Theorem [Fundamental Theorem of Statistical Learning]

Let $\mathcal{H}$ be a hypothesis class from $\mathcal{X}$ to $\boxed{\{0,1\},}$ and the loss function is a 0-1 loss, then the following are equivalent.

1. $\mathcal{H}$ has the uniform convergent property
2. Any ERM rule is a successful agnostic PAC learner of $\mathcal{H}$.
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 finite VC-dimension.

we already learned that $1 \to 2$ (lecture 3)

$2 \to 3$ trivial

$3 \to 4$ trivial

$2 \to 5$ trivial

$2 \to 6$: $\sim 6 \to \sim 2$ No-free-lunch-theorem.

we are left to prove $6 \to 1$

Theorem [Fundamental theorem. Quantitative version]

$VCdim(\mathcal{H}) = d$, then

1. $\mathcal{H}$ has uniform convergence property with
$$C_1 \cdot \frac{d + \log 1/\delta}{\varepsilon^2} \le m_{\mathcal{H}}^{UC}(\varepsilon, \delta) \le C_2 \cdot \frac{d + \log 1/\delta}{\varepsilon^2}$$
2. $\mathcal{H}$ is agnostic PAC learnable with
$$C_1 \cdot \frac{d + \log 1/\delta}{\varepsilon^2} \le m_{\mathcal{H}}(\varepsilon, \delta) \le C_2 \frac{d + \log 1/\delta}{\varepsilon^2}$$
3. $\mathcal{H}$ is PAC learnable with.
$$C_1 \cdot \frac{d + \log 1/\delta}{\varepsilon} \le m_{\mathcal{H}}(\varepsilon, \delta) \le C_2 \cdot \frac{d \log 1/\varepsilon + \log 1/\delta}{\varepsilon}$$

$C_1 = \frac{\ }{\varepsilon}$ $=$ ε

proof later.

---

proof strategy for fundamental theorem of statistical learning.

① [Sauer's lemma]

polynomial exponential

If $VCdim(\mathcal{H}) = d$, then for any $C \subseteq \mathcal{X}$,

the effective size of $\mathcal{H}$ restricted to C, i.e., $|\mathcal{H}_C|$ is only $O(|C|^d) \ll 2^{|C|}$ worst-case

big-O notation

$|\mathcal{H}_C| = O(f(|C|, d))$ means that $\exists a > 0$ s.t. $\leq a \cdot f(|C|, d)$ for large enough d

Definition. (Restriction of $\mathcal{H}$ to C)

$\mathcal{H}$ is class of functions $\mathcal{X} \to \{0,1\}$, and $C = (c_1, \dots, c_m) \subseteq \mathcal{X}$,

the Restriction of $\mathcal{H}$ to C

$$\mathcal{H}_C \triangleq \{ \tilde{h}: C \to \{0,1\},\ c_i \mapsto h(c_i) \mid h \in \mathcal{H} \}$$

equivalently we can represent $\tilde{h}$ by a vector $(h(c_1), \dots, h(c_m))$.

② If $|\mathcal{H}_C|$ grows polynomially in $|C|$, then we have uniform Convergence Property.

Definition (Uniform Convergence)

A hypothesis class $\mathcal{H}$ has the uniform convergence property,

if $\boxed{\exists\, m_{\mathcal{H}}^{UC}}$ s.t. for every ε, δ and every $\mathcal{D}$, if $m \geq m_{\mathcal{H}}^{UC}(\varepsilon, \delta)$

the S is ε-representative with probability at least $1-\delta$.

$|L_D(h) - L_S(h)| < \varepsilon.\ \forall h \in \mathcal{H}.$

Definition [Growth Function]

The Growth Function $\tau_{\mathcal{H}}: \mathbb{N} \to \mathbb{N}$, is defined as

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

which is the number of different functions from C of size m to $\{0,1\}$ that can be obtained by restricting $\mathcal{H}$ to C.

* If $m \leq VCdim(\mathcal{H}) = d$, then by definition of VCdim, $\tau_{\mathcal{H}}(m) = 2^m$.

largest m s.t. $\tau_{\mathcal{H}}(m) = 2^m$.

* For $m > d$, m grows polynomially in m.

---

① Lemma [Sauer-Shelah-Perles]

If $VCdim(\mathcal{H}) = d < \infty$, then for all m, $\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$.

m choose i, $\frac{m!}{i!(m-i)!}$

$\rightarrow d \geq m$, $\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{m} \binom{m}{i} = 2^m$ ← exp in m

$\rightarrow d \ll m$, $\tau_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^d$ ← polynomial in m

Stirling's formula.

