Learning Infinite Classes I.

In this lecture and the next one we explore PAC learning when the hypothesis class has infinite cardinality. We will ultimately prove the "fundamental theorem of learning" next time and see that one can learn infinite hyp classes as long as the "VC dimension" is finite. The notion of VC dimension is introduced in next lecture but this one prepares the ground.

We start with a recap of what has been done till now:

- Model of PAC learning and definition
- Uniform convergence property.
- Learning finite classes with ERM.
- No free lunch thm.

Notations:

$\mathcal{X}$ domain set $x \in \mathcal{X}$.

$\mathcal{Y}$ label set $y \in \mathcal{Y}$, $\underbrace{\mathcal{Z} = \mathcal{X} \times \mathcal{Y}}$.

unknown sample distr: sample set

$$\mathcal{D}(x, y) = \mathcal{D}(x)\, \mathcal{D}(y/x).$$

$$S = \{z_1 \dots z_m\} = \{x_1 y_1, x_2 y_2, \dots, x_m y_m\}$$

$|S| = m$

$\mathcal{H} \ni h \quad h : \mathcal{X} \to \mathcal{Y}$ "hypothesis"

$x \mapsto h(x) = y$

Loss fct $\ell(h, z) = \begin{cases} \mathbb{1}(y \neq h(x)) \\ (y - h(x))^2 \\ \dots \end{cases}$

True loss / risk $L_{\mathcal{D}}(h) = \underset{z \sim \mathcal{D}}{\mathbb{E}} \left(\ell(h, z) \right)$

Empirical loss $L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i)$.

Definition. $\mathcal{H}$ is agnostic PAC learnable w.r.t ℓ and Z, if given any $\epsilon > 0$, $\delta > 0$, $\exists\ m_{\mathcal{H}}(\epsilon, \delta) \in \mathbb{N}$ and an algorithm or rule $A(S) : \mathcal{X} \to \mathcal{Y}$ such that $\forall\ \mathcal{D}$ if $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ we have

$$\mathbb{P}_{S \sim \mathcal{D}^m} \left\{ L_{\mathcal{D}}(A(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \right\} \geq 1 - \delta$$

Definition : $\mathcal{H}$ is simply PAC learnable if above holds for just realizable distributions, that is $\forall\ \mathcal{D}$ such that $\exists\ f$ with $L_{\mathcal{D}}(f) = 0$.

(in other words $\min L_{\mathcal{D}} = 0$ and $\mathbb{P}\left\{ L_{\mathcal{D}}(A(S)) \leq \epsilon \right\} \geq 1 - \delta$.

Definition : We say $\mathcal{H}$ satisfies the uniform convergence property if for $m > m_{\mathcal{H}}^{UC}(\epsilon, \delta)$:

$$\mathbb{P}\left\{ \forall h \in \mathcal{H} : |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon \right\} \geq 1 - \delta$$

$$\mathbb{P}\left\{ \max_{h \in \mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon \right\} \geq 1 - \delta.$$

Theorem. Let $|\mathcal{H}| < +\infty$ (finite $\mathcal{H}$) and let loss fct be bounded (say $0 \leq \ell \leq 1$). Then $\mathcal{H}$ has the uniform convergence property with $m_{\mathcal{H}}^{UC}(\epsilon, \delta) = \left\lceil \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2} \right\rceil$

Corollary: A finite $\mathcal{H}$ is agnostic PAC learnable with $A(S) = ERM(S)$ with sample complexity $m_{\mathcal{H}}^{UC}(\epsilon, \delta)$, i.e. $|S| = m \geq m_{\mathcal{H}}^{UC}$.

Remark $ERM(S) = \operatorname{argmin} L_S(h)$.

