THEORY OF MACHINE LEARNING

LECTURE 7

VC DIMENSION, FUNDAMENTAL THEOREM (CONTD.)

LAST WEEK

PAC: Probably Approximately Correct.

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

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

- **Observation.** If training data happened to be a representative sample, ERM is an agnostic learning algorithm

  If S is representative, h automatically generalizes.

- **Observation 2.** For a finite hypothesis class, a *random* sample of size ~ log |H| is representative (for any distribution $\mathcal{D}$ over a potentially infinite X.)

  w.h.p. ($\geq 1-\delta$)

  $\frac{\log(|H|/\delta)}{\varepsilon^2}$

- Chernoff + Union bound

WHAT ABOUT INFINITE CLASSES?

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

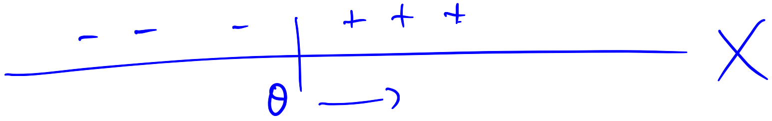
  - How do you prove that a "small" sample is ε-rep. when $\mathcal{H}$ is infinite?

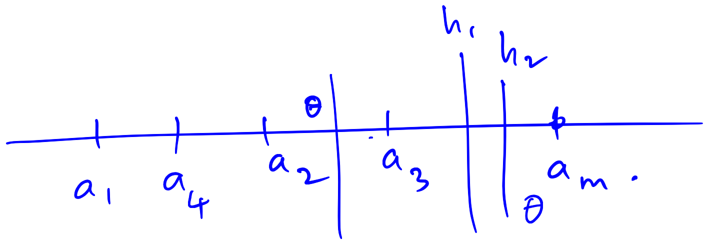
- "Growth function" of a class: total number of distinct ways in with H can label a set of m points (# distinct "sign patterns")

  m

  $|S| = m$

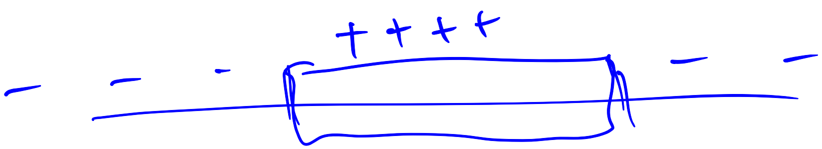
- H is the class of l.t.f. in 1D: (m+1)
- H is the class of intervals in 1D: O(m^2)
- H is the class of axis-parallel rectangles? messy, but O(m^4)
- H is the class of "convex polygons" in 2D: 2^m (exponential)

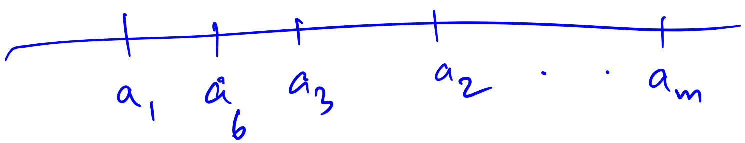




$\tau_H(m)$: max # of distinct classifications for a set S of size m

$(m+1)$

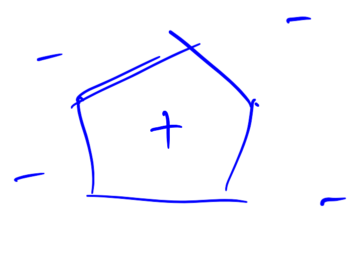


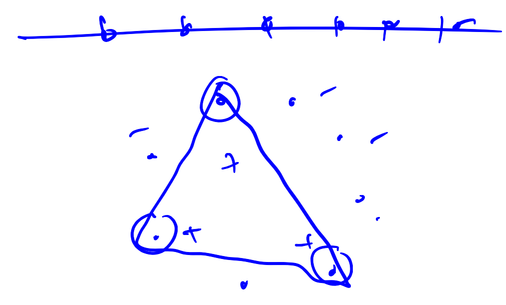


$\sim O(m^2)$ distinct classifications.