② Theorem [for uniform Convergence] for every $D, \delta, h \in \mathcal{H}$

$$|L_D(h) - L_S(h)| \le \frac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\delta \cdot \sqrt{2m/2}} \quad \text{w.p } 1-\delta.$$

Proof of fundamental theorem of statistical learning.>
using ① Sauer's Lemma, we get from ②, for $m > d$

$$|L_D(h) - L_S(h)| \le \frac{4 + \sqrt{d \log(\frac{2em}{d})}}{\delta \sqrt{2m/2}}$$

for simplicity let $\sqrt{d \log \frac{2em}{d}} \ge 4$,

$$\le \frac{2}{\delta}\sqrt{\frac{2d\log(\frac{2em}{d})}{m}} \overset{\text{want}}{\le} \varepsilon$$

$$\to \frac{8d\log(\frac{2em}{d})}{(\delta\varepsilon)^2} \le m$$

$$\to m \ge \underbrace{4 \cdot \frac{8d}{(\delta\varepsilon)^2}\log\left(\frac{2d}{(\delta\varepsilon)^2}\right) + \frac{16d\log(\frac{2e}{d})}{(\delta\varepsilon)^2}}$$ is sufficient.

this gives a loose bound on $m_{\mathcal{H}}(\varepsilon,\delta)$, but sufficient to prove (Agnostic) PAC learnability.

Proof of ②> claim: $\mathbb{E}_S\left[\sup_{h\in\mathcal{H}} |L_D(h) - L_S(h)|\right] \le \frac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\sqrt{m/2}}$

$\Rightarrow$ Markov's ineq: $P\left(\sup_{h\in\mathcal{H}}|L_D(h) - L_S(h)| > \varepsilon\right) \le \frac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\varepsilon\sqrt{m/2}} \le \delta.$

for $\varepsilon \ge \frac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\sqrt{m/2}\cdot\delta}$

$$\mathbb{E}_S\left[\sup_{h\in\mathcal{H}} |L_D(h) - L_S(h)|\right] = \mathbb{E}_{S=\{(x_i,y_i)\}_{i=1}^m}\left[\sup_{h\in\mathcal{H}}\left|\mathbb{E}_{S'=\{(x_i',y_i')\}_{i=1}^m}[L_{S'}(h)] - L_S(h)\right|\right]$$

$L_D(h) = \mathbb{E}_{S'}[L_{S'}(h)]$

Jensen's ineq. $\Rightarrow$ $\le \mathbb{E}_S\left[\sup_{h\in\mathcal{H}} \mathbb{E}_{S'}|L_{S'}(h) - L_S(h)|\right]$

Convex f.

$\mathbb{E}[f(z)] \ge f(\mathbb{E}[z])$ $\quad$ $|z|$ $\quad$ max{linear fs}

$$\le \mathbb{E}_S\left[\mathbb{E}_{S'}\left[\sup_{h\in\mathcal{H}}|L_{S'}(h) - L_S(h)|\right]\right]$$

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

$z_i' = (x_i', y_i')$

$\ell(h,(z_i')) - \ell(h,z_i) \overset{\text{dist.}}{=} \sigma_i \cdot (\ell_h(z_i') - \ell_h(z_i))$, $\forall \sigma_i \in \{+1,-1\}$

z_i', z_i are iid $\overset{\text{dist}}{=} \sigma_i \cdot (\ell_h(z_i') - \ell_h(z_i))$, $\sigma_i \sim \begin{cases} +1 & \text{w.p } \frac{1}{2} \\ -1 & \frac{1}{2}\end{cases}$

$$\Rightarrow (*) = \mathbb{E}_{\sigma=(\sigma_1,\dots,\sigma_m)}\left[\mathbb{E}_{S,S'}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{m}\sum_{i=1}^m \sigma_i(\ell_h(z_i') - \ell_h(z_i))\right| \,\middle|\, \sigma\right]\right]$$

linearity of expectation $\Rightarrow$
$$= \mathbb{E}_{S,S'}\left[\underbrace{\mathbb{E}_{\sigma=(\sigma_1,\dots,\sigma_m)}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{m}\sum_{i=1}^m \sigma_i(\ell_h(z_i') - \ell_h(z_i))\right| \,\middle|\, S,S'\right]}_{S,S' \text{ is fixed, only } \mathcal{H}_C \text{ matters } C = S\cup S',\ |C|\le 2m}\right]$$

$$= \mathbb{E}_\sigma\left[\max_{h\in\mathcal{H}_C}\left|\underbrace{\frac{1}{m}\sum_{i=1}^m \sigma_i(\ell_h(z_i') - \ell_h(z_i))}_{\theta_i,\ \mathbb{E}\theta_i = 0,\ -1\le\theta_i\le 1}\right| \,\middle|\, S,S'\right]$$

