Mechine Learning HW2

1.

$\underset{\text{coin}}{\mu_m}, \underset{\text{no coin}}{1-\mu_m}, \quad m = ((t-1) \bmod M) + 1, \quad N_m = m - 1 + (t-1)$

We define μ_m is $E_{out}(h)$, $\frac{C_m}{N_m}$ is $E_{in}(h)$, $\varepsilon = \sqrt{\frac{\ln t - \frac{1}{2}\ln\delta}{N_m}}$

$\therefore$ By Hoeffding's inequality,

$$P(\mu - \nu > \varepsilon) \le e^{-2\varepsilon^2 N}$$

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

$$= e^{\ln(\delta t^{-2})}$$

$$= \delta t^{-2}$$

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

#

2.

For all slot machines, By Hoeffdings inequality,

$$P\left(\mu_m > \frac{C_m}{N_m} + \sqrt{\frac{\ln t + \ln M - \frac{1}{2}\ln\delta}{N_m}}\right) \le \sum_{t=1}^{\infty}\sum_{m=1}^{M} e^{-2\left(\frac{\ln\frac{tM}{\delta^{\frac{1}{2}}}}{N_m}\right)N_m}$$

$$= \left(\sum_{t=1}^{\infty} \frac{\delta t^{-2}}{M^2}\right) \cdot M$$

$$= \delta \cdot \underbrace{\frac{\pi^2}{6M}}_{} \le \delta \quad \left(\frac{\pi^2}{6M} < 1 \because M \text{夠大}\right)$$

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

Hence, we know the opposite statement

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

#

3.

special lottery game

- A: all even orange, all odd green
- B: all even green, all odd orange
- C: 1~4 orange, 5~8 green
- D: 1~4 green, 5~8 orange

5 tickets

→ all green

Divided into 6 cases, suppose each kind of ticket has same quantity

① 5同

② 4同1異

③ 3同2同

④ 3同2異

⑤ 2同2同1異

⑥ 2同3異

total : $4^5 = 1024$

For ① 5同 (ex : A,A,A,A,A)

$\therefore n(①) = 4$

For ② 4同1異

ex: AAAAB → X

C → ✓ $\frac{5!}{4!} = 5$

D → ✓ $\frac{5!}{4!} = 5$

ex: BBBBA → X

C → ✓ $\frac{5!}{4!} = 5$

D → ✓ $\frac{5!}{4!} = 5$

$\therefore n(②) = 40$

ex: CCCCA → ✓ $\frac{5!}{4!} = 5$

B → ✓ $\frac{5!}{4!} = 5$

D → X

ex: DDDDA → ✓ $\frac{5!}{4!} = 5$

B → ✓ $\frac{5!}{4!} = 5$

C → X

For ③ 3同2同

ex: AAABB → X

CC → ✓ $\frac{5!}{2!3!}=10$

DD → ✓ $\frac{5!}{2!3!}=10$

ex: BBBAA → X

CC → ✓ $\frac{5!}{2!3!}=10$

DD → ✓

∴ n(③) = 80

ex: CCCAA → ✓ $\frac{5!}{2!3!}=10$

BB → ✓ $\frac{5!}{2!3!}=10$

DD → X

ex: DDDAA → ✓ $\frac{5!}{2!3!}=10$

BB → ✓

CC → X $\frac{5!}{2!3!}=10$

For ④, 3同2異

ex: AAABC → X

CD → X

BD → X

ex: BBBAC → X

AD → X

CD → X

ex: CCCAB → X

BD → X

AD → X

ex: DDDAB → X

BC → X

AC → X

$\boxed{\text{Small conclusion:}}$

If we pick three different kind of letters in 'A', 'B', 'C', 'D' ⇒ we can't form a pure green solution.

