Machine Learning Theory (CS 6783)

Lecture 3 : Statistical Learning

1 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 $\mathcal{X} = \{\}$) and $\mathcal{Y} = \{\pm 1\}$. We receive ± 1 valued samples $y_1, \ldots, y_n \in \{\pm 1\}$ drawn iid from Bernoullis distribution with parameter p (ie. Y is $+1$ with probability p and -1 with probability $1-p$). Our loss function is the zero-one loss function $\ell(y', y) = \mathbb{1}_{\{y' \neq y\}}$. 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 1. *For the problem above, one can bound the minimax rate as:*

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

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

Proof. Now note that :

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

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

$$\begin{aligned} L_p(\hat{y}) - \inf_{f \in \{\pm 1\}} L_p(f) &= |1-2p|\, \mathbb{1}_{\{\hat{y} \neq \mathrm{sign}(2p-1)\}} \\ &\leq \epsilon + \, \mathbb{1}_{\{\hat{y} \neq \mathrm{sign}(2p-1)\}}\, \mathbb{1}_{\{|1-2p| > \epsilon\}} \end{aligned}$$

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

$$\begin{aligned} L_p(\hat{y}) - \inf_{f \in \{\pm 1\}} L_p(f) &\leq \epsilon + \, \mathbb{1}_{\{\mathrm{sign}\left(\frac{1}{n}\sum_{t=1}^n y_t\right) \neq \mathrm{sign}(2p-1)\}}\, \mathbb{1}_{\{|1-2p| > \epsilon\}} \\ &\leq \epsilon + \, \mathbb{1}_{\{\mathrm{sign}\left(\frac{1}{n}\sum_{t=1}^n y_t\right) \neq \mathrm{sign}(2p-1) \& |1-2p| > \epsilon\}} \\ &\leq \epsilon + \, \mathbb{1}_{\{\left|\frac{1}{n}\sum_{t=1}^n y_t - (2p-1)\right| > \epsilon\}} \end{aligned}$$

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

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

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

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

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 :

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

□

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.

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,:

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

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*

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

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.

2.1 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 3. *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 2 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

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

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

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

Hence by union bound :

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

Plugging the above into Equation 1 we conclude that,

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

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

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

□

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 !

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

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

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,

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

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

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

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

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

Taking union bound we have,

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

Hence we conclude that

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

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,

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

□

The above claim provides us an intuition for MDL principle, the MDL learning rule picks the hypothesis in $\mathcal{F}$ as follows :

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

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,

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

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

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

In other words,

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

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

$$\begin{align}
\mathbb{E}_S\left[L_D(\hat{y}_{\mathrm{mdl}})\right] &\le \mathbb{E}_S\left[\frac{1}{n}\sum_{t=1}^{n} \ell(f_D(x_t), y_t) + \sqrt{\frac{\log(n/\pi^2(f_D))}{n}}\right] + \frac{4}{\sqrt{n}} \notag\\
&= L_D(f_D) + \sqrt{\frac{\log(n/\pi^2(f_D))}{n}} + \frac{4}{\sqrt{n}} \notag\\
&= \inf_{f\in\mathcal{F}} L_D(f) + \sqrt{\frac{\log(n/\pi^2(f_D))}{n}} + \frac{4}{\sqrt{n}} \tag{2}\\
& \tag{3}
\end{align}$$

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.

4 Universal Vs Uniform Learning

Why does Equation 2 not contradict No Free Lunch Theorems ? We saw in previous class that even for simple binary classification problems (even in realizable setting) if $|\mathcal{X}| > 2n$ then $\mathcal{F} = \mathcal{Y}^{\mathcal{X}}$ is not learnable. However the bound in Equation 2 says that even for very rich classes like any countably infinite class, by picking appropriate prior distribution π one can get that for any distribution D, $\mathbb{E}_S\left[L_D(\hat{y})\right] - \inf_{f\in\mathcal{F}} L_D(f)$ goes to 0.

Of course there is no contradiction here. While bound in Equation 2 does say that we can learn so that the expected loss of our hypothesis converges to the expected loss of the optimal predictor $f_D \in \mathcal{F}$, it does not converge at a uniform rate for over distributions D. That is the number of samples required to obtain accuracy ϵ for any distribution D depends on this distribution D. This type of learnability we call universal but not uniform learnability. The statement is true for every distribution but not at uniform rate. This as opposed to uniform learnability is when we get a uniform rate for any distribution D which is what is measured by the minimax rate $\mathcal{V}_n^{\text{stat}}(\mathcal{F})$. While universal learnability is rather nice, it is not really satisfactory. This is because we can't really guarantee how many samples we need to reach a given accuracy but only that eventually we will.

5 Sneak Peek

In todays class we asw how to get learnability for finite hypothesis classes. We also defined MDL strategy that provides non-uniform rates for (countably) infinite classes. However the analysis was unsatisfactory as the rates don't tell us as much in the form of a guarantee in that we can only say eventually accuracy improves to better than set ϵ.

In next class we well see how to truly deal with infinite classes. We will go over the technique of symmetrization and learn about the much acclaimed VC dimension.