Given $f: X \to \{0,1\}$, and a sample $\{(x_i, y_i)\}_{i=1}^n$

Define

Empirical Loss : $\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(f(x_i) \neq y_i)$

True Loss : $R(f) = \mathbb{E}_{P_{XY}}[\mathbb{1}\{f(x) \neq y\}]$

Definition A hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^X$ is PAC-learnable if there is a function $m_{\mathcal{H}} : [0,1]^2 \to \mathbb{N}$ and a learning algorithm A with the following property: For every $\varepsilon, \delta \in (0,1)$ and every choice of P_{XY}, when running the algorithm on $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ iid examples from $\mathcal{D}$, the algorithm returns $h \in \mathcal{H}$ s.t. w/ probability $> 1-\delta$

$$R(h) \leq \min_{h' \in \mathcal{H}} R(h') + \varepsilon$$

returned by A

best classifier in $\mathcal{H}$

Q: What hypothesis classes $\mathcal{H}$ are PAC-learnable?

- $|H| < \infty$ — PAC learnable

↳ finite

- A : $\hat{h} = \min_{h' \in \mathcal{H}} \hat{R}_S(h')$: ERM

- HW 3 ex 8

size of the class

$$R(\hat{h}) \leq \min_{h' \in \mathcal{H}} R(h) + \sqrt{\frac{\log |H| / \delta}{m}}$$

samples

Sketch ⇝ by Hoeffding

$\forall h \in \mathcal{H}$ $\quad \hat{R}_S(h) - R(h) \leq \sqrt{\frac{\log(1/\delta)}{m}}$

$$\rightarrow \sqrt{\frac{\log(|H|/\delta)}{m}} < \varepsilon$$

$$\rightarrow m_{\mathcal{H}}(\varepsilon, \delta) = \frac{\log(|H|/\delta)}{\varepsilon^2}$$

Q: Is $\mathcal{H} = \mathcal{Y}^{X}_{=\{0,1\}}$, all functions, PAC-learnable?

$\longrightarrow$ NO

No Free Lunch Theorem

input

Consider $m \in \mathbb{N}$, X s.t. $|X| \geq 2m$, any algorithm A which outputs $A(S)$, given $S \subseteq X$, $|S| = m$. Then

sample size s.t.

there exists $f : X \to [0,1]$

w/ probability $\geq \frac{1}{7}$

$$R_{P_f}(A(S)) \geq \frac{1}{8}$$

randomness from sample

Corollary $\mathcal{Y}^X$ is not PAC-Learnable

Given $f : X \to \{0,1\}$ let

$$P_f((x,y)) = \begin{cases} 1/|X| & y = f(x) \\ 0 & y \neq f(x) \end{cases}$$

<u>Proof</u> WLOG $|X| = 2m$.

$|\mathcal{Y}^X| = 2^{2m} \to$ uniform dist on $\mathcal{Y}^X$

<u>Will show</u> ✱

$$\mathbb{E}_{f\sim \mathcal{Y}^X} \mathbb{E}_{S\sim P_f^m} R(A(S)) \geq \frac{1}{4}$$

sample of size m labeled by f

$\to \exists f \in \mathcal{Y}^X$, $\mathbb{E}_{S\sim P_f^m} R(A(S)) \geq \frac{1}{4}$

Markov's inequality

<u>Prove ✱.</u>

$$\mathbb{E}_{f\sim \mathcal{Y}^X} \mathbb{E}_{S\sim P_f^m} \mathbb{E}_{x\sim P_f} \left[\mathbb{1}\{A(S)(x) \neq f(x)\} \right]$$

$$= \mathbb{E}_{S} \; \mathbb{E}_{f \sim y^x} \; \mathbb{E}_{x \sim P_f} \left[\mathbb{1}\{A(S)(x) \neq f(x)\} \right]$$

$$= \cancel{\mathbb{E}_{S} \; \mathbb{E}_{f} \left[\mathbb{1}(A(S)(x) \neq f(x) \mid x \in S \right] \mathbb{P}(x \in S)}$$

$$+ \; \mathbb{E}_{S} \; \mathbb{E}_{f} \left[\mathbb{1}(A(S)(x) \neq f(x) \mid x \notin S \right] \mathbb{P}(x \notin S)$$

$$\geq \frac{1}{2} \cdot \frac{1}{2}$$

$$= \frac{1}{4}$$

• $h \in \mathcal{H}$, if $\mathcal{H}$ is PAC-Learnable

$$\frac{1}{8} \le R(h) = \underbrace{\min_{h' \in \mathcal{H}} R(h')}_{\text{approximation error}} + \underbrace{R(h) - \min_{h' \in \mathcal{H}} R(h')}_{\text{estimation error}}$$

→ w/ enough samples $< \varepsilon$

• Finite Classes

$$R(h) \lesssim \underbrace{\min_{h' \in \mathcal{H}} R(h')}_{\text{bias}} + \sqrt{\frac{\overbrace{\log(|\mathcal{H}|/\delta)}^{\text{Complexity}}}{m}} \to 0$$

$m \uparrow$

What if $|\mathcal{H}| = \infty$?