∴ n(④) = 0

For ⑤, 2同2同1異, by small conclusion in ④,

we know there are 3 different kind of letters in ⑤

$\therefore n(⑤) = 0$

For ⑥, 2同3異, again, by small conclusion in ④

We know there are 4 different kind of letters in ⑥

$\therefore n⑥ = 0$

$\therefore$ In summary, by case ①②③④⑤⑥,

$$P(\text{pure green}) = \frac{n(①)+n②+n(③)+n(④)+n(⑤)+n(⑥)}{4^5}$$

$$= \frac{4+40+80}{1024} = \frac{124}{1024} = \frac{31}{256}$$

#

4. Similarly, from same method in problem 3,

We have pure green 2's with following combinations

DDDDD, 1

BBBBB, 1

BBBBD, $\frac{5!}{4!} = 5$

DDDDB, $\frac{5!}{4!} = 5$

BBBDD, $\frac{5!}{2!3!} = 10$

DDDBB $\frac{5!}{2!3!} = 10$

$$P(\text{pure green 2's}) = \frac{1+1+5+5+10+10}{1024}$$

$$= \frac{32}{1024} = \frac{1}{32}$$ #

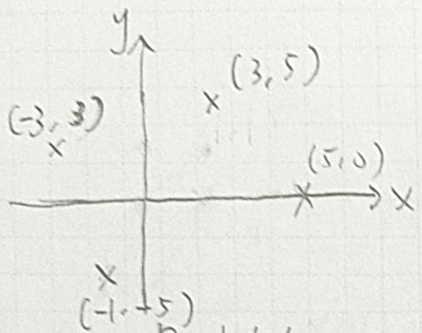
5\.

Let 4 input vectors are $\vec{x_1} = (5, 0)$

$\vec{x_2} = (3, 5)$

$\vec{x_3} = (-3, 3)$

$\vec{x_4} = (-1, -5)$



By labeling 4 vectors with 'negative rectangle' hypothesis,

labels for (x_1, x_2, x_3, x_4)	rectangle contains point (vectors)
$(-1, -1, -1, -1)$	x_1, x_2, x_3, x_4
$(-1, -1, -1, +1)$	x_1, x_2, x_3
$(-1, -1, +1, -1)$	x_1, x_2, x_4
$(-1, +1, -1, -1)$	x_1, x_3, x_4
$(+1, -1, -1, -1)$	x_2, x_3, x_4
$(-1, -1, +1, +1)$	x_1, x_2
$(+1, -1, -1, +1)$	x_2, x_3
$(+1, +1, -1, -1)$	x_3, x_4
$(-1, +1, -1, +1)$	x_1, x_3
$(+1, -1, +1, -1)$	x_2, x_4
$(-1, +1, +1, -1)$	x_1, x_4
$(-1, +1, +1, +1)$	x_1
$(+1, -1, +1, +1)$	x_2
$(+1, +1, -1, +1)$	x_3
$(+1, +1, +1, -1)$	x_4
$(+1, +1, +1, +1)$	None

$\therefore$ We have $16 = 2^4$ dichotomies, hence

$\Rightarrow$ The VC dimension of the hypothesis set is no less than 4 #

6.

$s \in \{-1, 1\}$

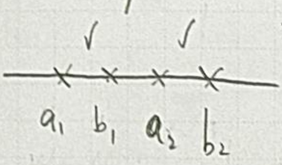
$a_m < b_m$, for $1 \leq m \leq M$

$b_m < a_{m+1}$, for $1 \leq m \leq M-1$

$s, a_1, b_1, a_2, b_2, a_3, b_3, \ldots, a_M, b_M$

$$h_{(s,a,b)}(x) = \begin{cases} s & , \text{if } a_m \leq x \leq b_m \text{ for some } 1 \leq m \leq M \\ -s & , \text{otherwise} \end{cases}$$

Consider simple case $M=2$ first,



① Let 4 points x_1, x_2, x_3, x_4, if we want to labels $+1, +1, -1, -1$, we put x_1, x_2 within the intervals $[a_1, b_1]$ or $[a_2, b_2]$, x_3, x_4 outside the intervals.

② Consider alternate labels, $+1, -1, +1, -1$
We can put x_1 to $[a_1, b_1]$, x_2 out of the interval, and the x_3 to $[a_2, b_2]$, x_4 out of interval.

Similarly, we know other permutations of $+1, -1$ for 4 points works

$\therefore$ When $M=2$, H can shatter 4 points

However, as we consider 5 points, we still can achieve alternating $+1, -1, +1, -1, +1$, when we take $s=-1$, $\therefore$ we can shatter 5 points

However, when we have 6 points, we can't put each points in different intervals ∴ Can't be shattered ∴ for $M=2$, $d_{vc}=5$

Hence, We know the 'alternating label' will cause the points 'can't' be shattered.

