8. No Free Lunches

Wednesday, January 19, 2022 2:14 PM

§5 "Bias Complexity Tradeoff" [SS]

① (today) Beyond selecting features, we need some prior knowledge still

∄ universal learner, i.e. no learner can succeed at all tasks

Do not interpret this as evidence to "give up". Very limited, technical scope

② (next class)

Split risk into Approx. Error / bias + Estimation Error

Approx. Error / bias: $\mathcal{H}$ not expressive enough

Estimation Error: m not large enough (small $\mathcal{H}$ helps here)

§5.1 No Free Lunch theorems

Thm "No Free Lunch" (binary classification, 0-1 loss function)

Let A be any learning algorithm, with output $h_S = A(S)$, domain X, $Y = \{0,1\}$, $|S| = m$ and $|X| \geq 2m$ (eg. $|X| = \infty$). Then ∃ distribution D over $X \times Y$ such that

① $\exists f: X \to Y$ w/ $L_D(f) = 0$ — benchmark/comparison: "It's a fair problem"

② w.p. $\geq 1/7$ (over $S \sim D^m$), $L_D(h_S) \geq 1/8$

i.e., cannot take $\delta \to 0$ nor $\varepsilon \to 0$

proof (a bit subtle but useful for later)

Let $C \subseteq X$ have size $2m$ — set, not sequence, so no duplicates

We observe S of size m, so we'll observe no more than $1/2$ of C

(and this gives room for our adversary to pick tricky D)

For D,

first, can ignore D on $X \setminus C$

Let f be an oracle function, $f: C \to Y := \{0,1\}$

Since $|C| = 2m$, $|Y| = 2$, there are $T := 2^{2m}$ such oracle functions, which we enumerate $\{f_i\}_{i=1}^{T}$

Each f_i is compatible with only some distributions,

i.e., need $D(y \mid x) = f_i(x)$ (cond.)

and among all compatible distr., choose the one w/ a uniform marginal (marginal $D_x(x) = \mathbb{E}_y D((x,y) \mid y)$)

and for f_i, call this distr. D_i

i.e., $\forall i \in [T]$, $D_i((x,y)) = \begin{cases} 1/|C| & \text{if } y = f_i(x) \\ 0 & \text{else} \end{cases}$

So by design,

$$L_{D_i}(f_i) = 0$$

Which $i \in [T]$ to choose?

Our "adversary" goes second: we choose algo A

which returns $h_S = A(S)$, then adversary chooses $i \in [T]$

such that $\mathbb{E}_{S \sim D_i^m} L_{D_i}(h_S) \geq 1/4$ (*)

So with $D = D_i$, use $\mathbb{E}_{S \sim D^m} L_D(h_S) \geq 1/4$

$\Rightarrow \mathbb{P}[L_D(h_S) \geq 1/8] \geq 1/7$

} not completely trivial
Markov ineq or similar
See Exerc. 5.1

To show (*), want

$$\max_{i \in [T]} \mathbb{E}_{S \sim D_i^m} L_{D_i}(h_S) \geq 1/4$$

Fixing i, note that since D_x is uniform, for $m=1$, $\{x_1\}$, every value of x_1 is equally likely... and $m \geq 1$, $\{x_1, \dots, x_m\}$, all sequences (duplicates OK) are equally likely.

Drawing m elements (w/ replacement) from $2m$ elements of C gives $k = (2m)^m$ such sequences (all equally likely).

From $(x_1, \dots, x_m)$ we get one $S = ((x_1, y_1), \dots, (x_m, y_m))$ since $y = f_i(x)$ (via how we defined D_i)

Enumerate all these datasets as S_j^i, $j \in [k]$ $i \in [T]$

Still fixing $i \in [T]$

$$\mathbb{E}_{S \sim D_i^m} L_{D_i}(h_S) = \frac{1}{k} \sum_{j=1}^{k} L_{D_i}(h_{S_j^i}) \quad \text{since uniform}$$

Now deal with i. We'll use max $\geq$ avg

$$\max_{i\in[T]} \mathop{\mathbb{E}}_{S\sim D_i^m} L_{D_i}(h_S) \geq \frac{1}{T}\sum_{i=1}^{T}\left(\frac{1}{K}\sum_{j=1}^{K} L_{D_i}(h_{S_j^i})\right)$$

$$= \frac{1}{K}\sum_{j=1}^{K}\frac{1}{T}\sum_{i=1}^{T} L_{D_i}(h_{S_j^i})$$

avg $\geq$ min

$$\geq \min_{j\in[K]}\frac{1}{T}\sum_{i=1}^{T} L_{D_i}(h_{S_j^i}) \quad (**)$$

Fix $j\in[K]$ for the moment.

