THEORY OF MACHINE LEARNING

LECTURE 6

VC DIMENSION, FUNDAMENTAL THEOREM

LAST LECTURE

Coreset: strong coresets.

- Representative sample: for a hyp class H and distribution D over X, S is called "representative" if

  training error

  for all $h \in H$, | (avg error on S)(h) - risk_D (h) | $\leq \epsilon$ test error.

- **Observation.** If training data happened to be a representative sample, ERM gives a hypothesis with good generalization. (you will be close to the risk of the best $h \in \mathcal{H}$.) (Question of ERM being efficient is orthogonal…)

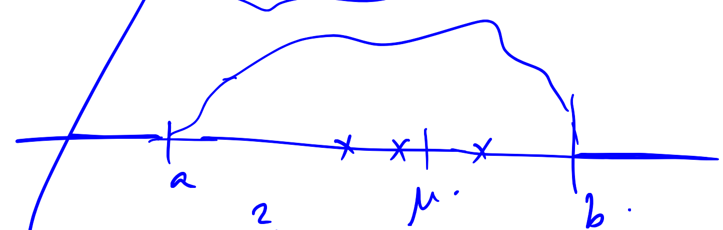
- **Observation 2.** For a finite hypothesis class, a *random* sample of size ~ log |H| is representative

- Proof using "concentration" inequality (Chernoff/Hoeffding)

CONCENTRATION BOUND

- **Chernoff bound (Hoeffding).** Suppose $X_1, X_2, \ldots X_n$ are n iid samples from a distribution with mean μ and support [a, b]. Then we have

$$\Pr\left[\,\left|\frac{1}{n}(X_1+\cdots+X_n)-\mu\right|>\epsilon\,\right]\le 2\exp\left(-\frac{\epsilon^2 n}{(a-b)^2}\right)$$



- Note: exponential dependence on n

"large deviation bounds"

to make RHS $< \delta \iff e^{-\frac{\varepsilon^2 n}{(a-b)^2}} < \frac{\delta}{2}$

$$\frac{\varepsilon^2 n}{(a-b)^2} > \log\left(\frac{2}{\delta}\right) \iff n > \frac{(a-b)^2}{\varepsilon^2}\cdot\log\left(\frac{2}{\delta}\right).$$

FINITE CLASSES ARE LEARNABLE

$\mathcal{H}$: finite hyp. class : $H = \{h_1, h_2, \ldots, h_M\}$.

- Claim: for any X and distribution D over it, a sample of size $O\left(\frac{1}{\epsilon^2}\log\frac{2|H|}{\delta}\right)$ is representative with prob. at least $1-\delta$

  M

  $\{\forall h \in \mathcal{H},$
  $|\text{error on sample} - \text{error on } \mathcal{D}| \le \epsilon\}$
  with prob. $\ge 1-\delta$.

- Proof: write 'm' for the sample size
  - First look at a single $h \in H$
  - Prob. that |sample error(h) – risk(h)| > ϵ can be viewed as an application of Chernoff bound!

    $[0,1]$

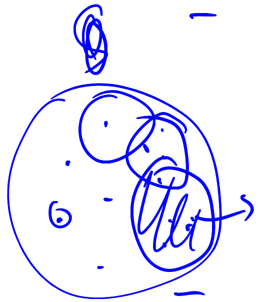
    $X_1 = \begin{cases} 1 & \text{if } h \text{ is incorrect on sample \#1} \\ 0 & \text{if } h \text{ is correct.} \end{cases}$

    $X_2 = \{ \ldots$

  - Gives a bound $2e^{-\epsilon^2 m} < \frac{\delta}{|H|}$
  - Union bound to prove that Pr[diff > ϵ for some h] < δ

- Let us say that S is "bad" for h if

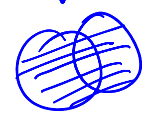
$$|\text{sample error on } S(h) - \text{risk}(h)| > \varepsilon.$$



$$\text{Prob}(S \text{ is bad for some given } h) \leq \frac{\delta}{|H|}$$

(& this is true for ~~each~~ every $h \in \mathcal{H}$)

(when we sample a random S)