⇒ In general case 'M', we can shatter $2M+1$ points

- For any arrangement of 1s, −1s on the $2M+1$ points we can always position our intervals such that 'alternating labels' works. That is, we can place our points in consecutive interval (such as $[a_m, b_m]$, $[b_m, a_{m+1}]$) with label +1 and label −1 alternatively. and this can be done M times, which means we can put $2M+1$ points into different intervals with proper $S=+1 \vee S=-1$ ∴ $2M+1$ points can be shattered.

⇒ Consider $2M+2$ point,

Similarly, we place $x_1 \sim x_{2M+1}$ points in $2M+1$ intervals which cause alternating +1 and −1 with proper $S=1 \vee S=-1$. However, we still have one point remaining, which can't be inserted into new interval.

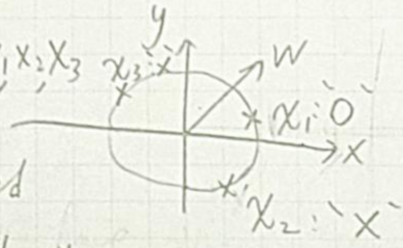
That is the dichotomies may end with $[+1,-1,+1,\dots,\underbrace{+1,+1}]$ $\vee [+1,-1,+1,\dots,\underbrace{-1,-1}]$, $[-1,+1,-1,\dots,\underbrace{+1,+1}]$, $[-1,+1,-1,\dots,\underbrace{-1,-1}]$ which means that $2M+2$ points can't be shattered.

Therefore, the VC dimension of H is '2M+1' #

7.

2D perceptron: $H_0 = \{h : h(x) = sign(w_1x_1, w_2x_2)\}$

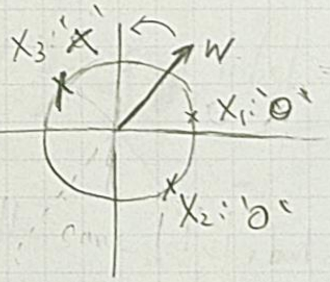
Consider 3 points x_1, x_2, x_3



the problem can be viewed as a hypothesis W which will rotate around '3' points, as the W reached each of x_1, x_2, x_3, it will change the value from +1 to -1, it means it will generate a new dichotomy.

initially, we choose x_1: 'O', x_2: 'O', x_3: 'X', rotate W 'counter-clockwisely.'

	x_1	x_2	x_3
initialization	O	O	X
reach x_3,	O	O	O
reach x_2,	O	X	O
reach x_1	X	X	O
reach x_3	X	X	X
reach x_2	X	O	X
~~reach x_1~~	~~O~~	~~O~~	~~X~~ (∵ repeated)



Hence we know if we have 3 points with W, each of them will be reach 2 times, it will generate 6 dichotomies at most ∴ $m_H(3) = 6$

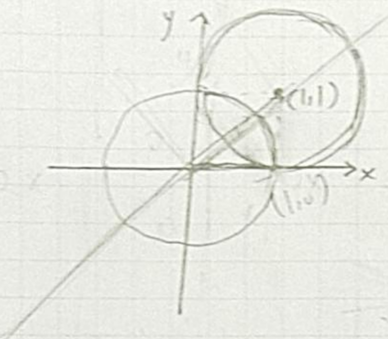
⇒ For N points, as each of $x_1, x_2, \cdots x_N$ will be reached 2 times, which will generate '2N' dichotomies at most ⇒ $m_H(N) = 2N$ #

8.

$H = H_0 \cup H_1$

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

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



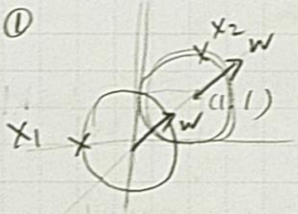
$$m_H(N) = m_{H_0}(N) + m_{H_1}(N) - m_{H_0 \cap H_1}(N)$$

For $m_{H_0}(N)$ and $m_{H_1}(N)$, by problem 7, we know

$$m_{H_0}(N) = m_{H_1}(N) = 2N$$

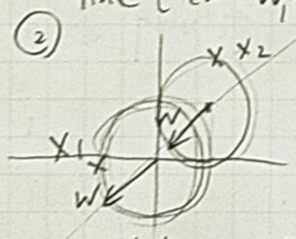
For $m_{H_0 \cap H_1}(N)$, suppose we have more than 2 inputs $x_1, x_2, \dots$ that is $N \geq 2$, we know only 4 directions of w are possible for generating hypothesis both satisfied H_0 and H_1.

