Introduction to Statistical Learning: The Fundamental Theorem of PAC Learning

Julien Béguinot, Aslan Tchamkerten

Télécom Paris

The following theorem is called the non quantitative fondamental theorem of PAC learning.

Theorem 1 (Theorem 6.7 in UML Textbook)**.** *Let $\mathcal{H}$ be an hypothesis class of functions from $\mathcal{X}$ to $\{0,1\}$. Let the loss function be the $0/1$ loss. The following assertions are equivalents:*

1. *$\mathcal{H}$ as the uniform convergence property*
2. *Any ERM rule 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 rule is a successful PAC learner for $\mathcal{H}$*
6. *$\mathcal{H}$ has a fuinite VC dimension.*

In this session we prove the non trivial step $6 \to 1$. This will conclude the equivalence between PAC learnability and finite VC dimension while you only proved that finite VC dimension was a necessary condition for PAC learnability in class.

Definition 1 (Growth Function)**.** Let $\mathcal{H}$ be a hypothesis class. Then the *growth function* of $\mathcal{H}$ denoted $\tau_{\mathcal{H}} : \mathbb{N} \mapsto \mathbb{N}$, is defined as

$$\tau_{\mathcal{H}}(m) \triangleq \max_{C \subset \mathcal{X} : |C| = m} |\mathcal{H}_{|C}|. \tag{1}$$

Recall that when $\mathcal{H}$ is a set of function from $\mathcal{X}$ to $\mathcal{Y}$ then $\mathcal{H}_{|C}$ is the set of functions induced by the restrictions of the function of $\mathcal{H}$ to $C \subseteq \mathcal{X}$.

Exercice 1. *We proceed step by step.*

- *(Warmup) Show that if* $\mathrm{VCdim}(\mathcal{H}) = d < +\infty$ *then for any* $m \leq d, \tau_{\mathcal{H}}(m) = 2^m$.
- *We show by* strong induction *over m that for any $C = \{c_1, \ldots, c_m\}$ then $\forall \mathcal{H}$, $|\mathcal{H}_{|C}| \leq |\{B \subseteq C : \mathcal{H}$ shatters $B\}|$.*
  - *Show it holds for $m = 1$.*
  - *We assume that for any set of size $k < m$ the property holds. Let $C = (c_1, \ldots, c_m)$ be of size m and $C' = (c_2, \ldots, c_m)$. We define the sets*

$$Y_0 \triangleq \{(y_2, \ldots, y_m) : (0, y_2, \ldots, y_m) \in \mathcal{H}_C \text{ or } (1, y_2, \ldots, y_m) \in \mathcal{H}_C\}$$

and

$$Y_1 \triangleq \{(y_2, \ldots, y_m) : (0, y_2, \ldots, y_m) \in \mathcal{H}_C \text{ and } (1, y_2, \ldots, y_m) \in \mathcal{H}_C\}.$$

Let $\mathcal{H}' = \{h \in \mathcal{H} : \exists h' \in \mathcal{H}, h(c_1) = 1 - h'(c_1)$ *andi for* $i = 2, \ldots, m, h(c_i) = h'(c_i)$. *Explain why* $|\mathcal{H}_C| = |Y_0| + |Y_1|$. *Explain why* $Y_0 = \mathcal{H}_{C'}$ *and* $Y_1 = \mathcal{H}'_{C'}$. *Explain why the proposition applies to* C *using thye inductive hypothesis.*

- *Deduce that Sauer's Lemma holds:*

  Lemma 1 (Sauer-Shelah-Perles)**.** *Let* $\mathcal{H}$ *be a hypothesis with finite VC dimension* d. *Then for any* $m > d$,

$$\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq \left(\frac{em}{d}\right)^d. \tag{2}$$

- *Explain why*

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|\right] = \mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} |\mathbb{E}_{S'\sim\mathcal{D}^m} L_{S'}(h) - L_S(h)|\right] \tag{3}$$

- *Prove that*

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|\right] \leq \mathbb{E}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\left|\sum_{i=1}^{m}(l(h,z_i') - l(h,z_i))\right|\right]. \tag{4}$$

  You may use Jensen's inequality which states that $\mathbb{E}[\phi(X)] \geq \phi(\mathbb{E}[X])$ *when* ϕ *is convex.*

- *Explain why for every* $\sigma \in \{\pm 1\}^m$ *we have*

$$\mathbb{E}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\left|\sum_{i=1}^{m}(l(h,z_i') - l(h,z_i))\right|\right] = \mathbb{E}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i(l(h,z_i') - l(h,z_i))\right|\right]. \tag{5}$$

- *Concludes that if* σ *is sampled uniformly then*

$$\mathbb{E}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\left|\sum_{i=1}^{m}(l(h,z_i') - l(h,z_i))\right|\right] = \mathbb{E}_\sigma\mathbb{E}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i(l(h,z_i') - l(h,z_i))\right|\right]. \tag{6}$$

- *Fix* S, S' *and explain why*

$$\sup_{h\in\mathcal{H}} \frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i(l(h,z_i') - l(h,z_i))\right| = \max_{h\in\mathcal{H}_C} \frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i(l(h,z_i') - l(h,z_i))\right| \tag{7}$$

  where C *corresponds to the instance in* S, S'.

- *Let* $\theta_h \triangleq \frac{1}{m}\sum_{i=1}^{m}\sigma_i(l(h,z_i') - l(h,z_i))$. *Explain why for every* $\rho > 0$,

$$\mathbb{P}(|\theta_h| > \rho) \leq 2\exp(-2m\rho^2). \tag{8}$$

  You may use Hoeffding's inequality.

- *Show that this implies in turn that*

$$\mathbb{P}(\max_{h\in\mathcal{H}_C} |\theta_h| > \rho) \leq 2|\mathcal{H}_C|\exp(-2m\rho^2). \tag{9}$$

- *We admit the following technical lemma:*

  Lemma 2. *Let X be a random variable and $x' \in \mathbb{R}$ be a scalar and assume there exists $a > 0$ and $nb \geq e$ such that for all $t \geq 0$, $\mathbb{P}(|X - x'| > t) \leq 2be^{-\frac{t^2}{a^2}}$. Show that $\mathbb{E}[|X - x'|] \leq a(2 + \sqrt{\log(b)})$.*

  Prove that

$$\mathbb{E}_{S \sim \mathcal{D}^m}\left[\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|\right] \leq \frac{4 + \sqrt{\log \tau_{\mathcal{H}}(2m)}}{\sqrt{2m}}. \tag{10}$$

- *Concludes that for every $\mathcal{D}$ and every $\delta \in (0, 1)$ then with probability of at least $1 - \delta$ over $S \sim \mathcal{D}^m$ we have,*

$$|L_{\mathcal{D}}(h) - L_S(h)| \leq \frac{4 + \sqrt{\log \tau_{\mathcal{H}}(2m)}}{\delta\sqrt{2m}}. \tag{11}$$

- *Show that if the VC dimension is finite equal to d then*

$$m_{\mathcal{H}}^{\mathrm{UC}}(\epsilon, \delta) \leq 4\frac{16d}{(\epsilon\delta)^2} \log\left(\frac{16d}{(\epsilon\delta)^2}\right) + \frac{16d\log(2e/d)}{(\epsilon\delta)^2}. \tag{12}$$

  You may use the fact that for $a \geq 1$ and $b > 0$. Then $x \geq 4a\log(2a) + 2b \implies x \geq a\log(x) + b$.

- *Conclude the proof of Theorem 1.*