Machine Learning Theory (CS 6783)

Lecture 4 : Statistical Learning

1 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 1. *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

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

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

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

This completes the proof. □

- The question of whether 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.

1.1 Finite Class

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

Claim 2. *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*

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

Proof. By Claim 1 we have that

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

Now note that

$$\begin{aligned}
\mathcal{V}_n^{\text{stat}}(\mathcal{F}) &\leq \mathbb{E}_S \left[\sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[\ell(f(x), y)\right] - \frac{1}{n} \sum_{t=1}^n \ell(f(x_t), y_t) \right| \right] \\
&= \mathbb{E}_S \left[\mathbb{1}_{\left\{\sup_{f \in \mathcal{F}} \left| \mathbb{E}[\ell(f(x),y)] - \frac{1}{n}\sum_{t=1}^n \ell(f(x_t),y_t) \right| \leq \epsilon\right\}} \sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[\ell(f(x), y)\right] - \frac{1}{n} \sum_{t=1}^n \ell(f(x_t), y_t) \right| \right] \\
&\quad + \mathbb{E}_S \left[\mathbb{1}_{\left\{\sup_{f \in \mathcal{F}} \left| \mathbb{E}[\ell(f(x),y)] - \frac{1}{n}\sum_{t=1}^n \ell(f(x_t),y_t) \right| > \epsilon\right\}} \sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[\ell(f(x), y)\right] - \frac{1}{n} \sum_{t=1}^n \ell(f(x_t), y_t) \right| \right] \\
&\leq \epsilon + \mathbb{E}_S \left[\mathbb{1}_{\left\{\sup_{f \in \mathcal{F}} \left| \mathbb{E}[\ell(f(x),y)] - \frac{1}{n}\sum_{t=1}^n \ell(f(x_t),y_t) \right| > \epsilon\right\}} \sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[\ell(f(x), y)\right] - \frac{1}{n} \sum_{t=1}^n \ell(f(x_t), y_t) \right| \right] \\
&\leq \epsilon + 2P\left(\sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[\ell(f(x), y)\right] - \frac{1}{n} \sum_{t=1}^n \ell(f(x_t), y_t) \right| > \epsilon \right) \qquad (1)
\end{aligned}$$

Now note that for any fixed $f \in \mathcal{F}$, by Hoeffding bound,

$$P\left(\left| \mathbb{E}\left[\ell(f(x), y)\right] - \frac{1}{n} \sum_{t=1}^n \ell(f(x_t), y_t) \right| > \epsilon \right) \leq 2\exp\left(-\frac{\epsilon^2 n}{2}\right)$$

Hence by union bound :

$$P\left(\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x),y)\right]-\frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right|>\epsilon\right)\leq 2|\mathcal{F}|\exp\left(-\frac{\epsilon^2 n}{2}\right)$$

Plugging the above into Equation 1 we conclude that,

$$\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x),y)\right]-\frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right|\right]\leq\epsilon+4|\mathcal{F}|\exp\left(-\frac{\epsilon^2 n}{2}\right)$$

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

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

□

Thus we see that for any finite class $\mathcal{F}$, the minimax rate is in fact $O^*\left(\sqrt{\log|\mathcal{F}|/n}\right)$. It is easy to in fact show that the rate is order $\sqrt{\frac{\log|\mathcal{F}|}{n}}$, that is without the extra $\log n$. Think about how to show this !

2 MDL bound (Occam's Razor Principle)

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.

Claim 3. *For any countable set $\mathcal{F}$, any fixed distribution π on $\mathcal{F}$,*

$$\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left\{\left|L_D(f)-\frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right|-\sqrt{\frac{\log(n/\pi^2(f))}{n}}\right\}\right]\leq\frac{4}{\sqrt{n}}$$

Proof. The basic idea is to use Hoeffding bound along with union bound as before, but instead of using same ϵ for every $f\in\mathcal{F}$ in Hoeffding bound, we use f specific $\epsilon(f)$. We shall specify the exact form of $\epsilon(f)$ later. For now note that, since the losses are bounded by 1,

$$\sup_{f\in\mathcal{F}}\left\{\left|L_D(f)-\frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right|-\epsilon(f)\right\}\leq 0+2\,\mathbb{1}_{\left\{\sup_{f\in\mathcal{F}}\left\{\left|L_D(f)-\frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right|-\epsilon(f)>0\right\}\right\}}$$

