Recap: lecture 3 (chapter 4)

Hoeffding's ineq. $a \le \theta_i \le b$ (bounded), $\mathbb{E}[\theta_i] = \mu$,

$$\mathbb{P}\left(\left| \frac{1}{m}\sum_{i=1}^{m} \theta_i - \mu \right| > \varepsilon \right) \le 2e^{-\frac{2m\varepsilon^2}{(b-a)^2}}$$

Concentration of measure

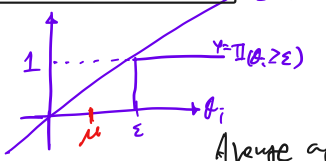
The more you know about the R.V. θ_i, the tighter concentration you can prove.

① [least information, Markov's inequality]

If $\theta_i \ge 0$ w.p.1 then $\mathbb{P}(\theta_i \ge \varepsilon) \le \frac{\mu}{\varepsilon}$ (no concentration, why? tight?)

and $\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m}\theta_i \ge \varepsilon\right) \le \frac{\mu}{\varepsilon} = O(1)$

Proof: $\mathbb{P}(\theta_i \ge \varepsilon) = \mathbb{E}[\mathbb{I}_{(\theta_i \ge \varepsilon)}]$ (indicator) $= \int_{\varepsilon}^{\infty} f_{\theta_i}(t)\,dt$



$\theta_i \ge 0 \Rightarrow \le \mathbb{E}\left[\frac{1}{\varepsilon}\theta_i\right] = \frac{\mu}{\varepsilon}$

Average age = 30, $m = 1{,}000{,}000$

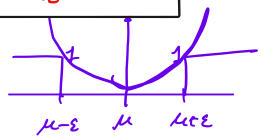
Q: How many over 70? $\frac{3}{4}$.

② [2nd order statistics, Chebyshev's inequality]

If $\mathrm{Var}(\theta_i) \le \sigma^2$ then $\mathbb{P}(|\theta_i - \mu| \ge \varepsilon) \le \frac{\sigma^2}{\varepsilon^2}$

and $\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| \ge \varepsilon\right) \le \frac{\sigma^2}{m\varepsilon^2} = O\left(\frac{1}{m}\right)$ (variance $\times \frac{1}{m}$)

Proof: $\mathbb{P}(|\theta_i - \mu| \ge \varepsilon) = \mathbb{E}[\mathbb{I}_{(\theta_i \le \mu-\varepsilon \text{ or } \theta_i \ge \mu+\varepsilon)}]$



Guided by the fact that Var $\le \sigma^2$: $\le \mathbb{E}\left[\frac{1}{\varepsilon^2}(\theta_i - \mu)^2\right] \le \frac{\sigma^2}{\varepsilon^2}$

---

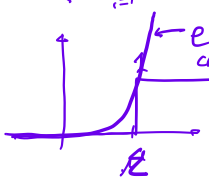
③ [bounded domain, Hoeffding's inequality]

If $a \le \theta_i \le b$ then $\mathbb{E}\left[e^{\lambda(\theta_i - \mu)}\right] \le e^{\frac{\lambda^2(b-a)^2}{8m^2}}$ (**)

and $\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| \ge \varepsilon\right) \le 2\cdot\exp\left(-\frac{2m\varepsilon^2}{(b-a)^2}\right) = O(e^{-m\varepsilon^2})$

Generic Recipe

$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu \ge \varepsilon\right) = \mathbb{E}\left[\mathbb{I}_{\left(\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu \ge \varepsilon\right)}\right]$



$\le \mathbb{E}\left[e^{\lambda\left(\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right) - \varepsilon}\right]$

$\le e^{-\lambda\varepsilon}\cdot\prod_{i=1}^{m}\mathbb{E}\left[e^{\lambda\frac{\theta_i - \mu}{m}}\right]$ ⟹

$\le e^{-\lambda\varepsilon}\prod_{i=1}^{m} e^{\frac{\lambda^2(b-a)^2}{8m^2}}$ ⟸

opt over λ → $\le \exp\left(-\lambda\varepsilon + \frac{\lambda^2(b-a)^2}{8m}\right)$

$-\varepsilon + \frac{2\lambda(b-a)^2}{8m} = 0$ → $\le \exp\left(\frac{-2m\varepsilon^2}{(b-a)^2}\right)$

$\lambda^* = \frac{4\varepsilon m}{(b-a)^2}$

do the same for $\frac{1}{m}\sum\theta_i < \mu - \varepsilon$

Special to Hoeffding's

(**) $\mathbb{E}\left[e^{\frac{\lambda(\theta_i - \mu)}{m}}\right] \le e^{\frac{\lambda^2(b-a)^2}{8m^2}}$