Theorem : No Free Lunch. (labeling problems)

Let A be a fixed algo or rule. Let S be sample set with $m \leq \frac{|\mathcal{X}|}{2}$ (so for infinite $\mathcal{X}$ this is true for any integer $m \in \mathbb{N}$). Then $\exists\, \mathcal{D}$ such that

a) we are in realizable case : $\exists f$ with $L_{\mathcal{D}}(f) = 0$

b) $\underset{S \sim \mathcal{D}^m}{\mathbb{P}} \left\{ L_{\mathcal{D}}(A(S)) \geq \frac{1}{8} \right\} \geq \frac{1}{7}$.

Proof idea : we considered $C \subset \mathcal{X}$ with $|C| = 2m$ and all possible 2^{2m} labeling fcts $C \to \{0,1\}$

$f_1, f_2, \dots f_{2^{2m}}$ and $\mathcal{D}_i(x,y) = \underbrace{\mathcal{D}(x)}_{\frac{1}{|C|}} \underbrace{\mathcal{D}(y|x)}_{\delta_{y, f(x)}}$

and showed that a) and b) hold for at least one such fct.

Corollary: Take $\mathcal{X}$ to be infinite and let $\mathcal{H} : \mathcal{X} \to \{0,1\}$ be all possible labeling fcts. This hypothesis class is not learnable.

Proof: We proceed by contradiction. Assume $\mathcal{H}$ is PAC learnable. Fix $\epsilon > 0$, $\delta > 0$. There must exist A and $m_{\mathcal{H}}(\epsilon, \delta)$ s.t $\forall \mathcal{D}$ with realizable property $L_{\mathcal{D}}(f) = 0$ (some f) we have

$$\mathbb{P}\{ L_{\mathcal{D}}(A(S)) \leq \epsilon \} \geq 1 - \delta$$

for $m \geq m_{\mathcal{H}}(\epsilon, \delta)$.

But by the No Free Lunch Theorem when $|\mathcal{X}| > 2m$ (which is the case here for any $m \in \mathbb{N}$) there exists $\mathcal{D}$ with $\mathbb{P}\{ L_{\mathcal{D}}(A(S)) \geq \frac{1}{8} \} \geq \frac{1}{7}$.

Thus

$$\mathbb{P}\{ L_{\mathcal{D}}(A(S)) \geq \tfrac{1}{8} \} + \mathbb{P}(L_{\mathcal{D}}(A(S)) \leq \tfrac{1}{8}) \geq \tfrac{1}{7} + 1 - \delta$$

> 1 for $\delta < \frac{1}{7}$

$\Rightarrow$ contradiction. ∎

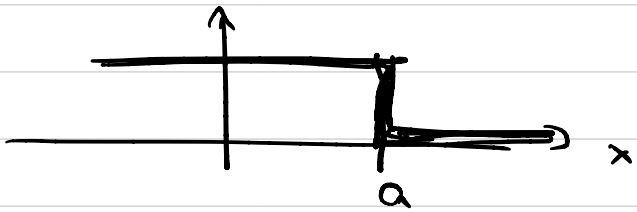
Learnability of infinite classes.

Now we come to infinite classes. By the previous corrolary we see that if an hypothesis class $\mathcal{H}$ is too big it is not PAC learnable.

However we have seen in the exercise sessions that threshold functions:

$$\mathcal{H} = \{ h_a : h_a(x) = \mathbb{1}(x \leq a), a \in \mathbb{R} \}$$

are PAC learnable.



we have $M_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\log 2/\delta}{\epsilon} \right\rceil$.

We will ultimately show that if some "effective size of $\mathcal{H}$" is finite $\mathcal{H}$ is agnostic PAC learnable with ERM.

Definition. Growth rate of $\mathcal{H}$.

The growth rate of an hypothesis class is by definition:

$$\tau_{\mathcal{H}}(m) = \max_{C \subset \mathcal{X} : |C| = m} |\mathcal{H}_C|$$

where $\mathcal{H}_C$ is the restriction of $\mathcal{H}$ to the set C:

$\neq$

We restrict the functions $h : \mathcal{X} \to \mathcal{Y}$ to the arguments in C. So we consider $h : C \to \mathcal{Y}$. This forms set $\mathcal{H}_C$.

