Boosting

- Boosting is a method for constructing good predictors from mediocre predictors

Definition: Let $\gamma \in (0, 1/2)$
A hypothesis class $H \subseteq Y^X$ is called **γ-weakly PAC learnable** iff there exists an algorithm A such that for any $\delta \in (0,1)$ there exists sample size m such that for any distribution D over X and any target function $h \in H$

$$\text{err}_{D,h}[A(S)] < \frac{1}{2} - \gamma$$

$\leq \nu_{i,h}[h(x) \neq y] \leq \frac{1}{2} - \gamma$

where $S = ((x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m))$ is $y_i = h(x_i)$ and $x_1, x_2, \ldots, x_m$ is an i.i.d. sample from D.

- Definition is almost identical to PAC learnability. The main difference is that ε was replaced by $1/2 - \gamma$.
- The order of quantifiers is the same as in PAC learning.

Theoretical question:

- Is H PAC learnable if H is γ-weakly PAC learnable?

- The answer is yes!

- The proof constructs a boosting algorithm that uses <u>weak learning algorithm</u> as a black box.

- We will allow learners that receive a <u>weighted</u> data set (D, S) where $S \subseteq (X \times Y)$ and D is distribution over S.

- D is represented as a set of weights $w_1, w_2, \ldots, w_{|S|}$ where $w_i \geq 0$ and
$$\sum_{i=1}^{|S|} w_i = 1$$

- The notion of sample error is

extended

$$\widehat{err}_{S,D}(h) = \sum_{i=1}^{|S|} w_i \, \mathbb{1}[h(x_i) \neq y_i]$$

Ada Boost

Parameters:
- Number of rounds T
- Weak learning algorithm

Input: Sample $S \in (X \times \{+1, -1\})^*$

- Set $m = |S|$
- $D_1 = \left(\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m}\right) \in \mathbb{R}^m$

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

- Call weak learner

$$h_t = WL(S, D_t)$$

labeled sample

sample weights

- Compute weighted sample error of h_t

$$\varepsilon_t = \sum_{i=1}^{m} D_{t,i} \cdot \mathbb{1}[y_i \neq h_t(x_i)]$$

- Compute

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

- Update

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

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

Output: classifier $H: X \to \{+1, -1\}$

Output: classifier $H: \mathcal{X} \to \{-1, +1\}$

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

weighted majority vote

Theorem: Let $\gamma_t = 1/2 - \varepsilon_t$.
The classifier h output by AdaBoost satisfies

$$\widehat{\text{err}}_S(h) \leq \prod_{t=1}^{T} \sqrt{1 - 4\gamma_t^2} \leq \exp\left(-2 \sum_{t=1}^{T} \gamma_t^2\right)$$

Proof:

- Let $F(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$

- We have

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

$$= D_{1,i} \frac{\exp\left(-y_i \sum_{t=1}^{T} \alpha_t h_t(x_i)\right)}{\prod_{t=1}^{T} Z_t}$$

- Note that $h(x) = \text{sign}\left(F(x)\right)$
- Claim

$$\mathbb{1}[h(x) \neq y] \leq e^{-yF(x)}$$

- If $h(x) = y$, the inequality is trivial.
- If $h(x) \neq y$ then $yF(x) \leq 0$ and thus $e^{-yF(x)} \geq 1$.

Thus $e^{-yF(x)} \geq \mathbb{1}[h(x) \neq y]$.

- Thus

$$\widehat{err}_S(h) = \sum_{i=1}^{m} D_{1,i} \cdot \mathbb{1}[h(x_i) \neq y_i]$$

$$\leq \sum_{i=1}^{m} D_{1,i} \; e^{-y_i F(x_i)}$$

$$= \sum_{i=1}^{m} D_{T+1,i} \prod_{t=1}^{T} z_t$$

$$= \prod_{t=1}^{T} z_t$$

- Finally

$$z_t = \sum_{i}^{m} D_{t,i} \; e^{-\alpha_t y_i h_t(x_i)}$$

$$\overline{i=1}$$

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

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

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

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

$$= 2\sqrt{\left(\tfrac{1}{2}-\gamma_t\right)\left(\tfrac{1}{2}+\gamma_t\right)}$$

$$= \sqrt{1-4\gamma_t^2}$$

$$\leq e^{-2\gamma_t^2} \qquad (1-x \leq e^{-x})$$

■

- Suppose that all predictors outputted by the weak learning algorithm lie in some class $H \subseteq \{+1,-1\}^X$.

