

Machine Learning Theory (CS 6783)

Lecture 11 : Wrapping-up Supervised Statistical Learning

1 Recap

1. For any statistical learning problem we have,

$$\mathbb{E}_S \left[L_D(\hat{y}_{\text{erm}}) - \inf_{f \in \mathcal{F}} L_D(f) \right] \leq \frac{2}{n} \mathbb{E}_S \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^n \epsilon_t \ell(f(x_t), y_t) \right] = 2 \mathbb{E}_S \left[\hat{\mathcal{R}}_S(\ell \circ \mathcal{F}) \right]$$

2. For any L -Lipchitz loss

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

3. Covering : V is an ℓ_p -cover of $\mathcal{F}$ on $x_1, \dots, x_n$ at scale β if

$$\forall f \in \mathcal{F}, \exists \mathbf{v} \in V \text{ s.t. } \left(\frac{1}{n} \sum_{t=1}^n |f(x_t) - \mathbf{v}[t]|^p \right)^{1/p} \leq \beta$$

$$\mathcal{N}_p(\mathcal{F}, \beta; x_1, \dots, x_n) = \min\{|V| : V \text{ is an } \ell_p\text{-cover of } \mathcal{F} \text{ on } x_1, \dots, x_n \text{ at scale } \beta\}$$

4. Pollard bound:

$$\mathbb{E}_S \left[L_D(\hat{y}_{\text{erm}}) - \inf_{f \in \mathcal{F}} L_D(f) \right] \leq 2 \mathbb{E}_S \left[\hat{\mathcal{R}}_S(\mathcal{F}) \right] \leq 2 \inf_{\beta > 0} \left\{ \beta + \sqrt{\frac{\log \mathcal{N}_1(\mathcal{F}, \beta; x_1, \dots, x_n)}{n}} \right\}$$

5. Dudley Integral bound:

$$\hat{\mathcal{R}}_S(\mathcal{F}) \leq \hat{D}_S(\mathcal{F}) := \inf_{\alpha > 0} \left\{ 4\alpha + 12 \int_\alpha^1 \sqrt{\frac{\log \mathcal{N}_2(\mathcal{F}, \beta; x_1, \dots, x_n)}{n}} d\beta \right\}$$

2 Sudakov's Theorem and Partial Converse

Theorem 1. *There is a universal constant $c > 0$ such that*

$$\hat{\mathcal{R}}_S(\mathcal{F}) \geq \frac{c}{\log n} \sup_{\alpha > 0} \alpha \sqrt{\frac{\log \mathcal{N}_2(\mathcal{F}, \alpha, x_1, \dots, x_n)}{n}}$$

The above theorem (paraphrased) is due to Sudakov. We shall not go over its proof.

Theorem 2.

$$\frac{c}{12 \log^2 n} \left(\mathcal{D}_S(\mathcal{F}) - \frac{4}{n} \right) \leq \hat{\mathcal{R}}_S(\mathcal{F}) \leq \mathcal{D}_S(\mathcal{F})$$

Proof. We already showed that $\hat{\mathcal{R}}_S(\mathcal{F}) \leq \mathcal{D}_S(\mathcal{F})$. Now on the other hand, we have

$$\mathcal{D}_S(\mathcal{F}) = \inf_{\alpha > 0} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^1 \sqrt{\log(\mathcal{N}_2(\mathcal{F}, \delta, n))} d\delta \right\}$$

However by Sudakov's theorem we have that for any $\delta > 0$, we have

$$\sqrt{\frac{\log \mathcal{N}_2(\mathcal{F}, \delta, x_1, \dots, x_n)}{n}} \leq \frac{\log n \hat{\mathcal{R}}_S(\mathcal{F})}{c \delta}$$

Using this,

$$\begin{aligned} \mathcal{D}_S(\mathcal{F}) &\leq \inf_{\alpha > 0} \left\{ 4\alpha + \frac{12}{c} \log n \hat{\mathcal{R}}_S(\mathcal{F}) \int_{\alpha}^1 \frac{1}{\delta} d\delta \right\} \\ &= \inf_{\alpha > 0} \left\{ 4\alpha + \frac{12}{c} \log n \log(1/\alpha) \hat{\mathcal{R}}_S(\mathcal{F}) \right\} \end{aligned}$$

Picking $\alpha = \frac{1}{n}$ we conclude that $\mathcal{D}_S(\mathcal{F}) \leq \frac{4}{n} + \frac{12}{c} \log^2 n \hat{\mathcal{R}}_S(\mathcal{F})$

□

3 Lower Bounds on Supervised Learning for $\mathcal{Y} \subset \mathbb{R}$

Basic idea : To show lower bound, we pick $k \cdot n$ points $x_1, \dots, x_{kn}$ and signs $\epsilon_1, \dots, \epsilon_{kn}$. The signs are not revealed to the learner. We use the uniform distribution over the kn pairs of instances as the distribution D . That is $D = \text{Unif}\{(x_1, \epsilon_1), \dots, (x_{kn}, \epsilon_{kn})\}$. Learner can even know this fact, only learner does not get to see the ϵ_t 's before hand. Now we sample n points from this distribution and provide this to the learner. Clearly the learner sees at most n labels and so on the the remaining $kn - n$ points learner has no way to predict anything meaningful. The rest is simply massaging the math.

