· Project Version 1 due today

· Theory HW 1 due Thursday

Lecture 13

Recap

Concentration of measure

① $\theta \geq 0$, $\mathbb{E}\theta = \mu$ $\longrightarrow$ $\mathbb{P}(\theta > \varepsilon) \leq \frac{\mu}{\varepsilon}$

② $\mathbb{E}\theta = \mu$, $\mathrm{Var}\theta = \sigma^2$ $\longrightarrow$ $\mathbb{P}(|\theta - \mu| > \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$

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

③ $\mathbb{E}\theta = \mu$, $a \leq \theta \leq b$ $\longrightarrow$ $\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \varepsilon\right) \leq 2 \cdot e^{-\frac{2m\varepsilon^2}{(b-a)^2}}$

④ $\mathbb{E}\theta_i = p_i$, $\theta_i \in \{0,1\}$ $\longrightarrow$ $\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m}\theta_i > (1+\delta)\mu\right) \leq e^{-\frac{\delta^2}{2+\frac{2}{3}\delta}\cdot\mu m}$

$\mathbb{E}\frac{1}{m}\sum_{i=1}^{m}\theta_i = \frac{1}{m}\sum_{i=1}^{m}p_i = \mu$

Recap.

lecture 10. (ch 2&3)

- PAC learning ERM ← realizable $f \in H$, $y \in \{0,1\}$, $|H| < \infty$
- technique: union bound

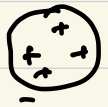
$$m_H = \frac{\log\frac{|H|}{\delta}}{\varepsilon}$$

lecture 11. (ch 4)

- Agnostic PAC learning ERM ← non-realizable, any bounded loss, $|H| < \infty$
- technique: Hoeffding's ineq.

$$m_H = \frac{\log\frac{|H|}{\delta}}{\varepsilon^2}$$

Q3.3



$H = \{h_r : r \geq 0\}$, $h_r = \mathbb{I}_{\{\|x\| \leq r\}}$, $|H| = \infty$

$$m_H = \frac{\log\frac{1}{\delta}}{\varepsilon}$$ → How?

Chapter 5. No free lunch theorem.

The reason Q.3.3 has low sample complexity is that we had a prior knowledge.

- Realizable: [ground truth D, f has a circular boundary
  H is a collection of circular decision boundries.

fundamental Question:

- is such prior knowledge necessary for PAC learning? YES.

Learning Problem $\Rightarrow$ Goal: minimize $L_D(h)$

- fix X, Y
- Nature chooses D_{xy}
- Learner observes $S = \{(x_i, y_i)\}_{i=1}^m \sim D$
- Applies Alg over H.
- Outputs $h \in H$.

Q. Can there be a universal Algorithm that is PAC learnable for all D?

Q. Is Prior Knowledge necessary for PAC learning?

- lec 10: realizable w.r.t. H: $|H| < \infty$
- lec 11: $\min_{h \in H} L_D(h)$ is small, $|H| < \infty$
- Q 3.3: realizable w.r.t Parametric family H

Q. How much prior knowledge is good to assume?

a lot : learner knows $D \rightarrow h^*$: Bayes optimal Predictor.

learner knows a class { realizable, $H \subseteq$ ⊕, $|H| < \infty$

No : learner knows nothing → [No free lunch theorem]

No-free-lunch theorem

Let Alg be any learning algorithm

$Y = \{0, 1\}$

any X s.t. $|X| \geq 2 \cdot m$

Then exists a distribution D over $X \times Y$ and H s.t.

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

② $P_S(L_D(Alg(S)) \geq \frac{1}{8}) \geq \frac{1}{7}$

Negative result → Constructive proof (sketch)

since $|X| \geq 2m$, let $C \subseteq X$ s.t. $|C| = 2m$

we observe random m samples in X.

Let H be all possible ways to label C s.t. $|H| = 2^{2m}$

$= \{h_0, h_1, \cdots, h_{2^{2m}-1}\}$ ← * Very high degrees of freedom in assigning y's.

