Machine Learning Theory (CS 6783)

Lecture 4 : Statistical Learning

1 Recap

Last class we showed that

$$\mathcal{V}_n^{\text{stat}}(\mathcal{F}) \le \sup_D \mathbb{E}_S \left[\sup_{f \in \mathcal{F}} \left\{ \mathbb{E}\left[\ell(f(x), y)\right] - \frac{1}{n} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right\} \right]$$

This was using the Empirical Risk Minimizer (ERM)

1. When $|\mathcal{F}| < \infty$, using the above we showed that

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

2. For countably infinite class we showed MDL bound and the algorithm based on this bound.

3. However the learning rate was not uniform over $\mathcal{F}$

Can we get rate uniform over $\mathcal{F}$ for infinite classes $\mathcal{F}$?

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

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

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

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

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,

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

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. $\square$

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 $f_1 = -1, f_2 = -1+\epsilon, \ldots, f_{N_\epsilon - 1} = 1-\epsilon, f_{N_\epsilon} = 1$.

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

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

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

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

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

3 Symmetrization and Rademacher Complexity

$$
\begin{aligned}
&\mathbb{E}_S\left[L_D(\hat{y}_{\mathrm{erm}})\right] - \inf_{f\in\mathcal{F}} L_D(f) \\
&\qquad \le \mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left\{\mathbb{E}\left[\ell(f(x),y)\right] - \frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right\}\right] \\
&\qquad \le \mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}\ell(f(x'_t),y'_t) - \frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right\}\right] \\
&\qquad = \mathbb{E}_{S,S'}\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}\epsilon_t(\ell(f(x'_t),y'_t) - \ell(f(x_t),y_t))\right\}\right] \\
&\qquad \le 2\mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\ell(f(x_t),y_t)\right\}\right] \\
&\qquad =: \mathcal{R}_n(\mathcal{F})
\end{aligned}
$$

Where in the above each ϵ_t is a Rademacher random variable that is $+1$ with probability $1/2$ and -1 with probability $1/2$. The above is called Rademacher complexity of the loss class $\ell \circ \mathcal{F}$. In general Rademacher complexity of a function class measures how well the function class correlates with random signs. The more it can correlate with random signs the more complex the class is.

Example : linear predictor/loss, d dimensions

$$
\begin{aligned}
\mathbb{E}_S\left[L_D(\hat{y})\right] - \inf_{f\in\mathcal{F}} L_D(f) &\le 2\mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\ell(f(x_t),y_t)\right\}\right] \\
&= 2\mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{\mathbf{f}:\|\mathbf{f}\|_2\le 1}\left\{\frac{1}{n}\sum_{t=1}^{n}\epsilon_t y_t \mathbf{f}^\top \mathbf{x}_t)\right\}\right] \\
&= \frac{2}{n}\mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{\mathbf{f}:\|\mathbf{f}\|_2\le 1}\left\{\mathbf{f}^\top\left(\sum_{t=1}^{n}\epsilon_t y_t \mathbf{x}_t\right)\right\}\right] \\
&= \frac{2}{n}\mathbb{E}_S\mathbb{E}_\epsilon\left[\left\|\sum_{t=1}^{n}\epsilon_t y_t \mathbf{x}_t\right\|\right]_2 \\
&\le \frac{2}{n}\mathbb{E}_S\sqrt{\mathbb{E}_\epsilon\left[\left\|\sum_{t=1}^{n}\epsilon_t y_t \mathbf{x}_t\right\|_2^2\right]} \\
&= \frac{2}{n}\mathbb{E}_S\sqrt{\mathbb{E}_\epsilon\left[\sum_{t=1}^{n}\|\epsilon_t y_t \mathbf{x}_t\|_2^2 + \sum_{i,j:i\neq j}\epsilon_i y_i \mathbf{x}_i \epsilon_j y_j \mathbf{x}_j\right]} \\
&= \frac{2}{n}\mathbb{E}_S\sqrt{\sum_{t=1}^{n}\|y_t \mathbf{x}_t\|_2^2} \le \frac{2}{\sqrt{n}}
\end{aligned}
$$

