昆山杜克大学
DUKE KUNSHAN UNIVERSITY

Fundamental Theorem of Statistical Learning

STATS 303 Statistical Machine Learning

Spring 2022

Lecture 22

Fundamental Theorem of Statistical Learning

Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0,1\}$ and let the loss function be 0-1 loss. Then the following statements are equivalent:

1. $\mathcal{H}$ has UCP
2. Any ERM is a successful agnostic PAC learner for $\mathcal{H}$
3. $\mathcal{H}$ is agnostic PAC learnable
4. $\mathcal{H}$ is PAC learnable
5. Any ERM is a successful PAC learner for $\mathcal{H}$
6. $\mathcal{H}$ has a finite VC-dimension

Idea of proof:

- If $\mathrm{VCdim}(\mathcal{H}) = d$, then the "effective size" of $\mathcal{H}$ is small.

- Small "effective size" $\Rightarrow$ UCP

Sauer's Lemma

Def. (growth function) Let $\mathcal{H}$ be a hypothesis class. The growth function of $\mathcal{H}$, denoted by

$$\tau_{\mathcal{H}} : \mathbb{N} \to \mathbb{N},$$

is defined by

$$\tau_{\mathcal{H}}(m) = \max_{\substack{C \subset \mathcal{X}: \\ |C| = m}} |\mathcal{H}_C|$$

Remark: If $\mathrm{VCdim}(\mathcal{H}) = d$, then

$$m \le d \quad \Rightarrow \quad \tau_{\mathcal{H}}(m) = 2^m.$$

Lemma (Sauer–Shelah–Perles) Let $\mathcal{H}$ be a hypothesis class with $\mathrm{VCdim}(\mathcal{H}) \le d < \infty$. Then for all $m \in \mathbb{N}$,

$$\tau_{\mathcal{H}}(m) \le \sum_{i=0}^{d} \binom{m}{i}$$

Remark: In particular, if $m \ge d+1$, then this implies

$$\tau_{\mathcal{H}}(m) \le \left(\frac{em}{d}\right)^{d}$$

following Stirling's approximation.

Proof of Sauer's Lemma:

We prove a stronger claim:

$$\begin{cases} \text{For any } C = \{c_1, \cdots, c_m\} \subset \mathcal{X}, \text{ for any } \mathcal{H}, \\ |\mathcal{H}_C| \le \left| \{ B \subset C : \mathcal{H} \text{ shatters } B \} \right| \end{cases}$$

If we can prove , then if $\mathrm{VCdim}(\mathcal{H}) \le d$, then no set with size larger than d can be shattered by $\mathcal{H}$.

Therefore, $\left| \{ B \subset C : \mathcal{H} \text{ shatters } B \} \right| \le \underbrace{\sum_{i=0}^{d} \binom{m}{i}}_{\text{total \# of subsets with size } \le d}$.

To prove ☺, we use mathematical induction.

- For $m = 1$, $C = \{c_1\}$. There are two cases:
  - $|\mathcal{H}_C| = 1$. Only the empty set $\emptyset$ is shattered by $\mathcal{H}$.
  - $|\mathcal{H}_C| = 2$. Both $\emptyset$ and C are shattered by $\mathcal{H}$.

  In both cases, $|\mathcal{H}_C| = |\{B \subset C : B \text{ is shattered by } \mathcal{H}\}|$.

  Therefore, ☺ holds for $m = 1$.

- Assume ☺ holds for all sets of size $< m$. We are going to prove that ☺ also holds for m.

  Fix $\mathcal{H}$. Let $C = \{c_1, \cdots, c_m\}$.