①



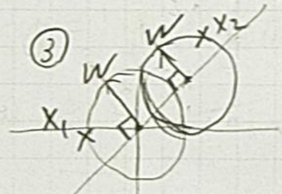
W is parallel to center-to-center line (ex : $w_1 // (1,1)$)

②

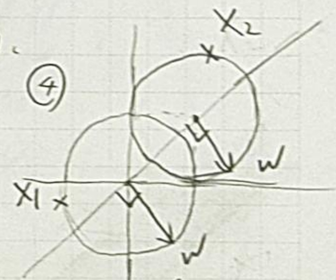


W is pallelel to center-to-center line in opposite direction (ex : $w_2 // (-1,-1)$)

③



④



W is perpendicular to center-to-center line with (ex : $w_3 // (-1,1)$, $w_4 // (1,-1)$)

∴ By ①②③④, we know

$$m_{H_0 \cap H_1}(N) = 4$$

$\therefore m_{H_0 \cup H_1}(N) = m_{H_0}(N) + m_{H_1}(N) - m_{H_0 \cap H_1}(N)$

$= 2N + 2N - 4$

$= 4N - 4 \quad (\text{for } N \geq 2)$

For $N=1$, $m_{H_0 \cap H_1}(N) = 2$ (just 1 input $\Rightarrow$ 2 dichotomy at most.)

$\therefore m_{H_0 \cup H_1}(N) = m_{H_0}(N) + m_{H_1}(N) - m_{H_0 \cap H_1}(N)$

$= 2N + 2N - 2$

$= 4N - 2 \quad (\text{for } N = 1)$

$\therefore$ for $N=1 \Rightarrow m_{H_0 \cup H_1}(N) = 4 \cdot 1 - 2 = 2 = 2^1$

$\therefore$ shattered

for $N=2 \Rightarrow m_{H_0 \cup H_1}(N) = 4 \cdot 2 - 4 = 4 = 2^2$

$\therefore$ shattered.

for $N=3 \Rightarrow m_{H_0 \cup H_1}(N) = 4 \cdot 3 - 4 = 8 = 2^3$

$\therefore$ shattered.

for $N=4 \Rightarrow m_{H_0 \cup H_1}(N) = 4 \cdot 4 - 4 = 12 < 2^4 = 16$

$\therefore$ NOT shattered

$\therefore$ VC dimension of "H" $d_{vc}(H) = 3$ #

9. $h_{(s,\theta)}(x) = s \cdot \text{sign}(x-\theta)$

x is generated uniformly $[-1,1]$

$y = \text{sign}(x)$ with noise flipping the sign 10% of the time

Consider 4 cases,

case 1 : $x > \theta$, $s=1$

$\because h_{(s,\theta)} = 1$, when $x > \theta$, it will satisfies $h_{(s,\theta)} = 1$

if $\theta > 0 \Rightarrow P_{\text{picking } x} = \dfrac{\text{len}([\theta,1])}{\text{len}([-1,1])} = \dfrac{1-|\theta|}{1-(-1)} = \dfrac{1-|\theta|}{2}$

$\because \theta > 0 \therefore x > 0$, when the misclassification occurs

sign(x) flips to -1 with probabity 0.1

$$\therefore P = 0.1\left(\frac{1-|\theta|}{2}\right) \quad —— ①$$

if $\theta < 0 \Rightarrow$ $P_{\text{picking } x} = \dfrac{\text{len}([\theta,0]+[0,1])}{\text{len}[-1,1]}$

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

if x is picking from $[\theta,0] \Rightarrow \text{sign}(x) < 0$

$$\therefore P = 0.9\frac{|\theta|}{2} \quad —— ②$$ ($\because$ misclassified)

if x is picking from $[0,1] \Rightarrow \text{sign}(x) > 0$

$$\therefore P = 0.1 \cdot \frac{1}{2} \quad —— ③$$

Case 2. $x > \theta$, $s = -1$

$\therefore h_{(s,\theta)} = -1$, when $x < 0 \Rightarrow$ satisfied $\mathrm{sign}(x) = -1$
$= h_{(s,\theta)}$

if $\theta < 0$, $P_{\text{picking } x} = \dfrac{\mathrm{len}([\theta, 0] + [0, 1])}{\mathrm{len}[-1, 1]}$