A, B.

$$\Rightarrow \text{Prob}(S \text{ is good for ALL } h \in \mathcal{H}) \geq 1 - \delta.$$

$$\Pr(A \vee B) = \Pr(A) + \Pr(B) - \underbrace{\Pr(A \wedge B)}_{\geq 0} \leq \Pr(A) + \Pr(B)$$

$$\Pr(S \text{ is good } \forall h \in \mathcal{H}) = 1 - \underbrace{\Pr(\exists h \in \mathcal{H} \text{ s.t. } S \text{ is bad for } h)}.$$

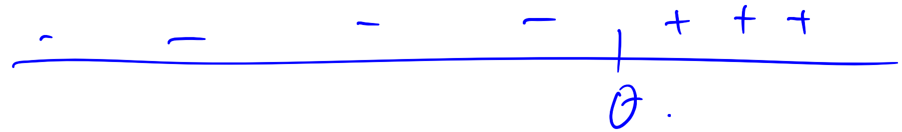
<u>Want to claim:</u> $\Pr[\exists h \in \mathcal{H} \text{ s.t. } S \text{ is bad for } h] \leq \delta$.

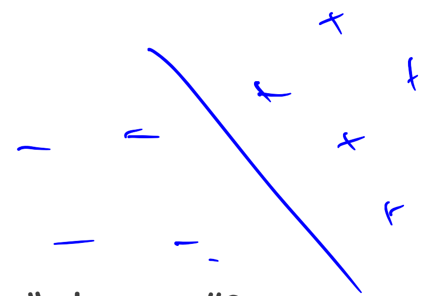
$$/\!/$$

$$\Pr(S \text{ is bad for } h_1 \vee S \text{ is bad for } h_2 \vee \dots \vee S \text{ is bad for } h_M)$$

$$\leq \Pr(\underbrace{S \text{ is bad for } h_1}_{< \delta/|H|}) + \Pr(S \text{ is bad for } h_2) + \dots \leq \delta.$$

WHAT ABOUT INFINITE CLASSES?

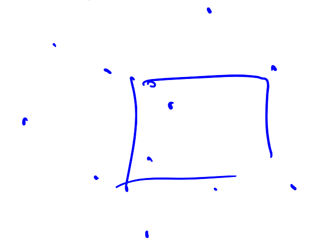




- Note: if sample is representative, we are good!

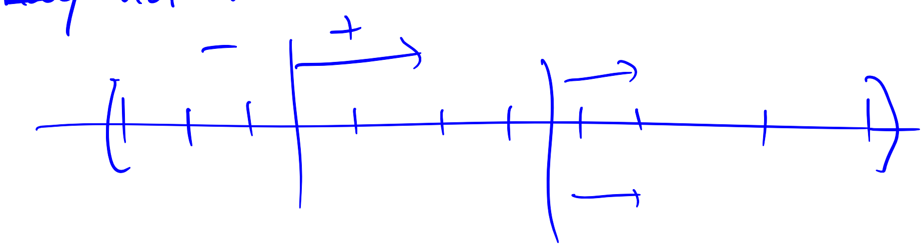
  (modulo inefficiency of ERM)

- What if we can divide hypotheses into finitely many "classes"?



- Example of threshold functions on a line

→ If $|X|$ is finite, then the fact that there are infinitely many ~~classes~~ hypotheses ~~may~~ does not matter!



GROWTH FUNCTION OF A CLASS

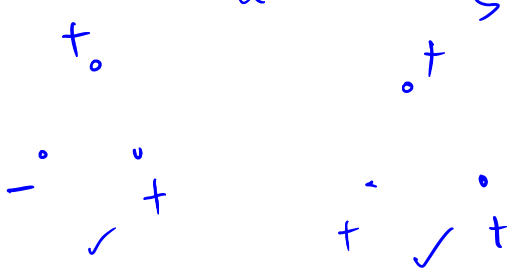
Obsn: What matters isn't the # of distinct hypotheses, it's the # of ways in which the hypotheses classify points of the domain.