For $\mathcal{Y} = \{0, 1\}$ and $\mathcal{H}$ = labeling functions this amounts to consider only those labelings of points $C = \{c_1, c_2, \ldots, c_m\}$ for $|C| = m$.

Examples.

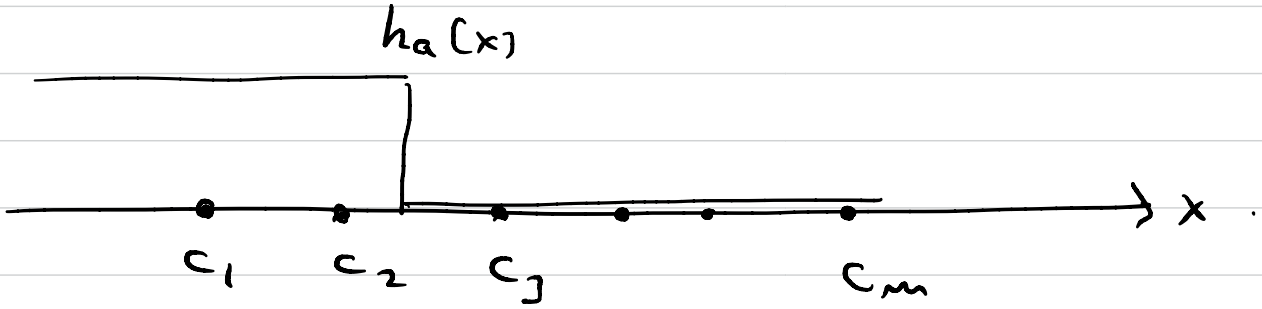
- $\mathcal{H}$ = all possible labeling sets

$\Rightarrow$ $\mathcal{H}_C$ = all possible labeling sets of $c_1 \dots c_m$

and $|\mathcal{H}_C| = 2^m$ so $\tau_{\mathcal{H}}(m) = 2^m$

The growth rate is exponential.

- Threshold sets $h_a(x) = \mathbb{I}(x \le a)$.



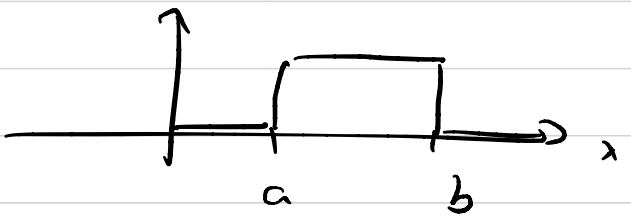
possible labelings are

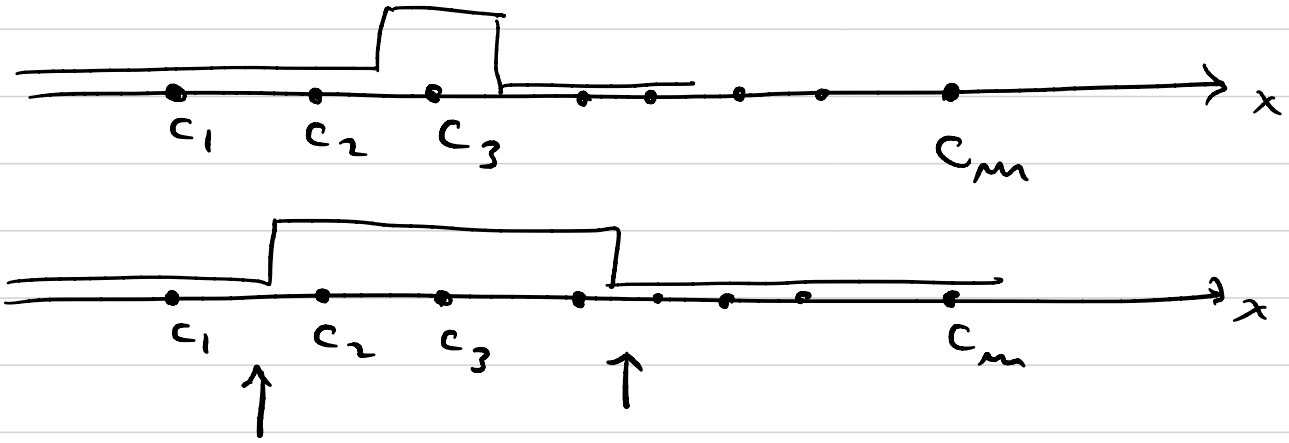
0 0 ... 0
1 0 ... 0
1 1 ... 0
1 1 1 ... 0
1 1 1 ... 1

