Machine Learning HW2

B11901174 傅啟恩

1. By one-sided Hoeffding's inequality,

$$P(\mu > \nu + \epsilon) \le \exp(-2\epsilon^2 N) \quad \text{, where in this case,}$$

$\nu = \frac{C_m}{N_m}$ sampled prob., real probability $\mu = \mu_m$, choose $\epsilon = \sqrt{\frac{\ln t - \frac{1}{2}\ln\delta}{N_m}}$

$$\exp(-2\epsilon^2 N) = \exp\left(-2 \cdot \frac{\ln\frac{t}{\sqrt{\delta}}}{\cancel{N_m}} \cancel{N_m}\right)$$

$$= \exp\left(\ln\frac{\delta}{t^2}\right) = \frac{\delta}{t^2} = \delta t^{-2}$$

$$\Leftrightarrow P\left(\mu_m > \frac{C_m}{N_m} + \sqrt{\frac{\ln t - \frac{1}{2}\ln\delta}{N_m}}\right) \le \delta t^{-2}$$

#

2.

$$\sum_{\substack{t=1 \\ M+1}}^{\infty} \sum_{m=1}^{M} \exp\left(-2 \cdot \frac{\ln\frac{tM}{\sqrt{\delta}}}{N_m} N_m\right)$$

$$= \sum_{\substack{t=1 \\ M+1}}^{\infty} \frac{\delta t^{-2}}{M^2} \cdot M \le \delta M^{-2}\frac{\pi^2}{6} \le \delta$$

for $m = 1, 2, \dots, M$, $t = M+1, M+2 \dots \infty$

$$P\left(\mu_m \le \frac{C_m}{N_m} + \sqrt{\frac{\ln t - \frac{1}{2}\ln\delta + \ln M}{N_m}}\right) \ge 1 - \delta$$

#

3. For each ticket, if all tickets A/C or A/D or B/C or B/D (4 kinds) would cause there's some "green"

$\therefore$ the probability would be $\left(\frac{1}{2}\right)^5 \times 4$ kinds $- \left(\frac{1}{4}\right)^5 \times 4$

$$= \frac{31}{256}$$

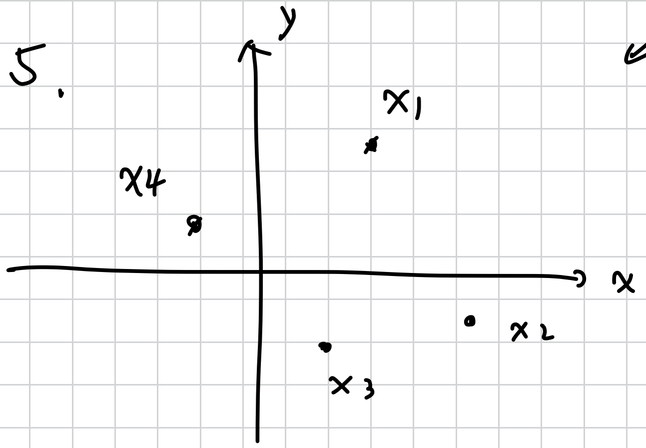
#

4. P(five tickets contain 5 green 2's) would be

all tickets choose B or D

$= \left(\frac{1}{2}\right)^5 = \frac{1}{32}$ #

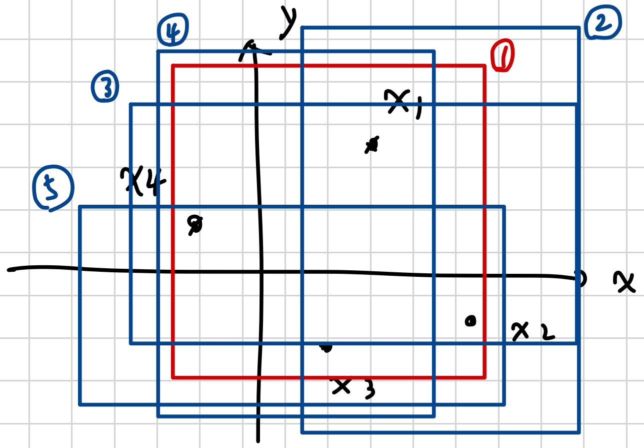
5.