$= \dfrac{|\theta|}{2} + \dfrac{|1|}{2}$

if x is picking from $[\theta, 0]$ $\therefore$ $\mathrm{sign}(x) = -1$ (satisfied)

$\therefore$ We need misclassified probability 0.1 flipping to $+1$ $\therefore$ $P = 0.1 \cdot \dfrac{|\theta|}{2}$ —— ④

x is picking from $[0, 1] \Rightarrow \mathrm{sign}(x) = 1$ (misclassified)

$\therefore P = 0.9 \cdot \dfrac{1}{2}$ —— ⑤

if $\theta > 0$

$P_{\text{picking } x} = \dfrac{\mathrm{len}[\theta, 1]}{\mathrm{len}[-1, 1]} = \dfrac{1 - |\theta|}{2} = \dfrac{1 - |\theta|}{2}$

we know $\mathrm{sign}(x) = +1$, misclassified

$\therefore P = 0.9 \dfrac{1 - |\theta|}{2}$ —— ⑥

Case 3 $x<\theta$, $s=1$

$\because h_{(s,\theta)} = -1$, when $x<0 \Rightarrow \text{sign}(x) = -1$ (satisfied) $= h_{(s,\theta)}$

if $\theta > 0$, $P_{\text{picking } x} = \dfrac{\text{len}([-1,0]+[0,\theta])}{\text{len}[-1,1]}$

$= \dfrac{1}{2} + \dfrac{|\theta|}{2}$

if x is picking from $[-1,0]$ $\because \text{sign}(x) = -1$ (satisfied)

$\because$ We need misclassified probability 0.1 flipping to `+1` $\therefore P = 0.1 \cdot \frac{1}{2}$ —— ⑦

if x is picking from $[0,\theta]$ $\because \text{sign}(x)=1$ (misclassified)

$P = 0.9 \dfrac{|\theta|}{2}$ —— ⑧

if $\theta < 0$, $P_{\text{picking } x} = \dfrac{\text{len}[-1,\theta]}{\text{len}[-1,1]} = \dfrac{-|\theta|-(-1)}{2} = \dfrac{1-|\theta|}{2}$

with $x<0 \Rightarrow \text{sign}(x) = -1$ (classified)

We need misclassified probability

$\therefore P = 0.1 \dfrac{1-|\theta|}{2}$ —— ⑨

case 4. $x < \theta$, $s = -1$

$\because h_{(s,\theta)}(x) = 1$ $\therefore$ when $x > 0 \Rightarrow sign(x) = 1$ (classified) $= h_{(s,\theta)}$

if $\theta > 0$, $P_{picking\ x} = \dfrac{len([-1,0],[0,\theta])}{len[-1,1]}$

$= \dfrac{1}{2} + \dfrac{|\theta|}{2}$

if x is picking from $[-1,0]$ $\therefore sign(x) = -1$ (misclassfied)

$\therefore P = 0.9 \cdot \dfrac{1}{2}$ —— 10

if x is pick from $[0,\theta]$ $\therefore sign(x) = 1$ (classified)

$\therefore P = 0.1 \dfrac{|\theta|}{2}$ —— 11

if $\theta < 0$, $P_{picking\ x} = \dfrac{len[-1,\theta]}{len[-1,1]} = \dfrac{1-|\theta|}{2}$

we know $x < 0$, $sign(x) = -1$ ($\therefore$ misclassified)

$\therefore P = 0.9 \dfrac{1-|\theta|}{2}$ —— 12

In summary,

We see $s=1$, by 1 2 3 7 8 9

$P = 0.1\left(\dfrac{1-|\theta|}{2}\right) + 0.9\dfrac{|\theta|}{2} + 0.1\dfrac{1}{2} + 0.1\dfrac{1}{2} + 0.9\dfrac{|\theta|}{2} + 0.1\dfrac{1-|\theta|}{2}$