  Denote $C' = \{c_2, \cdots, c_m\}$

In addition, define the following two sets:

$$Y_0 = \{ \overbrace{(y_2, \cdots, y_m)}^{\text{"labels"}} : \ (0, y_2, \cdots, y_m) \in \mathcal{H}_C \text{ or } (1, y_2, \cdots, y_m) \in \mathcal{H}_C \}$$

$$Y_1 = \{ (y_2, \cdots, y_m) : \ (0, y_2, \cdots, y_m) \in \mathcal{H}_C \text{ and } (1, y_2, \cdots, y_m) \in \mathcal{H}_C \}$$

Claim: $|\mathcal{H}_C| = |Y_0| + |Y_1|$

The reason holds is the following:

If $(y_1, y_2, \cdots, y_m) \in \mathcal{H}_C$, then either

$$(y_2, \cdots, y_m) \in Y_0 \text{ but } (y_2, \cdots, y_m) \notin Y_1$$

or

$$(y_2, \cdots, y_m) \in Y_0 \cap Y_1 .$$

In the former case, $(\underset{\substack{\uparrow \\ \text{either 0 or 1.}}}{\bullet}, y_2, \cdots, y_m) \in \mathcal{H}_C$; in the latter case, $(0, y_2, \cdots, y_m) \in \mathcal{H}_C$ and $(1, y_2, \cdots, y_m) \in \mathcal{H}_C$.

Therefore, $|Y_0| + |Y_1| = |Y_0 - Y_1| + 2|Y_1| = |\mathcal{H}_C|$

Moreover, $Y_0 = \mathcal{H}_{C'}$.

From induction,

$$|Y_0| = |\mathcal{H}_{C'}| \le \left| \{ B \subset C' : \mathcal{H} \text{ shatters } B \} \right|$$

$$= \left| \{ B \subset C : c_1 \notin B \text{ and } \mathcal{H} \text{ shatters } B \} \right|$$

Next, define $\mathcal{H}' \subset \mathcal{H}$ to be

$$\mathcal{H}' := \{ h \in \mathcal{H} : \text{there exists } h' \in \mathcal{H} \text{ s.t. } (1 - h'(c_1), h'(c_2), \dots, h'(c_m)) = (h(c_1), h(c_2), \dots, h(c_m)) \}$$

That is, $\mathcal{H}'$ is the hypothesis class whose members come in pairs that differ at c_1 only.

Obviously, $\mathcal{H}'$ shatters $B \subset C'$

$$\iff \mathcal{H}' \text{ shatters } B \cup \{c_1\}.$$

Also, $Y_1 = \mathcal{H}'_{C'}$

By induction,

$$|Y_1| = |\mathcal{H}'_{C'}| \le \left|\{B \subset C' : \mathcal{H}' \text{ shatters } B\}\right|$$

$$= \left|\{B \subset C : \mathcal{H}' \text{ shatters } B \cup \{c_1\}\}\right|$$

$$\le \left|\{B \subset C : \mathcal{H} \text{ shatters } B \text{ and } c_1 \in B\}\right|$$

Bringing  and  together, we have

$$|Y_0| + |Y_1| \le \left|\{B \subset C : c_1 \notin B \text{ and } \mathcal{H} \text{ shatters } B\}\right| + \left|\{B \subset C : \mathcal{H} \text{ shatters } B \text{ and } c_1 \in B\}\right|$$

$$= \left|\{B \subset C : \mathcal{H} \text{ shatters } B\}\right|$$

Noting that $|\mathcal{H}_C| = |Y_0| + |Y_1|$ (by ), we are done with the proof.

UCP for classes with small "effective size"

Thm* Let $\mathcal{H}$ be a hypothesis class and $\tau_{\mathcal{H}}$ be its growth function. Then for any distribution $\mathcal{D}$ and any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$|L_{\mathcal{D}}(h) - L_S(h)| \leq \frac{4 + \sqrt{\log \tau_{\mathcal{H}}(2m)}}{\delta \sqrt{2m}}.$$

Suppose Thm* holds. For $m > d$, $\tau_{\mathcal{H}}(2m) \leq \left(\frac{2em}{d}\right)^d$.