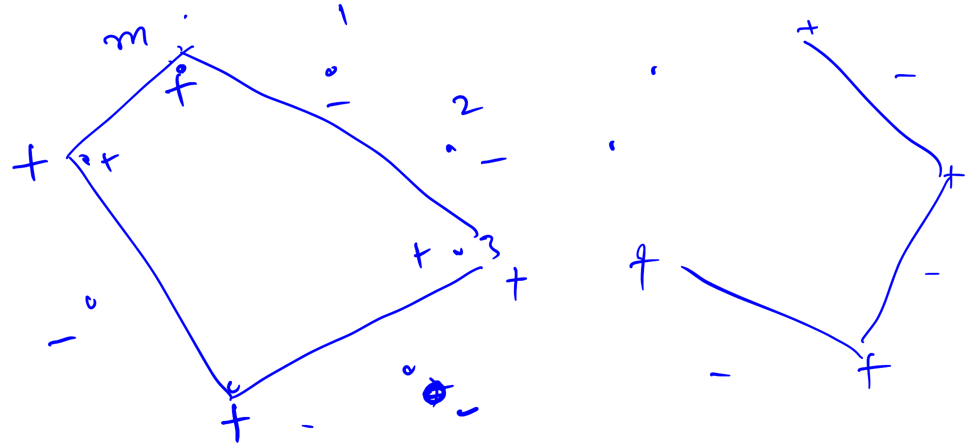
Convex polygons in 2D:

$\tau_{\mathcal{H}}(m) = 2^m$.





given any potential classification, we can obtain it via a ~~convex~~ hypothesis in ~~our~~ the class of convex polygons.



$$\tau_{\mathcal{H}}(m) = 2^m.$$

<u>Very rough statement:</u> if $\mathcal{H}$ has the property that any $h \in \mathcal{H}$ is determined by $\sim d$ "parameters", then the growth fn $\sim n^d$.

TODAY

- "Small" growth function => hypothesis class is learnable!

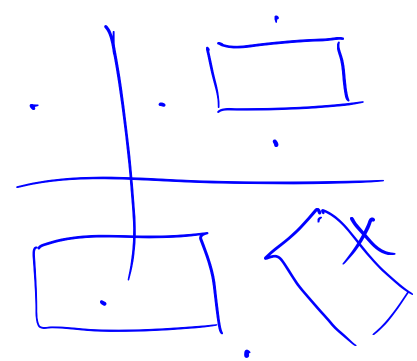
  (a "small" sample is ε-representative whp) .

- How to bound growth function? VC dimension & Sauer's lemma (attributed to Sauer and Shelah, VC)

  if VC dim = d, growth function $\leq O(n^d)$

- defn of ε-rep was "deterministic".

$\mathcal{H}$: axis parallel rectangles



LEARNABILITY IN TERMS OF THE GROWTH FUNCTION

- Theorem: Suppose $\tau_H(m)$ be the growth function of a hypothesis class H. Then for any X, D, if we take a sample S of size *m*, with prob. 1-δ,

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

- In other words, if 'm' is chosen so that RHS < ϵ, theorem implies that random sample of size m is ϵ-representative

Recall: defn of ε-rep: had an ε on the RHS. So if we choose m s.t. $\dfrac{4+\sqrt{\log \tau_H(2m)}}{\delta\sqrt{2m}} \le \varepsilon$, then Theorem implies that a random sample of size m is ε-rep; with prob. $\geq 1-\delta$.

EXAMPLES

Want: $\boxed{\dfrac{4+\sqrt{\log(2m+1)}}{\delta\cdot\sqrt{2m}}} \le \varepsilon$

$\dfrac{\log l+1}{\varepsilon^2}$.

Suffices to set $m = \boxed{\dfrac{4}{(\varepsilon\delta)^2}\cdot\log\left(\dfrac{1}{\varepsilon\delta}\right)}$.

$\boxed{m \sim \dfrac{O_\delta(1)}{\varepsilon^2}\cdot\log\left(\dfrac{1}{\varepsilon}\right)}$

- H is the class of l.t.f. in 1D: (m+1)
- H is the class of intervals in 1D: O(m^2)

$\frac{1}{3}$ $\frac{2}{3}$.

- H is the class of axis-parallel rectangles? messy, but O(m^4) → same bound, bigger constant.
- H is the class of "convex polygons" in 2D: 2^m (exponential)

If growth fn $\le m^{d}$,

$\sqrt{\log(2m)^4} = \sqrt{4\log(2m)}$

Obs 1: $\longrightarrow$ If $\tau_H(m) \le O(m^d)$, then to make

$$\frac{4 + \sqrt{\log(\tau_H(2m))}}{\delta \cdot \sqrt{2m}} \le \varepsilon,$$

we just need to set

$$m \geq \frac{16 \cdot d}{(\varepsilon\delta)^2} \log\left(\frac{d}{\varepsilon\delta}\right)$$

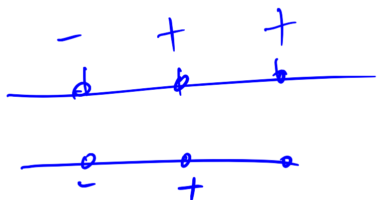
Obs 2: $\longrightarrow$ If $\tau_H(m) = (1.5)^m$, then ~~say we~~ the condition $2^{\sqrt{m}}$