$\mathbb{E}\left[e^{\frac{\lambda(\theta_i - \mu)}{m}}\right] \le \max_{[a,b]} \mathbb{E}\left[e^{\frac{\lambda(\theta_i - \mu)}{m}}\right]$

convexity of exp(·) → $\le \max_{P_a, P_b} P_a e^{\frac{\lambda(a-\mu)}{m}} + P_b e^{\frac{\lambda(b-\mu)}{m}}$

$P_a + P_b = 1$, $aP_a + bP_b = \mu$. $\le e^{-\frac{\lambda\mu}{m}} \cdot \max_{P_a, P_b} P_a e^{\frac{\lambda a}{m}} + P_b e^{\frac{\lambda b}{m}}$

$\le e^{-\frac{\lambda\mu}{m}} \cdot \exp\left[\frac{\lambda a P_a + \lambda b P_b}{m} + \frac{(b-a)^2\lambda^2}{8m^2}\right]$ ($\lambda\mu/m$ cancels)

$\le \exp\left(\frac{(b-a)^2\lambda^2}{8m^2}\right)$

* Chernoff's Bound

$\theta_1 \cdots \theta_m$ independent but not identically distributed Bernoulli dist. $\theta_i \begin{cases} 1 & \text{w.p. } p_i \\ 0 & 1-p_i \end{cases}$

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^m \theta_i > (1+\delta)\frac{1}{m}\sum_{i=1}^m p_i\right) \le e^{-\underbrace{c(\delta)}_{\doteq (1+\delta)\lg(1+\delta)-\delta \ge \frac{\delta^2}{2+\frac{2}{3}\delta}} \cdot \underbrace{\sum_{i=1}^m p_i}_{O(m)}} \le e^{-\frac{\delta^2}{2+\frac{2}{3}\delta}\sum_{i=1}^m p_i}$$

(multiplicative; 2 phases)

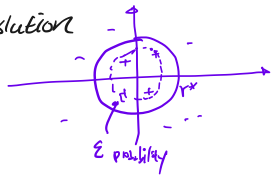
Q3.3 $\mathcal{X}=\mathbb{R}^2$, $\mathcal{Y}=\{0,1\}$, $\mathcal{H}=\{h_r : r\in\mathbb{R}_+\}$

$h_r(x) = \mathbb{I}_{\{\|x\|\le r\}}$

Prove that $\mathcal{H}$ is PAC learnable (assuming realizability), with

$$m_{\mathcal{H}}(\varepsilon,\delta) \le \left\lceil \frac{\lg 1/\delta}{\varepsilon} \right\rceil$$

solution



$\forall X, D, f_{r^*}$, define $H_{Bad} = \{r : r \le \tilde{r}, D(\{x : \tilde{r} \le \|x\| \le r^*\}) = \varepsilon\}$.

Algo: return smallest circle that includes all +.

$\mathbb{P}_S(\text{Alg} < P) \le (1-\varepsilon)^m \le e^{-\varepsilon m} \le \delta$.

for $m \ge \frac{1}{\varepsilon}\lg\frac{1}{\delta}$

Recap 2.

lecture 2 (ch 2, 3): PAC learning. ← Realizable, $\mathcal{Y}=\{0,1\}$, $|\mathcal{H}|<\infty$

Technique: Union Bound.

lecture 3 (ch 4): Agnostic PAC learning

$$L_D(h) \le \min_{h'\in\mathcal{H}} L_D(h') + \varepsilon \quad \text{w.p. } 1-\delta$$

Technique: concentration of measure.

Chapter 5.

the reason above problem has low sample complexity for PAC learning is that we had a prior knowledge [ground truth D has a circular decision boundary; $\mathcal{H}$ is collection of circular decision boundaries.

Q. is such prior knowledge necessary? yes.

Learning Problem, $\mathcal{X}, \mathcal{Y}$ fixed

Q. Can there be a universal algorithm Alg that is PAC learnable for all problem instances $(\mathcal{X}, \mathcal{Y}, D)$?

Q. is prior knowledge necessary for PAC learning?

Parametric family D → Realizable w.r.t. $\mathcal{H}$ (Rec 2) → $\min_{h\in\mathcal{H}} L_D(h)$ is small (implicitly Rec 3) → should use Agnostic PAC, otherwise.

