No Free Lunch

STATS 303 Statistical Machine Learning

Spring 2022

Lecture 19

no free lunch

universal learner

- In previous discussion, we assume that there is a hypothesis class $\mathcal{H}$ which serves as the search space for our model h.
- We then find the ERM $h_S \in \text{argmin}_{h \in \mathcal{H}} L_S(h)$
- $\mathcal{H}$ is a prior belief, determined by the task.
- Is this prior belief necessary? Is it possible to have a universal learner that works for any task? Specifically, is there an algorithm that outputs a low-risk h as long as it receives a large number of training data?

universal learner

More specifically, does there exist a learning algorithm A and a training set size m, such that:

- for every distribution $\mathcal{D}$, if A receives m i.i.d. examples from $\mathcal{D}$, there is a high chance it outputs a predictor h with a low risk?

This is impossible ☹

no free lunch (NFL)

Theorem (No-Free-Lunch)

Let A be any learning algorithm for the task of binary classification w.r.t. the 0-1 loss over a domain $\mathcal{X}$. Let m, the size of the training set, be any number with $m < |\mathcal{X}|/2$. Then, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:

1. There exists a function $f: \mathcal{X} \to \{0,1\}$ with $L_{\mathcal{D}}(f) = 0$;
2. With probability of at least $\frac{1}{7}$ over the choice of $S \sim \mathcal{D}^m$,we have $L_{\mathcal{D}}(h) \geq \frac{1}{8}$ where $h = A(S)$ is the output of the algorithm.

TLDR version: "Any algorithm will fail for some reasonable data distribution."

no free lunch (NFL)

Theorem (No-Free-Lunch)

Let A be any learning algorithm for the task of binary classification w.r.t. the 0-1 loss over a domain $\mathcal{X}$. Let m, the size of the training set, be any number with $m < |\mathcal{X}|/2$. Then, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:

1. There exists a function $f: \mathcal{X} \rightarrow \{0,1\}$ with $L_{\mathcal{D}}(f) = 0$;
2. With probability of at least $\frac{1}{7}$ over the choice of $S \sim \mathcal{D}^m$,we have $L_{\mathcal{D}}(h) \geq \frac{1}{8}$ where $h = A(\mathrm{S})$ is the output of the algorithm.

Wordier version: “Every learner fails on some task, though the task can be successfully learned by another learner.”

no free lunch (NFL)

Theorem (No-Free-Lunch)

Let A be any learning algorithm for the task of binary classification w.r.t. the 0-1 loss over a domain $\mathcal{X}$. Let m, the size of the training set, be any number with $m < |\mathcal{X}|/2$. Then, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:

1. There exists a function $f: \mathcal{X} \rightarrow \{0,1\}$ with $L_{\mathcal{D}}(f) = 0$;
2. With probability of at least $\frac{1}{7}$ over the choice of $S \sim \mathcal{D}^m$,we have $L_{\mathcal{D}}(h) \geq \frac{1}{8}$ where $h = A(S)$ is the output of the algorithm.

These numbers are not important, and we can change them to other numbers, say 1/16 and 1/5.

proof of NFL

- Let's proof the no-free-lunch theorem (NFL)!
- Consider $C \subset \mathcal{X}$ such that $|C| = 2m$.
- There are $T = 2^{2m}$ possible function that maps from C to $\{0,1\}$. Let's call them $f_1, \cdots, f_T$.
- For each $i \in [T] = \{1, \cdots, T\}$, define $\mathcal{D}_i$ to be the following distribution:

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

proof of NFL

Claim: For every algorithm A that receives a training set of m examples from $C \times \{0,1\}$,

$$\max_{i \in [T]} \mathbb{E}_{S \sim \mathcal{D}_i^m}\left[L_{\mathcal{D}_i}(A(S))\right] \geq \frac{1}{4}$$

This means that for every algorithm A' that receives a training set of m examples from $\mathcal{X} \times \{0,1\}$, there exists a function $f: \mathcal{X} \to \{0,1\}$ and a distribution $\mathcal{D}$ such that $L_{\mathcal{D}}(f) = 0$ and

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

chosen to be f_j where j is argmax

chosen to be $\mathcal{D}_j$ accordingly.

proof of NFL

Claim: For every algorithm A that receives a training set of m examples from $C \times \{0,1\}$,

$$\max_{i \in [T]} \mathbb{E}_{S \sim \mathcal{D}_i^m}\left[L_{\mathcal{D}_i}(A(S))\right] \geq \frac{1}{4}$$

This means that for every algorithm A' that receives a training set of m examples from $\mathcal{X} \times \{0,1\}$, there exists a function $f: \mathcal{X} \to \{0,1\}$ and a distribution $\mathcal{D}$ such that $L_{\mathcal{D}}(f) = 0$ and

$$\mathbb{E}_{S \sim \mathcal{D}^m}[L_{\mathcal{D}}(A'(S))] \geq \frac{1}{4}$$

If we can prove this, then $\Rightarrow P\left(L_{\mathcal{D}}(A'(S)) \geq \frac{1}{8}\right) \geq \frac{1}{7}$

why $\Rightarrow P\left(L_{\mathcal{D}}(A'(S)) \geq \frac{1}{8}\right) \geq \frac{1}{7}$?

