i \frac{w^T x_i}{||w||}$$