Observed $(x_1,\dots,x_m)\in C$ (m or fewer distinct items) $\quad |C| = 2m$

so we didn't observe $\{v_1,\dots,v_p\}\subseteq C$, $p\geq m$

$$L_{D_i}(h) := \mathop{\mathbb{E}}_{(x,y)\sim D} \ell(h,(x,y))$$

$$= \mathop{\mathbb{E}}_{x\sim D_x} \ell(h,(x,f_i(x)))$$

$$= \mathop{\mathbb{E}}_{x\sim D_x} \mathbb{1}_{h(x)\neq f_i(x)} \quad \text{since 0-1 loss}$$

$$= \frac{1}{2m}\sum_{x\in C} \mathbb{1}_{h(x)\neq f_i(x)}$$

$$\geq \frac{1}{2m}\sum_{r=1}^{p} \mathbb{1}_{h(v_r)\neq f_i(v_r)} \quad \text{since loss is non-negative}$$

$$\geq \frac{1}{2p}\sum_{r=1}^{p} \mathbb{1}_{h(v_r)\neq f_i(v_r)} \quad \text{since } p\geq m$$

Looking back at (**),

$$\frac{1}{T}\sum_{i=1}^{T} L_{D_i}(h_{S_j^i}) \geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{2p}\sum_{r=1}^{p} \mathbb{1}_{h_j^i(v_r)\neq f_i(v_r)}$$

$$= \frac{1}{2p}\sum_{r=1}^{p}\frac{1}{T}\sum_{i}^{T} \mathbb{1}_{h_j^i(v_r)\neq f_i(v_r)} \quad \text{swap sum}$$

examine this (***)

Fix r for the moment

Look at all f_i, $i\in[T]$, $T = 2^{2m}$ (an even number)

All functions from C to $\{0,1\}$

Ex $C=\{a,b,c\}$ where $c = v_r$ (unobserved)

$Y=\{0,1\}$ so $T = 2^3$ functions

	f(a)	f(b)	f(c)	
i=1	0	0	0	} pair
2	0	0	1	
3	0	1	0	} pair
4	0	1	1	
5	1	0	0	} pair
6	1	0	1	
7	1	1	0	} pair
8	1	1	1	

Write all T indices as $T/2$ pairs (i, i')

where $f_i(a) = f_{i'}(a)$

$f_i(b) = f_{i'}(b)$

$f_i(c) \neq f_{i'}(c)$

or more generally, $f_i(x) = f_{i'}(x) \quad \forall x \in C \setminus \{v_r\}$

$(*) f_i(x) \neq f_{i'}(x) \quad x = v_r$

(think of binary notation, v_r index as least significant bit)

back to our term (***)

paired up

$$\frac{1}{T}\sum_{i}^{T} \mathbb{1}_{h(v_r) \neq f_i(v_r)} = \frac{1}{T}\left(\sum_{i=1}^{T/2} \mathbb{1}_{h_j^i(v_r) \neq f_i(v_r)} + \sum_{i'=1}^{T/2} \mathbb{1}_{h_j^{i'}(v_r) \neq f_{i'}(v_r)} \right)$$

$(*)$ implication $f_i(v_r) = 0$ and $f_{i'}(v_r) = 1$

or $f_i(v_r) = 1$ and $f_{i'}(v_r) = 0$.

Now, is $h_j^i = h_j^{i'}$?

At first, not necessarily since ones is based on data S_j^i, the other on $S_j^{i'}$

recall we've enumerated all of these, so not random.

where i controls the labels via f_i

But, based on our pairs (i, i'),

f_i and $f_{i'}$ agree everywhere but v_r and v_r is unobserved. So $S_j^i = S_j^{i'}$ hence

$h_j^i = h_j^{i'}$

$$(\star\star\star) = \frac{1}{T}\left(\sum_{i=1}^{T/2} \mathbb{1}_{h_j^i(v_r) \neq f_i(v_r)} + \sum_{i'=1}^{T/2} \mathbb{1}_{h_j^{i'}(v_r) \neq f_{i'}(v_r)} \right)$$

(different: i vs i'; same: j)

so exactly 1 term is 1, the other 0

$$= \frac{1}{T}\left(\sum_{i=1}^{T/2} 1 \right) = 1/2$$

Put it altogether:

$$\max_{i \in [T]} \mathop{\mathbb{E}}_{S \sim D_i^m} L_{D_i}(h_S) \geq \min_{j \in [K]} \frac{1}{T} \sum_{i=1}^{T} L_{D_i}(h_{S_j^i}) \quad (\star\star)$$

$$\geq \min_{j \in [K]} \frac{1}{2p} \sum_{r=1}^{p} \underbrace{\frac{1}{T} \sum_{i} \mathbb{1}_{h_j^i(v_r) \neq f_i(v_r)}}_{= 1/2 \quad (\star\star\star)}$$

no more j dependence

$$= \frac{1}{2p} \sum_{r=1}^{p} \frac{1}{2}$$

$$= 1/4 \qquad \square$$

Implications of No Free Lunch

If $|X| = \infty$ and $\mathcal{H} = \{$ all functions from X to $Y\}$ then our reference f is in $\mathcal{H}$, so... $\mathcal{H}$ isn't PAC learnable.

(Recall: PAC learnable, $\exists A$ s.t. $\forall \varepsilon, \delta \in (0,1)$, $\exists m$ s.t. if $|S| \geq m$ iid samples $\forall$ distr. D, then w.p. $\geq 1-\delta$, $L_D(h_S) \leq \varepsilon + \min_{h \in \mathcal{H}} L_D(h_S)$ $(h_S := A(S))$)

So we need $\mathcal{H} \subsetneq Y^X$, i.e., need some prior knowledge to pick $\mathcal{H}$ (can't avoid inductive bias)