HW2 is out   Due 1^{st} Oct

Recap

Introduction to Boosting

→ if realizability assumption holds, and $\mathcal{H}$ is PAC-learnable,

$$\Rightarrow \exists\ m_{\mathcal{H}}: \mathbb{R}^{+2} \to \mathbb{N} \text{ s.t.}$$

for any sample size $m \geq m_{\mathcal{H}}(\delta, \varepsilon)$, an ERM h will have

$$R_S(h) < \varepsilon \quad \text{with } \Pr \text{ (over } S \sim \mathcal{D}^m) \geq 1-\delta.$$

e.g.

$$R(h) \leq \hat{R}_S(h) + 2\mathrm{Rad}_S(\mathcal{H}) + 3\sqrt{\frac{\log 2/\delta}{2m}}$$

$|R(h) - \hat{R}_S(h)|$: * generalization gap

* ~~excess risk~~

(Excess risk is $R(h^*) - R(h)$) where h^* is Bayes optimal discriminant

by showing that

$$\Pr_{S \sim \mathcal{D}^m}\left(|R(h) - \hat{R}_S(h)| > \varepsilon\right)$$

→ $\mathcal{H}$ is large & complex, ERM problem can have ↑ comp complexity.

→ Boosting: if h_S is an ERM over a "simpler" class $\mathcal{H}W$, then can we use $\{ERM_{\mathcal{H}W}\}$ to reduce empirical & generalization error?

→ What is $\mathcal{H}W$?
An ERM on $\mathcal{H}W$ has "error" better than a "random guess".

Assume we are given

$$S = \{(x_i, y_i)\}_{i=1}^m$$

$$\ell(\underset{(x,y)}{z}, h) = \mathbb{1}_{\{y h(x) < 0\}}$$

ERM problem over $\mathcal{H}W$ satisfies

(i) $\underset{h \in \mathcal{H}W}{\operatorname{argmin}} \hat{R}_S(h)$ has ↓ comp complexity than $\underset{h \in \mathcal{H}}{\operatorname{argmin}} \hat{R}_S(h')$

(ii) $\sum_{i=1}^m P(i) = 1$ for distributions P on S,

$$\underline{R_{S,P}(h)} = \sum_{i=1}^m P(i)\, \ell(z_i, h) \text{ is small.}$$

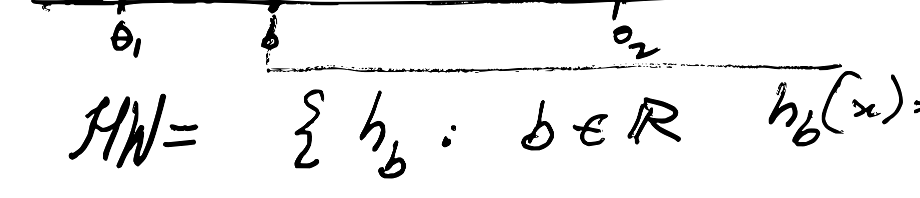
$$< \frac{1}{2} - \gamma \quad \text{with } \gamma > 0.$$

h which is an ERM over such a class of weak rules $\mathcal{H}W$ is called a "weak learner".

Example

Binary classification in 1D over linear class:

$$h(x) = \frac{\omega x}{|\omega|} + \frac{b}{|\omega|} = x + b$$



$$\mathcal{H}W = \{ h_b : b \in \mathbb{R} \quad h_b(x) = x + b \}$$

True labeling function h belongs to

$$\mathcal{H} = \{ h_{\theta_1, \theta_2} : \theta_1, \theta_2 \in \mathbb{R} \text{ s.t.}$$

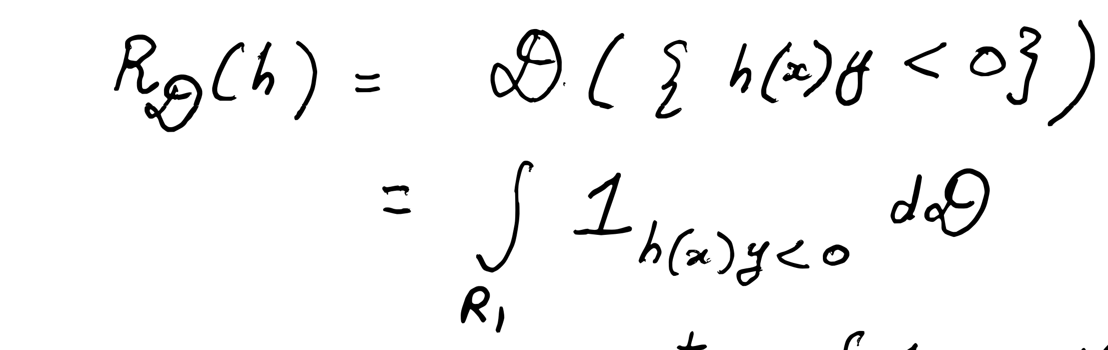
$$h_{\theta_1 \theta_2}(x) = \begin{cases} 1 & x < \theta_1 \\ -1 & \theta_1 \leq x \leq \theta_2 \\ 1 & x > \theta_2 \end{cases} \}$$

if $h \in \mathcal{H}W$, we do not expect to label arbitrarily well and with arbitrarily high prob.