In summary, for $s \in \{1, -1\}$

When we pick $s=1$ (with probability $\frac{1}{2}$), by ①②③⑦⑧⑨

$$\therefore P_{s=1} = \frac{1}{2}(① + ② + ③ + ⑦ + ⑧ + ⑨)$$

$$= \frac{1}{2}\left(0.1\left(\frac{1-|\theta|}{2}\right) + 0.9\frac{|\theta|}{2} + 0.1\frac{1}{2} + 0.1\frac{1}{2} + 0.9\frac{|\theta|}{2} + 0.1\left(\frac{1-|\theta|}{2}\right)\right)$$

$$= \frac{1}{2}(0.2 + 0.8|\theta|) = 0.1 + 0.4|\theta|$$

$$= 0.5 - 0.4s + 0.4 \cdot s \cdot |\theta| \Big|_{s=1}$$

When we pick $s=-1$ (with probability $\frac{1}{2}$), by ④⑤⑥⑩⑪⑫

$$\therefore P_{s=-1} = \frac{1}{2}(④ + ⑤ + ⑥ + ⑩ + ⑪ + ⑫)$$

$$= \frac{1}{2}\left(0.1\frac{|\theta|}{2} + 0.9\frac{1}{2} + 0.9\left(\frac{1-|\theta|}{2}\right) + 0.9 \cdot \frac{1}{2} + 0.1\frac{|\theta|}{2} + 0.9\left(\frac{1-|\theta|}{2}\right)\right)$$

$$= \frac{1}{2}(1.8 - 0.8|\theta|) = 0.9 - 0.4|\theta|$$

$$= 0.5 - 0.4s + 0.4 \cdot s \cdot |\theta| \Big|_{s=-1}$$

By generalizing the solution,

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

#

13. Without any anchor points, As given by Cover's theorem, for perceptrons in R^d is $m_H(N) = 2\sum_{i=0}^{d}\binom{N-1}{i}$

With K anchor points,

If we fix a perceptron to pass through a single anchor point, this can be though of as reducing the dimension of our perceptron by 1 ∴ if we have k anchor points, it reduce our dimension by 'k'

∴ It will behave as 'd-k' dimension in space.

∴ the growth function for perception in $\tilde{H}$ should be equivalent to $\in R^{d-k}$

By Cover's Theorem,

$$m_{\tilde{H}}(N) = 2\sum_{i=1}^{d-k}\binom{N-1}{i}$$

#

<Proof> When we fix a perceptron to pass through an anchor point a in R^d, which means $W^T a = 0$.

Hence, this effectively reduce the effective dimension of weight space 'w' by 1.

Furthermore, since perceptron's behaviors are determined by their weights, we can conclude that perceptron's effective dimension also reduced by 1.

By Repeating this statement for each 'k' anchor points, the perceptron's effective dimension is reduced by 'k'

$\therefore$ Perceptron $\tilde{H}$ have growth function which is effectively equivalent to perceptrons in R^{d-k},

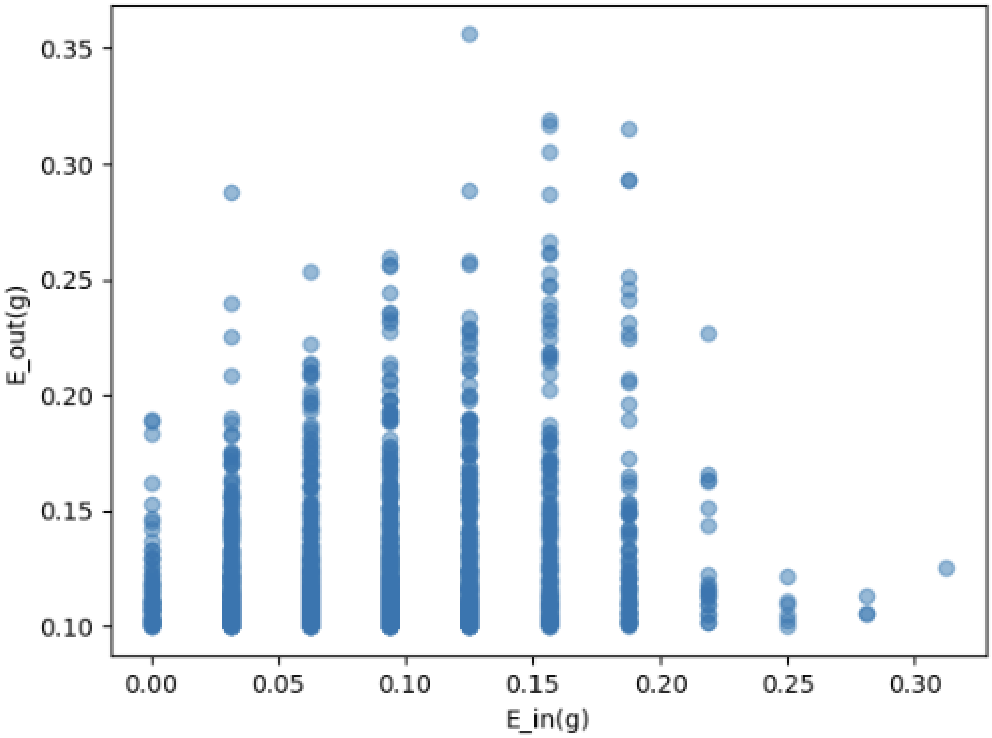
which gives $m_{\tilde{H}}(N) = 2\sum_{i=0}^{d-k}\binom{N-1}{i}$