$\Big\}$ $|\mathcal{H}_C| = m+1$.

$\Rightarrow$ $\tau_{\mathcal{H}}(m) = m+1$ linear growth rate.

- Rectangles $h_{a,b}(x) = \mathbb{1}(a \le x \le b)$





$\sim M$ choices for one threshold

$\sim M$ choices for other threshold.

$\Rightarrow$ $\tau_{de}(M) \sim M^2$

squared growth rate.

Final lesson: Growth rate can be exponential or polynomial ect...

We will see ultimately that pol growth rate means a class is learnable.

Main Theorem: Fix $\delta > 0$. For any $\mathcal{D}$ we have:

$$P_{S \sim \mathcal{D}^m}\left\{ \sup_{h \in \mathcal{H}} \left| L_{\mathcal{D}}(h) - L_S(h) \right| \leq \frac{4 + \log \tau_{\mathcal{H}}(2m)}{\delta \sqrt{2m}} \right\} \geq 1 - \delta$$

Corollary. Let $\tau_{\mathcal{H}}(m)$ has polynomial growth rate, specifically $\tau_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^d$

then $\mathcal{H}$ has the uniform convergence property and is therefore PAC learnable by the ERM rule.

Remark: Next time we give a combinatorial caracterization of $\mathcal{H}$ with pol growth rate and will see that $d = VCdim(\mathcal{H})$.

Sketch of proof of corollary:

for large enough m and some constant C:

$$m \geq C \frac{d + \log(1/\delta)}{\epsilon^2}$$

$$\Rightarrow \frac{\sqrt{4 + d \log\left(\frac{em}{d}\right)}}{\delta\sqrt{2m}} \leq \epsilon$$

Thus for $m \geq C \frac{d + \log 1/\delta}{\epsilon^2}$ we have

$$\mathbb{P}\left\{ \sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq \frac{4 + \sqrt{d \log \frac{em}{d}}}{\delta\sqrt{2m}} \right\}$$

$$\leq \mathbb{P}\left\{ \sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon \right\}$$

Thus $$\mathbb{P}\left\{ \sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon \right\} \geq 1 - \delta$$

which means emf conv prop with $m_{\mathcal{H}}(\epsilon, \delta) = C \frac{d + \log 1/\delta}{\epsilon^2}$.

■

Proof of Main Theorem.

We will show that

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|\right] \leq \frac{4+\sqrt{\log \tau_{\mathcal{H}}(2m)}}{\sqrt{2m}}$$

Then by Markov's inequality:

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \geq \frac{4+\sqrt{\log \tau_{\mathcal{H}}(2m)}}{\delta\sqrt{2m}}\right]$$

$$\leq \frac{\mathbb{E}\left[\sup_{h\in\mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|\right]}{\frac{4+\sqrt{\log \tau_{\mathcal{H}}(2m)}}{\delta\sqrt{2m}}}$$

$$\leq \delta .$$

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left\{ \sup_{h\in H} |L_{\mathcal{D}}(h) - L_S(h)| \right\} = (*).$$

$L_{\mathcal{D}}(h)$: $\mathbb{E}_{S'\sim\mathcal{D}^m}(L_{S'}(h))$ by definition.