- Maximum number of "possible classifications" of an input of size m

Growth function of a hypothesis class $\mathcal{H}$, over domain X:

$$\tau_{\mathcal{H}}(m) := \max_{\substack{|S|=m \\ S \subseteq X}} \left\{ \begin{array}{l} \text{\# of distinct ways in which} \\ \text{hypotheses in } \mathcal{H} \text{ classify } S. \end{array} \right\}.$$

$$\leq 2^m$$



Two classifications are distinct if they differ on even one example.

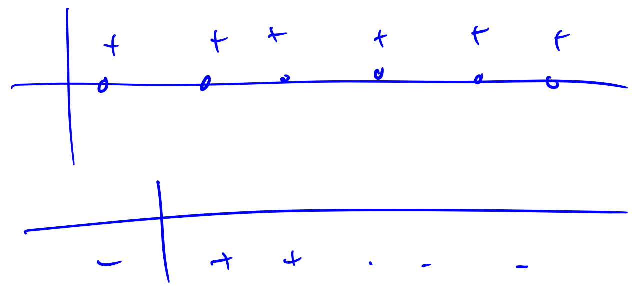
$S = (x_1, x_2, \ldots, x_m)$

$h(S) = (+, -, -, \ldots) \rightarrow$ sign pattern.

Qn: What is the growth function of LTFs, over $X = \mathbb{R}$?

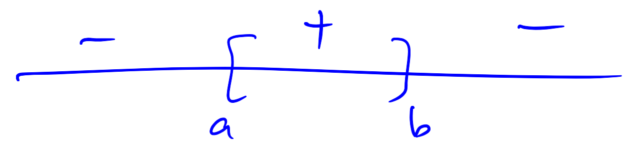
$\mathcal{H}$

$$\tau_{\mathcal{H}}(m) = \max_{\substack{|S| = m \\ S \subseteq \mathbb{R}}} \{\#\text{ distinct ways in which } S \text{ is classified by hypotheses in } \mathcal{H}\}$$



$$\tau_{\mathcal{H}}(m) = m+1.$$

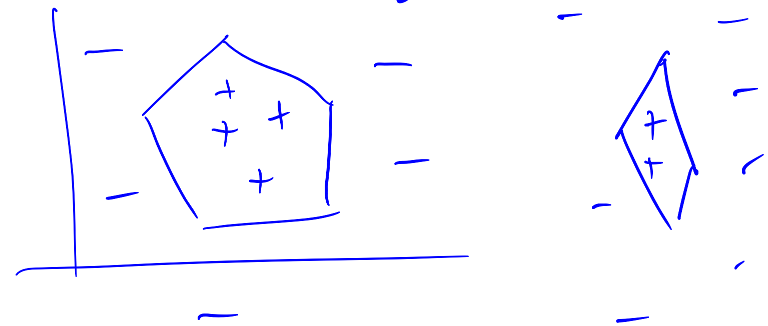
Qn: What is the growth function of intervals on the $\mathbb{R}$ line?



Simple exercise: $\tau_{\mathcal{H}}(m) \leq O(m^2)$.

Qn: Consider $\mathcal{H} = \{ \text{convex polygons in } \mathbb{R}^2 \}$.

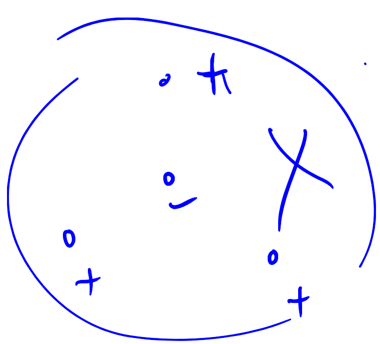
What is $\tau_{\mathcal{H}}(m)$?

