1 No Free Lunch Theorem

The more expressive the class $F$ is, the larger is $\mathcal{V}_n^{PAC}(F)$, $\mathcal{V}_n^{NR}(F)$ and $\mathcal{V}_n^{stat}(F)$. The no free lunch theorem says that if $F = \mathcal{Y}^X$ the set of all function, then there is not convergence of minimax rates.

**Proposition 1.** If $|X| \geq 2n$ then,

$$\mathcal{V}_n^{PAC}(\mathcal{Y}^X) \geq \frac{1}{4}$$

**Proof.** Consider $D_X$ to be the uniform distribution over $2n$ points. Also let $f^* \in \mathcal{Y}^X$ be a random choice of the possible $2^{2n}$ function on these points. Now if we obtain sample $S$ of size at most $n$, then

$$\mathcal{V}_n^{PAC}(\mathcal{Y}^X) = \inf_y \sup_{D_X,f^* \in F} \mathbb{E}_{S:|S|=n} [\mathbb{P}_{x \sim D_x} (\hat{y}(x) \neq f^*(x))]$$

$$\geq \inf_y \mathbb{E}_{f^*} \left[ \mathbb{E}_{S:|S|=n} \left[ \frac{1}{2n} \sum_{j=1}^{2n} 1\{\hat{y}(x_j) \neq f^*(x_j)\} \right] \right]$$

$$= \inf_y \mathbb{E}_{f^*} \left[ 1 + \mathbb{E}_{i_1,\ldots,i_n \sim \text{Unif}[2n]} \left[ \sum_{j \notin \{i_1,\ldots,i_n\}} 1\{\hat{y}(x_j) \neq f^*(x_j)\} \right] \right]$$

$$= \frac{1}{2n} \inf_y \mathbb{E}_{i_1,\ldots,i_n \sim \text{Unif}[2n]} \left[ \mathbb{E}_{f^*} \left[ \sum_{j \notin \{i_1,\ldots,i_n\}} 1\{\hat{y}(x_j) \neq f^*(x_j)\} \right] \right]$$

But outside of sample $S$, on each $x$, $f^*(x)$ can be $\pm 1$ with equal probability. Hence,

$$\mathcal{V}_n^{PAC}(\mathcal{Y}^X) \geq \frac{1}{2n} \inf_y \mathbb{E}_{i_1,\ldots,i_n \sim \text{Unif}[2n]} \left[ \mathbb{E}_{f^*} \left[ \sum_{j \notin \{i_1,\ldots,i_n\}} 1\{\hat{y}(x_j) \neq f^*(x_j)\} \right] \right] \geq \frac{1}{2n} \frac{n}{2} = \frac{1}{4}$$

This shows that we need some restriction on $F$ even for the realizable PAC setting. We cannot learn arbitrary set of hypothesis, there is no free lunch.
2 Example 0: Coin Flips

We consider as a warmup example, the simplest statistical learning/prediction problem. That of learning coin flips! Let us consider the case where we don’t receive any input instance (or \( X = \{ \} \)) and \( Y = \{ \pm 1 \} \). We receive \( \pm 1 \) valued samples \( y_1, \ldots, y_n \in \{ \pm 1 \} \) drawn iid from Bernoulli distribution with parameter \( p \) (i.e. \( Y \) is +1 with probability \( p \) and -1 with probability \( 1 - p \)).

Our loss function is the zero-one loss function \( \ell(y', y) = 1_{\{y' \neq y\}} \). Recall that our goal in statistical learning is to minimize \( L_p(\hat{y}) - \inf_{f \in \{\pm 1\}} L_p(f) \). (Effectively our only choice of \( \mathcal{F} \) for this problem is the set of constant mappings, \( \mathcal{F} = \{ \pm 1 \} \)).

Claim 2. For the problem above, one can bound the minimax rate as:

\[
\mathcal{V}_{stat}^n (\mathcal{F}) \leq \sqrt{\log n/n}
\]

The prediction rule that enjoys the above bound is \( \hat{y} = \text{sign} \left( \frac{1}{n} \sum_{t=1}^n y_t \right) \).