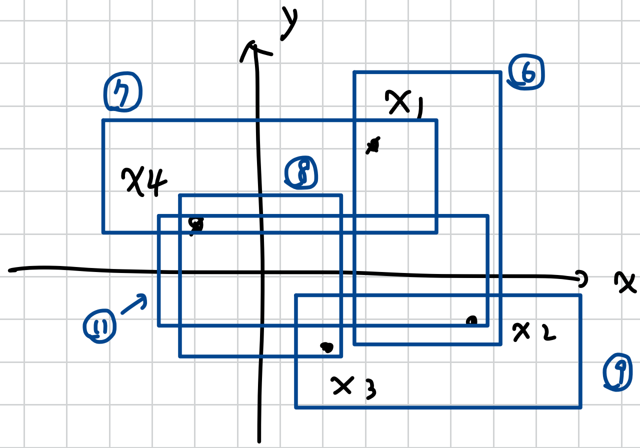


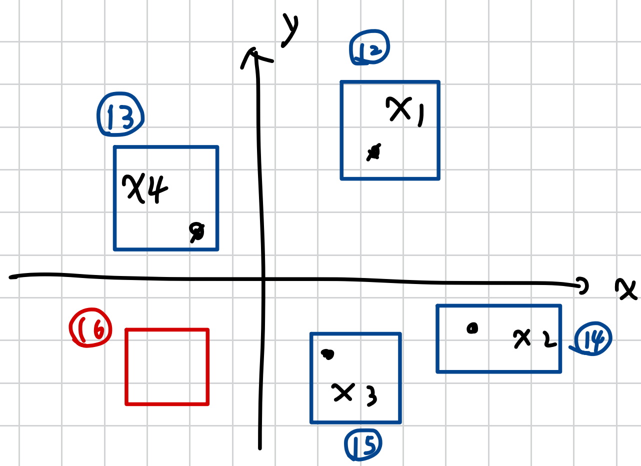
↙ If x_1, x_2, x_3, x_4 are mapped to R^2 like this,

Choose "o"	as -1	← in the rect	
"x"	as $+1$	← outside	"

	x_1	x_2	x_3	x_4
1	o	o	o	o
2	o	o	o	x
3	o	o	x	o
4	o	x	o	o
5	x	o	o	o
6	o	o	x	x
7	o	x	x	o
8	x	x	o	o
9	x	o	o	x
10	o	x	o	x
11	x	o	x	o
12	o	x	x	x
13	x	x	x	o
14	x	o	x	x
15	x	x	o	x
16	x	x	x	x





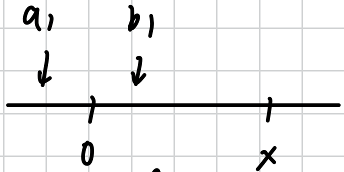


⇒ for 4-point vectors x_1, x_2, x_3, x_4, there exists some x_1, x_2, x_3, x_4 that can be shattered by the hypothesis. ($N=4$ IS NOT the break point)

⇐ The VC dimension is no less than 4 !
(= break point − 1)

6.

