Chapter 5.

the reason above problem has low sample complexity for PAC learning is that we had a prior knowledge:
- ground truth D has a circular decision boundary
- $\mathcal{H}$ is collection of circular decision boundaries.

Q. is such prior knowledge necessary? yes.

Learning Problem, $\mathcal{X}, \mathcal{Y}$ fixed

NATURE D $\xrightarrow{S}$ LEARNER $\mathcal{H}$, Algo $\Downarrow$ h.

$D \rightarrow$ Goal $\leftarrow h$

Goal: $L_D(h) \downarrow$

Q. Can there be a universal algorithm Algo that is PAC learnable for all problem instances D?

Q. is prior knowledge necessary for PAC learning?
- Parametric family D
- Realizable w.r.t. $\mathcal{H}$ (lec 2)
- $\min_{h \in \mathcal{H}} L_D(h)$ is small (implicitly, lec 3) → should use Agnostic PAC, otherwise.

How much prior knowledge is good to assume?

a lot
- Learner knows D, h^* → $\mathcal{H} = \{h^*\}$ achieves minimum Risk
  e.g. Bayes Optimal
- Learner knows a class
  - lecture 2: realizable finite $\mathcal{H}$ → PAC learnable $m \approx \frac{\lg \frac{|\mathcal{H}|}{\delta}}{\varepsilon}$
  - lecture 3: e.g. $\mathcal{H} = \{h_r : \mathbb{I}_{\{x \le r\}}\}$ → PAC learnable $m \approx \frac{\lg \frac{1}{\delta}}{\varepsilon^2}$
    non-realizable finite $\mathcal{H}$ → Agnostic PAC learnable $m \approx \frac{\lg \frac{|\mathcal{H}|}{\delta}}{\varepsilon^2}$
- Learner knows nothing: no-free-lunch Theorem.

No

No-free-lunch theorem

Theorem. Let Algo be any learning algorithm. $\mathcal{Y} = \{0,1\}$, any $\mathcal{X}$, $m \le \frac{|\mathcal{X}|}{2}$

Then exists a distribution D over $\mathcal{X} \times \mathcal{Y}$ s.t.

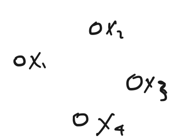
1. $\exists f : \mathcal{X} \to \{0,1\}$ with $L_D(f) = 0$
2. $P_S\left(L_D(\text{Algo}(S)) \ge \frac{1}{8}\right) \ge \frac{1}{7}$

Constructive Proof ⇓ Negative Result.

proof sketch: $|\mathcal{X}| \ge 2m$, $C \subseteq \mathcal{X}$, and $|C| = 2m$, and we observe random S of m samples.

$m = 2$
$|C| = 4$, $|\mathcal{H}| = 16$

$h_0 (0000)$
$h_1 (0001)$
$h_2 (0010)$
$\vdots$



$|\mathcal{H}| = 2^{2m}$ all possible ways to label.
$= \{h_0, h_1, \dots h_{2^{2m}-1}\}$

for any Algo, $\exists D$ s.t. one cannot do better than random guess on $\frac{1}{2}$ of the samples not observed.

$\rightarrow$ $\underset{S}{\mathbb{E}}\left[\underbrace{L_D(A(S))}_{\text{positive R.V}}\right] \ge \frac{1}{4}$ → claim by the Markov ineq.

$\frac{1}{4}$: $\frac{1}{2}$ of samples error $\frac{1}{2}$.

Corollary. $\mathcal{X}$ is infinite & $\mathcal{H}$ is all possible functions from $\mathcal{X}$ to $\{0,1\}$ then $\mathcal{H}$ is not PAC learnable.

Q. How do we choose the right hypothesis class?

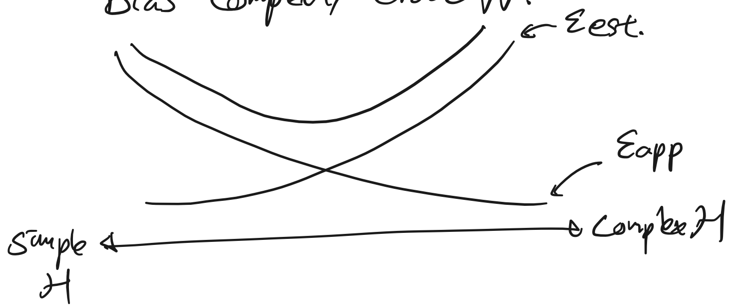
$$\underset{\text{ERM}}{L_D(h_S)} = \underbrace{\left(L_D(h_S) - \min_{h' \in \mathcal{H}} L_D(h')\right)}_{\varepsilon_{est} = \text{Estimation error}} + \underbrace{\min_{h' \in \mathcal{H}} L_D(h')}_{\varepsilon_{app} = \text{approximation}}$$

· goes down with $m \uparrow$ error

· depends on complexity $\mathcal{H}$ · does not depend on m

· property of $\mathcal{D}, \mathcal{H}$. Complex

Bias-Complexity tradeoff.