Proof. Now note that:

\[
L_p(\hat{y}) - \inf_{f \in \{\pm 1\}} L_p(f) = \mathbb{E}_{y \sim p} \left[ 1_{\{y \neq \hat{y}\}} \right] - \min_{f \in \{\pm 1\}} \mathbb{E}_{y \sim p} \left[ 1_{\{f \neq y\}} \right] = p \, 1_{\{ \hat{y} \neq 1 \}} + (1 - p) \, 1_{\{ \hat{y} \neq -1 \}} - \min\{p, 1 - p\}
\]

Now if \( \hat{y} = \text{sign}(2p - 1) \) then \( p \, 1_{\{ \hat{y} \neq 1 \}} + (1 - p) \, 1_{\{ \hat{y} \neq -1 \}} = \min\{p, 1 - p\} \) and in this case \( L_p(\hat{y}) - \min_{f \in \{\pm 1\}} L_p(f) = 0 \). On the other hand, if \( \hat{y} = \text{sign}(2p - 1) \), then \( p \, 1_{\{ \hat{y} \neq 1 \}} + (1 - p) \, 1_{\{ \hat{y} \neq -1 \}} = \max\{p, 1 - p\} \) and so \( L_p(\hat{y}) - \min_{f \in \{\pm 1\}} L_p(f) = \, |2p - 1| \). Hence combining the two cases we conclude that

\[
L_p(\hat{y}) - \inf_{f \in \{\pm 1\}} L_p(f) = |1 - 2p| \, 1_{\{ \hat{y} \neq \text{sign}(2p - 1) \}} \leq \epsilon + \, 1_{\{ \hat{y} \neq \text{sign}(2p - 1) \}} \, 1_{\{|1 - 2p| > \epsilon\}}
\]

Now the prediction strategy (really the only sensible deterministic strategy) we consider is: \( \hat{y} = \text{sign} \left( \frac{1}{n} \sum_{t=1}^n y_t \right) \). Hence,

\[
L_p(\hat{y}) - \inf_{f \in \{\pm 1\}} L_p(f) \leq \epsilon + \, 1_{\{ \text{sign} \left( \frac{1}{n} \sum_{t=1}^n y_t \right) \neq \text{sign}(2p - 1) \}} \, 1_{\{|1 - 2p| > \epsilon\}} \leq \epsilon + \, 1_{\{|\frac{1}{n} \sum_{t=1}^n y_t - (2p - 1)| > \epsilon\}}
\]

The reason for the last statement is that if \( |2p - 1| > \epsilon \), then for \( \text{sign} \left( \frac{1}{n} \sum_{t=1}^n y_t \right) \neq \text{sign}(2p - 1) \) it has to at least be that \( \frac{1}{n} \sum_{t=1}^n y_t \) is away from \( 2p - 1 \) by at least \( \epsilon \). (think about the picture on the real line). Hence taking expectation over sample \( S \) we conclude that

\[
\mathbb{E}_S \left[ L_p(\hat{y}) - \inf_{f \in \{\pm 1\}} L_p(f) \right] \leq \epsilon + P \left( \left| \frac{1}{n} \sum_{t=1}^n y_t - (2p - 1) \right| > \epsilon \right)
\]

2
However note that $\mathbb{E}[y] = 2p - 1$ and so by applying Hoeffding’s inequality, we have that for any $\epsilon > 0$,

$$P \left( \left| \frac{1}{n} \sum_{i=1}^{n} y_n - 2p + 1 \right| > \epsilon \right) \leq 2 \exp(-n\epsilon^2/2)$$

Hence,

$$\mathbb{E}_S \left[ L_p(\hat{y}) - \inf_{f \in \{\pm 1\}} L_p(f) \right] \leq \epsilon + 2 \exp(-n\epsilon^2/2) \leq 3 \sqrt{\frac{\log n}{n}}$$

Where we set $\epsilon = \sqrt{\log n/n}$. The above bound we proved for the specific strategy in the claim. This of course implies that the minimax value is bounded as:

$$\mathcal{V}^{\text{stat}}_n(\mathcal{F}) \leq \sqrt{\log n/n}$$

Things to try out for fun:

- Show $1/\sqrt{n}$ rate for this problem.
- Think about high probability version for the problem.

What can we learn from this:

- Algorithm: pick hypothesis minimizing error on sample
- Notice the CLT/concentration inequality popping in to the analysis.