No-Free-Lunch

Theorem: Let A be any learning algorithm for binary classification ($\mathcal{Y}=\{0,1\}$) with respect to the 0-1 loss and domain X. Then, there exists a distribution, D, over $X \times \{0,1\}$ such that, if running A on $m < \frac{|X|}{2}$ samples, we have

(1) $\exists f: X \to \{0,1\}$ with $L_D(f)=0$

(2) $D^m\left(\left\{S: L_D(A(S)) \geq \frac{1}{8}\right\}\right) \geq \frac{1}{7}$

The distribution defines the learning task".

(Regarding the theorem, ERM on $H = \{f\}$ would yield a hypothesis with $L_D(f) = 0$.)

<u>Proof</u>: let $\boxed{C \subset X \text{ with } |C| = 2m}$

We have $T = 2^{2m}$ possible functions from C to $\{0,1\}$.

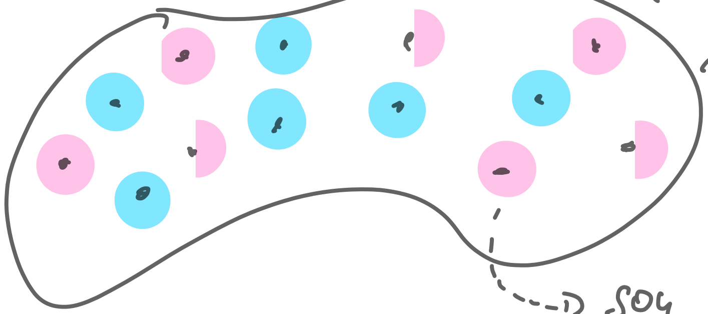
$$2^{2m} \left\{ \begin{array}{l} f_1 \\ \vdots \\ f_i \\ \vdots \\ f_T \end{array} \right.$$

We take for each $i \in \{1, \dots, T\}$

$$\mathcal{D}_i(\{(x,y)\}) = \begin{cases} 1/|C|, & \text{if } y = f_i(x) \\ 0, & \text{else} \end{cases}$$

Note that $\mathcal{D}_i$ is a distribution over $C \times \{0, 1\}$.

$2m$



this is the labeling for one specific choice of f_i = label 1 = label 0

say $f_i(x) = 1$, then $(x, 1)$ has mass $\frac{1}{|C|}$

$(x, 0)$ has mass 0

Obviously, $L_{\mathcal{D}_i}(f_i) = 0$ for all $i \in \{1, \dots, T\}$.

<u>What do we need to show?</u> We need to show that if A receives m instances in S from $C \times \{0,1\}$ and returns $A(S): C \to \{0,1\}$, then

$$\max_{i \in \{1,\dots,T\}} \mathbb{E}_{S \sim D_i^m} \left[L_{D_i}(A(S)) \right] \geq \frac{1}{4} \qquad (*)$$

If this holds, there would be a D_i (i.e., the D_i where the maximum is attained) such that for any learning algorithm receiving m samples, we have

$$\mathbb{E}_{S \sim D^m} \left[L_D(A(S)) \right] \geq \frac{1}{4}$$

↳ this is exactly the one D_i we found in $(*)$

Why is (X) enough?

Markov's inequality does not really help here, as

$$\mathbb{P}[\ldots \geq \ldots] \leq \frac{\mathbb{E}[\ldots]}{\ldots}$$

But, we can easily derive

$$\mathbb{P}[z \geq 1-a] \geq \frac{\mathbb{E}[z] - (1-a)}{a}$$

Now, if $a = \frac{7}{8} \Rightarrow 1-a = \frac{1}{8}$ and $\mathbb{E}[z] \geq \frac{1}{4}$

$$\Rightarrow \mathbb{P}\left[z \geq \frac{1}{8}\right] \geq \frac{\frac{1}{4} - \left(1-\frac{7}{8}\right) 8}{7} = \frac{1}{7}$$

if we have $|C| = 2m$, we have $(2m)^m$ possible training sets of size m. let $k = (2m)^m$. If we enumerate, we get

$$S_1, \ldots\ldots, S_k$$

if S_j is labeled by f_i, we write S_j^i. In general, we have

$$\mathop{\mathbb{E}}_{S \sim D_i^m}\left[L_{D_i}(A(S))\right] = \frac{1}{k} \cdot \sum_{j=1}^{k} L_{D_i}\left(A(S_j^i)\right)$$

lets take the max over $i \in \{1, \ldots, T\}$

$$\max_{i \in \{1,\ldots,T\}} \frac{1}{k} \cdot \sum_{j=1}^{k} L_{D_i}\left(A(S_j^i)\right) \geq \frac{1}{T} \cdot \sum_{i=1}^{T} \frac{1}{k} \sum_{j=1}^{k} L_{D_i}\left(A(S_j^i)\right)$$

(because max ≥ avg.)