*unstructured

$m=2$, $|H| = 2^{2 \cdot m} = 16$

$|C| = 2m$

X $H = \{(y_1, y_2, y_3, y_4)$

x_1 0000, 0001,

0010, 0011,

x_2 ⋮ }

x_3 x_4

for any algorithm that outputs $h \in \mathcal{H}$,

$\exists D$ s.t. one cannot do better than random guess on $\frac{1}{2}$ of C not observed.

$\rightarrow \mathbb{E}_S [L_D(A_{f_0}(S))] \geq \frac{1}{4}$

$L_D(A_{f_0}(S))$: positive R.V.

$\frac{1}{4}$: $\frac{1}{2}$ of samples have error $\frac{1}{2}$

$\rightarrow$ This implies the claim by tighter Markov inequality.

$0 \leq \theta \leq 1$.

Corollary. If $|\mathcal{X}| = \infty$ & $\mathcal{H}$ 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?

$$L_D(h_S) = \left(L_D(h_S) - \min_{h' \in \mathcal{H}} L_D(h') \right) + \min_{h' \in \mathcal{H}} L_D(h')$$

h_S: ERM

$\left(L_D(h_S) - \min_{h' \in \mathcal{H}} L_D(h') \right)$: ε_{est} = estimation Error

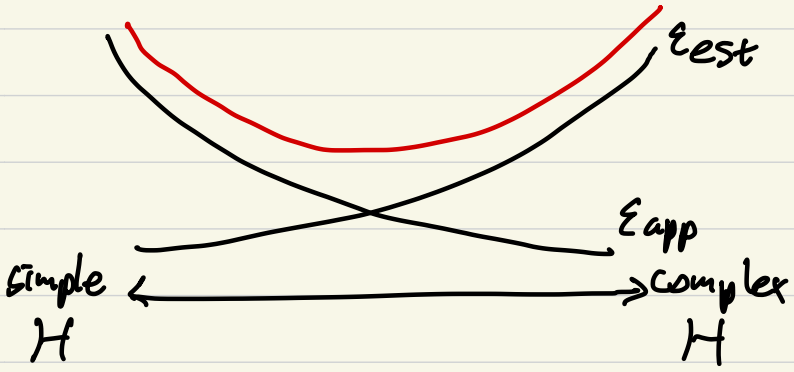
$\min_{h' \in \mathcal{H}} L_D(h')$: ε_{app} = approximation Error

If $m \uparrow$: $\varepsilon_{est} \downarrow$

If complexity of $\mathcal{H} \uparrow$: $\varepsilon_{est} \uparrow$

no change $\varepsilon_{app} \downarrow$

Bias-Complexity tradeoff



Chapter 6. VC-dimensions

fundamental Question
- what class $\mathcal{H}$ is PAC learnable?
[Vladmir Vapnik & Alexey Chervonenkis, 1970]

Definition. [Restriction of $\mathcal{H}$ to $C \subseteq \mathcal{X}$]

$\mathcal{H}$ is class of functions: $\mathcal{X} \to \{0,1\}$, and
$C = (c_1, c_2, \cdots, c_m) \subseteq \mathcal{X}$, then
the restriction of $\mathcal{H}$ to C is

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

$\tilde{h}$ can equivalently represented by an m-dim vector
$\tilde{h} = (\tilde{h}(c_1), \tilde{h}(c_2), \cdots, \tilde{h}(c_m))$

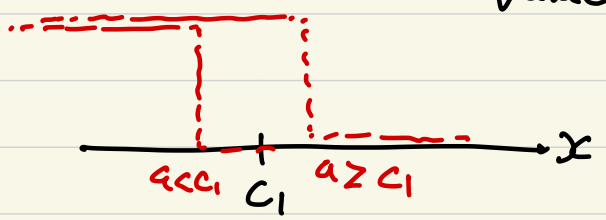
Definition. [Shattering]

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

example> $\mathcal{H} = \{ h_a : \mathbb{R} \to \mathbb{R} = \mathbb{I}\{x \le a\} \}$ is class of all threshold functions.