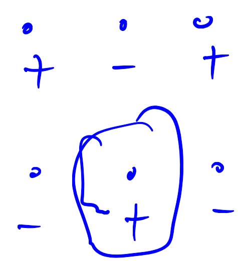


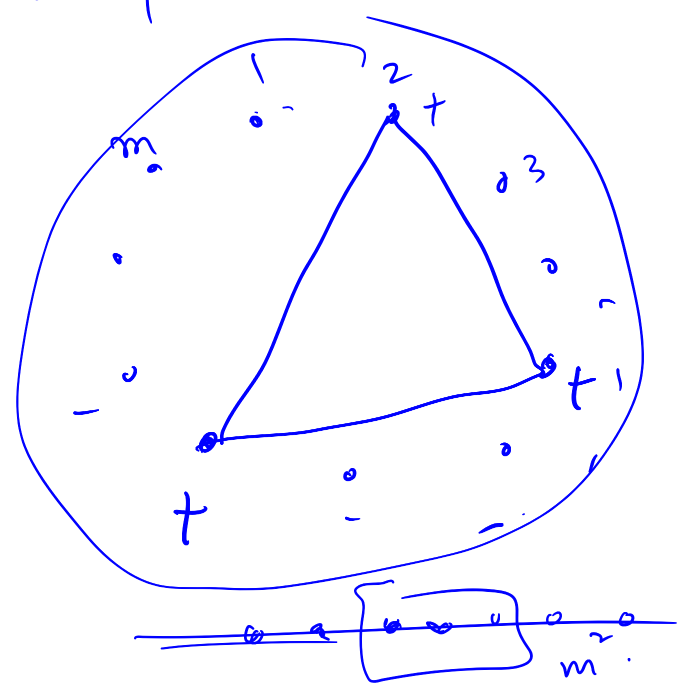
2^m?

To show $\tau_{\mathcal{H}}(m) = 2^m$, you must show one set of m pts s.t. all sign patterns on those points are possible.

$m=4$







$\tau_H(m) = 2^m$.

$S \subseteq \{m\}$

LEARNABILITY IN TERMS OF THE GROWTH FUNCTION

- finite hyp classes: $O(\log |H|)$ samples suffice for PAC-learning

- Theorem: Suppose $\tau_H(m)$ is an upper bound on the total number of distinct "classifications" (or "sign patterns") possible for any sample of size *m*. Then for any X, D, if we take a sample S of size *m*, we have, with prob. 1-δ,

$$\sup_{h \in H} |err(h, S) - \underbrace{err(h, D)}_{risk_D(h)}| \leq \boxed{\frac{4 + \sqrt{\log \tau_H(2m)}}{\delta\sqrt{2m}}}$$

how big should m be so that: $\dfrac{4 + \sqrt{\log(2m+1)}}{\delta \cdot \sqrt{2m}} < \varepsilon$

say $m = \dfrac{4}{\varepsilon^2 \delta^2} \cdot \log^2\left(\dfrac{1}{\varepsilon\delta}\right)$

HOW TO BOUND GROWTH FUNCTION?

- Shattering.

- VC dimension.

SAUER-SHELAH LEMMA (VAPNIK-CHERVONENKIS)

- **Lemma.** Let H be a hypothesis class of finite VC dimension d. Then for every *m*, we have:

$$\tau_H(m) \leq \binom{m}{0} + \binom{m}{1} + \cdots + \binom{m}{d}$$

- Much better than exponential, for m large
- Proof by a clever inductive argument

FUNDAMENTAL THEOREM OF (STAT) LEARNING THEORY

- Theorem: The following statements are *equivalent:*
  - Class H is PAC learnable
  - Class H is *agnostically* PAC learnable
  - Class H has finite VC dimension

- Implies that if H has infinite VC dimension, it is **not** PAC learnable! (same proof as no-free-lunch theorem)

LEARNABILITY IN TERMS OF THE GROWTH FUNCTION

- Theorem: Suppose $\tau_H(m)$ is an upper bound on the total number of distinct "classifications" (or "sign patterns") possible for any sample of size m. Then for any X, D, if we take a sample S of size m, we have, with prob. 1-δ,

$$\sup_{h \in H} |err(h,S) - err(h,D)| \leq \frac{4 + \sqrt{\log \tau_H(2m)}}{\delta \sqrt{2m}}$$

OUTLINE -- THE TWO SAMPLE TRICK

- Want to show that a random sample is ϵ-representative
- Take sample S, define event:

  A = Pr [sample is not representative]
- Way to "test" if S is not representative?
  - "Cross validation"
- Define new event S, S'
- "Swapping"