For any ERM $h \in \mathcal{H}W$, for any $\mathcal{D}$,

$$R_{\mathcal{D}}(h) = \underset{z \sim \mathcal{D}}{\mathbb{E}}\, \ell(z, h) < \frac{1}{3}$$

Picture proof:



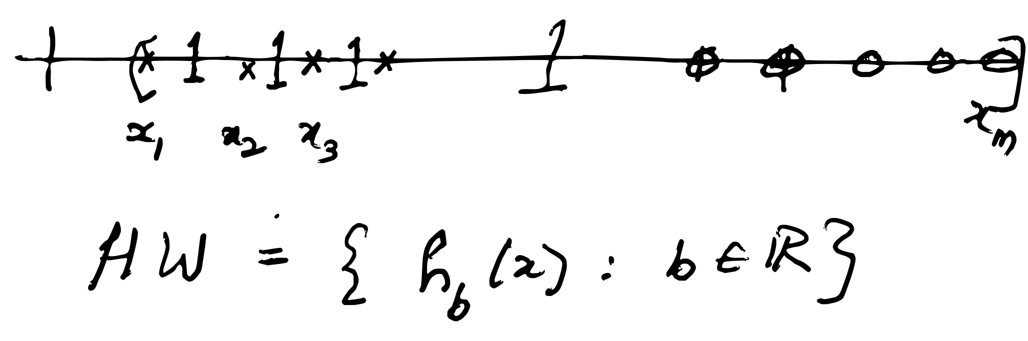
$$R_{\mathcal{D}}(h) = \mathcal{D}(\{ h(x) y < 0 \})$$

$$= \int_{R_1} \mathbb{1}_{h(x) y < 0}\, d\mathcal{D} + \int_{R_2} \mathbb{1}_{h(x) < 0}\, d\mathcal{D}$$

$$= \int_{R_1 \cup R_2} d\mathcal{D}$$

$$< \frac{1}{3} \quad \text{with high probability}$$

If efficient weak learning is possible for $ERM_{\mathcal{H}W}$, can you combine ERMs on $\mathcal{H}W$ to lower empirical error & generalization error



$$\mathcal{H}W = \{ h_b(x) : b \in \mathbb{R} \}$$

Move b in between x_i, x_{i+1} to make empirical error as small as possible.

AdaBoost

Freund Shapire 1995

Input : $S = \{(x_i, y_i)\}_{i=1}^{m}$

A: algorithm to solve ERM over $\mathcal{H}W$

T : max rounds

Output : if ERM over $\mathcal{H}W$ is f_t at time t, then $\mathrm{sgn}\left(\sum_{t \le T} f_t w_t\right)$

P_t : discrete prob. dis. over S at t.

$P_1 = 1/m$

Algorithm

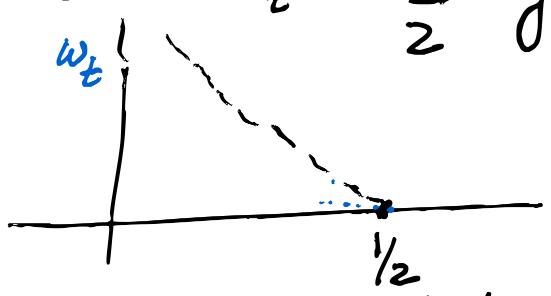
for $t = 1 : T$

→ 1. Invoke A. to get f_t

→ 2. $\varepsilon_t = R_{P_t}(f_t)$

$$= \sum_{i=1}^{m} P_t(i)\, \mathbb{1}_{\{f_t(x_i) y_i < 0\}}$$

→ 3. Set $w_t = \frac{1}{2} \log\left(\frac{1}{\varepsilon_t} - 1\right)$

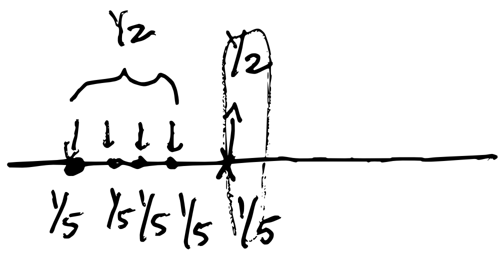


(use w.l. on diff. distributions over S)

→ 4. $P_{t+1}(i) = \dfrac{P_t(i) \times e^{-w_t y_i f_t(x_i)}}{Z_t}$

(Z_t: normalization const).

(hard examples are ↑ weights)



5. $h_t = h_{t-1} + w_t f_t$

Return $\mathrm{sgn}(h_T)$

Remark

How to decide how many samples to use for each ERM?

Suppose we had sample complexity $m_{\mathcal{H}}(\varepsilon, \delta) \Rightarrow$

$$m \geq m_{\mathcal{H}}(\varepsilon, \delta)$$

then with probability at most δ, ERM will fail

$(\varepsilon_t = \frac{1}{2} - \gamma_t, \ \gamma_t > 0)$

If we run AdaBoost for T steps, with $\Pr_{S \sim \mathcal{D}^m} \geq 1 - \delta T \longrightarrow$

Boosting reduces training error:

$$\hat{R}_S(h_T) \leq \prod_t 2\sqrt{\varepsilon_t(1-\varepsilon_t)} \quad A$$

$$= \prod_t \sqrt{1-4\gamma_t^2} \quad B \quad \left(\varepsilon_t = \frac{1}{2} - \gamma_t\right)$$

$$\gamma_t > 0$$

$$\leq e^{-2\sum_t \gamma_t^2} \quad C$$

$B \to C$

<u>Proof</u>:

$$\varepsilon_t = \frac{1}{2} - \gamma_t$$

$$D_{t+1}(i) = \frac{e^{-y_i h_t(x_i)}}{Z_{t+1}}$$

$$Z_{t+1} = \sum_{i=1}^{m} e^{-y_i h_t(x_i)}$$

Boosting generalization error:

$$R(h) \leq \hat{R}_S(h) + C\sqrt{\frac{Td}{m}}$$

d: VC dimension of weak hypothesis space

Boosting the margins

$$\rho(h; x) = y\,h(x)$$

$$R(h) \leq \hat{R}_{S,\rho}(h) + \frac{C}{\rho}\sqrt{\frac{d}{m}}$$