Therefore, with probability at least $1-\delta$,

$$|L_{\mathcal{D}}(h) - L_S(h)| \leq \frac{4 + \sqrt{d \log (2em/d)}}{\delta \sqrt{2m}}$$

For simplicity, let's assume $4 \leq \sqrt{d \log (2em/d)}$, which can be done by choosing a large m.

Then

$$|L_{\mathcal{D}}(h) - L_S(h)| \leq \frac{\sqrt{2 d \log (2em/d)}}{\delta \sqrt{m}}$$

Let's choose m s.t.

$$\frac{\sqrt{2\, d \log(2em/d)}}{\delta \sqrt{m}} \leq \varepsilon$$

That is,
$$\sqrt{m} \geq \frac{\sqrt{2d \log m + 2d \log(2e/d)}}{\varepsilon \delta}$$

or
$$m \geq \frac{2d \log m}{(\varepsilon \delta)^2} + \frac{2d \log(2e/d)}{(\varepsilon \delta)^2}$$

<u>Lemma</u>. Let $a \geq 1$ and $b > 0$. Then

$$x \geq 4a \log(2a) + 2b \Rightarrow x \geq a \log(x) + b$$

According to this lemma, we only need to take

$$m \geq 4 \frac{2d}{(\varepsilon \delta)^2} \log\left(\frac{4d}{(\varepsilon \delta)^2}\right) + 2 \frac{2d \log(2e/d)}{(\varepsilon \delta)^2}.$$

Suppose we sample this many data points. Then with probability at least $1-\delta$, $|L_{\mathcal{D}}(h) - L_S(h)| \leq \varepsilon$.

Therefore, we have the UCP.

Proof of Thm*

First note that it suffices to prove

$$\mathbb{E}_{\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|\right] \le \frac{4+\sqrt{\log \tau_{\mathcal{H}}(2m)}}{\sqrt{2m}}$$

If we have proved , then Markov Inequality implies

$$\mathbb{P}_{\mathcal{D}^m}\left(\sup_{h\in\mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| > \frac{4+\sqrt{\log \tau_{\mathcal{H}}(2m)}}{\delta\sqrt{2m}}\right)$$

$$\le \frac{\mathbb{E}_{\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|\right]}{\frac{4+\sqrt{\log \tau_{\mathcal{H}}(2m)}}{\delta\sqrt{2m}}} \le \frac{\frac{4+\sqrt{\log \tau_{\mathcal{H}}(2m)}}{\sqrt{2m}}}{\frac{4+\sqrt{\log \tau_{\mathcal{H}}(2m)}}{\delta\sqrt{2m}}}$$

$$= \delta$$

That is, with probability at least $1-\delta$,

$$\sup_{h\in\mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \le \frac{4+\sqrt{\log \tau_{\mathcal{H}}(2m)}}{\delta\sqrt{2m}}$$

To prove , write

$$L_{\mathcal{D}}(h) = \mathbb{E}_{S'\sim\mathcal{D}^m}\left[L_{S'}(h)\right]$$

Then the LHS of , $\mathbb{E}_{\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}\left|L_{\mathcal{D}}(h) - L_S(h)\right|\right]$, becomes

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}\left|\mathbb{E}_{S'\sim\mathcal{D}^m}\left[L_{S'}(h)\right] - L_S(h)\right|\right].$$

Note that $\left|\mathbb{E}_{S'\sim\mathcal{D}^m}\left[L_{S'}(h)\right] - L_S(h)\right|$

$$\leq \mathbb{E}_{S'\sim\mathcal{D}^m}\left|L_{S'}(h) - L_S(h)\right|$$

Also, $\sup_{h\in\mathcal{H}}\mathbb{E}_{S'\sim\mathcal{D}^m}\left|L_{S'}(h) - L_S(h)\right|$

$$\leq \sup_{h\in\mathcal{H}}\mathbb{E}_{S'\sim\mathcal{D}^m}\sup_{h\in\mathcal{H}}\left|L_{S'}(h) - L_S(h)\right|$$

$$= \mathbb{E}_{S'\sim\mathcal{D}^m}\sup_{h\in\mathcal{H}}\left|L_{S'}(h) - L_S(h)\right|$$