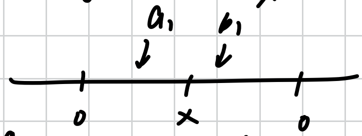
for $N = 2$



$M = 1$

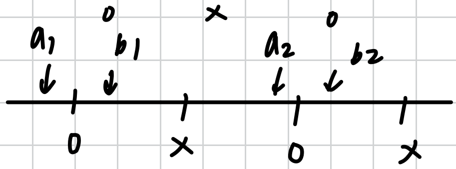
交錯排列的 dichotomy will need the most "M"s

$N = 3$



$M = 1$ at least

$N = 4$



$M = 2$

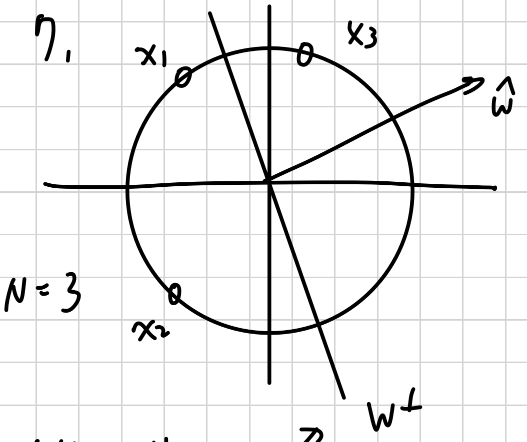
⋮

for N inputs, there exists at least $\lfloor N/2 \rfloor$ Non-adjacent intervals that marked as "O" or "X" as shown above, so we need $a_1, b_1, a_2, b_2 \dots a_M, b_M$, where $M \geq \lfloor \frac{N}{2} \rfloor$ to shatter the hypothesis. Therefore, the break point of the hypothesis is 2^*M+2, the VC Dimension =

$2^*M+2-1 = \underline{2^*M+1}$ #

(for $M=1$, hypothesis cannot shatter $2^*M+2 = 4$ inputs data because it would contain o x o x case.)

7.



$H_0: \{h: h(x) = sign(w_1 x_1 + w_2 x_2)\}$

Think as $\hat{w} = (w_1, w_2)$

$\hat{x} = (x_1, x_2)$

$h(x) = sign(\hat{w}^T \hat{x})$

(We call the $\vec{\ }$ normal vector of w "$w^{\perp}$")

We can know that $w^{\perp}$ separate the $\mathbb{R}^2$ into 2 sides, one that causes $sign(\hat{w}^T\hat{x}) > 0$ with $\hat{x}$ falls on the side, the other causes $sign(\hat{w}^T\hat{x}) \leq 0$

x_1	x_2	x_3	
×	×	o	Initial state
o	×	o	touch x_1
o	o	o	x_2
o	o	×	x_3
×	o	×	x_1
×	×	×	x_2
~~×~~	~~×~~	~~o~~	x_3

table 1

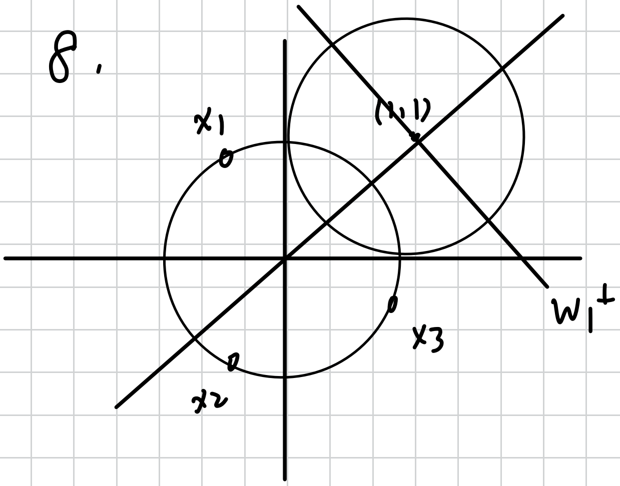
$\Rightarrow$ In this figure shown above, we suppose at Initial state, x_1 = "x", x_2 = "x", and x_3 = "o", as we rotate the $w^{\perp}$ by counterclockwise, we will then create different "dichotomies" that satisfy the hypothesis (for some $\hat{w}$)

After rotating 360°, $w^{\perp}$ will touch 2^*N points, causes $2^*N + 1 \underbrace{-1}_{\text{repeat}}$

$= 2^*N$ dichotomies

$$\Rightarrow \underline{m_H(N) = 2^*N} \ \#$$

8.



$H_0 : \{ h : h_0(x) = \text{sign}(w_1 x_1 + w_2 x_2) \}$

$H_1 : \{ h : h_1(x) = \text{sign}(w_1(x_1 - 1) + w_2(x_2 - 1)) \}$

From problem 7, $N=3$ inputs has VC Dimension $= 2$, $m_H(3) = 2*3 = 6$ for origin-passing perceptrons.

If we consider $H = H_0 \cup H_1$, we can create lines ($W_1^{\perp}$) such that append 2 dichotomies ($x_1 = x_2 = x_3 =$ "o" / "x"),

$\Leftrightarrow$ there exists some x_1, x_2, x_3 such that hypothesis H can shatter the case ($2*3 + 2 = 8 = 2^{(N=3)}$)

$\Rightarrow$ VC Dimension of H is at least 3.

