Machine Learning Theory (CS 6783)

Lecture 3 : No Free Lunch Theorem, ERM, Uniform Convergence and MDL Principle

1 No Free Lunch Theorem

The more expressive the class $\mathcal{F}$ is, the larger is $\mathcal{V}_n^{PAC}(\mathcal{F}), \mathcal{V}_n^{NR}(\mathcal{F})$ and $\mathcal{V}_n^{stat}(\mathcal{F})$. The no free lunch theorem says that if $\mathcal{F} = \mathcal{Y}^{\mathcal{X}}$, then, there is no convergence of minimax rates.

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

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

Proof. Consider D_X to be the uniform distribution over $2n$ points. Also let $f^* \in \mathcal{Y}^{\mathcal{X}}$ be a random choice of the possible 2^{2n} function on these points. Let $\epsilon_j = f^*(x_j)$, note that due to our random choice of f^*, ϵ_j is $+1$ with probability $1/2$ and -1 w.p. $1/2$. Now if we obtain sample S of size n, then

$$\begin{aligned}
\mathcal{V}_n^{PAC}(\mathcal{Y}^{\mathcal{X}}) &= \inf_{\hat{\mathbf{y}}} \sup_{D_X, f^* \in \mathcal{F}} \mathbb{E}_{S:|S|=n} \left[\mathbb{P}_{x \sim D_x} \left(\hat{\mathbf{y}}(x) \neq f^*(x) \right) \right] \\
&\geq \inf_{\hat{\mathbf{y}}} \mathbb{E}_{f^*} \left[\mathbb{E}_{S:|S|=n} \left[\mathbb{P}_{x \sim D_x} \left(\hat{\mathbf{y}}(x) \neq f^*(x) \right) \right] \right] \\
&= \inf_{\hat{\mathbf{y}}} \mathbb{E}_{\epsilon_1, \ldots, \epsilon_{2n}} \left[\mathbb{E}_{S:|S|=n} \left[\frac{1}{2n} \sum_{j=1}^{2n} \mathbb{1}_{\{\hat{\mathbf{y}}(x_j) \neq \epsilon_j\}} \right] \right] \\
&= \inf_{\hat{\mathbf{y}}} \mathbb{E}_{\epsilon_1, \ldots, \epsilon_{2n}} \left[\mathbb{E}_{i_1, \ldots, i_n \sim \text{Unif}[2n]} \left[\frac{1}{2n} \sum_{j=1}^{2n} \mathbb{1}_{\{\hat{\mathbf{y}}(x_j; (x_{i_1}, \epsilon_{i_1}), \ldots, (x_{i_n}, \epsilon_{i_n})) \neq \epsilon_j\}} \right] \right] \\
&\geq \frac{1}{2n} \inf_{\hat{\mathbf{y}}} \mathbb{E}_{\epsilon_1, \ldots, \epsilon_{2n}} \left[\mathbb{E}_{i_1, \ldots, i_n \sim \text{Unif}[2n]} \left[\sum_{j \notin \{i_1, \ldots, i_n\}} \mathbb{1}_{\{\hat{\mathbf{y}}(x_j; (x_{i_1}, \epsilon_{i_1}), \ldots, (x_{i_n}, \epsilon_{i_n})) \neq \epsilon_j\}} \right] \right] \\
&= \frac{1}{2n} \inf_{\hat{\mathbf{y}}} \mathbb{E}_{i_1, \ldots, i_n \sim \text{Unif}[2n]} \left[\mathbb{E}_{\epsilon_1, \ldots, \epsilon_{2n}} \left[\sum_{j \notin \{i_1, \ldots, i_n\}} \mathbb{1}_{\{\hat{\mathbf{y}}(x_j; (x_{i_1}, \epsilon_{i_1}), \ldots, (x_{i_n}, \epsilon_{i_n})) \neq \epsilon_j\}} \right] \right] \\
&= \frac{1}{2n} \inf_{\hat{\mathbf{y}}} \mathbb{E}_{i_1, \ldots, i_n \sim \text{Unif}[2n]} \left[\mathbb{E}_{\epsilon_1, \ldots, \epsilon_{2n}} \left[\sum_{j \notin \{i_1, \ldots, i_n\}} \mathbb{E}_{\epsilon_j} \left[\mathbb{1}_{\{\hat{\mathbf{y}}(x_j; (x_{i_1}, \epsilon_{i_1}), \ldots, (x_{i_n}, \epsilon_{i_n})) \neq \epsilon_j\}} \right] \right] \right] \\
&= \frac{1}{2n} \mathbb{E}_{i_1, \ldots, i_n \sim \text{Unif}[2n]} \left[\sum_{j \notin \{i_1, \ldots, i_n\}} \frac{1}{2} \right] = \frac{1}{4n} \mathbb{E}_{i_1, \ldots, i_n \sim \text{Unif}[2n]} \left[2n - |\{i_1, \ldots, i_n\}| \right] \geq \frac{1}{4}
\end{aligned}$$

□

This shows that we need some restriction on $\mathcal{F}$ even for the realizable PAC setting. We cannot learn arbitrary set of hypothesis, there is no free lunch.