.... $$= \frac{1}{k} \cdot \sum_{j=1}^{k} \frac{1}{T} \sum_{i=1}^{T} L_{D_i}(A(S_j^i))$$

$$\geq \min_{j \in \{1,\dots,k\}} \frac{1}{T} \cdot \sum_{i=1}^{T} L_{D_i}(A(S_j^i))$$

<u>Lets fix one $j \in \{1, \dots, k\}$:</u>

Our $S_j^i|_x$ contains some $(x_1 \dots x_m)$ and we do not see $(v_1, \dots, v_p)$ with $p \geq m$ (because $x_1 \dots x_m$ are iid from D_i). therefor the $\geq$

Hence, for <u>any</u> $h : C \to \{0,1\}$, it holds that

$$L_{D_i}(h) = \frac{1}{2m} \cdot \sum_{c \in C} \mathbb{1}_{h(c) \neq f_i(c)} \geq \frac{1}{2m} \cdot \sum_{k=1}^{p} \mathbb{1}_{h(v_k) \neq f_i(v_k)}$$

$$\geq \frac{1}{2p} \cdot \sum_{k=1}^{p} \mathbb{1}_{h(v_k) \neq f_i(v_k)}$$

$$\geq \frac{1}{2} \cdot \frac{1}{p} \cdot \sum_{k=1}^{p} \mathbb{1}_{h(v_k) \neq f_i(v_k)}$$

This holds for any $h : C \to \{0,1\} \Rightarrow$ it also holds for the hypothesis returned by A.

$$\Rightarrow \frac{1}{T} \cdot \sum_{i=1}^{T} L_{D_i}\left(A\left(S_j^i\right)\right) \geq \frac{1}{T} \cdot \sum_{i=1}^{T} \frac{1}{2p} \cdot \sum_{k=1}^{p} \mathbb{1}_{h(v_k) \neq f_i(v_k)}$$

as it holds for any h!

(remember that we fixed j!)

$$= \frac{1}{2} \cdot \frac{1}{p} \cdot \sum_{k=1}^{p} \frac{1}{T} \sum_{i=1}^{T} \mathbb{1}_{A(S_j^i)(v_k) \neq f_i(v_k)}$$

$$\geq \frac{1}{2} \cdot \min_{k \in \{1,\dots,p\}} \frac{1}{T} \cdot \sum_{i=1}^{T} \mathbb{1}_{A(S_j^i)(v_k) \neq f_i(v_k)}$$

Lets fix r for a moment!

Remember that $T = 2^{2m}$ $(f_1, \dots, f_T)$.

<u>Ex:</u> $m=1$, $T = 2^2 = 4$

	X_1	V_1
f_1	0	0
f_2	0	1
f_3	1	0
f_4	1	1

f_1 and f_2 only differ on V_1!
f_3 and f_4 only differ on V_1!

we partition the functions into (f_1, f_2) and $(f_3, f_4) \Rightarrow$ 2 disjoint sets; in general we have $T/2$ disjoint sets if we follow this construction.

For a pair $(f_i, f_{i'})$ we have $(S_j^i, S_j^{i'})$ and $S_j^i = S_j^{i'}$!
(they only differ on the V_r's).

$$\Rightarrow \mathbb{1}_{A(S_j^i)(v_r) \neq f_i} + \mathbb{1}_{A(S_j^{i'})(v_r) \neq f_i} = 1$$

We can now write $\frac{1}{T} \sum_{i=1}^{T} \mathbb{1}_{A(S_j^i)(v_r) \neq f_i}$ as a sum over our disjoint sets. So, we get

$$\frac{1}{T} \sum_{i=1}^{T} \dots = \frac{1}{\cancel{T}} \frac{\cancel{T}}{2} = \frac{1}{2}$$

<u>Collecting sub-results</u> gives

$$\max_{i \in \{1,\dots,T\}} \mathbb{E}_{S \sim D_i^m} \left[L_{D_i}(A(S)) \right] \geq \min_{j \in \{1,\dots,k\}} \frac{1}{T} \cdot \sum_{i=1}^{T} L_{D_i}(A(S_j^i))$$

$$\geq \min_{j \in \{1,\dots,k\}} \frac{1}{2} \cdot \min_{r \in \{1,\dots,p\}} \underbrace{\frac{1}{T} \cdot \sum_{i=1}^{T} \mathbb{1}_{A(S_j^i)(v_r) \neq f_i}}_{1/2}$$

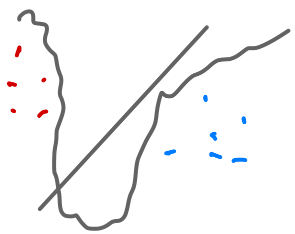
$$\geq \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

10

Corollary: Let X be a domain of infinite size and let H be the class of all possible functions from X to $\{0,1\}$. Then, H is not PAC learnable.

Side-remark:

BIAS – COMPLEXITY TRADEOFF



$$L_D(\overset{\text{ERM output}}{h_S}) = \varepsilon_{approx.} + \varepsilon_{est.}$$

$$= \min_{h' \in H} L_D(h') + L_D(h_S) - \varepsilon_{approx}$$

$$= \underbrace{\min_{h' \in H} L_D(h')}_{\text{quantifies the approximation error}} + \underbrace{L_D(h_S) - \min_{h' \in H} L_D(h')}_{\text{estimation error}}$$

$\longleftrightarrow$

TRADEOFF