#

HTML hw2 solution

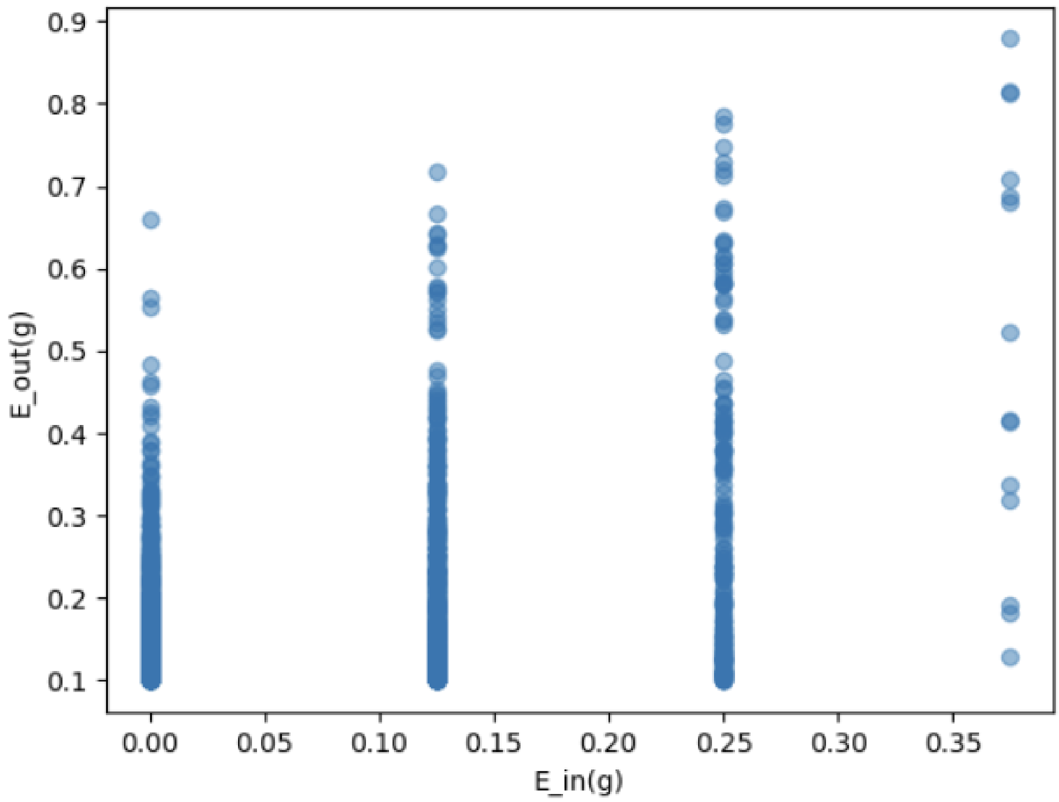
姓名: 謝銘倫, 系級:電機三, 學號:B10502166

10.



Median of E_{in}(g) - E_{out}(g) = -0.0389

11.