Hence, taking expectation w.r.t. sample we have that

$$\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left\{\left|L_D(f)-\frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right|-\epsilon(f)\right\}\right]\leq 2P\left(\sup_{f\in\mathcal{F}}\left\{\left|L_D(f)-\frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right|-\epsilon(f)>0\right\}\right)$$

By Hoeffding inequality, for any fixed $f \in \mathcal{F}$

$$P\left(\left|\mathbb{E}\left[\ell(f(x),y)\right] - \frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right| - \epsilon(f) > 0\right) \leq 2\exp\left(-\frac{\epsilon^2(f)n}{2}\right)$$

Taking union bound we have,

$$P\left(\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x),y)\right] - \frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right| - \epsilon(f) > 0\right) \leq \sum_{f\in\mathcal{F}} 2\exp\left(-\frac{\epsilon^2(f)n}{2}\right)$$

Hence we conclude that

$$\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left\{\left|L_D(f) - \frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right| - \epsilon(f)\right\}\right] \leq 4\sum_{f\in\mathcal{F}}\exp\left(-\frac{\epsilon^2(f)n}{2}\right)$$

For the prior choice of π of distribution over set $\mathcal{F}$, let us use

$$\epsilon(f) = \sqrt{\frac{\log(n/\pi^2(f))}{n}}$$

Hence we can conclude that,

$$\begin{aligned}\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left\{\left|L_D(f) - \frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right| - \sqrt{\frac{\log(n/\pi^2(f))}{n}}\right\}\right] &\leq 4\sum_{f\in\mathcal{F}}\exp\left(-\frac{\epsilon^2(f)n}{2}\right) \\ &\leq \frac{4\sum_f \pi(f)}{\sqrt{n}} = \frac{4}{\sqrt{n}}\end{aligned}$$

□

The above claim provides us an intuition for MDL principle, 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 staying close to prior π. Why is this learning rule appealing ?

Let us use the claim above to analyze the learning rule. Note that from the above claim, we have that,

$$\mathbb{E}_S\left[L_D(\hat{y}_{\text{mdl}}) - \frac{1}{n}\sum_{t=1}^{n}\ell(\hat{y}_{\text{mdl}}(x_t),y_t) - \sqrt{\frac{\log(n/\pi^2(\hat{y}_{\text{mdl}}))}{n}}\right] \leq \frac{4}{\sqrt{n}}$$

By definition of $\hat{y}_{\text{mdl}}$ we can conclude that

$$\mathbb{E}_S\left[L_D(\hat{y}_{\text{mdl}}) - \inf_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}\ell(\hat{y}_{\text{mdl}}(x_t),y_t) + \sqrt{\frac{\log(n/\pi^2(\hat{y}_{\text{mdl}}))}{n}}\right\}\right] \leq \frac{4}{\sqrt{n}}$$

In other words,

$$\mathbb{E}_S\left[L_D(\hat{y}_{\text{mdl}})\right] \leq \mathbb{E}_S\left[\inf_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)+\sqrt{\frac{\log(n/\pi^2(f))}{n}}\right\}\right]+\frac{4}{\sqrt{n}}$$

Let $f_D = \operatorname*{argmin}_{f\in\mathcal{F}} L_D(f)$, replacing the infimum above we conclude that

$$\begin{aligned}
\mathbb{E}_S\left[L_D(\hat{y}_{\text{mdl}})\right] &\leq \mathbb{E}_S\left[\frac{1}{n}\sum_{t=1}^{n}\ell(f_D(x_t),y_t)+\sqrt{\frac{\log(n/\pi^2(f_D))}{n}}\right]+\frac{4}{\sqrt{n}} \\
&= L_D(f_D)+\sqrt{\frac{\log(n/\pi^2(f_D))}{n}}+\frac{4}{\sqrt{n}} \\
&= \inf_{f\in\mathcal{F}} L_D(f)+\sqrt{\frac{\log(n/\pi^2(f_D))}{n}}+\frac{4}{\sqrt{n}} \qquad (2)
\end{aligned}$$

(3)

Thus with the above bound, even for countably infinite $\mathcal{F}$ we can get bounds on $\mathbb{E}_S\left[L_D(\hat{y})\right] - \inf_{f\in\mathcal{F}} L_D(f)$ that decreases with n.