Machine Learning Theory (CS 6783)

Lecture 5 : Statistical Learning

1 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 1. *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

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

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.

2 Universal Vs Uniform Learning

For the MDL algorithm we saw that,

$$\mathbb{E}_S\left[L_D(\hat{y}_{\text{mdl}})\right] \leq \inf_{f\in\mathcal{F}} L_D(f) + \sqrt{\frac{\log(n/\pi^2(f_D))}{n}} + \frac{4}{\sqrt{n}} \tag{3}$$

$$\tag{4}$$

where $f_D = \underset{f\in\mathcal{F}}{\operatorname{argmin}}\, L_D(f)$.

Why does Equation 1 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 1 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 1 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.

3 Infinite Hypothesis Class : first attempt

As a first attempt, one can think of approximating the function class to desired accuracy by a finite number of representative elements. We call this a point-wise cover.

Definition 1. *We say that set $\mathcal{F}_\epsilon = \{\tilde{f}_1, \ldots, \tilde{f}_N\}$ is an ϵ point-wise cover for function class $\mathcal{F}$ if $\forall f \in \mathcal{F}$ there exists $i \in [N]$ s.t.*

$$\sup_{x,y} |\ell(f(x), y) - \ell(\tilde{f}_i(x), y)| \leq \epsilon$$

Further define $N(\epsilon)$ to be the smallest N such that there exists an ϵ cover of $\mathcal{F}$ of cardinality at most N.

Claim 2. *For any function class $\mathcal{F}$, we have that*

$$\mathcal{V}_n^{\text{stat}}(\mathcal{F}) \leq \inf_{\epsilon>0}\left\{4\epsilon + \sqrt{\frac{\log N(\epsilon)}{n}}\right\}$$

Proof. Let $\mathcal{F}_\epsilon = \{\tilde{f}_1, \ldots, \tilde{f}_{N(\epsilon)}\}$ be an ϵ cover for the function class $\mathcal{F}$. Further for every $f \in \mathcal{F}$, let $i(f)$ correspond to the index of the element in $\mathcal{F}_\epsilon$ that is ϵ close to that f. Now note that,

$$\begin{aligned}
&\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left\{\mathbb{E}\left[\ell(f(x),y)\right] - \frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right\}\right] \\
&\qquad \le \mathbb{E}_S\left[\sup_{i\in[N_\epsilon]}\left\{\mathbb{E}\left[\ell(\tilde{f}_i(x),y)\right] - \frac{1}{n}\sum_{t=1}^{n}\ell(\tilde{f}_i(x_t),y_t)\right\}\right] \\
&\qquad\quad + \mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x),y)\right] - \frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t) - \mathbb{E}\left[\ell(\tilde{f}_{i(f)}(x),y)\right] + \frac{1}{n}\sum_{t=1}^{n}\ell(\tilde{f}_{i(f)}(x_t),y_t)\right|\right] \\
&\qquad \le \sqrt{\frac{\log N(\epsilon)}{n}} + 4\epsilon
\end{aligned}$$

where the first term in the last inequality is by using the finite class bound and the second term is by using the definition of ϵ cover as $\tilde{f}_{i(f)}$ is ϵ close to f. Since choice of ϵ was arbitrary we can take the infimum over choices of ϵ to conclude the proof. □

Example : linear predictor, absolute loss, 1 dimension
$f(x) = f \cdot x, \quad \mathcal{F} = \mathcal{X} = [-1,1], \quad \mathcal{Y} = [-1,1], \quad \ell(y',y) = |y - y'|$

$N_\epsilon = \frac{2}{\epsilon}$, Cover given by $f_1 = -1, f_2 = -1+\epsilon, \ldots, f_{N_\epsilon - 1} = 1-\epsilon, f_{N_\epsilon} = 1$.

$$V_n^{\text{stat}}(\mathcal{F}) \le \sqrt{\frac{\log n}{n}}$$

Example : linear predictor/loss, d dimensions
$f(x) = \mathbf{f}^\top \mathbf{x}$. $\mathcal{F} = \mathcal{X} = \{\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\|_2 \le 1\}$. $\mathcal{Y} = [-1,1]$. $\ell(y',y) = y \cdot y'$

$N_\epsilon = \Theta\left(\frac{2}{\epsilon}\right)^d$

$$V_n^{\text{stat}}(\mathcal{F}) \le \sqrt{\frac{d \log n}{n}}$$

Example : thresholds
$f(x) = \text{sign}(f - x), \quad \mathcal{F} = \mathcal{X} = [-1,1], \quad \mathcal{Y} = \{-1,1\}, \quad \ell(y',y) = \mathbb{1}_{\{y \ne y'\}}, \quad N_\epsilon = \infty$ for any $\epsilon < 1$.