Model selection

$\mathcal{H}_1, \cdots \mathcal{H}_k$

$\hat{h}$.

w.h.p.

$$err(\hat{h}, D) \leq \min_i \left(err(h_i^*, D) + O\left(\sqrt{\frac{\ln|\mathcal{H}_i|}{m}}\right)\right)$$

1. validation

$\hat{h}_1$ → V → $\hat{h}$

$\hat{h}_k$ → V

Idea 2: Structural risk minimization

(penalized ERM)

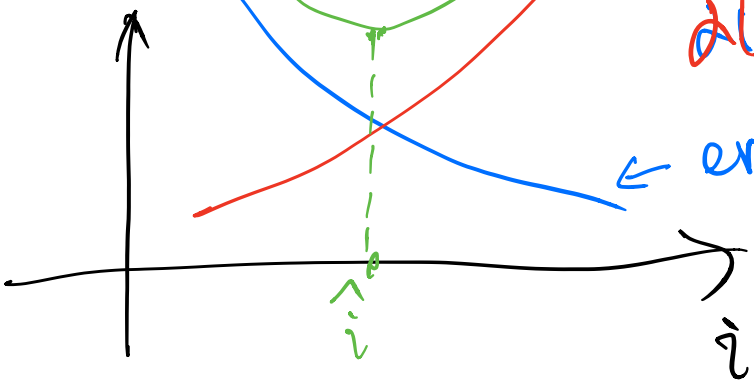
$$\hat{i} = \text{argmin} \left[err(\hat{h}_i, S) + \sqrt{\frac{\ln \frac{2k|\mathcal{H}_i|}{\delta}}{2m}} \right]$$

$err(h_i^*, D)$ ↓ $\quad \mathcal{L}(\mathcal{H}_i, m)$

$i \in \{1, \dots, k\}$

$\text{err}(\hat{h}_i, D)$

output $\hat{h} = \hat{h}_{\hat{i}}$

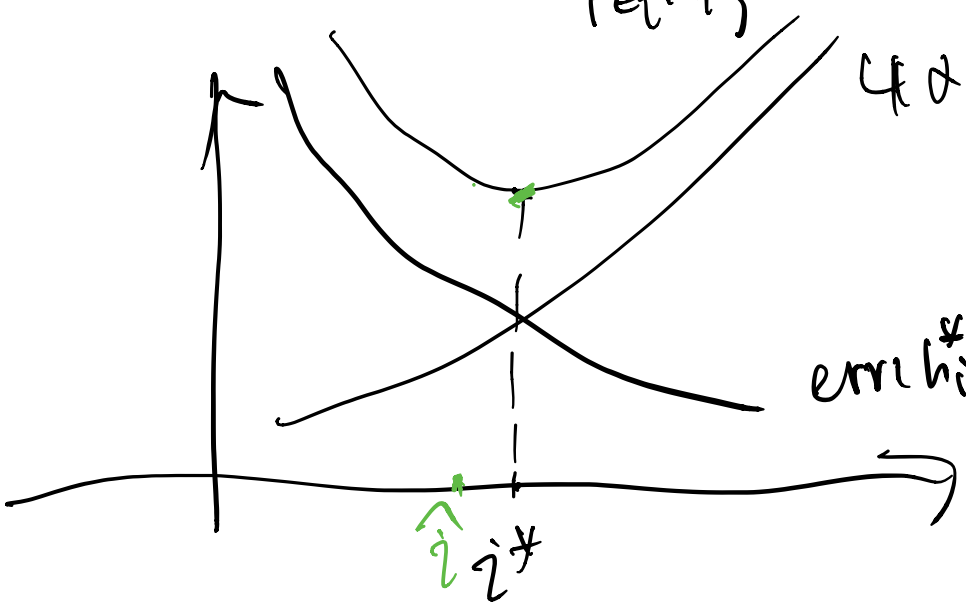


$\phi(H_i, m)$

$\text{err}(\hat{h}_i, S)$

Ex: $H_1 \subseteq H_2 \subseteq \cdots \subseteq H_k$

$$\text{err}(\hat{h}, D) \leq \min_{i \in \{1,\dots,k\}} \text{err}(h_i^*, D) + 4 \cdot \phi(H_i, m) \quad (*)$$



oracle inequality

$$i^* = \arg\min_i \left(\text{err}(h_i^*, D) + 4\phi(H_i, m) \right)$$

$\text{err}(\hat{h}, D)$ competitive w/ $\text{err}(h_{i^*}, D)$ w/o knowing i^*.