We shall consider the absolute loss $\ell(y', y) = |y - y'|$. However similar analysis can be extended to other commonly used supervised learning losses (called margin losses) like all ℓ_p losses, logistic loss, hinge loss etc.

Lemma 3. For any class $\mathcal{F} \subset [-1, 1]^{\mathcal{X}}$ and for any $k \in \mathbb{N}$,

$$\mathcal{V}_n^{\text{proper}}(\mathcal{F}) \geq \mathcal{R}_{kn} - \frac{1}{k} \mathcal{R}_n(\mathcal{F}) \quad \text{and} \quad \mathcal{V}_n^{\text{improper}}(\mathcal{F}) \geq \mathcal{R}_{kn} - \frac{1}{k}$$

Proof.

$$\begin{aligned} &\inf_{\hat{y}} \sup_D \mathbb{E}_S \left[L_D(\hat{y}) - \inf_{f \in \mathcal{F}} L_D(f) \right] \\ &\geq \inf_{\hat{y}} \sup_{x_1, \dots, x_{kn}} \inf_{\epsilon_1, \dots, \epsilon_{kn}} \mathbb{E}_{S \sim \text{Unif}\{(x_1, \epsilon_1), \dots, (x_{kn}, \epsilon_{kn})\}} \left[\frac{1}{kn} \sum_{t=1}^{kn} |\hat{y}_S(x_t) - \epsilon_t| - \inf_{f \in \mathcal{F}} \frac{1}{kn} \sum_{t=1}^{kn} |f(x_t) - \epsilon_t| \right] \\ &\geq \sup_{x_1, \dots, x_{kn}} \inf_{\hat{y}} \inf_{\epsilon_1, \dots, \epsilon_{kn}} \mathbb{E}_{S \sim \text{Unif}\{(x_1, \epsilon_1), \dots, (x_{kn}, \epsilon_{kn})\}} \left[\frac{1}{kn} \sum_{t=1}^{kn} |\hat{y}_S(x_t) - \epsilon_t| - \inf_{f \in \mathcal{F}} \frac{1}{kn} \sum_{t=1}^{kn} |f(x_t) - \epsilon_t| \right] \end{aligned}$$

For any $y' \in [-1, 1]$, $|y' - \epsilon_t| = 1 - y'\epsilon_t$ and so,

$$\begin{aligned}
&= \sup_{x_1, \dots, x_{kn}} \inf_{\hat{y}} \mathbb{E}_{\epsilon_1, \dots, \epsilon_{kn}} \mathbb{E}_{S \sim \text{Unif}\{(x_1, \epsilon_1), \dots, (x_{kn}, \epsilon_{kn})\}} \left[\frac{1}{kn} \sum_{t=1}^{kn} -\epsilon_t \hat{y}_S(x_t) - \inf_{f \in \mathcal{F}} \frac{1}{kn} \sum_{t=1}^{kn} -\epsilon_t f(x_t) \right] \\
&= \sup_{x_1, \dots, x_{kn}} \left\{ \inf_{\hat{y}} \mathbb{E}_S \mathbb{E}_\epsilon \left[\frac{1}{kn} \sum_{t=1}^{kn} -\epsilon_t \hat{y}_S(x_t) \right] - \mathbb{E}_\epsilon \left[\inf_{f \in \mathcal{F}} \frac{1}{kn} \sum_{t=1}^{kn} -\epsilon_t f(x_t) \right] \right\} \\
&= \sup_{x_1, \dots, x_{kn}} \left\{ \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{kn} \sum_{t=1}^{kn} \epsilon_t f(x_t) \right] - \sup_{\hat{y}} \mathbb{E}_S \mathbb{E}_\epsilon \left[\frac{1}{kn} \sum_{t=1}^{kn} \epsilon_t \hat{y}_S(x_t) \right] \right\}
\end{aligned}$$