<u>Definition</u> The shattering coefficient of $\mathcal{H}$ is

$$S(\mathcal{H}, n) = \max_{x_1, \dots, x_n} \left| \{ (f(x_1), \dots f(x_n)) : f \in \mathcal{H} \} \right|$$

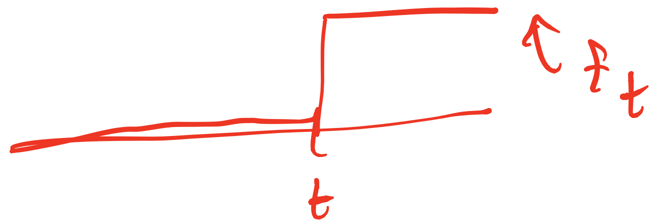
bounded above by 2^n

<u>Definition</u> The VC-dim of $\mathcal{H}$ is the maximum $k \in \mathbb{N}$ s.t.

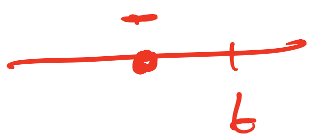
$$S(\mathcal{H}, k) = 2^k$$

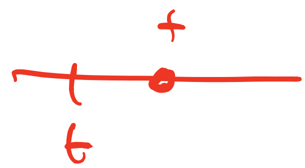
$\rightarrow$ get all possible labelings of k points.

