CSC 588: Machine learning theory **Spring 2021**

Lecture 10: Lower bound of sample complexities of VC classes

Lecturer: Chicheng Zhang *Scribe: Xiaolan Gu*

1 Lower bounds for statistical learning

In previous lectures, it was shown that if we have $O(\frac{d}{\epsilon^2})$ number of training examples, then ERM has excess error less than ϵ. In this lecture, let's consider the opposite region: what if we only have $O(d)$ examples. Note that we will discuss learnability as a property of hypothesis class $\mathcal{H}$ only.

Definition 1. *$\mathcal{H}$ is said to be agnostic PAC learnable if there exists an algorithm $\mathcal{A}$ and a sample complexity function $f(\cdot,\cdot)$ such that for any distribution D, for any $\epsilon, \delta > 0$, if $m \geq f(\epsilon, \delta)$, then with probability $1-\delta$ over the draw of m training examples i.i.d. from D,*

$$\mathrm{err}(\mathcal{A}(S), D) - \min_{h' \in \mathcal{H}} \mathrm{err}(h', D) \leq \epsilon$$

where $\mathcal{A}(S) = \hat{h}$.

Definition 2. *$\mathcal{H}$ is said to be **(realizable)** PAC learnable if there exists an algorithm $\mathcal{A}$ and a sample complexity function $f(\cdot,\cdot)$ such that for any distribution D **realizable by** $\mathcal{H}$, for any $\epsilon, \delta > 0$, if $m \geq f(\epsilon, \delta)$, then with probability $1-\delta$ over the draw of m training examples i.i.d. from D,*

$$\mathrm{err}(\mathcal{A}(S), D) \leq \epsilon$$

Finite VC dimension $\Rightarrow$ uniform convergence $\Rightarrow$ ERM sample complexity $O(\frac{d}{\epsilon^2}) \Rightarrow \mathcal{H}$ is PAC learnable
The following theorem shows that PAC learnable $\Rightarrow$ Finite VC dimension

Theorem 3. *Given a $\mathcal{H}$ such that $VC(\mathcal{H}) \geq d$. If the number of trainning examples $m \leq \frac{d}{2}$, then for any algorithm $\mathcal{A}$, there exists a distribution D realizable by $\mathcal{H}$*

$$\mathbb{E}_{S \sim D^m} \mathrm{err}(\mathcal{A}(S), D) \geq \frac{1}{4} \tag{1}$$

Remark 1. Eq. (1) also implies that

$$\mathbb{P}_{S \sim D^m}\left(\mathrm{err}(\mathcal{A}(S), D) > \frac{1}{8}\right) \geq \frac{1}{8} \tag{2}$$

showing that $\mathcal{A}$ does not $(\epsilon = \frac{1}{8}, \delta = \frac{1}{9})$-PAC learn $\mathcal{H}$ with $m \leq \frac{d}{2}$ examples. The reason is if $\mathcal{A}$ $(\epsilon = \frac{1}{8}, \delta = \frac{1}{9})$-PAC learn $\mathcal{H}$ with $m \leq \frac{d}{2}$ example, then

$$\mathbb{P}_{S \sim D^m}\left(\mathrm{err}(\mathcal{A}(S), D) > \frac{1}{8}\right) \leq \frac{1}{9}$$

which contradicts with (2). We can show (1) $\Rightarrow$ (2) by the fact that for any random variable $X \in [0,1]$ with $\mathbb{E}[X] \geq \frac{1}{4}$, then $\mathbb{P}(X > \frac{1}{8}) \geq \frac{1}{8}$. The proof is shown as follows

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}[X\mathbf{1}(X \leq \frac{1}{8})] + \mathbb{E}[X\mathbf{1}(X \in (\frac{1}{8}, 1])] \leq \frac{1}{8} + \mathbb{E}[\mathbf{1}(X > \frac{1}{8})] \\ \Rightarrow \quad \mathbb{P}[X > \frac{1}{8}] &= \mathbb{E}[\mathbf{1}(X > \frac{1}{8})] \geq \mathbb{E}[X] - \frac{1}{8} \geq \frac{1}{4} - \frac{1}{8} = \frac{1}{8} \end{aligned}$$

Remark 2. $VC(\mathcal{H}) = \infty \Rightarrow \mathcal{H}$ is not PAC learnable, because $VC(\mathcal{H}) = \infty$ implies that $\forall m, \forall \mathcal{A}, \exists D$ realizable by $\mathcal{H}$

$$\mathbb{P}_{S\sim D^m}\left(\mathrm{err}(\mathcal{A}(S), D) > \frac{1}{8}\right) > \frac{1}{9}$$

Proof of Theorem 3. We can rewrite the problem as minimax lower bound

$$\min_{\mathcal{A}} \max_{D:\text{realizable by } \mathcal{H}} \mathbb{E}_{S\sim D^m} \mathrm{err}(\mathcal{A}(S), D) \geq \frac{1}{4} \tag{3}$$