4 Infinite $\mathcal{F}$: Binary Classes and Growth Function

First let us simplify the Rademacher complexity for binary classification problem. Note that for binary classification problem where $\mathcal{Y} \in \{\pm 1\}$, the loss can be rewritten as $\ell(y', y) = \mathbb{1}_{\{y \neq y'\}} = \frac{1 - y \cdot y'}{2}$. Hence

$$\begin{aligned} 2\mathbb{E}_S \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{t=1}^n \epsilon_t \ell(f(x_t), y_t) \right\} \right] &= 2\mathbb{E}_S \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{t=1}^n \epsilon_t \frac{1 - f(x_t) \cdot y_t}{2} \right\} \right] \\ &= \mathbb{E}_S \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n \epsilon_t y_t f(x_t) \right] \end{aligned}$$

Now consider the inner term in the expectation above, ie. $\mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n \epsilon_t y_t f(x_t) \right]$. Note that given any fixed choice of $y_1, \ldots, y_n \in \{\pm 1\}$, $\epsilon_1 y_1, \ldots, \epsilon_n y_n$ are also Rademahcer random variables. Hence for the binary classification problem,

$$2\mathbb{E}_S \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{t=1}^n \epsilon_t \ell(f(x_t), y_t) \right\} \right] = \mathbb{E}_S \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n \epsilon_t f(x_t) \right]$$

In the above statement we moved from Rademacher complexity of loss class $\ell \circ \mathcal{F}$ to the Rademacher complexity of the function class $\mathcal{F}$ for binary classification task. This is a precursor to what we will refer to as contraction lemma which we will show later.

Why is the introduction of Rademacher averages important ? To analyze the term, $\mathbb{E}_S \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n \epsilon_t f(x_t) \right]$ consider the inner expectation, that is conditioned on sample consider the term $\mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n \epsilon_t f(x_t) \right]$. Note that $\frac{1}{n} \sum_{t=1}^n \epsilon_t f(x_t)$ is still average of 0 mean random variables and we can apply Hoeffding bound for each fixed $f \in \mathcal{F}$ individually. Now $\mathcal{F}$ might be an infinite class, but, conditioned on input instances $x_1, \ldots, x_n$, one can ask, what is the size of the projection set

$$\mathcal{F}_{|x_1, \ldots, x_n} = \{f(x_1), \ldots, f(x_n) : f \in \mathcal{F}\}$$

For any binary class $\mathcal{F}$, first note that this set can have a maximum cardinality of 2^n however it could be much smaller. In fact we can have,

$$\mathbb{E}_S \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n \epsilon_t f(x_t) \right] = \mathbb{E}_S \mathbb{E}_\epsilon \left[\sup_{\mathbf{f} \in \mathcal{F}_{|x_1, \ldots, x_n}} \frac{1}{n} \sum_{t=1}^n \epsilon_t \mathbf{f}[t] \right] \leq \mathbb{E}_S \left[\sqrt{\frac{\log |\mathcal{F}_{|x_1, \ldots, x_n}|}{n}} \right]$$

where the last step is using Proposition 2 which we shall prove in a bit. Now one can define the growth function for a hypothesis class $\mathcal{F}$ as follows.

$$\Pi_{\mathcal{F}}(n) = \sup\{|\mathcal{F}_{|x_1, \ldots, x_n}| : x_1, \ldots, x_n \in \mathcal{X}\}$$

Hence we conclude that

$$\mathcal{V}_n(\mathcal{F}) \leq \sqrt{\frac{\log \Pi_{\mathcal{F}}(n)}{n}}$$

Example : thresholds

What does the growth function of the class of threshold function look like ?

Well sort any given n points in ascending order, using thresholds, we can get at most $n+1$ possible labeling on the n points. Hence $\Pi_{\mathcal{F}}(n) = n+1$. From this we conclude that for the learning thresholds problem,

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