---

chapter 6.

realizable + finite $\mathcal{H} \longrightarrow$ PAC learnable (lec 2)

finite $\mathcal{H} \longrightarrow$ Agnostic PAC learnable (lec 3).

realizable + infinite $\mathcal{H} = \{h_r : \mathbb{I}_{(\|x\| \le r)}\} \longrightarrow$ PAC learnable (lec 4)

$C \subset \mathcal{X}, |C| \ge 2m$, $\mathcal{H}$ contains all possible 2^{2m} labeling $\longrightarrow$ no learning (lec 4)

Q. what class $\mathcal{H}$ is PAC learnable? Vladimir Vapnik & Alexey Chervonenkis 1970

Definition. (Restriction of $\mathcal{H}$ to C)

$\mathcal{H}$ is class of functions $\mathcal{X} \to \{0,1\}$, and $C = (c_1, \cdots, c_m) \subseteq \mathcal{X}$,

the $\underbrace{\text{Restriction of } \mathcal{H} \text{ to } C}$

$$\mathcal{H}_C \triangleq \left\{ \tilde{h} : \begin{array}{c} C \to \{0,1\} \\ c_i \mapsto h(c_i) \end{array} \;\middle|\; h \in \mathcal{H} \right\}$$

equivalently we can represent $\tilde{h}$ by a vector $(h(c_1), \cdots, h(c_m))$.

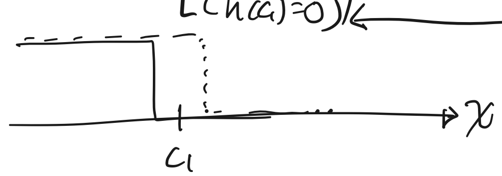
Definition. (Shattering)

A hypothesis class $\mathcal{H}$ shatters a set $C \subseteq \mathcal{X}$ if the restriction of $\mathcal{H}$ to C is the set of all functions from C to $\{0,1\}$. That is $|\mathcal{H}_C| = 2^{|C|}$.

Example. $\mathcal{H} = \{h_a : \mathbb{R} \to \mathbb{R} = \mathbb{I}_{(x \le a)}\}$ class of all threshold functions.

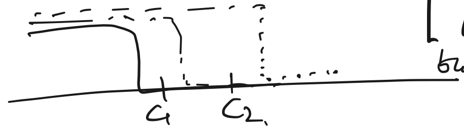
① $C = \{c_1\}$

$$\mathcal{H}_C = \begin{cases} (h(c_1) = 1) & \longleftarrow h_{c_1+\beta} \\ (h(c_1) = 0) & \longleftarrow h_{c_1-\beta} \end{cases}$$



$\therefore$ $\mathcal{H}$ shatters $C = \{c_1\}$

② $C = \{c_1, c_2\}$, $\mathcal{H}_C = \begin{cases} (0,0) & \longleftarrow h_{c_1-\beta}(\cdot) \\ (1,0) & \longleftarrow h_{\frac{c_1+c_2}{2}}(\cdot) \end{cases}$



$[\,(1,0) \leftarrow h_{\frac{c_1+c_2}{2}}(\cdot)$
but not $(0,1)$

$\therefore$ $\mathcal{H}$ does not shatter $C=\{c_1, c_2\}$

Corollary (no-free-lunch thm restated)

for some $\mathcal{H}$, $\mathcal{X}$, $\mathcal{Y}=\{0,1\}$, if $\exists C \subseteq \mathcal{X}$, $|C|=2m$ & shattered by $\mathcal{H}$,

then for any learning algorithm, $\exists \mathcal{D}$ & $h \in \mathcal{H}$ s.t.

$L_{\mathcal{D}}(h)=0$ but $\mathbb{P}(L_{\mathcal{D}}(A(S)) \geq \frac{1}{8}) \leq \frac{1}{7}$.

Definition (VC-dimension)

The VC-dimension of $\mathcal{H}$, $\triangleq \mathrm{VCdim}(\mathcal{H})$, is the maximal size of $C \subseteq \mathcal{X}$ that can be shattered by $\mathcal{H}$.

→ no set of size $|C|+1$ can be shattered.

Example. need to show that $\exists C$ of $|C|=d$ shattered say,

& $\forall C$ of $|C|=d+1$, not shattered.

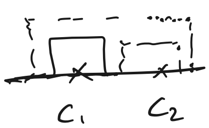
Example (Threshold: $\mathcal{H}=\{\mathbb{1}(x<a)\}$)

$\exists$ $C=\{c_1\}$ can be shattered.

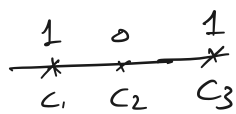
all $C=\{c_1,c_2\}$ can not be shattered.

$\mathrm{VCdim}(\mathcal{H})=1$.

Example (Intervals: $\mathcal{H}=\{\mathbb{1}(a<x<b): a\leq b \in \mathbb{R}\}$)



existence of shattered set

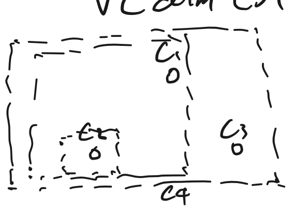


all set not shattered.

$\mathrm{VCdim}(\mathcal{H})=2$.

Example (axis aligned rectangles: $\mathcal{H}=\{\mathbb{1}(a_1\leq x_1\leq b_1, a_2\leq x_2\leq b_2): a_1\leq b_1, a_2\leq b_2\}$)

$\mathrm{VCdim}(\mathcal{H})=4$



existence of shattered set.

all set of 5 not shattered.

Example (finite classes: $|\mathcal{H}|<\infty$)

shattering requires at least $2^{|C|}$ hypothesis. so

$$2^{|C|} \le |\mathcal{H}|$$

$\rightarrow \quad |C| \le \log_2 |\mathcal{H}|$

$\rightarrow \quad \text{VCdim} \le \log_2 |\mathcal{H}|$ → tight in the worst case
→ VCdim can be much smaller.
e.g. finite & threshold.

Example (number of parameters).
above examples have # parameters match VCdim,
Not always true: $\mathcal{H} = \{h_\theta : \theta \in \mathbb{R}\}$, $\text{VCdim}(\mathcal{H}) = \infty$ [HW1 Problem 4]

$$h_\theta = \lceil 0.5 \sin(\theta x) \rceil$$