2 Empirical Risk Minimization and The Empirical Process

One algorithm/principle/ learning rule that is natural for statistical learning problems is the Empirical Risk Minimizer (ERM) algorithm. That is pick the hypothesis from model class $\mathcal{F}$ that best fits the sample, or in other words,:

$$\hat{\mathbf{y}}_{\text{erm}} = \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t)$$

Claim 2. *For any $\mathcal{Y}$, $\mathcal{X}$, $\mathcal{F}$ and loss function $\ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$ (subject to mild regularity conditions required for measurability), we have that*

$$\begin{aligned}
\mathcal{V}_n^{\text{stat}}(\mathcal{F}) &\leq \sup_D \mathbb{E}_S \left[L_D(\hat{\mathbf{y}}_{\text{erm}}) - \inf_{f \in \mathcal{F}} L_D(f) \right] \\
&\leq \sup_D \mathbb{E}_S \left[\sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[\ell(f(x), y)\right] - \frac{1}{n} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right| \right]
\end{aligned}$$

Proof. Note that

$$\begin{aligned}
&\mathbb{E}_S \left[L_D(\hat{\mathbf{y}}_{\text{erm}}) \right] - \inf_{f \in \mathcal{F}} L_D(f) \\
&\qquad = \mathbb{E}_S \left[L_D(\hat{\mathbf{y}}_{\text{erm}}) \right] - \inf_{f \in \mathcal{F}} \mathbb{E}_S \left[\frac{1}{n} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right] \\
&\qquad \leq \mathbb{E}_S \left[L_D(\hat{\mathbf{y}}_{\text{erm}}) - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right] \\
&\qquad \leq \mathbb{E}_S \left[L_D(\hat{\mathbf{y}}_{\text{erm}}) - \frac{1}{n} \sum_{t=1}^{n} \ell(\hat{\mathbf{y}}_{\text{erm}}(x_t), y_t) \right]
\end{aligned}$$

since $\hat{\mathbf{y}}_{\text{erm}} \in \mathcal{F}$, we can pass to upper bound by replacing with supremum over all $f \in \mathcal{F}$ as

$$\begin{aligned}
&\leq \mathbb{E}_S \sup_{f \in \mathcal{F}} \left[\mathbb{E}\left[\ell(f(x), y)\right] - \frac{1}{n} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right] \\
&\leq \mathbb{E}_S \left[\sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[\ell(f(x), y)\right] - \frac{1}{n} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right| \right]
\end{aligned}$$

This completes the proof. □

- The question of whether the problem is learnable can now be understood by studying if, uniformly over class $\mathcal{F}$ does average converge to expected loss ?
- For bounded losses, for any fixed $f \in \mathcal{F}$, the difference of average loss and expected loss for a given $f \in \mathcal{F}$ goes to 0 by Hoeffding bound.
- The difference of average loss and expected loss is an empirical process indexed by class $\mathcal{F}$.

2.1 Example Finite Class

For now and most of the course we will assume that the loss function ℓ is bounded in magnitude by 1. Under this assumption, for any finite $\mathcal{F}$ we have the following bound.

Proposition 3. *For any loss function bounded by* 1*,*

$$\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x),y)\right]-\frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right|\right]\leq 8\sqrt{\frac{\log(n\,|\mathcal{F}|)}{n}}$$

Proof. For each $f\in\mathcal{F}$, define the random variable, Z^f as $\ell(f(x),y)$ where $(x,y)\sim D$. Notice that $\mathbb{E}\left[Z^f\right]=\mathbb{E}\left[\ell(f(x),y)\right]$ and $Z_1^f,\ldots,Z_n^f$ are iid copies of Z^f. Hence by Hoeffding's inequality, for any $\epsilon>0$,

$$P_S\left(\left|\mathbb{E}\left[Z^f\right]-\frac{1}{n}\sum_{t=1}^{n}Z_t^f\right|>\epsilon\right)\leq 2\exp\left(-\frac{n\epsilon^2}{2}\right)$$

Hence, by union bound,

$$P_S\left(\max_{f\in\mathcal{F}}\left|\mathbb{E}\left[Z^f\right]-\frac{1}{n}\sum_{t=1}^{n}Z_t^f\right|>\epsilon\right)\leq 2|\mathcal{F}|\exp\left(-\frac{n\epsilon^2}{2}\right)$$

Hence,

$$\begin{aligned}\mathbb{E}_S\left[\max_{f\in\mathcal{F}}\left|\mathbb{E}\left[Z^f\right]-\frac{1}{n}\sum_{t=1}^{n}Z_t^f\right|\right]&\leq\epsilon P_S\left(\max_{f\in\mathcal{F}}\left|\mathbb{E}\left[Z^f\right]-\frac{1}{n}\sum_{t=1}^{n}Z_t^f\right|\leq\epsilon\right)+2P_S\left(\max_{f\in\mathcal{F}}\left|\mathbb{E}\left[Z^f\right]-\frac{1}{n}\sum_{t=1}^{n}Z_t^f\right|>\epsilon\right)\\&\leq\epsilon+2P_S\left(\max_{f\in\mathcal{F}}\left|\mathbb{E}\left[Z^f\right]-\frac{1}{n}\sum_{t=1}^{n}Z_t^f\right|>\epsilon\right)\\&\leq\epsilon+4|\mathcal{F}|\exp\left(-\frac{n\epsilon^2}{2}\right)\end{aligned}$$

Picking $\epsilon=\sqrt{\frac{\log(n|\mathcal{F}|^2)}{n}}$ we can conclude the statement. □

3 Infinite Hypothesis Class : first attempt, MDL

We saw how one can get bounds for the case when $\mathcal{F}$ has finite cardinality. How about the case when $\mathcal{F}$ has infinite cardinality ? To start with, let us consider the case when $\mathcal{F}$ is a countable set. One thing we can do is to try to be smarter with the application of union bound and Hoeffding bound applied in the analysis of the finite case.

MDL Algorithm: The MDL learning rule picks the hypothesis in $\mathcal{F}$ as follows :

$$\hat{y}_{\text{mdl}}=\operatorname*{argmin}_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)+3\sqrt{\frac{\log(n/\pi^2(f))}{n}}$$

Interpretation : minimize empirical error while also ensuring that the hypothesis we pick has a large prior π. Why is this learning rule appealing ?