- The classifier output by AdaBoost is of the form

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

- Thus it lies in class

$$H_T = \left\{ \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x) \right) : h_1, \ldots, h_T \in H \right\}$$

- In order to analyze generalization error of AdaBoost we need to upper bound

$$\Pi_{H_T}(m) \quad \left(\text{or } VC(H_T) \right)$$

<u>Lemma:</u> Let $H \subseteq \{+1, -1\}^X$ and $d = VC(H_T)$. Let $m \geq \max\{T, d\}$

Then

$$\Pi_{H_T}(m) \leq \left(\frac{em}{T} \right)^T \left(\frac{em}{d} \right)^{dT}$$

<u>Proof:</u>

- Fix a sample S of size m.
- We need to show

$$\left|\Pi_{H_T}(S)\right| \leq \left(\frac{em}{T}\right)^T \left(\frac{em}{d}\right)^{dT}$$

- We know

$$\left|\Pi_H(S)\right| \leq \left(\frac{em}{d}\right)^d$$

- We can thus consider representatives $H' \subseteq H$ where $|H'| \leq \left(\frac{em}{d}\right)^d$ and

$$\Pi_H(S) = \Pi_{H'}(S)$$

- A behavior of $h \in H$ on S is determined by choice of $h_1, \ldots, h_T \in H$) and the coefficients $\alpha_1 \ldots \alpha_T$.
- There are at most

$$\left(\frac{em}{d}\right)^{dT}$$

sequences $h_1, h_2, \ldots, h_T \in H$ of representatives of length T

- Given a sequence $h_1 \ldots h_T \in H$, we can represent any $x \in X$ by the vector

$$\left(h_1(x), h_2(x), \ldots, h_T(x)\right)$$

- A behavior $h \in H$ can be thought of as hyperplane in dimension T operating on $(h_1(x), \ldots, h_T(x))$.
- For a fixed $h_1 \ldots h_T \in H'$, let
$$S' = \{(h_1(x), \ldots, h_T(x)) : x \in S\}$$
- Then there at most
$$\left(\frac{em}{T}\right)^T$$
behaviors of the hyperplanes on S.
- Multiplying $\left(\frac{em}{T}\right)^T$ and $\left(\frac{em}{d}\right)^{dT}$

(i) (ii)

together we get the result. ∎

Lemma: If $VC(H) = d \geq 3$ and $T \geq 3$

$$VC(H_T) \leq T(d+1)\left(3\log(T(d+1)) + 2\right)$$

Proof:

- Consider a set S shattered by H_T of size m.
- Suppose $m > T(d+1)\left(3\log(T(d+1)) + 2\right)$
- Previous lemma implies

$$2^m \leq \left(\frac{em}{T}\right)^T \left(\frac{em}{d}\right)^{dT}$$

- Since $d, T \geq 3 > e$

$$\left(\frac{em}{T}\right)^T \left(\frac{em}{d}\right)^{dT} \leq m^{d(T+1)}$$

- $2^m \leq m^{d(T+1)}$
- This inequality fails for $m > T(d+1)\left(3\log(T(d+1)) + 2\right)$

Lemma: Let $\gamma \in (0, \frac{1}{2})$.

Suppose $\varepsilon_1, \varepsilon_2, \varepsilon_3, \ldots \leq \frac{1}{2} - \gamma$.

If $T > \frac{\ln|S|}{2\gamma^2}$ then

$$\widehat{err}_S(h) = 0.$$

Proof:

- After T rounds

$$\widehat{err}_S(h) \leq e^{-2\gamma^2 T}$$

- If $T > \frac{\ln|\mathcal{H}|}{2\gamma^2}$ then

$$\widehat{err}_S(h) < e^{-\ln|\mathcal{H}|} = \frac{1}{|\mathcal{H}|}$$

- Since $<$ is strict inequality and $\widehat{err}_S(h)$ is an integer multiple of $\frac{1}{|\mathcal{H}|}$,

$$\widehat{err}_S(h) = 0$$

∎

- So AdaBoost implements an

ERM for class H_T