THEORY OF MACHINE LEARNING

LECTURE 5

PAC MODEL, VC DIMENSION

LAST WEEK

Goal of statistical learning: given some iid examples from $\mathcal{D}$ + labels, learn a good h.

Unknown distn $\mathcal{D}$

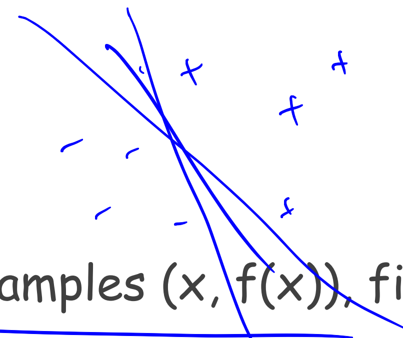
h ; $risk_{\mathcal{D}}(h) = \Pr_{x \sim \mathcal{D}}[h(x) \neq label(x)]$.

- Motivation: do we really need to restrict the hypothesis/concept class *before starting* learning? – yes! (No free lunch theorem)
- (PAC Learning): Learnability of a concept class H over domain X
  - Informally, for any $f \in H$ and any distribution D over X, given examples of the form (x, f(x)), with $x \sim \mathcal{D}$ we can learn a hypothesis 'h' such that Risk_D (h) is < ϵ, with high prob. ($(1-\delta)$, for some parameter δ)
  - Sample size only function of H, ϵ, δ (not distribution) $\mathcal{D}$.
  - Learned hypothesis need not belong to H (improper learning)

(realizable setting).

- (Agnostic): f need not belong to H → want the risk to be $\leq \min_{h' \in \mathcal{H}} risk_{\mathcal{D}}(h') + \varepsilon$.

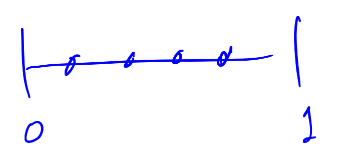
GENERIC ALGORITHM



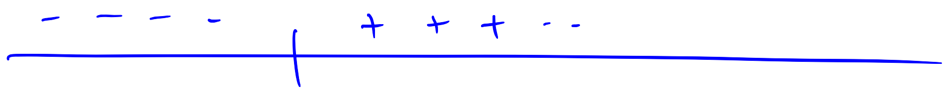
- Empirical risk minimization: given the samples (x, f(x)), find hypothesis $h \in H$ that minimizes the total error on the samples

  S

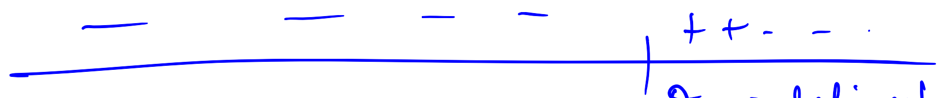
  - most natural algorithm == minimize training error
- How to do it efficiently? (learning half spaces – can do it in the realizable case ...)
  - Don't care for now… maybe brute force over hypothesis class?
- When does it work? (outputs h with small gen. error).
  - If sample is "representative" of distribution --- for *every* hypothesis in class, error on sample ~= error on distribution (i.e. risk)



$\mathcal{H}$: ~~set~~ all linear threshold functions in 1-D.

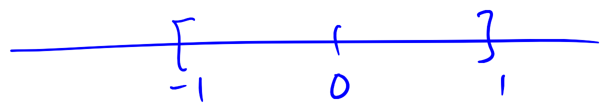


$X = \mathbb{R}$.



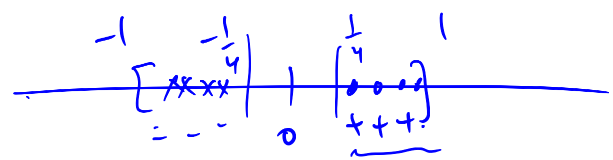
$\theta \to$ defines $h(\cdot)$.

$\mathcal{D}$:



Unif$(-1, 1)$.

goal: learn with error (risk bound) ≤ 0.1



15 samples.