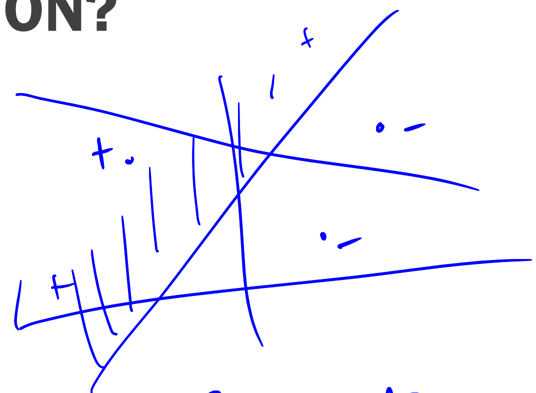
$$\frac{4 + \sqrt{\log(\tau_H(2m))}}{\delta\sqrt{2m}} \le \varepsilon \iff \frac{4 + \sqrt{(2m)\log(1.5)}}{\delta \cdot \sqrt{2m}} \le \varepsilon .$$

$$LHS \simeq \frac{1}{\delta}.$$

If growth fn is exp., then a sample cannot be guaranteed to be representative whp.

HOW TO BOUND GROWTH FUNCTION?



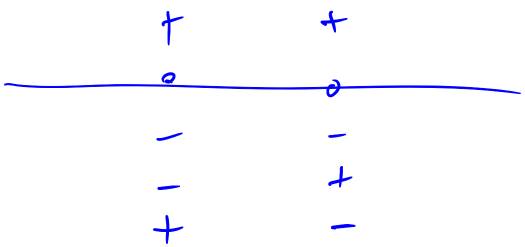


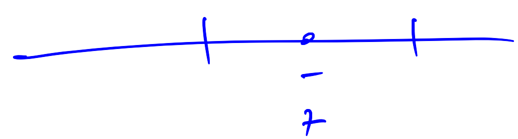
- **Shattering & VC dimension.**

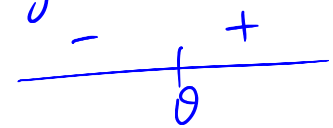
Shattering: given a hyp. class $\mathcal{H}$ and a set S of points, we say that $\mathcal{H}$ shatters S if every possible classification on S can be achieved by some $h \in \mathcal{H}$.

VC-dimension: VC-dimension of a class $\mathcal{H}$ of functions over X is ~~the~~ defined as the size of the largest set $S \subseteq X$ that is shattered by $\mathcal{H}$.

LTF:





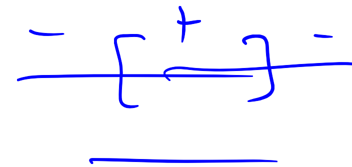


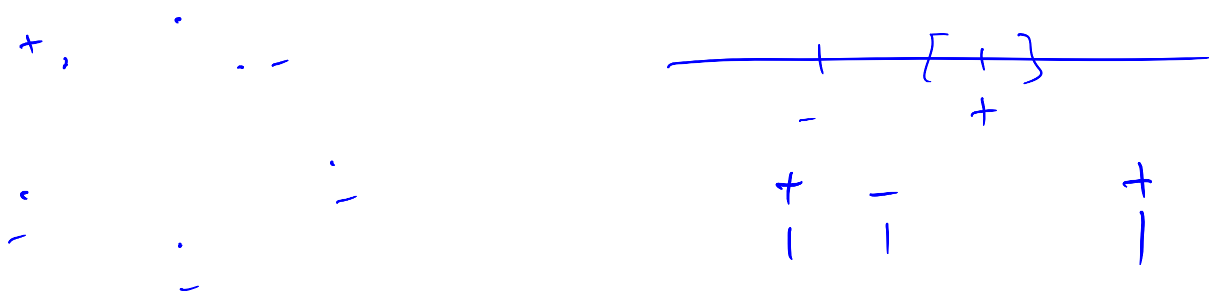
What about $\mathcal{H} := \{$ intervals on the line (where label is + inside & − outside $\}$.

in order to show VC-dim($\mathcal{H}$) = d

✓ 1) ~~[illegible]~~ ∃ a set of size d ~~that~~ can be shattered → $d = 2$

✓ 2) No set of size $(d+1)$ can be shattered! →





No set of 3 points can be shattered...

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 $\tau_H(m)$ be the growth function of a hypothesis class H. Then for any X, D, if we take a sample S of size m, with prob. 1-δ,

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

- In other words, if 'm' is chosen so that RHS < ϵ, theorem implies that random sample of size m is ϵ-representative

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"