How much prior knowledge is good to assume?

a lot

- Learner knows D, h^* → $\mathcal{H}=\{h^*\}$ achieves minimum risk (e.g. knows Bayes Optimal)
- Learner knows a class
  - lecture 2: realizable finite $\mathcal{H}$ → PAC learnable $m \approx \frac{\lg\frac{|\mathcal{H}|}{\delta}}{\varepsilon}$
  - lecture 3: e.g. $\mathcal{H}=\{h_r : \mathbb{I}_{\text{sphere}(r)}\}$ → PAC learnable $m \approx \frac{\lg\frac{1}{\delta}}{\varepsilon}$
  - non-realizable finite $\mathcal{H}$ → Agnostic PAC learnable $m \approx \frac{\lg\frac{|\mathcal{H}|}{\delta}}{\varepsilon^2}$
- Learner knows nothing: no-free-lunch theorem.

NO

No-free-lunch theorem

Theorem. Let Alg be any learning algorithm. $\mathcal{Y}=\{0,1\}$, any $\mathcal{X}$, $m \le \frac{|\mathcal{X}|}{2}$

Then exists a distribution D over $\mathcal{X}\times\mathcal{Y}$ s.t.

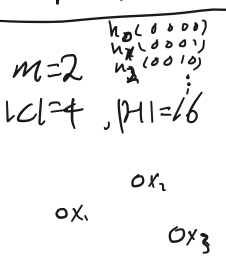
1. $\exists f : \mathcal{X} \to \{0,1\}$ with $L_D(f)=0$

Constructive Proof ⇒ Negative Result.

2. $P_S(L_D(A_{lgo}(S)) \geq 1/8) \geq 1/7$

proof sketch: $|\mathcal{X}| \geq 2m$, $C \subseteq \mathcal{X}$, and $|C| = 2m$, , and we observe random S of m samples.

$m=2$, $|C|=4$, $|\mathcal{H}|=16$



$|\mathcal{H}| = 2^{2m}$ all possible (ways to) labels.

$= \{h_0, h_1, \dots h_{2^{2m}-1}\}$

for any A_{lgo}, $\exists$ ① s.t. one cannot do better than random guess on $1/2$ of the samples not observed.

→ $\mathop{\mathbb{E}}_{S}[L_D(A(S))] \geq 1/4$ → claim by the Markov inequality.

$L_D(A(S))$: positive R.V.; $1/4$: $1/2$ of samples error $1/2$.

Corollary. $\mathcal{X}$ is infinite & $\mathcal{H}$ is all possible functions from $\mathcal{X}$ to $\{0,1\}$ then $\mathcal{H}$ is not PAC learnable.

Q. How do we choose the right hypothesis class?

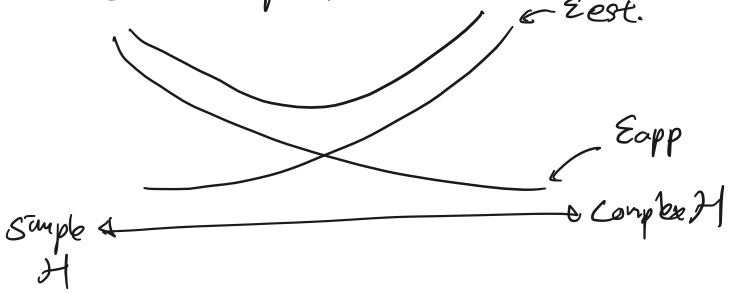
$$L_D(h_S) = \underbrace{\left(L_D(h_S) - \min_{h \in \mathcal{H}} L_D(h)\right)}_{\varepsilon_{est}} + \underbrace{\min_{h' \in \mathcal{H}} L_D(h')}_{\varepsilon_{app}}$$

h_S: ERM

ε_{est} = Estimation error
- goes down with m
- depends on complexity $\mathcal{H}$

ε_{app} = approximation error
- does not depend on m
- property of D, $\mathcal{H}$. Complex

Bias-Complexity Tradeoff.



Chapter 6.

realizable + finite $\mathcal{H}$ → PAC learnable (lec 2)

finite $\mathcal{H}$ → Agnostic PAC learnable (lec 3).

realizable + infinite $\mathcal{H} = \{h_r : \mathbb{I}_{(|x| \leq r)}\}$ → PAC learnable (lec 4)

$C \subset \mathcal{X}$, $|C| \geq 2m$, $\mathcal{H}$ contains all possible 2^{2m} labeling → no learning (lec 4)

Q. what class $\mathcal{H}$ is PAC learnable? Vladimir Vapnik & Alexey Chervonenkis 1970

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

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

$c_i \mapsto h(c_i)$

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

Definition. (Shattering)

A hypothesis class $\mathcal{H}$ shatters a set $C \subseteq X$ if the restriction of $\mathcal{H}$ to C is the set of all functions from C to $\{0,1\}$. That is $|\mathcal{H}_C| = 2^{|C|}$.

