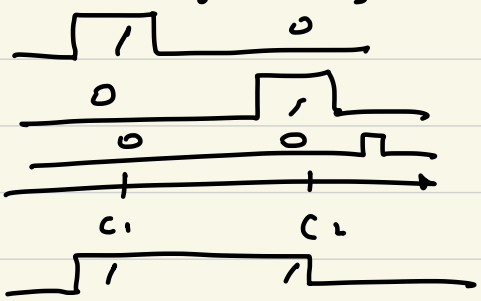
Lecture 14. (chapter 6)

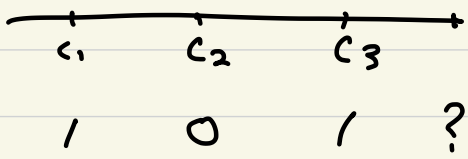
Recap: [VC-dimension of a hypothesis class $\mathcal{H}$]

The VC-dimension of $\mathcal{H}$, denoted by $VCdim(\mathcal{H})$, is the maximal size of $C \subseteq \mathcal{X}$ that can be shattered by $\mathcal{H}$.

* need to show that ① $\exists C$ s.t. $|C|=d$ shattered by $\mathcal{H}$
② $\forall C$ s.t. $|C|=d+1$ not shattered by $\mathcal{H}$

ex) $\mathcal{H}=\{\mathbb{I}_{[a\le x\le b]} \mid a,b\in\mathbb{R}\}$ shatters $|C|=2$ but does not shatter $\forall C$, $|C|=3$ $\Rightarrow VCdim = 2$.





1 0 1 ?

Theorem [Fundamental Theorem of Statistical Learning]

Let $\mathcal{H}$ be a hypothesis class from $\mathcal{X}$ to $\{0,1\}$, 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 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 is a successful PAC learner of $\mathcal{H}$
6. $\mathcal{H}$ has a finite VC-dimension.

we already learned that $1 \to 2$ (lecture 12)
$2 \to 3$, $3 \to 4$, $2 \to 5$ trivial
$2 \to 6$: $\sim 6 \to \sim 2$ (lecture 13. No-free-lunch thm).

left to show $6 \to 1$.

* we focus on binary classification.
  Also holds for other tasks like regression with quadratic loss.
  But exists example where PAC learnable but not uniformly convergent.
  chapter 13 exercise 2. [Uniform Convergence $m \propto \log d$
  [learnability $m \propto 1$: indep(d)

Theorem [Fundamental Theorem, Quantitative Version]
If $VCdim(\mathcal{H}) = d$, then

1. $\mathcal{H}$ has Uniform Convergence property with
$$C_1 \cdot \frac{d + \log 1/\delta}{\varepsilon^2} \leq m_{\mathcal{H}}^{U.C.}(\varepsilon, \delta) \leq C_2 \cdot \frac{d + \log 1/\delta}{\varepsilon^2}$$

2. $\mathcal{H}$ is Agnostic PAC learnable with
$$C_1 \cdot \frac{d + \log 1/\delta}{\varepsilon^2} \leq m_{\mathcal{H}}(\varepsilon, \delta) \leq C_2 \cdot \frac{d + \log 1/\delta}{\varepsilon^2}$$

3. $\mathcal{H}$ is PAC learnable with
$$C_1 \cdot \frac{d + \log 1/\delta}{\varepsilon} \leq m_{\mathcal{H}}(\varepsilon, \delta) \leq C_2 \cdot \frac{d \log 1/\varepsilon + \log 1/\delta}{\varepsilon}$$

proof later.

proof sketch of fundamental theorem of learning.

① [Sauer's lemma]

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 $O(|C|^d) \ll 2^{|C|}$

polynomial — worst-case exponential in $|C|$

*Recap [Restriction of $\mathcal{H}$ to $C = \{c_1, \cdots, c_m\} \subseteq \mathcal{X}$]

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

$$\tilde{h} = (\tilde{h}(c_1), \tilde{h}(c_2), \cdots, \tilde{h}(c_m))$$

$$= (0 \quad 1 \quad 0 \cdots 0)$$

② If $|\mathcal{H}_C|$ grows polynomially in $|C|$,
then we have <u>uniform convergence</u>: $\forall m \geq m_{\mathcal{H}}^{U.C}(\varepsilon, \delta)$

$$\mathbb{P}_S\left(|L_D(h) - L_S(h)| < \varepsilon, \forall h \in \mathcal{H} \right) \geq 1 - \delta.$$

In lecture 11, we used Höffding's to show $|\mathcal{H}| < \infty \to$ uniform convergence.