Now define $J \subset [2n]$ as, $J_S = \{i : (x_i, \epsilon_i) \in S\}$. Notice that for any $i \in J_S^c$, because $\hat{y}_S$ is only a function of sample S , we have $\mathbb{E}_S [\mathbb{E}_{\epsilon_i} [\epsilon_i \hat{y}_S(x_i)]] = \mathbb{E}_S [\mathbb{E}_{\epsilon_i} [\epsilon_i] \hat{y}_S(x_i)] = 0$. Hence :

$$\begin{aligned}
\mathcal{V}_n^{\text{stat}}(\mathcal{F}) &\geq \sup_{x_1, \dots, x_{kn}} \left\{ \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{kn} \sum_{t=1}^{kn} \epsilon_t f(x_t) \right] - \frac{1}{kn} \sup_{\hat{y}} \mathbb{E}_S \mathbb{E}_\epsilon \left[\sum_{t \in J} \epsilon_t \hat{y}_S(x_t) \right] \right\} \\
&\geq \sup_{x_1, \dots, x_{kn}} \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{kn} \sum_{t=1}^{kn} \epsilon_t f(x_t) \right] - \frac{1}{kn} \sup_{x_1, \dots, x_{kn}} \sup_{\hat{y}} \mathbb{E}_S \mathbb{E}_\epsilon \left[\sum_{t \in J} \epsilon_t \hat{y}_S(x_t) \right] \\
&= \mathcal{R}_{kn}(\mathcal{F}) - \frac{1}{kn} \sup_{x_1, \dots, x_n} \sup_{\hat{y}} \mathbb{E}_\epsilon \left[\sum_{t=1}^n \epsilon_t \hat{y}(x_t) \right]
\end{aligned}$$

Now if we consider minimax rates with respect to only *proper learning algorithms*, that is $\hat{y}_S \in \mathcal{F}$, then

$$\begin{aligned}
\mathcal{V}_n^{\text{stat}}(\mathcal{F}) &\geq \mathcal{R}_{kn}(\mathcal{F}) - \frac{1}{kn} \sup_{x_1, \dots, x_n} \sup_{\hat{y}} \mathbb{E}_\epsilon \left[\sum_{t=1}^n \epsilon_t \hat{y}(x_t) \right] \\
&\geq \mathcal{R}_{kn}(\mathcal{F}) - \frac{1}{kn} \sup_{x_1, \dots, x_n} \mathbb{E}_\epsilon \left[\sup_{\hat{y} \in \mathcal{F}} \sum_{t=1}^n \epsilon_t \hat{y}(x_t) \right] \\
&= \mathcal{R}_{kn}(\mathcal{F}) - \frac{1}{k} \mathcal{R}_n(\mathcal{F})
\end{aligned}$$

On the other hand if we consider *improper learning algorithms* as well, then

$$\mathcal{V}_n^{\text{stat}}(\mathcal{F}) \geq \mathcal{R}_{kn}(\mathcal{F}) - \frac{1}{kn} \sup_{x_1, \dots, x_n} \sup_{\hat{y}} \mathbb{E}_\epsilon \left[\sum_{t=1}^n \epsilon_t \hat{y}(x_t) \right] \geq \mathcal{R}_{kn}(\mathcal{F}) - \frac{1}{k}$$

□

Using $k = 2$, in the above, we get that for proper learning algorithms, $\mathcal{V}_n^{\text{stat}}(\mathcal{F}) \geq \mathcal{R}_{2n}(\mathcal{F}) - \frac{1}{2} \mathcal{R}_n(\mathcal{F})$. If $\mathcal{R}_n(\mathcal{F}) = \Theta(n^{-p})$ for some $p \geq 2$ then, from this we conclude that if we consider minimax rate for proper learning,

$$\mathcal{V}_n^{\text{stat}}(\mathcal{F}) \geq 0.29 \mathcal{R}_{2n}(\mathcal{F})$$

On the other hand if we consider improper learning as well, if $\mathcal{R}_n(\mathcal{F}) = \Omega(n^{-1/p})$ then picking $k = 2n^{1/(p-1)}$, in the lower bound above for improper learning we can conclude that,

$$\mathcal{V}_n^{\text{stat}}(\mathcal{F}) \geq \Omega \left(n^{-\frac{1}{p-1}} \right)$$

4 Putting It All Together

Theorem 4. *For any real valued hypothesis class $\mathcal{F}$, and supervised statistical learning problem with absolute loss (also for squared loss, logistic loss,...), the following are equivalent :*

1. $\mathcal{F}$ is uniformly learnable ($\mathcal{V}_n^{\text{stat}}(\mathcal{F}) \rightarrow 0$)
2. $\mathcal{R}_n(\mathcal{F}) \rightarrow 0$
3. $\mathcal{D}_n(\mathcal{F}) \rightarrow 0$

Summary :

1. We have a crisp certificate for learnability for real valued supervised learning problems. Rates are tight for absolute loss, hinge loss and zero-one loss.
2. Any one of Rademacher complexity, covering numbers or fat-shattering dimension can provide to within log factors the optimal rates.