Lec 33

Recap: PAC learning

↳ $H = \{h_1, h_2, \cdots h_M\}$

↳ For final hypothesis: $g \in H$

↳ Generalization errors:

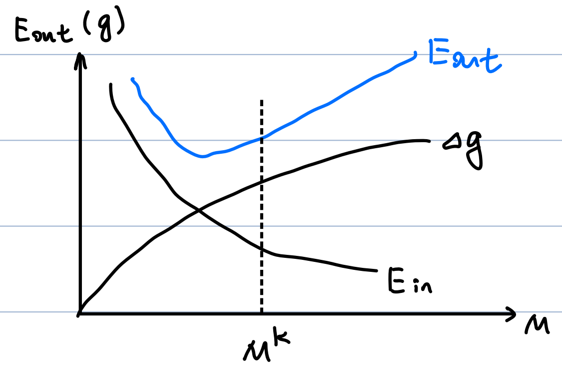
$$\Delta g = |E_{in}(g) - E_{out}(g)|$$

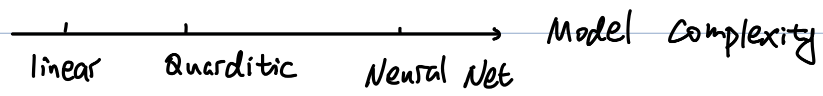
$$\Pr\{\Delta(g) \geq \varepsilon\} \leq \Pr\{\bigcup_{i=1}^{M} \{\Delta(h_i) \geq \varepsilon\}\}$$

$$\underset{\text{union binded}}{\leq} 2Me^{-2N\varepsilon^2}$$

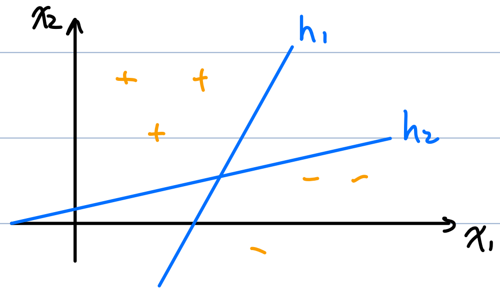
$\Rightarrow$ with prob. $\geq 1-\delta$

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2N}\log\frac{2M}{\delta}}$$





↳ Trouble: union bound is loose in general



h_1 and h_2 has same effect on $\mathcal{D}$.

1. How to get rid of union bound?

↳ Idea:

Consider arbitary hypo. h

$E_{in}(h) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{y_n \neq h(\underline{x}_n)\}$

depends on h only through $\{h(\underline{x}_1) \cdots h(\underline{x}_N)\}$

∴ Replace M with an "effective #. of hypothesis" $m_H(N)$

2. Dichotomy

↳ Let $\underline{x}_1, \cdots \underline{x}_N \in \mathbb{R}^d$ be fixed points

Let $h \in H$ be some hypo. $h: \mathbb{R}^d \to \{+1, -1\}$

↳ Def: Dichotomy vector

$(h(\underline{x}_1), h(\underline{x}_2) \cdots h(\underline{x}_N)) \in \{+1, -1\}^N$

↳ Def: Dichotomy set

$H(\underline{x}_1, \underline{x}_2, \cdots \underline{x}_N) = \{[h(\underline{x}_1), \cdots h(\underline{x}_N)] ; h \in H\}$

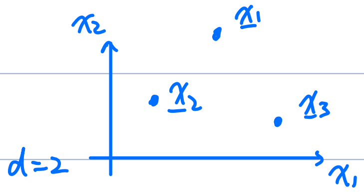
i.e. the collection of all binary vectors generated by H on $\underline{x}_1 \cdots \underline{x}_N$

- Note: there are no repeated element in a set

$\therefore |H(\underline{x}_1, \cdots \underline{x}_N)| \leq 2^N$

(No matter how large $M = |H|$ is!)

e.g.



perceptron learning $|H| = \infty$

(i.e. can draw ∞ No. of lines)

then $|H(\underline{x}_1, \underline{x}_2, \underline{x}_3)| \leq 8$

↳ Def: The hypothesis set $\mathcal{H}$ shatters $(\underline{x}_1 \cdots \underline{x}_N)$

if $|\mathcal{H}(\underline{x}_1, \underline{x}_2 \cdots \underline{x}_N)| = 2^N$

e.g.1 Linear classification in $d=2$

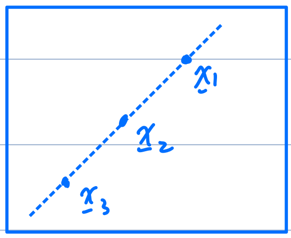
$\mathcal{H} \equiv$ all $\underline{w} \in \mathbb{R}^3$

$y = \text{sign}(w_0 + w_1 x_1 + w_2 x_2)$

$h(\underline{x}_1)$	$h(\underline{x}_2)$	$h(\underline{x}_3)$	can generat class. line?
+	+	+	✓
+	+	−	✓
+	−	+	✓
+	−	−	✓
⋮	⋮	⋮	⋮

∴ the linear classifier shutters the dataset

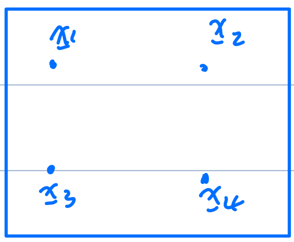
e.g.2. Colinear case:



$h(\underline{x}_1)$	$h(\underline{x}_2)$	$h(\underline{x}_3)$	can generat class. line?
+	+	+	✓
+	+	−	✓
+	−	+	✗
+	−	−	✓
⋮	⋮	⋮	⋮ (3✓ 1✗)

∴ $|\mathcal{H}(\underline{x}_1, \underline{x}_2, \underline{x}_3)| = 6$ ∴ Not shattering.

e.g. 3:



$|\mathcal{H}(\underline{x}_1, \underline{x}_2 \cdots \underline{x}_4)| = 14 < 2^4$

∴ Not shattering

↳ Def: <u>Growth function:</u>

$$m_H(N) = \max_{\underline{x}_1, \underline{x}_2, \cdots \underline{x}_N \in \mathbb{R}^d} |H(\underline{x}_1, \underline{x}_2, \cdots \underline{x}_N)|$$

all possible location of points

$\equiv$ Effective #. of hypothesis

e.g. For linear classification in $d=2$

$m_H(3) = 8$ $\quad$ $m_H(4) = 14$

means at most 14 hypo. if we have 4 data points.
worst case

· Note:

① $m_H(N) \le 2^N$

② A tighter bound can be found using the VC dimension.

↳ Def:

Let k be an int s.t. $m_H(k) < 2^k$,

then k is a <u>break point</u> of H.

e.g. For linear class. $d=2$, break point $k = 4$.

↳ def: <u>VC dimension:</u>

Let N be an int s.t $m_H(N) = 2^N$, and $m_H(N+1) < 2^{N+1}$

(i.e. N+1 is the first break point)

then the VC dimension of H is:

$d_{vc}(H) = N$