THEORY OF MACHINE LEARNING

LECTURE 8

FUNDAMENTAL THEOREM OF STATISTICAL ML, INTRO TO OPTIMIZATION

LAST LECTURE

- 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) | $\leq \epsilon$
- How to show that random sample is representative whp, for an *infinite* hypothesis class (Chernoff + Union bound fails)
- Growth function $\tau_H(m)$; small growth function => random sample is representative

  $\max_{\substack{|S|=m \\ S \subseteq X}}$ # distinct ways in which hypotheses in $\mathcal{H}$ classify S.
- Polynomial vs exponential!
- Shattering, VC dimension

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,

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

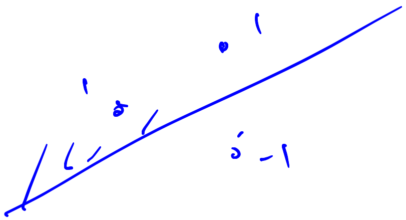
m

- If $\tau_H(m) \approx m^d$ for some parameter *d* then m ~ $\frac{d\log\left(\frac{d}{\epsilon}\right)}{\epsilon^2}$ makes the RHS < ϵ

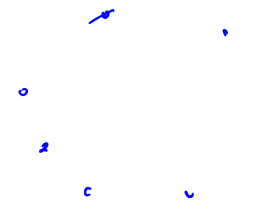
- if $\tau_H(m) = (1.5)^m$

random sample of size $\frac{d\log(d/\epsilon)}{\varepsilon^2}$ $(= m)$ is ε-representative w.p. ~ 0.9.

LAST LECTURE – SHATTERING AND VC DIMENSION

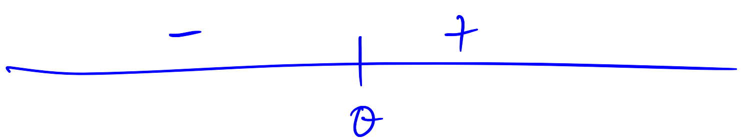


$\mathcal{H} = \{h : X \to \{\pm 1\}\}$; $S \subseteq X$.



- A hypothesis class H is said to shatter a set S if all possible classifications (all 2^|S| of them) can be obtained using hypotheses $h \in \mathcal{H}$.
- Intuitively for such a hyp class, giving the labels of a subset of S doesn't give any information about labels of other points!
- VC dimension: is the size of the largest set in X that can be shattered by H

  max { m : ∃ S of size m that is shattered }.
- Examples: VC dimension of 1-D LTFs, etc.



meta heuristic: VC-dim = # parameters used to describe $h \in \mathcal{H}$

SAUER-SHELAH LEMMA (VAPNIK-CHERVONENKIS)

(if VC-dimension $\leq d$, then $\tau_H(m) \leq O(m^d)$).

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

$d = 5$.

$$\tau_H(m) \leq \binom{m}{0} + \binom{m}{1} + \cdots + \binom{m}{d} \quad [\sim m^d].$$

- Much better than exponential, for m large

$(1.5)^m$ grows way faster than m^d.

- Proof by a clever inductive argument

$\binom{n}{k} = 0$ if $n < k$.

$X = \mathbb{R}$.

$\mathcal{H} = \{\text{sign}(p(x)) : p \text{ is a polynomial}\}$

PTF.

FUNDAMENTAL THEOREM OF (STAT) LEARNING THEORY

$X = \mathbb{R}$; $\mathcal{H}$: LTFs $= \{ \text{sign}(x-\theta) ; \theta \in \mathbb{R} \}$.

degree-2 PTFs :

- Theorem: The following statements are *equivalent:*
  - Class H is PAC learnable (recall the (ε, δ) - definition of PAC learning .)
  - Class H is *agnostically* PAC learnable
  - Class H has finite VC dimension

SS Lemma $\Rightarrow$ $\tau_H(m) \le O(m^d)$

Prev. theorem $\Rightarrow$ m-sized sample is ε-rep. w.p. $1-\delta$ if $m > \dfrac{d \cdot \log\left(\frac{d}{\varepsilon\delta}\right)}{(\varepsilon\delta)^2}$.

