Machine Learning Theory (CS 6783)

Lecture 6 : Symmetrization, Rademacher Complexity, Growth function

1 Recap

In an earlier lecture we proved that

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

Last class we tried to use the above for infinite classes by approximating the function class uniformly by a finite class with cardinality $N(\epsilon)$ at scale ϵ. Let us review a specific example:

Example : linear predictor/loss, d dimensions
$f(x) = \mathbf{f}^\top\mathbf{x}$. $\mathcal{F} = \mathcal{X} = \{\mathbf{v}\in\mathbb{R}^d : \|\mathbf{v}\|_2 \le 1\}$. $\mathcal{Y} = [-1,1]$. $\ell(y',y) = y\cdot y'$, $N_\epsilon = \Theta\left(\frac{2}{\epsilon}\right)^d$

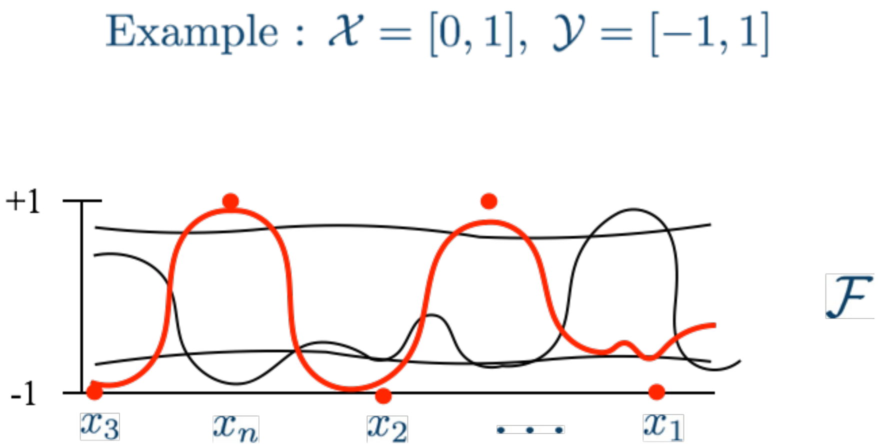
$$V_n^{\mathrm{stat}}(\mathcal{F}) \le \sqrt{\frac{d\log n}{n}}$$

Is this the best we can do? What if $d\to\infty$, in this case is the function class not learnable?

2 Symmetrization and Rademacher Complexity

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

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, dimension free bound

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

3 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

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

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.

4 Sneak Peek

Notice that $\Pi_{\mathcal{F}}(n) \leq 2^n$ for any binary function class $\mathcal{F}$ since there are at most 2^n possible ways to label n points. However it could be smaller. What we will see in the next class, is the notion of VC dimension. One of the fundamental quantities in learning theory.

VC dimension : size of largest set of input instances we can shattered

$$\text{VC}(\mathcal{F}) = \max\{d : \Pi_{\mathcal{F}}(d) = 2^d\}$$