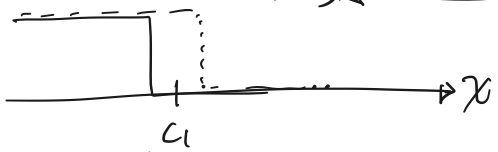
Example. $\mathcal{H} = \{h_a : \mathbb{R} \to \mathbb{R} = \mathbb{I}_{(x \le a)}\}$ class of all threshold functions.

① $C = \{c_1\}$

$\mathcal{H}_C = \{(h(c_1)=1), (h(c_1)=0)\}$

$h_{c_1+\beta}$

$h_{c_1-\beta}$



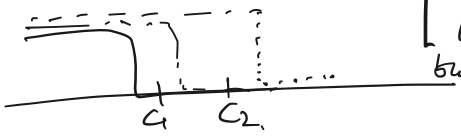
$\therefore$ $\mathcal{H}$ shatters $C = \{c_1\}$

② $C = \{c_1, c_2\}$, $\mathcal{H}_C = \{(0,0), (1,0), (1,1)\}$ but not $(0,1)$

$h_{\frac{c_1+c_2}{2}}(\cdot)$

$h_{c_1-\beta}(\cdot)$

$h_{\frac{c_1+\beta}{2}}(\cdot)$



$\therefore$ $\mathcal{H}$ does not shatter $C = \{c_1, c_2\}$

Corollary (No-free-lunch thm restated)

for some $\mathcal{H}$, X, $Y = \{0,1\}$, if $\exists C \subseteq X$, $|C| = 2m$ & shattered by $\mathcal{H}$,

then for any learning algorithm, $\exists D$ & $h \in \mathcal{H}$ s.t.

$L_D(h) = 0$ but $P(L_D(A(S)) \ge \frac{1}{8}) \le \frac{1}{7}$.

Definition (VC-dimension)

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

↳ no set of size $|C|+1$ can be shattered.

Example. need to show that $\exists C$ of $|C| = d$ shattered (say), & $\forall C$ of $|C| = d+1$, not shattered.

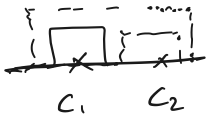
Example (Threshold: $\mathcal{H} = \{\mathbb{I}_{(x \le a)}\}$)

$\exists\, C = \{c_1\}$ can be shattered.

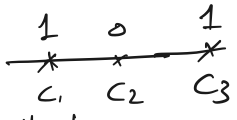
all $C=\{c_1,c_2\}$ can not be shattered.

$VCdim(\mathcal{H})=1$.

Example (Intervals: $H=\{\mathbb{I}_{(a<x<b)}: a\le b\in\mathbb{R}\}$)



existence of shattered set

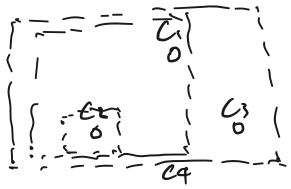


all set not shattered.

$VCdim(H)=2$.

Example (axis aligned rectangles: $H=\{\mathbb{I}_{(a_1\le x_1\le b_1,\ a_2\le x_2\le b_2)}: \begin{matrix}a_1\le b_1\\ a_2\le b_2\end{matrix}\}$

$VCdim(H)=4$



existence of shattered set.

all set of 5 not shattered.

Example (finite classes: $|\mathcal{H}|<\infty$)

shattering requires at least $2^{|C|}$ hypothesis. so

$$2^{|C|}\le|\mathcal{H}|$$

$$\rightarrow\quad |C|\le \log_2|\mathcal{H}|$$

$$\rightarrow\quad VCdim\le \log_2|\mathcal{H}|$$

↳ tight in the worst case
↳ VCdim can be much smaller.
cf. finite & threshold.

Example (number of parameters)

above examples have # parameters match VCdim,

Not always true: $\mathcal{H}=\{h_\theta:\theta\in\mathbb{R}\}$, $VCdim(H)=\infty$ [HW1 Problem]

$\|$

$[\,0.5\sin(\theta x)\,]$

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 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.

Theorem [Fundamental theorem. Quantitative version]
$VCdim(\mathcal{H}) = d$, then

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