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

[Note: to prove that VC-dim is infinite, you show that for any $m \in \mathbb{N}$, $\exists$ an S of size m that can be shattered. } .

$(x, \underline{h}(x))$

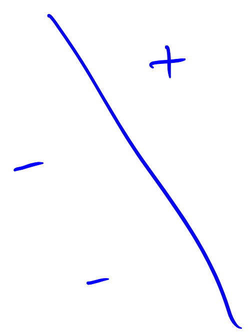
$\mathcal{H}$ is a given hyp. class.

PAC-learning ("realizable"): if true labeling function is some $h \in \mathcal{H}$, then we can find h' s.t. risk$(h') \leq \varepsilon$.

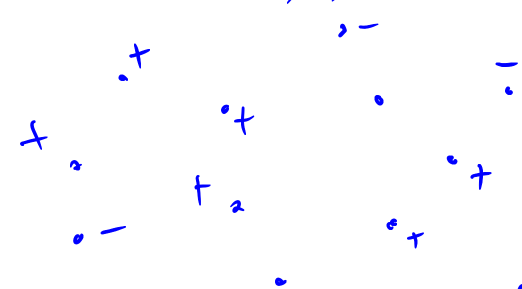
$\mathcal{H}$: give hyp. class.

PAC-learning (agnostic): ~~we~~ for any true label function f, we can find h' s.t. risk$(h') \leq \min_{h \in \mathcal{H}} \text{risk}(h) + \varepsilon$.

→ $\mathcal{H}$: linear separators in 2D.



$(x_1, f(x_1))$, $(x_2, f(x_2))$, ...



PAC-learnability of $\mathcal{H}$ implies that if we could solve ERM using $\frac{\log \frac{1}{\varepsilon}}{\varepsilon^2}$ pts, then error will be at most ε worse than best separator

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 – homework)

SOME IMPLICATIONS

- If H has infinite VC dimension, it is **not** PAC learnable! (same proof as no-free-lunch theorem – homework)
- ERM is all you need, assuming you have enough samples (proof of the theorem implies this.)
  - Doing ERM efficiently is a challenge (next section)
- Agnostic case usually as hard *as realizable case* (in terms of sample complexity.)
- **Caveat**. Learnability guarantees only apply to ERM, not (say) to an improper learner

if you perform "learning" and obtain h with 0 training error

$\mathcal{H}$ ..

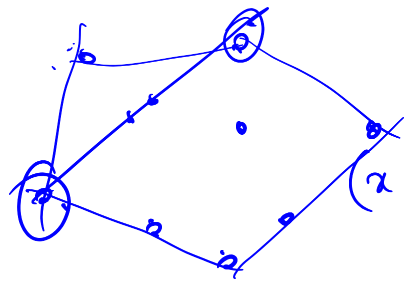
→ Most opt methods are not guaranteed to find optima. (because the ERM problem is NP-hard.).

→ Some settings in which they do. (convergence rates, etc.).

OPTIMIZATION

HOW TO SOLVE ERM EFFICIENTLY?

BASICS

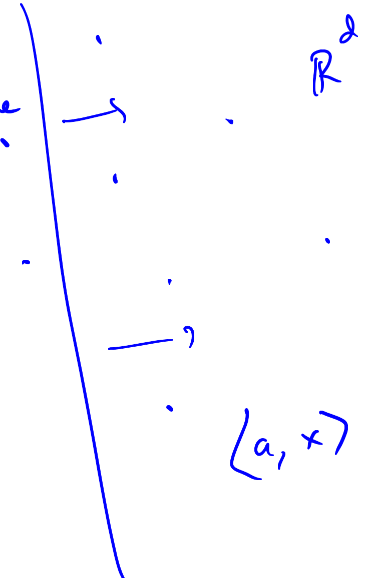


$x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ all points of the form $\sum_{i=1}^{n} \lambda_i x_i$, where

$(x, f(x))$

$x \in \mathbb{R}^d$.