Definition [Growth function]

The growth function $\tau_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ is defined as

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

* the number of different functions from C to $\{0,1\}$, restricting $\mathcal{H}$ to C.
* If $m \leq VCdim(\mathcal{H}) = d$, then by definition $\tau_{\mathcal{H}}(m) = 2^m$

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

* If $m > d$, $\tau_{\mathcal{H}}$ grows polynomially in m.

Precisely,

① Lemma [Sauer-Shelah-Perles]

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

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

→ $m \leq d$, $\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{m} \binom{m}{i} = 2^m$ ← exponential

→ $m \geq d+1$, $\tau_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^d$ ← polynomial

↑ Stirling's formula

② Theorem [for Uniform Convergence] for every $D, \delta, h \in H$

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

Proof of fundamental theorem >

Apply ① Sauer's lemma to ②, for $m > d$

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

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

$$\leq \frac{2\sqrt{d \log \frac{2em}{d}}}{\delta \sqrt{m/2}} \quad \overset{\text{want}}{\leq} \quad \varepsilon$$

→ $$\frac{8d \log\left(\frac{2em}{d}\right)}{\delta^2 \varepsilon^2} \leq m$$

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

this is loose bound but enough for PAC learnability

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

Markov's ineq. ⇒ $P_S\left(\sup_{h\in H} |L_D(h) - L_S(h)| > \varepsilon\right) \le \frac{4 + \sqrt{\log(\tau_H(2m))}}{\varepsilon \cdot \sqrt{m/2}}$

How to handle when $|H| = \infty$ → Symmetrization

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

$L_D(h) = \mathbb{E}_{S'}[L_{S'}(h)]$; $S = \{z_i\}_{i=1}^m = (x_i, y_i)$; $S' = \{z_i'\}_{i=1}^m$

Jensen's ineq.: any convex $f(\cdot)$, e.g. $f = |x|$: $f(\mathbb{E}[Z]) \le \mathbb{E}[f(Z)]$

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

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

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

Symmetrization trick: to retain randomness even for fixed S, S'

z_i, z_i' i.i.d

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

$$\overset{\text{dist}}{=} \sigma_i \cdot (\ell_h(z_i') - \ell_h(z_i)), \quad \sigma_i \sim \begin{cases} +1 & \text{b.p } 1/2 \\ -1 & 1/2 \end{cases}.$$

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

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

S, S' fixed, let $C = S \cup S'$, $|C| \le 2m$, only H_C matters in the supremum

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

σ_i is random, $\mathbb{E}\sigma_i = 0$, $-1 \le \sigma \le 1$.

Hoeffding's ineq. $\rightarrow$ $\mathbb{P}\left(\left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i \right| > \rho \right) \le 2 e^{-\frac{2m\rho^2}{4}}$

Union bound $\rightarrow$ $\mathbb{P}\left(\max_{h \in \mathcal{H}_C} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i \right| > \rho \right) \le |\mathcal{H}_C| \cdot \mathbb{P}\left(\left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i \right| > \rho \right)$

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

Sub-Gaussian tail (**) $\rightarrow$ bounded expectation

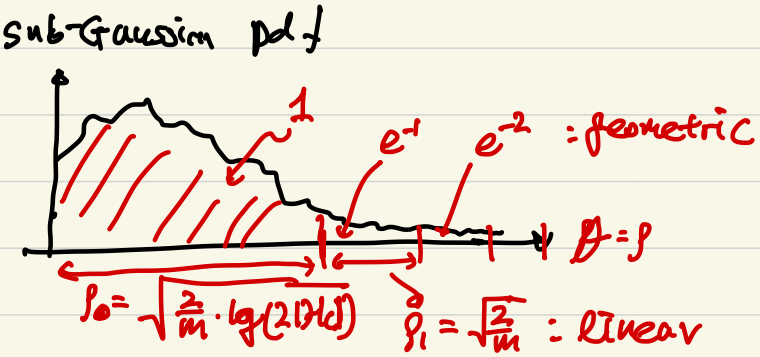
$$\mathbb{E}\left[\max_{h \in \mathcal{H}_C} \left| \frac{1}{m} \sum_{i=1}^{m} \left(l_h(z_i') - l_h(z_i) \right) \right| \right] \le \frac{4 + \sqrt{\log(|\mathcal{H}_C|)}}{\sqrt{m/2}}$$

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

(**) [lemma A.4 in the book]

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

Proof sketch >



$$\mathbb{E}[\theta] = \int_0^\infty \theta \cdot f(\theta) d\theta \le \rho_0 + \sum_{i=1}^{\infty} e^{-i} (\rho_0 + i \cdot \rho_1) \le \cdots \le 2\left[\sqrt{\frac{2}{m}} + \sqrt{\frac{2}{m} \cdot \log(2|\mathcal{H}_C|)} \right]$$