We only need to show the more general version: If a random variable X takes values in $[0,1]$, and $\mathbb{E}[X] \geq \frac{1}{4}$,then $P\left(X \geq \frac{1}{8}\right) \geq \frac{1}{7}$.

Markov inequality:

If $Y \geq 0$, then

$$\mathbb{P}(Y \geq a) \leq \frac{\mathbb{E}[Y]}{a}$$

because

$$\mathbb{E}[Y] = \int_0^\infty y\, p(y)\, dy$$

$$\geq \int_a^\infty y\, p(y)\, dy$$

$$\geq \int_a^\infty a\, p(y)\, dy = a\, \mathbb{P}(Y \geq a)$$

Proof:

$$P\left(X < \frac{1}{8}\right) = P\left(1 - X \geq \frac{7}{8}\right) \leq \frac{\mathbb{E}[1-X]}{7/8} \leq \frac{1-1/4}{7/8} = \frac{6}{7}$$

Therefore, $P\left(X \geq \frac{1}{8}\right) \geq 1 - \frac{6}{7} = \frac{1}{7}$. ☺

proof of NFL: only need to prove claim

Recall Claim: For every algorithm A that receives a training set of m examples from $C \times \{0,1\}$,

$$\max_{i \in [T]} \mathbb{E}_{S \sim \mathcal{D}_i^m}\left[L_{\mathcal{D}_i}(A(S))\right] \geq \frac{1}{4}$$

2m choices since $|C| = 2m$

$$\underbrace{\circ \ \circ \ \cdots \ \circ}_{m \text{ examples}}$$

- There are $k = (2m)^m$ possible sequences of m examples from C

  repetition allowed

- Denote them by $S_1, \cdots, S_k$
- Say $S_j = (x_1, \cdots, x_m)$. Denote $S_j^i = ((x_1, f_i(x_1)), \cdots, (x_m, f_i(x_m)))$
- If the distribution is $\mathcal{D}_i$,then the possible training sets are $S_1^i, S_2^i, \cdots, S_k^i$.

proof of NFL: only need to prove claim

- With the above notation,

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

- Only need to show: the max of the above is no less than ¼.

proof of NFL: only need to prove claim

- But

$$\max_{i\in[T]} \frac{1}{k}\sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i)) \geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{k}\sum_{j=1}^{k} L_{\mathcal{D}_i}\left(A\left(S_j^i\right)\right)$$

$$= \frac{1}{k}\sum_{j=1}^{k}\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}\left(A\left(S_j^i\right)\right)$$

$$\geq \min_{j\in[k]} \frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}\left(A\left(S_j^i\right)\right)$$

keep in mind that:

- i is the index for functions,
- j is the index for datasets

proof of NFL: only need to prove claim

- But

$$\max_{i\in[T]} \frac{1}{k}\sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i)) \geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{k}\sum_{j=1}^{k} L_{\mathcal{D}_i}\left(A\left(S_j^i\right)\right)$$

$$= \frac{1}{k}\sum_{j=1}^{k}\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}\left(A\left(S_j^i\right)\right)$$

$$\geq \min_{j\in[k]}\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}\left(A\left(S_j^i\right)\right)$$

We allow repetition. So we used $\leq m$ unique x's

- Fix some $j \in [k]$. Denote $S_j = (x_1, \cdots, x_m)$.
- Let $v_1, \cdots, v_p$ be the examples in C not appearing in S_j.
- Since $|C| = 2m$, we have $p \geq m$.
- Therefore, for every $h: C \to \{0,1\}$ and every i, we have

$$\mathbb{1}_A = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

$$L_{\mathcal{D}_i}(h) = \frac{1}{2m}\sum_{x\in C} \mathbf{1}_{[h(x)\neq f_i(x)]}$$

$$\geq \frac{1}{2m}\sum_{r=1}^{p} \mathbf{1}_{[h(v_r)\neq f_i(v_r)]}$$

$$\geq \frac{1}{2p}\sum_{r=1}^{p} \mathbf{1}_{[h(v_r)\neq f_i(v_r)]}$$

proof of NFL: only need to prove claim

Hence

$$\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}\left(A\left(S_j^i\right)\right) \geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{2p}\sum_{r=1}^{p}\mathbf{1}_{\left[A\left(S_j^i\right)(v_r)\neq f_i(v_r)\right]}$$

$$= \frac{1}{2p}\sum_{r=1}^{p}\frac{1}{T}\sum_{i=1}^{T}\mathbf{1}_{\left[A\left(S_j^i\right)(v_r)\neq f_i(v_r)\right]}$$

$$\geq \frac{1}{2}\min_{r\in[p]}\frac{1}{T}\sum_{i=1}^{T}\underbrace{\mathbf{1}_{\left[A\left(S_j^i\right)(v_r)\neq f_i(v_r)\right]}}$$