Ex 1 $\mathcal{H}_1 = \{ f_t = \mathbb{1}\{x \geq t\},\ x, t \in [0,1] \}$

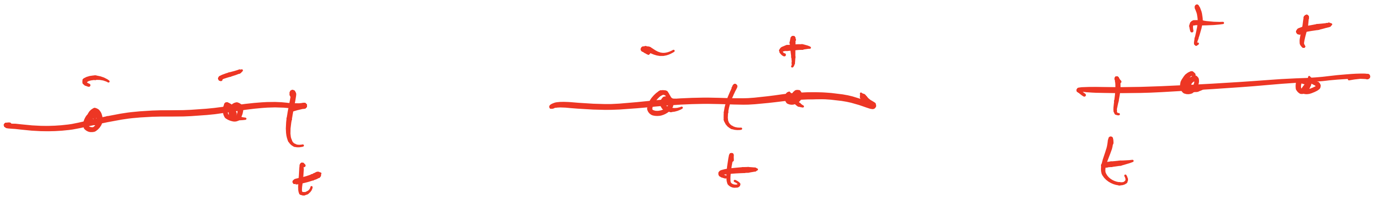


- $n = 1$ $\quad S(\mathcal{H}_1, 1) = 2$





- $n = 2$ $\quad S(\mathcal{H}_1, 2) = 3$



- n points $\quad S(\mathcal{H}_1, n) = n+1$

$$VC(\mathcal{H}_1) = 1$$

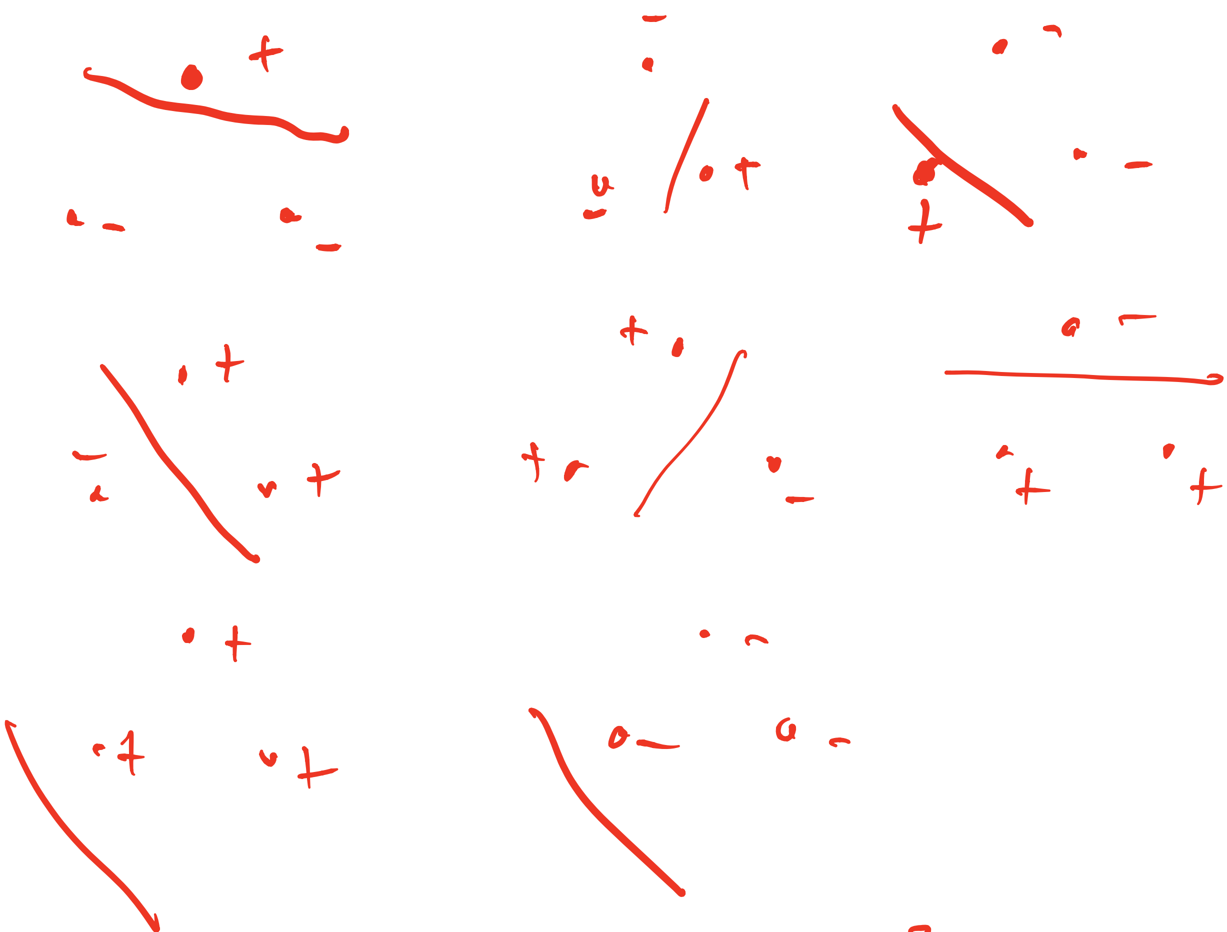
Ex 2 $X = \mathbb{R}^2$

$H_2 = \{$ hyperplane classifiers $\}$

$n = 1$

$S(H_2, 1) = 2$

$n = 3$



$S(H_2, 3) = 8 = 2^3$

$\underline{n = 4}$

$\leftarrow$ Can't get this

$$VC(\mathcal{H}_2) = 3$$

$\rightarrow$ In general

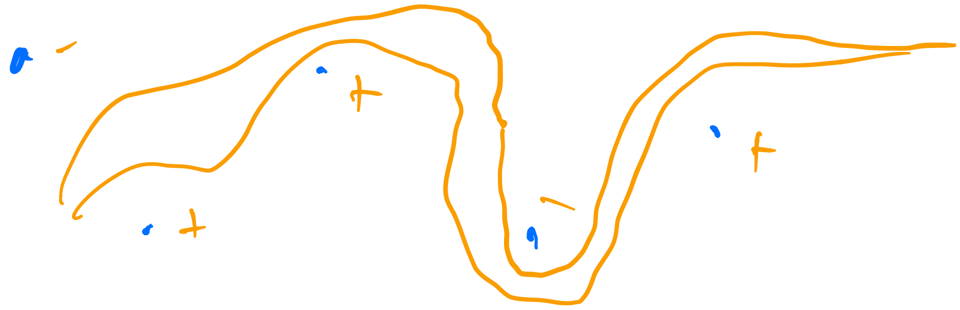
$$\mathcal{H}_d = \{\text{hyperplanes in } \mathbb{R}^d\}$$

$$VC(\mathcal{H}_d) = d + 1, \quad d \geq 2$$

Sauer's Lemma

$$S(\mathcal{H}, n) \leq (n+1)^{VC(\mathcal{H})}$$

$\rightarrow$ Effective # of classifiers on n points is same as # labelings



<u>Theorem</u> Vapnik–Chervonenkis

$$R(h) \leq \min_{h' \in \mathcal{H}} R(h') + \sqrt{\frac{\log (n+1)^{VC(\mathcal{H})} / \delta}{n}}$$

$$= \min_{h' \in \mathcal{H}} R(h') + \sqrt{\frac{VC(\mathcal{H}) \log n/\delta}{n}}$$

$\rightarrow$

$\rightarrow$ hyperplanes in d dimensions

$\rightsquigarrow$ $\dfrac{d \log(1/\varepsilon\delta)}{\varepsilon^2}$ samples

can learn a hyperplane

Fundamental Theorem of Learning

If $\mathcal{H}$ has finite VC dim

$\Downarrow \quad \Uparrow$

PAC Learnable