pf of (*):

$$\hat{h} = \operatorname{argmin}_{i \in \{1-k\}} \min_{h \in \mathcal{H}_i} \left(\operatorname{err}(h,S) + \sqrt{\frac{\ln \frac{|\mathcal{H}_i| \cdot k}{\delta}}{2m}} \right)$$

$\approx$ for $h \in \mathcal{H}_i$ w.h.p. $\operatorname{err}(h,D)$

step 1: w.p. $1-\delta$: for all i, for all $h \in \mathcal{H}_i$:

$$|\operatorname{err}(h,S) - \operatorname{err}(h,D)| \le \sqrt{\frac{\ln \frac{|\mathcal{H}_i| \cdot 2k}{\delta}}{2m}} \quad (\Delta_i)$$

fix $\mathcal{H}_i$. w.p. $1-\delta/k$. (Δ_i) true

$$P\left(\bigcap_i (\Delta_i)\right) \ge 1-\delta.$$

non-uniform concentration

given this happening:

$$\operatorname{err}(\hat{h},D) = \operatorname{err}(\hat{h}_{\hat{i}},D)$$

$$\le \operatorname{err}(\hat{h}_{\hat{i}},S) + \alpha(\mathcal{H}_{\hat{i}},m)$$

for all i

$$\le \operatorname{err}(\hat{h}_i,S) + \alpha(\mathcal{H}_i,m)$$

concentration within $\mathcal{H}_i$

$$\le \operatorname{err}(h_i^*,D) + 2\alpha(\mathcal{H}_i,m)$$

ERM analysis

$$\le \operatorname{err}(h_i^*,D) + 4\alpha(\mathcal{H}_i,m) \quad \blacksquare$$

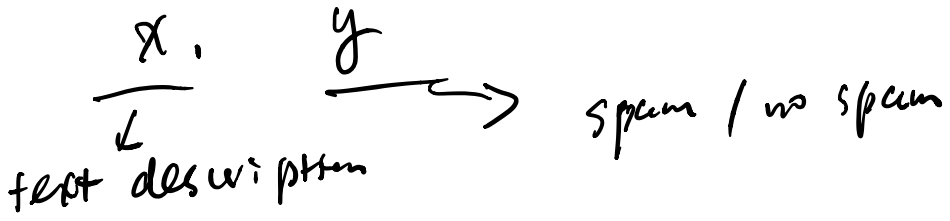
Remark: 1. $d(H, m)$ can be pessimistic in practice.

2. It may be possible to use more refined generalization error bounds.

Boosting

Motivation: combine weak classification rules to obtain strong ones.

Ex spam filtering



"free offer" → spam

"a million $"

Weak PAC learning: H. A. is a γ-weak PAC learner for H, if $\exists$ a fn $f: (0,1) \to \mathbb{N}$, for any $\delta > 0$, for any D realizable by H. w/ $m \geq f(\delta)$, A produces a classifier h. s.t. w.p. $1-\delta$:

$$err(h, D) \leq \frac{1}{2} - \gamma$$

0.01

$\mathcal{H}$ is said to be γ-weak-PAC learnable if there is a γ-weak-PAC learner for $\mathcal{H}$.

$\mathcal{H}$ PAC learnable $\Rightarrow$ weak PAC learnable

$\Leftarrow$

History:

1988: Kearns asked this question

1990: Schapire: boosting:

builds a PAC learner w/ black-box access to a weak PAC learner. based on recursion.

1990: Freund, boost by majority

combining the outputs of weak learners by a unweighted majority vote.

1997: Freund and Schapire: AdaBoost (no need to know γ)

Since: empirical success: XGBoost

AdaBoost:

given training examples: (x_1, y_1) -- (x_m, y_m)

Maintain a weighting on them, adjust weights. call weak learner.

Alg: Initialize $\left(D_1(i) = \frac{1}{m}\right)_{i=1}^{m}$.

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

$h_t \leftarrow$ B trained on weighted examples $(x_i, y_i, D_t(i))_{i=1}^{m}$.

weighted error: $\varepsilon_t = \underset{(x,y)\sim D_t}{P}\left(h_t(x) \neq y\right)$

$$= \sum_{i=1}^{m} I\left(h_t(x_i) \neq y_i\right) D_t(i).$$

$$\leq \frac{1}{2} - \gamma.$$

classifier wt: $\alpha_t = \frac{1}{2} \ln \frac{1-\varepsilon_t}{\varepsilon_t}$

$D_1(i) \propto 1$

$D_2(i) \propto e^{-\alpha_1 y_i h_1(x_i)}$

$D_3(i) \propto e^{-\alpha_1 y_i h_1(x_i) - \alpha_2 y_i h_2(x_i)}$

update wt on training examples.

$$D_{t+1}(i) = D_t(i) \cdot e^{-\alpha_t y_t h_t(x_i)} \Big/ Z_t$$

up weighting incorrect examples

Z_t: normalization factor.