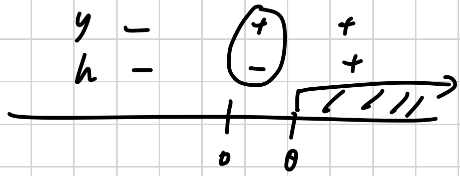
Now we consider $N=4$ points, because in PLA we know that $N=4$ cannot be shattered, and the set of PLA (perceptrons) lines ($W^{\perp}$) $\supset$ {origin-passing $\cup$ (1,1)-passing lines ($W^{\perp}$ in H)}

$\Rightarrow$ Thus, if PLA cannot shatter $N=4$, hypothesis H must also cannot shatter $N=4$

$\Rightarrow$ $\underline{\text{VC Dimension of } H = 4 - 1 = 3}$ #

9. $h_{s,\theta}(x) = s \cdot \text{sign}(x-\theta)$, $y = \text{sign}(x) + 10\%$ flip

case 1 : $s = +1$



$$E_{out}(h_{s,\theta}) = P(h_{s,\theta}(x) \neq y)$$

$$= 0.9 \times P(\text{sign}(x-\theta) \neq \text{sign}(x)) + 0.1 \times P(\text{sign}(x-\theta) = \text{sign}(x))$$

I. consider $\theta > 0$

$$P(\text{sign}(x-\theta) \neq \text{sign}(x)) = P(0 \leq x \leq \theta) = \frac{\theta}{2}$$

$$P(\text{sign}(x-\theta) = \text{sign}(x)) = 0.5 + P(x \geq \theta)$$

$$= 0.5 + 0.5(1-\theta)$$

$$= 1 - 0.5\theta$$

II. consider $\theta < 0$

$$P(\text{sign}(x-\theta) \neq \text{sign}(x)) = P(\theta \leq x \leq 0) = \frac{-\theta}{2}$$

$$P(\text{sign}(x-\theta) = \text{sign}(x)) = 0.5 + P(x \leq \theta)$$

$$= 0.5 + \frac{\theta+1}{2}$$

$$= 1 + 0.5\theta$$

conclusion

$\Rightarrow$ $P(\text{sign}(x-\theta) \neq \text{sign}(x)) = \frac{|\theta|}{2}$

$P(\text{sign}(x-\theta) = \text{sign}(x)) = 1 - 0.5|\theta|$

$$E_{out}(h_{s,\theta}) = 0.9 \times \frac{|\theta|}{2} + 0.1(1 - 0.5|\theta|)$$

$$= 0.45|\theta| + 0.1 - 0.05|\theta|$$

$$= 0.1 + 0.4|\theta|$$

case 2: $s = -1$

$$E_{out}(h_{s,\theta}) = 0.9 \cdot (1 - 0.5|\theta|) + 0.1 \times \frac{|\theta|}{2}$$

$$= 0.9 - 0.45|\theta| + 0.05|\theta|$$

$$= 0.9 - 0.4|\theta|$$

combine case 1+2

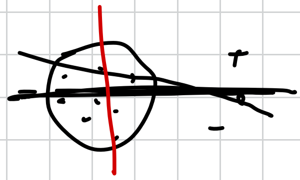
$\Rightarrow$

$$E_{out}(h_{s,\theta}) = 0.5 - 0.4s + 0.4s \cdot |\theta|$$

#

13.

Suppose $C(N,d)$ means the number of dichotomies for N points on $\mathbb{R}^d$.

By the Cover's Theorem in the reference, 

$C(N+1,d) = C(N,d) + C(N,d-1)$, which means

if we anchor one points

$$m_H(N) = 2\sum_{i=0}^{d-1} C^{N-1}_{i} \quad \text{for } k=1,\ \hat{a}_1 = \hat{0}$$