Therefore, $\mathbb{E}_{\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} \left| L_{\mathcal{D}}(h) - L_S(h) \right|\right]$

$$\leq \mathbb{E}_{S\sim\mathcal{D}^m} \mathbb{E}_{S'\sim\mathcal{D}^m} \sup_{h\in\mathcal{H}} \left| L_{S'}(h) - L_S(h) \right|$$

$$\leq \mathbb{E}_{S,S'\sim\mathcal{D}^m} \sup_{h\in\mathcal{H}} \frac{1}{m} \left| \sum_{i=1}^{m} \left[\ell(h, z_i') - \ell(h, z_i) \right] \right|$$

Let z_i be the samples in S

and z_i' be the samples in S'

Due to symmetry, we can change

$$\sum_{i=1}^{m} \left[\ell(h, z_i') - \ell(h, z_i) \right]$$

to $\sum_{i=1}^{m} \sigma_i \left(\ell(h, z_i') - \ell(h, z_i) \right)$

where each $\sigma_i \in \{\pm 1\}$.

$$= \mathbb{E}_{\sigma\sim\mathcal{U}_{\pm}^m} \mathbb{E}_{S,S'\sim\mathcal{D}^m} \sup_{h\in\mathcal{H}} \frac{1}{m} \left| \sum_{i=1}^{m} \sigma_i \left[\ell(h, z_i') - \ell(h, z_i) \right] \right|$$

$$= \mathbb{E}_{S,S'\sim\mathcal{D}^m} \mathbb{E}_{\sigma\sim\mathcal{U}_{\pm}^m} \sup_{h\in\mathcal{H}} \frac{1}{m} \left| \sum_{i=1}^{m} \sigma_i \left[\ell(h, z_i') - \ell(h, z_i) \right] \right|$$

Fix S, S'. Let C be the set of instances in S and S'.

Then

$$\mathbb{E}_{\sigma\sim U_\pm^m} \sup_{h\in\mathcal{H}} \frac{1}{m} \left| \sum_{i=1}^m \sigma_i \left[\ell(h, z_i') - \ell(h, z_i) \right] \right|$$

$$= \mathbb{E}_{\sigma\sim U_\pm^m} \sup_{h\in\mathcal{H}_C} \frac{1}{m} \left| \sum_{i=1}^m \sigma_i \left[\ell(h, z_i') - \ell(h, z_i) \right] \right|$$

Fix $h \in \mathcal{H}_C$. Denote $\theta_h = \frac{1}{m} \sum_{i=1}^m \sigma_i \left[\ell(h, z_i') - \ell(h, z_i) \right]$

Since $\mathbb{E}_{\sigma\sim U_\pm^m} [\theta_h] = 0$, by Hoeffding's Inequality,

$$\mathbb{P}(|\theta_h| > \rho) \le 2\exp(-2m\rho^2).$$

By a union bound,

$$\mathbb{P}\left(\max_{h\in\mathcal{H}_C} |\theta_h| > \rho \right) \le 2|\mathcal{H}_C| \exp(-2m\rho^2).$$

Therefore, following some Calculus,

$$\mathbb{E}_{\sigma\sim U_\pm^m}\left[\max_{h\in\mathcal{H}_C} |\theta_h| \right] \le \frac{4+\sqrt{\log(\mathcal{H}_C)}}{\sqrt{2m}}$$

$$\le \frac{4+\sqrt{\log \tau_{\mathcal{H}}(2m)}}{\sqrt{2m}}$$

Therefore, $\mathbb{E}_{\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} \left| L_{\mathcal{D}}(h) - L_S(h) \right| \right]$

$$\le \frac{4+\sqrt{\log \tau_{\mathcal{H}}(2m)}}{\sqrt{2m}}$$

Questions?

Reference

- *FTSL*
    - *[S-S] Ch 6.4*