Final classifier:

$$H_T(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right).$$

Analysis:

Thm: suppose for every t: $\varepsilon_t \le \frac{1}{2} - \gamma$,

then, $err(H_T, S) \le \exp(-2T \cdot \gamma^2)$

Pf: ① relate to exponential loss

② Adaboost optimizes exponential loss.

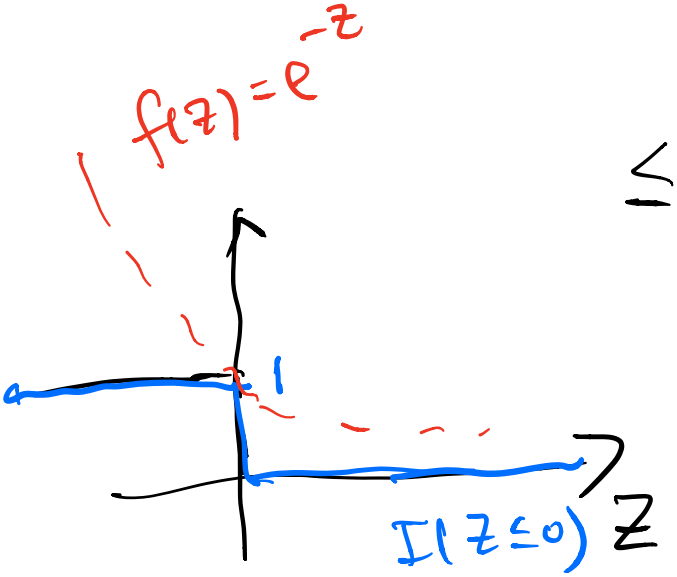
① : $err(H_T, S) = \frac{1}{m} \sum_{i=1}^{m} I(H_T(x_i) \neq y_i)$

$H_T = \text{sign}(f_T)$

$$f_t = \sum_{s=1}^{t} \alpha_s h_s(x)$$

$\downarrow z$

$$= \frac{1}{m} \sum_{i=1}^{m} I(y_i \cdot f_T(x_i) \le 0)$$



$$\le \frac{1}{m} \sum_{i=1}^{m} e^{-y_i f_T(x_i)}$$

$$= L_T$$ (cumulative exponential loss)

Show L_T is decreasing exponentially in T.

$$\frac{L_t}{L_{t-1}} = \frac{\frac{1}{m}\sum_{i=1}^{m} e^{-y_i f_t(x_i)}}{\frac{1}{m}\sum_{i=1}^{m} e^{-y_i f_{t-1}(x_i)}} \qquad (*)$$

$f_t(x_i) = f_{t-1}(x_i) + \alpha_t h_t(x_i)$

Observation 1:

$$D_t(i) \propto e^{-\sum_{s=1}^{t-1} \alpha_s y_i h_s(x_i)} = e^{-y_i f_{t-1}(x_i)}$$

$\Rightarrow \exists N_t$.

$$D_t(i) = e^{-y_i f_{t-1}(x_i)} / N_t .$$

plugging into $(*)$

$$= \frac{\sum_{i=1}^{m} N_t D_t(i) e^{-y_i \alpha_t h_t(x_i)}}{\sum_{i=1}^{m} N_t \cdot D_t(i)} \qquad \sum_{i=1}^{m} N_t \cdot D_t(i) = N_t$$

$$= Z_t .$$

$$\frac{L_t}{L_{t-1}} = Z_t$$

$$L_T = L_{t-1} \cdot Z_T$$

$$= L_{T-2} \cdot Z_{T-1} \cdot Z_T$$

$$= \dots$$

$$= \underbrace{L_0}_{=1} \cdot \prod_{t=1}^{T} Z_t$$

upper bounding Z_t:

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

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

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

$$\alpha_t = \frac{1}{2} \ln \frac{1-\varepsilon_t}{\varepsilon_t}$$

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

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

$$\varepsilon_t \le \frac{1}{2} - \gamma .$$

$$\le \sqrt{1 - 4\gamma^2} \le e^{-4\gamma^2}$$

$$\le e^{-2\gamma^2} .$$

$$\Rightarrow L_T \le \prod_{t=1}^{T} Z_t \le e^{-2T\gamma^2}$$ ✗.