$$= 2\sum_{i=0}^{2-1} C^{N-1}_{i} = 2 \times (C^{N-1}_{0} + C^{N-1}_{1})$$

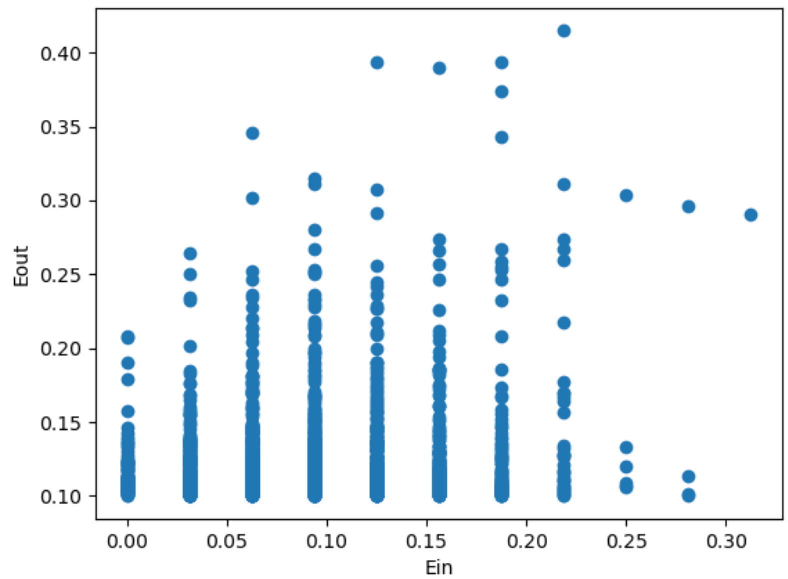
$$= 2 \times (1 + N - 1) = 2N$$

$\Rightarrow$ if we anchor k points $(0 \leq k < d)$

$$\boxed{\tilde{m_H}(N) = 2\sum_{i=0}^{d-k} C^{N-1}_{i}}$$

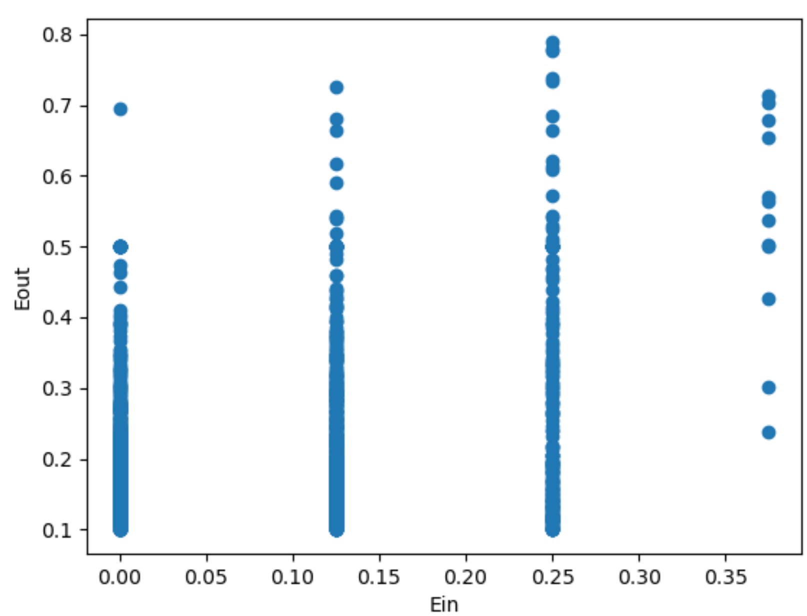
#

10. Median of Eout(g) - Ein(g): 0.03747820081411739



11. Median of Eout(g) - Ein(g): 0.12087622283412079

We saw that by reducing the size of x from 32 to 8, the number of different Ein become lower, and the range of Eout becomes larger from approximately (0, 0.45) to (0, 0.8), and also the median of Eout - Ein becomes 4 times larger than problem 10, which means Eout is more different from Ein in smaller dataset in decision stump.



12. Median of Eout(g) - Ein(g): 0.004120047286207934

By randomly chosen hs,θ as g, with s uniformly sampled from {−1,+1} and θ uniformly sampled from [−1,1], we saw that Ein has more possibilities, which range from (0, 1.0), and so does Eout, and the median of Eout - Ein becomes very small compared to problem 10, 11, which means Eout is very closed to Ein in each case. We can also see that Eout is higher as Ein goes higher, this means Eout and Ein share the same growing tendency.