① $C = \{c_1\}$

$\mathcal{H}_C = \{ 0 , 1 \}$
0: $h_a(c_1) = 0$ to $a < c_1$
1: $h_a(c_1) = 1$ $a \ge c_1$



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

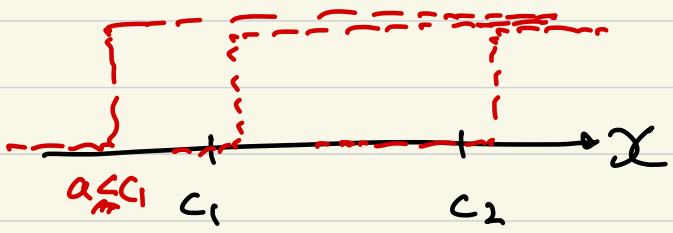
② $C = \{c_1, c_2\}$

$H = \{(0,0), (1,0), (1,1)\}$

$(h_a(c_1), h_a(c_2))$

$c_2 \le a$ $\quad$ $c_1 \le a < c_2$ $\quad$ $a < c_1$

$(1,1)$ is not possible



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

Corollary [No-free-lunch theorem restated]

for some H, X, $Y = \{0,1\}$,

if $\exists C \subseteq X$, $|C| = 2m$, and H shatters C,

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

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

Definition [VC-dimension]

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

↑ No set of size $VC_{dim}(H)+1$ can be shattered.

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

Example [Threshold : $\mathcal{H} = \{ \mathbb{I}(x \le a) \mid a \in \mathbb{R} \}$

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

$\forall C = \{c_1, c_2\}$ cannot be shattered.

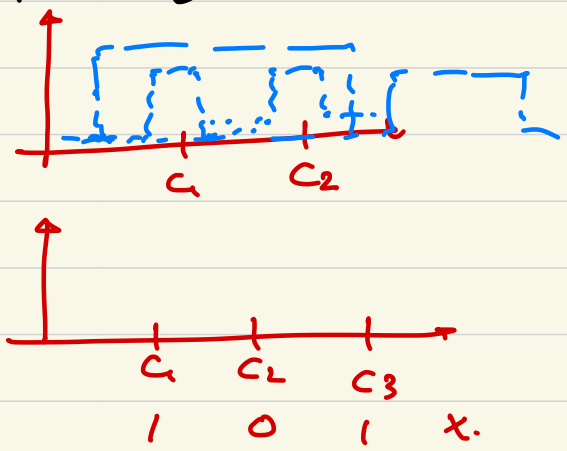
$\rightarrow VCdim(\mathcal{H}) = 1.$

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

$\exists C = \{c_1, c_2\}$ can be shattered

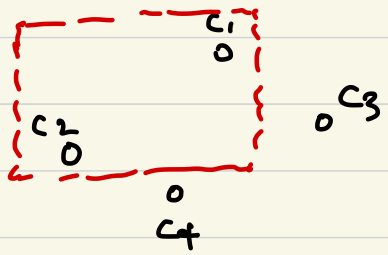
$\forall C = \{c_1, c_2, c_3\}$ cannot be shattered.

$\rightarrow VCdim(\mathcal{H}) = 2.$



Example [axis aligned rectangles : $\mathcal{H} = \{ \mathbb{I}_{\{ a_1 \le x_1 \le b_1,\ a_2 \le x_2 \le b_2 \}} \text{ for } a_1 < b_1,\ a_2 < b_2 \}$

$VCdim(\mathcal{H}) = 4$



any subset can be in a rectangle

however you place 5 points,
1 of them has to be "inside"
and cannot be a unique ○.

Example [finite class $|\mathcal{H}| < \infty$]

$VCdim \leq \log_2 |\mathcal{H}|$

Shattering requires at least $2^{|C|}$ hypothesis,

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

$$|C| \leq \log_2 |\mathcal{H}|$$

$$VCdim \leq \log_2 |\mathcal{H}|$$

Tight in the worst case; can be smaller like interval.

Example [number of parameters]

- above examples have # of parameters match VCdim.
- not always true: $\mathcal{H} = \{h_\theta : \theta \in \mathbb{R}\}$, $VCdim = \infty$ [Theory HW 2 Problem 4]

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