(feature vector)



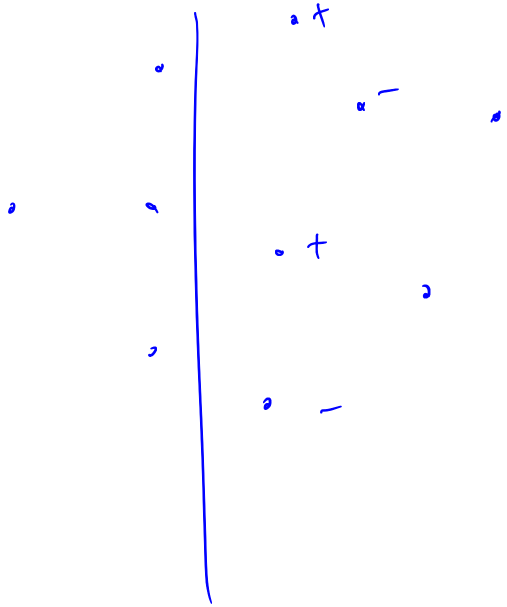
- **Linear classification**
- Linear classification – non realizable
  ↳ NP-hard.
- Loss functions ⟶ proxy for ERM.
- Convexity and convex optimization

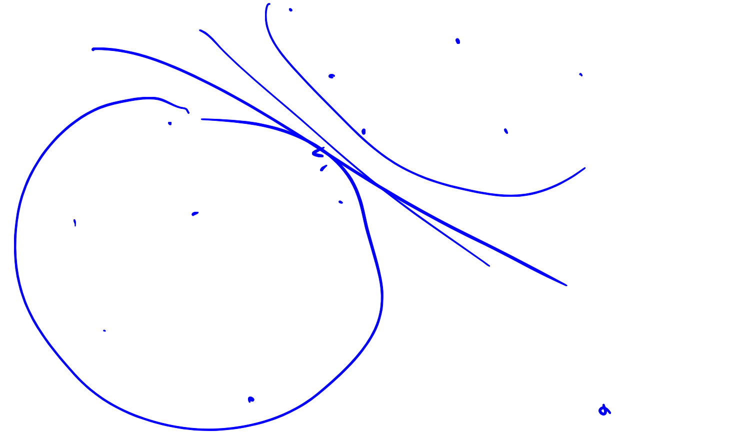
$\mathcal{H} = \left\{ h \text{ of the form } \operatorname{sign}(\langle a, x \rangle + b) \text{ for some } a \in \mathbb{R}^d, b \in \mathbb{R}. \right\}$

$x = (x_1, x_2, \ldots, x_d)$

$\text{VC-dim}(\mathcal{H}) = d+1$

→ VC theory tells us: $\sim \frac{d}{\varepsilon^2}$ examples, then we can find the "best linear classifier" for any $\mathcal{D}$, and any ground truth f.

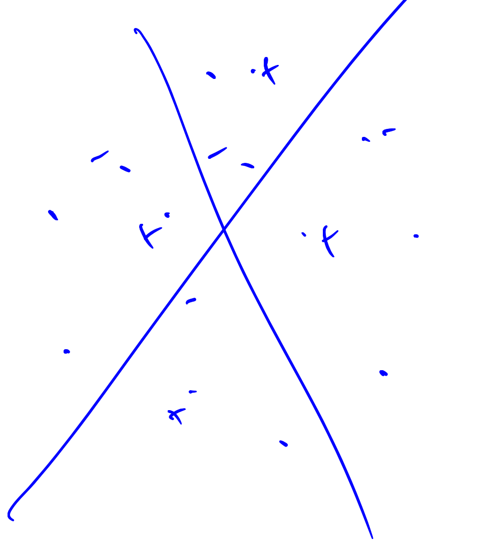




n $\lfloor \frac{d}{2} \rfloor$

n

Linear programming



label +

$\langle a, x^{(1)} \rangle + b \geq 0$

$\langle a, x^{(2)} \rangle + b < 0$ $\rightarrow$ label of $x^{(2)} < 0$.

$\vdots$

OPT error $= \varepsilon$

NP-hard to achieve error $< \frac{1}{2} - \delta$.