Hoeffding's inequality $\Rightarrow$ $P\left(\left|\frac{1}{m}\sum_{i=1}^m \theta_i\right| > \rho\right) \le 2\cdot e^{\frac{-2m\rho^2}{4}}$

Union Bound $\Rightarrow$ $P\left(\max_{h\in\mathcal{H}_C}\left|\frac{1}{m}\sum_{i=1}^m\theta_i\right| > \rho\right) \le 2|\mathcal{H}_C|\cdot P\left(\left|\frac{1}{m}\sum_{i=1}^m\theta_i\right| > \rho\right)$

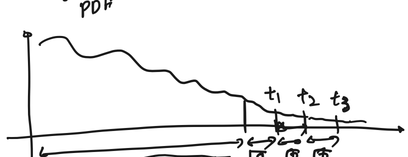
$$\le 2|\mathcal{H}_C|\cdot e^{-\frac{2m\rho^2}{4}}$$

$$(**) \Rightarrow \mathbb{E}\left[\max_{h\in\mathcal{H}_C}\left|\frac{1}{m}\sum_{i=1}^m \ell_h(z_i') - \ell_h(z_i)\right|\right] \le \frac{4 + \sqrt{\log(|\mathcal{H}_C|)}}{\sqrt{m/2}}$$

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

(**) $P(\theta > \rho) \le 2|\mathcal{H}_C|\cdot e^{-\frac{m}{2}\rho^2} \Rightarrow \mathbb{E}[\theta] \le \frac{4 + \sqrt{\log(|\mathcal{H}_C|)}}{\sqrt{m/2}}$ $\quad$ lemma A.4 book

proof>



$$\mathbb{E}[\theta] = \int_0^\infty \theta\cdot f(\theta)\,d\theta \le \sum_{i=1}^\infty t_i\, P(\theta > t_{i-1}) + \sqrt{\frac{2}{m}\cdot\log(2|\mathcal{H}_C|)}$$

$-(i-1+\sqrt{\log|\mathcal{H}_C|})^2$

$$\leq 2\sqrt{\tfrac{2}{m}} \cdot |\mathcal{H}_C| \cdot \sum_{i=1} \sqrt{\lg|\mathcal{H}_C| + i}\,)\, e \qquad +$$

$$\vdots$$

$$\leq 2\sqrt{\tfrac{2}{m}} + \sqrt{\tfrac{2}{m} \lg(2|\mathcal{H}_C|)}$$

---

proof of ① Sauer's Lemma >

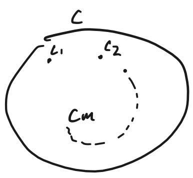
we will show that $\forall \mathcal{H}$

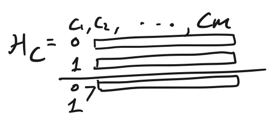
$$|\mathcal{H}_C| \underset{\text{n.t.s}}{\overset{(*)}{\leq}} |\{B \subseteq C : \mathcal{H} \text{ shatters } B\}| \overset{\text{def of VCdim}(\mathcal{H})=d}{\leq} \sum_{i=0}^{d} \binom{m}{i}$$

because largest set that can be shattered is $\underset{\text{size}}{\wedge}$ d.

proof of $(*)$ by Induction

- $m=1$: $\overset{\text{n.t.s}}{}$ $d=0 \longrightarrow |\mathcal{H}_C| \leq 1$. trivial since $d<1$. $\rightarrow$ does not shatter when $m=1$.
  $d=1 \longrightarrow |\mathcal{H}_C| \leq 2. = 2^m$
- suppose $(*)$ holds for sets $\overset{\text{size}}{}$ $\underset{|C'|=}{k} < m$, n.t.s. $(*)$ holds for set size $\underset{|C|=}{m}$.





$\forall B \subseteq C$ that $\mathcal{H}$ shatters B.

$$|\mathcal{H}_C| = \underset{\underset{\cancel{B}\ldots,}{\text{ass}\ \rotatebox{90}{$\geq$}}}{|Y_0|} + \underset{\underset{\{\qquad c_1 \in B,\ \mathcal{H} \text{ shatters } B\}}{\forall\ \wedge}}{|Y_1|}$$