Informal statement: # examples needed depends on both hypothesis class & distribution (also on ε, δ ..).

REPRESENTATIVE SAMPLE

[applies to realizable & agnostic cases)

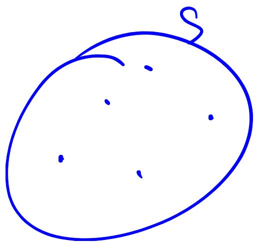
- Let H be a hypothesis class and X be an input space with a distribution D on it, and let f be a target function. Sample $S \subseteq X$ is said to be ϵ –"representative" if **for all** h in H, we have:

$$\boxed{\left| \frac{1}{|S|} \text{error}(S, h) - \text{risk}_D(h, \cancel{f}) \right| < \epsilon}$$

- If we happen to get a representative sample, ~~we have~~ ERM gives the desired bound on risk! (why?)

- Is a sample ε-representative "with high probability"?

Claim: If we perform ERM on a ε-representative sample S, we obtain a hyp. with risk ≤ (best-risk-in-H) + 2ε.



$\longrightarrow$ ERM h $\rightarrow \frac{1}{|S|}\,\text{error}(S,h) \approx_\varepsilon \text{risk}_\mathcal{D}(h) \rightarrow$ I

opt hypothesis $= h^*$, say. (hyp^n with the least risk w.r.t. $\mathcal{D}$.)

$\rightsquigarrow \frac{1}{|S|}\,\text{error}(S,h^*) \approx_\varepsilon \text{risk}_\mathcal{D}(h^*) \rightsquigarrow$ II

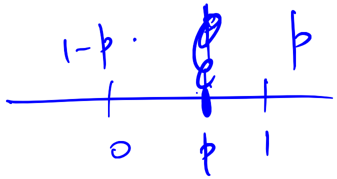
$$\text{error}(S,h) := \sum_{x\in S} \underbrace{\mathbb{1}}_{\text{indicator}}[h(x)\neq \text{label}(x)].$$

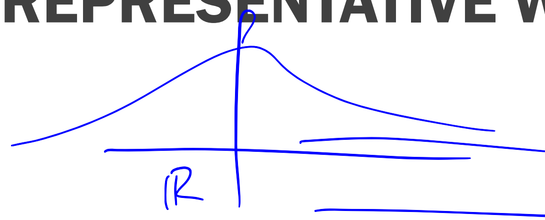
$$\text{risk}_\mathcal{D}(h) \underset{\text{from I}}{\le} \frac{1}{|S|}\cdot\text{error}(S,h)+\varepsilon \underset{\text{because of ERM}}{\le} \frac{1}{|S|}\cdot\text{error}(S,h^*)+\varepsilon$$

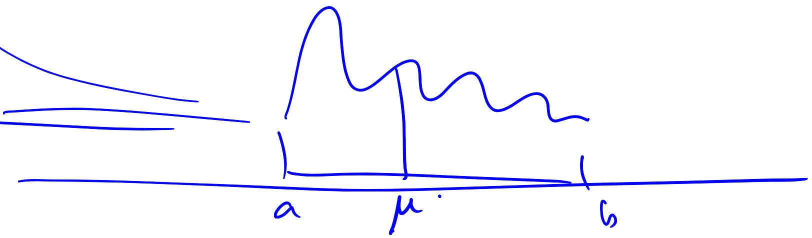
$$\underset{\text{from II}}{\le} \underbrace{\text{risk}_\mathcal{D}(h^*)}_{0} + \overbrace{(\varepsilon+\varepsilon)}^{2\varepsilon}.$$

(in the realizable case, you only get ε.).

RANDOM SAMPLE IS REPRESENTATIVE WHP!