$$\Rightarrow (*) = \mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_{h\in H} \left| \mathbb{E}_{S'\sim\mathcal{D}^m}\left(L_{S'}(h) - L_S(h)\right)\right|\right]$$

$$\leq \mathbb{E}_{S\sim\mathcal{D}^m} \sup_{h\in H} \mathbb{E}_{S'\sim\mathcal{D}^m} |L_{S'}(h) - L_S(h)|$$

$$\leq \mathbb{E}_{\substack{S\sim\mathcal{D}^m \\ S'\sim\mathcal{D}^m}} \sup_{h\in H} |L_{S'}(h) - L_S(h)|.$$

$$\left| \frac{1}{m} \sum_{i=1}^{m} \left(\ell(h, z_i') - \ell(h, z_i)\right) \right|$$

where we pick $2m$ samples according to $\mathcal{D}$

$$z_1 \; z_2 \; z_3 \; \dots \; z_m$$

$$z_1' \; z_2' \; z_3' \; \dots \; z_m'$$

This equivalent to

$$z_1 \quad z_2' \quad z_3 \quad \dots \quad z_m$$

$$z_1' \quad z_2 \quad z_3' \quad \dots \quad z_m'.$$

In general it is equivalent to exchange z_i & z_i' for each i. This means that the r.h.s of inequality is <u>equal</u> to

$$\mathbb{E}_{\substack{S \sim \mathcal{D}^m \\ S' \sim \mathcal{D}^m}} \sup_{h \in \mathcal{H}} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i \left(\ell(h, z_i) - \ell(h, z_i') \right) \right|$$

for any $(\sigma_1 \dots \sigma_m) \in \{-1, +1\}^m$.

We can further average (*) over $\sigma_1 \dots \sigma_m$ uniform and obtain

$$(*) \leq \mathbb{E}_{\sigma_1 \dots \sigma_m} \mathbb{E}_{\substack{S \sim \mathcal{D}^m \\ S' \sim \mathcal{D}^m}} \sup_{h \in \mathcal{H}} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i \left(\ell(h, z_i) - \ell(h, z_i') \right) \right.$$

$$(*) \leq \underset{\substack{S\sim\mathcal{D}^m \\ S'\sim\mathcal{D}^m}}{\mathbb{E}} \underbrace{\underset{\sigma_1\cdots\sigma_m}{\mathbb{E}} \sup_{h\in\mathcal{H}} \left| \frac{1}{m}\sum_{i=1}^{m} \sigma_i \left(\ell(h,z_i) - \ell(h,z_i')\right)\right|}$$

- for S and S' given "fixed"

  in this expression h see only

  the samples $\begin{pmatrix} x_1 \cdots x_m \\ x_1' \cdots x_m' \end{pmatrix} = C$

  Here $C \subset X$ and $|C| = 2m$.

  So even if $\mathcal{H}$ is infinite in fact

  this expression depends only on $\mathcal{H}_C$.

$\Rightarrow$

$$\underset{\sigma_1\cdots\sigma_m}{\mathbb{E}} \sup_{h\in\mathcal{H}} \left| \frac{1}{m}\sum_{i=1}^{m} \sigma_i \left(\ell(h,z_i) - \ell(h,z_i')\right)\right|$$

$$= \underset{\sigma_1\cdots\sigma_m}{\mathbb{E}} \sup_{h\in\mathcal{H}_C} \underbrace{\left| \frac{1}{m}\sum_{i=1}^{m} \sigma_i \left(\ell(h,z_i) - \ell(h,z_i')\right)\right|}$$

ϑ_h.

$$\mathbb{E}_{\sigma_1 \cdots \sigma_m}(\vartheta_h) = 0 \quad \text{since } \sigma_i \sim \pm 1 \text{ uniformly.}$$