Since f_i exhausts all possibilities as i goes from 1 to T, this is equal to 1 for half the time, and equal to 0 for the other half.

proof of NFL: only need to prove claim

Hence

$$\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}\left(A\left(S_j^i\right)\right) \geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{2p}\sum_{r=1}^{p}\mathbf{1}_{\left[A\left(S_j^i\right)(v_r)\neq f_i(v_r)\right]}$$

$$= \frac{1}{2p}\sum_{r=1}^{p}\frac{1}{T}\sum_{i=1}^{T}\mathbf{1}_{\left[A\left(S_j^i\right)(v_r)\neq f_i(v_r)\right]}$$

$$\geq \frac{1}{2}\min_{r\in[p]}\frac{1}{T}\sum_{i=1}^{T}\mathbf{1}_{\left[A\left(S_j^i\right)(v_r)\neq f_i(v_r)\right]} = \frac{1}{4}$$

proof of NFL: only need to prove claim

Recall Claim: For every algorithm A that receives a training set of m examples from $C \times \{0,1\}$,

$$\max_{i\in[T]} \mathbb{E}_{S\sim\mathcal{D}_i^m}\left[L_{\mathcal{D}_i}(A(S))\right] \geq \frac{1}{4}$$

- But we have showed:

$$\max_{i\in[T]} \mathbb{E}_{S\sim\mathcal{D}_i^m}\left[L_{\mathcal{D}_i}(A(S))\right] \geq \min_{j\in[k]} \frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}\left(A\left(S_j^i\right)\right) \geq \min_{j\in[k]} \frac{1}{4} = \frac{1}{4}$$

Therefore, we are done with the proof! We are done!

summary

Theorem (No-Free-Lunch)

Let A be any learning algorithm for the task of binary classification w.r.t. the 0-1 loss over a domain $\mathcal{X}$. Let m, the size of the training set, be any number with $m < |\mathcal{X}|/2$. Then, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:

1. There exists a function $f: \mathcal{X} \to \{0,1\}$ with $L_{\mathcal{D}}(f) = 0$;
2. With probability of at least $\frac{1}{7}$ over the choice of $S \sim \mathcal{D}^m$,we have $L_{\mathcal{D}}(h) \geq \frac{1}{8}$ where $h = A(S)$ is the output of the algorithm.

no restriction on $\mathcal{H}$ $\Rightarrow$ not PAC learable

Corollary

Let $\mathcal{X}$ be an infinite domain set and let $\mathcal{H}$ be the set of all functions from $\mathcal{X}$ to $\{0,1\}$. Then, $\mathcal{H}$ is not PAC learnable.

Proof: Apply NFL. Consider $\epsilon < \frac{1}{8}, \delta < \frac{1}{7}$. Done!

In other words, we cannot have a larger than $\frac{6}{7}$ probability that the error is smaller than $\frac{1}{8}$.

More specifically, to prove the corollary, we could simply take $\varepsilon = \frac{1}{9}$, $\delta = \frac{1}{8}$.

Assume by contradiction that $\mathcal{H}$ is PAC learnable.

Then there is an algorithm A and an integer $m = m(\varepsilon, \delta)$, such that,

- for any data distribution $\mathcal{D}$, if there is an f for which $L_{\mathcal{D}}(f) = 0$, then

$$\mathbb{P}_{\mathcal{D}}\left(L_{\mathcal{D}}(A(S)) \le \varepsilon\right) \ge 1 - \delta.$$

This means

$$\mathbb{P}_{\mathcal{D}}\left(L_{\mathcal{D}}(A(S)) \le \frac{1}{9}\right) \ge 1 - \frac{1}{8}$$

That is, $\mathbb{P}_{\mathcal{D}}\left(L_{\mathcal{D}}(A(S)) > \frac{1}{9}\right) < \frac{1}{8}$

$$\Downarrow$$

$$\mathbb{P}_{\mathcal{D}}\left(L_{\mathcal{D}}(A(S)) \ge \frac{1}{8}\right) < \frac{1}{8}$$

This contradicts with the conclusion of NFL which says $\mathbb{P}_{\mathcal{D}}\left(L_{\mathcal{D}}(A(S)) \ge \frac{1}{8}\right) \ge \frac{1}{7}$.

Questions?

Reference

- *No Free Lunch:*
  - *[S-S] Ch 5.1*