Define a family as distributions $\mathcal{P} = \{D_b : b \in \{\pm 1\}^d\}$. We want to show

$$\min_{\mathcal{A}} \mathbb{E}_{b\sim U(\pm 1)^d} \mathbb{E}_{S\sim D^m} \mathrm{err}(\mathcal{A}(S), D) \geq \frac{1}{4}$$

which implies (3). Find a set of unlabeled examples $z_1, \cdots, z_d$ shattered by $\mathcal{H}$ and define $D_b : \mathbb{P}(x = z_i, y = b_i) = \frac{1}{d}$ $(\forall i = 1, \cdots, d)$. Our first observation is all D_b's are realizable by $\mathcal{H}$. Denote $\hat{h} = \mathcal{A}(S)$. We are going to show

$$\forall \mathcal{A}, \quad \mathbb{E}_{b,S} \mathrm{err}(\hat{h}, D_b) \geq \frac{1}{4}$$

Given h, then $\mathrm{err}(h, D_b) = \sum_{i=1}^{d} \frac{1}{d} \mathbf{1}(h(z_i) \neq b_i)$. We want to show $\sum_{i=1}^{d} \mathbb{E}_{b,S} \mathbf{1}(h(z_i) \neq b_i) \geq \frac{d}{4}$ by showing

$$\mathbb{P}_{b,S}(h(z_1) \neq b_1) \geq \frac{1}{4} \tag{4}$$

and we also can show this for other i's. Denote unlabled sample set $S_x = \{x_1, \cdots, x_m\}$ are drawn i.i.d. from uniform($\{z_1, \cdots, z_d\}$), then $S = \{(x_1, y_1), \cdots, (x_m, y_m)\}$ are determined by S_x and b. Then

$$\begin{aligned} \mathbb{P}_{b,S}(h(z_1) \neq b_1) &\geq \mathbb{P}_{b,S_x}(h(z_1) \neq b_1, z_1 \notin S_x) \\ &= \mathbb{P}_{b,S_x}(h(z_1) \neq b_1 | z_1 \notin S_x) \cdot \mathbb{P}_{S_x}(z_1 \notin S_x) \end{aligned} \tag{5}$$

Note that

$$\mathbb{P}(z_1 \in S_x) = \mathbb{P}(z_1 \in \cup_i \{x_i\}) \leq \sum_{i=1}^{m} \mathbb{P}(z_1 = x_i) = \frac{m}{d} \leq \frac{1}{2}$$

which implies $\mathbb{P}_{S_x}(z_1 \notin S_x) \geq \frac{1}{2}$. On the other hand, conditioned on $z_1 \notin S_x$, $\hat{h}(z_1)$ is independent of b_1, then

$$\mathbb{P}_{b,S_x}(\hat{h}(z_1) \neq b_1 | z_1 \notin S_x) = \frac{1}{2}$$

Thus, (5) can be rewritten as $\mathbb{P}_{b,S}(\hat{h}(z_1) \neq b_1) \geq \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$, which finishes the proof of (4) and the whole proof. □

2 Review of what we learned

Definition 4. *$\mathcal{H}$ is said to satisfy the uniform convergence property if there exists a function $f_u : (0,1)^2 \to \mathbb{N}$ such that for any D, for any $\epsilon, \delta > 0$, if $m \geq f_u(\epsilon, \delta)$, then w.p. $1 - \delta$ over the draw of m i.i.d. training examples from D*

$$\forall h \in \mathcal{H}, \quad |\mathrm{err}(h, S) - \mathrm{err}(h, D)| \leq \epsilon$$

Theorem 5 (The fundamental theorem of statistical learning)**.** *The following statements are equivalent*

1. *$\mathcal{H}$ satisfies the uniform convergence property (Definition 4)*

2. $\mathcal{H}$ *is agnostic PAC learnable with ERM*

3. $\mathcal{H}$ *is agnostic PAC learnable*

4. $\mathcal{H}$ *is (realizable) PAC learnable*

5. $\mathcal{H}$ *has finite VC dimension*

Proof. By showing cycling implication.

- $1 \Rightarrow 2$: set the sample size to be greater than $f_u(\epsilon/2, \delta)$, then by definition $|\operatorname{err}(h, S) - \operatorname{err}(h, D)| \le \epsilon/2$ w.p. $1 - \delta$, which is the sufficient condition for the ERM to achieve excess error rate at most ϵ (as we discussed before)
- $2 \Rightarrow 3$: trivial
- $3 \Rightarrow 4$: seen before
- $4 \Rightarrow 5$: just proved (Theorem 3)
- $5 \Rightarrow 1$: last class of uniform convergence (by symmetrization, Rademacher random variables and Massart's Lemma)

□

Interpretation of finite VC dimension. For S which is a set of observations in the real-world, regard $\mathcal{H}$ as scientific theory. If the scientific theory is too complicated (i.e., $\mathcal{H}$ has infinite VC dimension), then there might not be a reliable way of using this theory to make future prediction with scientific outcomes.

3 Appendix: Exercises

Problem 1. Can we bound the following term using Massart's Lemma?

$$\mathbb{E}_{S,S' \sim D^m}\left[\sup_{f \in \mathcal{F}} \mathbb{E}_S f(Z) - \mathbb{E}_{S'} f(Z)\right] \tag{6}$$

Problem 2. Upper bound (6) by

$$\mathbb{E}_{S \sim D^m} \sup_{f \in \mathcal{F}} \mathbb{E}_S f(Z) + \mathbb{E}_{S' \sim D^m} \sup_{f \in \mathcal{F}} (-\mathbb{E}_{S'} f(Z))$$

without introducing Rademacher random variables. Will the proof still go through?