By Hoeffding's inequality:

$$\mathbb{P}_{\sigma_1 \cdots \sigma_m}\left(|\vartheta_h - \mathbb{E}_{\sigma_1 \cdots \sigma_m}(\vartheta_h)| > \varepsilon \right) \leq 2 e^{-\frac{2m\varepsilon^2}{4}}$$

$$\uparrow \quad 2 e^{-\frac{2m\varepsilon^2}{(b-a)^2}}$$

Remark $-1 \leq \vartheta_h \leq +1$

By the union bound:

$$\mathbb{P}_{\sigma_1 \cdots \sigma_m}\left(\sup_{h \in \mathcal{H}_c} |\vartheta_h| > \varepsilon \right) = \mathbb{P}\left\{ \bigcup_{h \in \mathcal{H}_c} \left(|\vartheta_h| > \varepsilon \right) \right\}$$

$$\leq |\mathcal{H}_c| \, 2 e^{-\frac{m\varepsilon^2}{2}}$$

This is finite here!

An analysis calculation shown later implies from here

$$\mathbb{E}_{\sigma_1 \cdots \sigma_m}\left[\sup_{h \in \mathcal{H}_c} |\vartheta_h| \right] \leq \frac{4 + \sqrt{\log |\mathcal{H}_c|}}{\sqrt{2m}}$$

Since this bound is even independent of S & S' we find (see bottom of page 16).

$$(*) \leq \frac{4 + \log |\mathcal{H}_C|}{\sqrt{2m}} \leq \frac{4 + \sqrt{\log \tau_{\mathcal{H}}(2m)}}{\sqrt{2m}}$$

(Recall $\tau_{\mathcal{H}}(m) = \max_{|C|=m} |\mathcal{H}_C|$ and here $|C| = 2m$)

∎

Lemma (needed on page 17).

$$\mathbb{P}(X > \rho) \leq a e^{-b\rho^2} \Rightarrow \mathbb{E}(X) \leq \frac{2 + \sqrt{\log a}}{\sqrt{b}}.$$

Proof. Let $X \geq 0$

$$\mathbb{E}(X) = \int_0^{+\infty} d\rho\, \rho\, p(\rho) = -\int_0^{+\infty} d\rho\, \rho \frac{d}{d\rho} \underbrace{\int_\rho^{+\infty} p(x)\, dx}_{\mathbb{P}(X \geq \rho)}$$

$$= \int_0^{+\infty} d\rho\, \mathbb{P}(X \geq \rho)$$

$$= \int_0^{\sqrt{\frac{\log a}{b}}} d\rho\, \mathbb{P}(X \geq \rho) + \int_{\sqrt{\frac{\log a}{b}}}^{+\infty} d\rho\, \mathbb{P}(X \geq \rho)$$

$$\leq \sqrt{\frac{\log a}{b}} + \underbrace{\int_{\sqrt{\frac{\log a}{b}}}^{+\infty} d\rho\, a e^{-b\rho^2}}_{\int_0^{+\infty} d\rho\, \underbrace{a e^{-b\left(\rho + \sqrt{\frac{\log a}{b}}\right)^2}}_{\leq e^{-b\rho^2} \text{ for } \rho > 0}}$$

check : $a\, e^{-b\left(\varepsilon + \sqrt{\frac{\log a}{b}}\right)^2} \leq e^{-b\varepsilon^2}$?

$$\Leftrightarrow \quad \log a - b\left(\varepsilon^2 + 2\varepsilon\sqrt{\frac{\log a}{b}} + \frac{\log a}{b}\right) \leq -b\varepsilon^2$$

$$\Leftrightarrow \quad -2\varepsilon b\sqrt{\frac{\log a}{b}} \leq 0 \qquad \checkmark$$

So we have

$$\mathbb{E}(X) \leq \sqrt{\frac{\log a}{b}} + \underbrace{\int_0^{+\infty} d\varepsilon\, e^{-b\varepsilon^2}}_{\sqrt{\frac{\pi}{b}} \leq \frac{2}{\sqrt{b}}}.$$

□