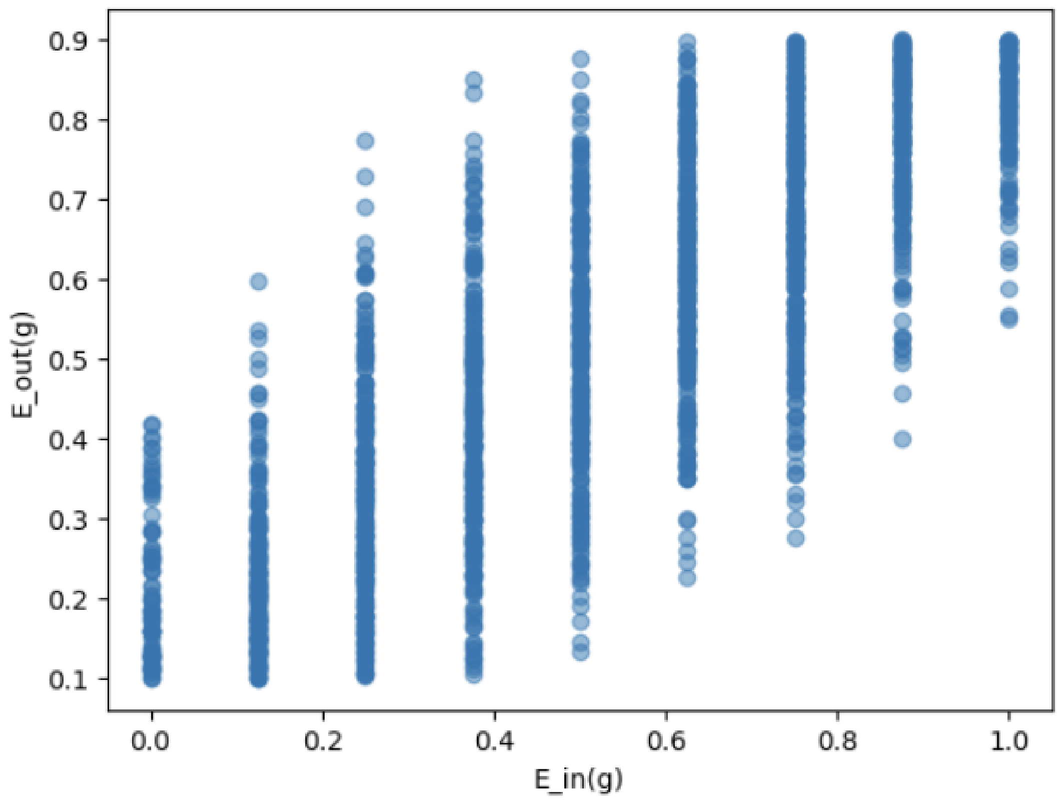


Median of E_{in}(g) - E_{out}(g) = -0.118

Findings:

1. As the data size changes from 32 to 8, the possible E_{in}(g) values are fewer than the one in problem10. The possible reason is that **greater variability with more data generates a wider range of E_{in} and E_{out} values.**
   Hence, The plot scatters more discretely than the plot in problem10.
2. The median of error E_{in} - E{out} is much larger than one in problem 10. The possible reason is that **a smaller sample size is more susceptible than noise and variability of data**

12.



Median of E_{in}(g) - E_{out}(g) = -0.0062

Findings:

1. The code in problem 12 runs faster than problem 11 because we just randomly choose s and θ without running the Decision Stump Algorithm which contains a single loop.
2. the distribution of the E_{in}(g) value in problem 12 varies from 0 to 1, but the distribution of the E_{in}(g) only varies in 0 to 0.4. the

possible reason is that **we choose the h_(s,θ) randomly without finding the minimum E_{in} like in problem 11, so the value of E_{in} scatters from 0 to 1 in problem 12.**

3. The scatter points in the plot of problem 12 distribute along the line y=x, which is equivalent to the line E_{in}(g)=E_{out}(g). Compared to the scatter plot in problem 11**,** the scatter points are distributed from the bottom to the top with high density with lower E_{out} value.
4. The median of error is much smaller in problem 12 than in problem 11. The possible reason is that **we choose the h_(s,θ) as g randomly which causes the scattering situation that is distributed along the line y=x (E_{in} = E_{out}).** Hence, after calculating the median of E_{in}- E_{out}, we get much smaller median error in problem 12.