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

(Very useful.)

$$\Pr\left[\left|\frac{1}{n}\left(X_1 + \cdots + X_n\right) - \mu\right| > \epsilon\right] \leq 2\exp\left(-\frac{\epsilon^2 n}{(a-b)^2}\right) \longrightarrow \frac{1}{n^2 \epsilon^4}$$

$< \delta$

- <u>Note</u>: exponential dependence on n

$$\Pr\left[\left|\text{sample avg} - \mu\right| > \varepsilon\right] \leq \text{exponentially small in \#samples.}$$

Suppose $(a-b)^2 = 1$

$$n \approx \frac{1}{\varepsilon^2} \cdot \log\left(\frac{2}{\delta}\right)$$

FINITE CLASSES ARE LEARNABLE

ε: accuracy.

δ: prob. parameter or confidence param.

$\mathcal{H}$: finite hypothesis class.

- Claim: for any X and distribution D over it, a sample of size $O\left(\frac{1}{\epsilon^2}\log\frac{|H|}{\delta}\right)$ is representative with prob. at least $1-\delta$
- Proof idea: first start with a single hypothesis $h \in H$; what is the probability that error on sample ~= error on D?

Want to show: $m = \boxed{4 \cdot \frac{1}{\varepsilon^2} \cdot \log\left(\frac{2|H|}{\delta}\right)}$.

A random sample with m examples is ε-representative with prob. $\geq 1-\delta$. $\rightarrow$ $\forall h \in \mathcal{H}$ $\left|\frac{1}{|S|}\text{error}(S,h) - \text{risk}_{\mathcal{D}}(h)\right| \leq \varepsilon$.

Fix some h; Claim: w.p. $\geq 1-\frac{\delta}{|H|}$ $\left|\frac{1}{|S|}\text{error}(S,h) - \text{risk}_{\mathcal{D}}(h)\right| \leq \varepsilon$.

ie, $\Pr\left[\left|\underbrace{\frac{1}{|S|}\text{error}(S,h)}_{\text{sample-based}} - \underbrace{\text{risk}_{\mathcal{D}}(h)}_{\text{distn.}}\right| > \varepsilon\right] \leq \frac{\delta}{|H|}$

h

$x_i \sim \mathcal{D}$.

Let $Y_i = \begin{cases} 1 & \text{if sample } h(x_i) \neq \text{label}(x_i) \\ 0 & \text{if } h(x_i) = \text{label}(x_i) \end{cases}$

$\Pr_{x \sim \mathcal{D}}\{h(x) \neq \text{label}(x)\}$.

$\parallel\!\!|$

$\text{risk}(h)$

$R(h)$

If we sample $x_i \sim \mathcal{D}$, how is Y_i distributed?

$$Y_i = \begin{cases} 1 & \text{w.p. } R(h) \\ 0 & \text{w.p. } 1 - R(h) \end{cases} \Rightarrow \mu \text{ of this distn} = R(h)$$

Chernoff [Hoeffding]: $\Pr\left\{\left|\frac{Y_1 + \dots + Y_m}{m} - \mu\right| > \varepsilon\right\} \leq 2 \cdot e^{-\frac{\varepsilon^2 m}{4}}$

$\leq \delta/|H|$

(plugging in $m = \frac{4}{\varepsilon^2}\log\left(\frac{2|H|}{\delta}\right)$

Union bound:

Want: some event ~~holds~~ occurs with prob. 99%.
$\mathcal{E}$.

$\Updownarrow$

Pr[event does not occur] $\leq 1\%$.

Say event = want A and B to hold. $\mathcal{E}$.

$\mathcal{E}$ not occurring $\equiv$ either A $\underbrace{\text{doesnt occur}}_{\leq \frac{1}{200}}$ or B doesnt $\underbrace{\text{occur}}_{\leq \frac{1}{200}}$.

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

GROWTH FUNCTION OF A CLASS

- For a class H and an input space X, we can define a notion of "growth function"

LEARNABILITY IN TERMS OF THE GROWTH FUNCTION

- Theorem: Suppose $\tau_H(m)$ is an upper bound on the total number of "distinct 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,

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

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: