Machine Learning Theory (CS 6783)

Lecture 3: No Free Lunch Theorem, ERM and Uniform Rates

1 Recap

PAC framework:

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

A problem is "PAC learnable" if $\mathcal{V}_n^{PAC} \to 0$. That is, there exists a learning algorithm that converges to 0 expected error as sample size increases.
Non-parametric Regression:

$$\mathcal{V}_n^{NR}(\mathcal{F}) := \inf_{\hat{\mathbf{y}}} \sup_{D_X, f^* \in \mathcal{F}} \mathbb{E}_{S:|S|=n} \left[\mathbb{E}_{x \sim D_X} \left[(\hat{\mathbf{y}}(x) - f^*(x))^2 \right] \right]$$

A statistical estimation problem is consistent if $\mathcal{V}_n^{NR} \to 0$.
Statistical learning:

$$\mathcal{V}_n^{stat}(\mathcal{F}) := \inf_{\hat{\mathbf{y}}} \sup_{D} \mathbb{E}_{S:|S|=n} \left[L_D(\hat{\mathbf{y}}) - \inf_{f \in \mathcal{F}} L_D(f) \right]$$

A problem is "statistically learnable" if $\mathcal{V}_n^{stat} \to 0$.
Statistical learning:

$$\mathcal{V}_n^{stat}(\mathcal{F}) := \inf_{\hat{\mathbf{y}}} \sup_{D} \mathbb{E}_{S:|S|=n} \left[L_D(\hat{\mathbf{y}}) - \inf_{f \in \mathcal{F}} L_D(f) \right]$$

A problem is "statistically learnable" if $\mathcal{V}_n^{stat} \to 0$.

Proposition 1. *For any class $\mathcal{F} \subset \{\pm 1\}^{\mathcal{X}}$,*

$$4\mathcal{V}_n^{PAC}(\mathcal{F}) \leq \mathcal{V}_n^{NR}(\mathcal{F}) \leq \mathcal{V}_n^{stat}(\mathcal{F})$$

and for any $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$,

$$\mathcal{V}_n^{NR}(\mathcal{F}) \leq \mathcal{V}_n^{stat}(\mathcal{F})$$

2 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 $\mathcal{F} = \mathcal{Y}^{\mathcal{X}}$ the set of all function, then there is not convergence of minimax rates.

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

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

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. Now if we obtain sample S of size at most n, then

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

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

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

□

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

This tells us that we need to restrict the set of models $\mathcal{F}$ we consider,

3 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,:

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

Claim 3. *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*

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

Proof. Note that

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

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

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

This completes the proof. □

- The question of whether minimax value converges to 0, or equivalently 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 $\mathcal{F}$. We study supremum (over $\mathcal{F}$) of these empirical processes. This is the main question of interest in empirical process theory.

3.1 Example: Finite Class

For now and for most of this course we shall assume that the loss ℓ is bounded by 1, that is $\sup_{y,y'\in\mathcal{Y}} |\ell(y',y)| \le 1$.

Claim 4. *Consider the case when the hypothesis $\mathcal{F}$ has finite cardinality, that is $|\mathcal{F}| < \infty$. For any loss ℓ bounded by 1, we have that*

$$\mathcal{V}_n^{\text{stat}}(\mathcal{F}) \le \sup_D \mathbb{E}_S\left[L_D(\hat{y}_{\text{erm}}) - \inf_{f\in\mathcal{F}} L_D(f)\right] \le 8\sqrt{\frac{\log n|\mathcal{F}|}{n}}$$

Proof. By Claim 3 we have that

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

Now let us define for every $f \in \mathcal{F}$, the random variable Z^f as $\ell(f(x), y)$ where $(x, y) \sim D$. Note that $\mathbb{E}\left[Z^f\right] = \mathbb{E}\left[\ell(f(x), y)\right]$ and $Z_1^f, \ldots, Z_n^f$ are n iid copies of Z^f. Hence by Hoeffding's inequality we have that:

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

Hence by union bound we have that

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

Hence,

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

Choosing $\epsilon = \sqrt{\log(n|\mathcal{F}|^2)/n}$ we get,

$$\mathbb{E}\left[\max_{f\in\mathcal{F}}\left|\mathbb{E}\left[Z^f\right] - \frac{1}{n}\sum_{t=1}^{n} Z_t^f\right|\right] \leq 8\sqrt{\frac{\log n|F|}{n}}$$

□

Hence for a finite class $\mathcal{F}$, using the ERM algorithm, one can achieve a rate of $O\left(\sqrt{\frac{\log n|F|}{n}}\right)$ (in fact the $\sqrt{\log n}$ factor can be shaved off with more careful analysis).

What about infinite class $\mathcal{F}$?