Chapter 6
Lower Bounds I

Ernest K. Ryu
Seoul National University

Machine Learning Theory
Spring 2024

No free lunch theorem

Consider the binary classification problem with 0-1 loss, and let $\mathcal{X}$ be infinite. Let A be an algorithm that takes in as input $\mathcal{D}_N = \{(X_1, Y_1), \ldots, (X_N, Y_N)\}$ and outputs a prediction function:

$$\hat{f}_{\mathcal{D}_N} = A(\mathcal{D}_N(p)).$$

So $\hat{f}_{\mathcal{D}_N}(x) \in \{-1, +1\}$ for all $x \in \mathcal{X}$. Let p a probability distribution on $\mathcal{X} \times \{-1, +1\}$. Then

$$\mathcal{R}_p[f] = \underset{(X,Y)\sim p}{\mathbb{P}} \big(f(X) \neq Y\big).$$

Theorem (No free lunch (NFL))

Let $\mathcal{P}$ denote the set of all probability distributions on $\mathcal{X} \times \{-1, +1\}$. For any $N > 0$ and any algorithm A

$$\sup_{p \in \mathcal{P}} \left\{ \underset{\mathcal{D}_N \sim p}{\mathbb{E}} \big[\mathcal{R}_p[A(\mathcal{D}_N)]\big] - \mathcal{R}_p^* \right\} \geq 1/2.$$

NFL: Corollaries

Corollary

Under the NFL assumptions, for any $N > 0$,

$$\inf_A \sup_{p \in \mathcal{P}} \left\{ \mathop{\mathbb{E}}_{\mathcal{D}_N \sim p} \left[\mathcal{R}_p[A(\mathcal{D}_N)] \right] - \mathcal{R}_p^* \right\} \geq 1/2.$$

So, the best algorithm cannot do better than chance ($1/2$ accuracy).

Corollary

Under the NFL assumptions, for any $N > 0$ any algorithm A, there is a $p \in \mathcal{P}$ such that

$$\mathop{\mathbb{E}}_{\mathcal{D}_N \sim p} \left[\mathcal{R}_p[A(\mathcal{D}_N)] \right] - \mathcal{R}_p^* \geq 1/2.$$

So, while an algorithm A can be good at some choices of p, it is not possible for A to be uniformly good for all $p \in \mathcal{P}$.

NFL: Interpretation

The proof of NFL is based on a fairly obvious argument:
If there are k pieces of information to learn (the sign of $r_1, \ldots, r_k$), you cannot possibly learn them with N data points if $k \gg N$. Since $|\mathcal{X}| = \infty$, it is possible to encode the k pieces of information into $p \in \mathcal{P}$.

The resolution to the NFL theorem is that $\mathcal{P}$ cannot be the set of arbitrary distributions. If $p(Y \mid X)$ depends, say, smoothly as a function of X, then we may be able to learn $p(Y \mid X)$ with N data points.

NFL: Proof

Proof. Let k be a positive integer. W.L.O.G., assume $\mathbb{N} \subset \mathcal{X}$. Given $r \in \{0,1\}^k$, we define the joint distribution $p(r)$ such that $\mathbb{P}(X = j, Y = r_j) = 1/k$ for $j \in \{1, \ldots, k\}$; that is, for X, we choose one of the first k elements of $\mathbb{N}$ uniformly at random, and then Y is selected deterministically as $Y = r_X$. Thus, $\mathcal{R}^*_{p(r)} = 0$ because there is a deterministic relationship.

Let

$$S(r) = \mathop{\mathbb{E}}_{\mathcal{D}_N \sim p} \left[\mathcal{R}_p[\hat{f}_{\mathcal{D}_N}]\right].$$

Note

$$\sup_{p \in \mathcal{P}} \left\{ \mathop{\mathbb{E}}_{\mathcal{D}_N \sim p} \left[\mathcal{R}_p[A(\mathcal{D}_N)]\right] - \mathcal{R}^*_p \right\} \geq \max_{r \in \{0,1\}^k} S(r),$$

since $\{p = p(r) \,|\, r \in \{0,1\}^k\} \subset \mathcal{P}$ and the RHS is a supremum over a smaller set.

NFL: Proof

The maximum of $S(r)$ over $r \in \{0,1\}^k$ is greater than the expectation of $S(r)$ for any probability distribution π on r, in particular the uniform distribution over $r \in \{0,1\}^k$ (each r_j being an independent unbiased Bernoulli variable). So

$$\max_{r \in \{0,1\}^k} S(r) \geq \mathop{\mathbb{E}}_{r \sim \pi} S(r) = \mathop{\mathbb{P}}_{\substack{r \sim \pi \\ p = p(r) \\ \mathcal{D}_N \sim p \\ (X,Y) \sim p}} \left(\hat{f}_{\mathcal{D}_N}(X) \neq Y \right) = \mathop{\mathbb{P}}_{\substack{r \sim \pi \\ p = p(r) \\ \mathcal{D}_N \sim p \\ X \sim p}} \left(\hat{f}_{\mathcal{D}_N}(X) \neq r_X \right),$$

because X is almost surely in $\{1, \dots, k\}$ and $Y = r_X$ almost surely.

NFL: Proof

Next, we have

$$\mathbb{E}_{r\sim\pi} S(r) = \mathbb{E}\left[\mathbb{P}(\hat{f}_{\mathcal{D}_N}(X) \neq r_X \mid X_1, \ldots, X_N, r_{X_1}, \ldots, r_{X_N})\right]$$

$$\geq \mathbb{E}\left[\mathbb{P}(\hat{f}_{\mathcal{D}_N}(X) \neq r_X \text{ and } X \notin \{X_1, \ldots, X_N\} \mid X_1, \ldots, X_M, r_{X_1}, \ldots, r_{X_N})\right]$$

$$= \mathbb{E}\left[\frac{1}{2}\mathbb{P}(X \notin \{X_1, \ldots, X_N\} | X_1, \ldots, X_N, r_{X_1}, \ldots, r_{X_N})\right],$$

because

$$\mathbb{P}(\hat{f}_{\mathcal{D}_N}(X) \neq r_X | X \notin \{X_1, \ldots, X_N\}, X_1, \ldots, X_N, r_{X_1}, \ldots, r_{X_N}) = 1/2$$

(the label $X = r_X$ has the same probability of being 0 or 1, given that it was not observed). Thus,

$$\mathbb{E}_{r\sim q} S(r) \geq \frac{1}{2}\mathbb{P}(X \notin \{X_1, \ldots, X_N\})$$

$$= \frac{1}{2}\mathbb{E}\left[\prod_{i=1}^{N} \mathbb{P}(X_i \neq X | X)\right] = \frac{1}{2}\left(1 - \frac{1}{k}\right)^N.$$

Finally, we let $k \to \infty$ to conclude the statement. $\square$

NFL: Significance

The NFL theorem ends up saying something fairly obvious and intuitive.

However, the NFL theorem is the first example of formalizing arguments for establishing complexity lower bounds. Further lower-bound results follow the overall rubric established by the NFL theorem and present non-obvious arguments and conclusions.