● Hoeffding Inequalities

1. Single Slot Machine

不難從形式中找到對應意義的符號：

$$N = N_m \,, \mu = \mu_m \,, \nu = \frac{c_m}{N_m}$$

$$\epsilon = \sqrt{\frac{logt - \frac{1}{2}log\delta}{N_m}}, \delta = t^2 exp(-2\epsilon^2 N_m)$$

只要從 $\delta = t^2 exp(-2\epsilon^2 N_m)$ 開始推導：

$$log(\delta t^{-2}) = -2\epsilon^2 N_m$$

$$\Rightarrow \epsilon = \sqrt{\frac{-\frac{1}{2}log(\delta t^{-2})}{N_m}} = \sqrt{\frac{logt - \frac{1}{2}log\delta}{N_m}}$$

代入由 δ 所推得的 ϵ 以及其他參數可以得到：

$$P(\mu > \nu + \epsilon) \leq exp(-2\epsilon^2 N)$$

$$\Rightarrow P\left(\mu_m > \frac{c_m}{N_m} + \sqrt{\frac{logt - \frac{1}{2}log\delta}{N_m}}\right) \leq \frac{\delta}{t^2}$$

2. All Slot Machines

Hoeffding Inequalities 告訴我們，對所有的 m：

$$P\left(\mu_m > \frac{c_m}{N_m} + \epsilon\right) \leq exp(-2\epsilon^2 N_m)$$

由於 M 跟 t 均大於 1，因此：

$$P\left(\mu_m > \frac{c_m}{N_m} + \epsilon\right) \leq M^2 t^2 exp(-2\epsilon^2 N_m)$$

令$\delta = M^2 t^2 exp(-2\epsilon^2 N_m)$：

$$log\frac{\delta}{M^2 t^2} = -2\epsilon^2 N_m \Rightarrow \epsilon = \sqrt{\frac{-\frac{1}{2}log\frac{\delta}{M^2 t^2}}{N_m}}$$

$$\Rightarrow \epsilon = \sqrt{\frac{logt + logM - \frac{1}{2}log\delta}{N_m}}$$

所以代回式子可以得到：

$$P\left(\mu_m > \frac{c_m}{N_m} + \sqrt{\frac{logt + logM - \frac{1}{2}log\delta}{N_m}}\right) \le \delta$$

所以反過來就可以推得：

$$P\left(\mu_m \le \frac{c_m}{N_m} + \sqrt{\frac{logt + logM - \frac{1}{2}log\delta}{N_m}}\right) \ge 1 - \delta$$

3. 抽抽樂

要有某些數字全都是綠色，可以知道 AB 這兩種獎券不可以同時被抽到，他們顏色的情形是互補的；CD 也是同樣道理。因此可以知道，這五抽裡面的獎券種類最多只可以包含兩種，並且只可以是，AC、AD、BC、BD 這四種組合，所以獎券總共有以下的取法數量：

$$4 \times 2^5$$

但是要注意，AC 跟 AD 的取法中，他們都包含「全都是 A」的取法，所以會多算一次，要扣除；BC 跟 BD、AC 跟 BC 還有 AD 跟 BD 也是同理：

$$4 \times 2^5 - 4$$

最後再除以全部的取法就可以得到機率：

$$\frac{4 \times 2^5 - 4}{4^5} = \frac{2^5 - 1}{4^4} = \frac{31}{256}$$

4. 抽抽樂-續

如果要五張券裡面的數字 2 都是綠色的，那這五張券只能是 B 或 D 這兩種獎券：

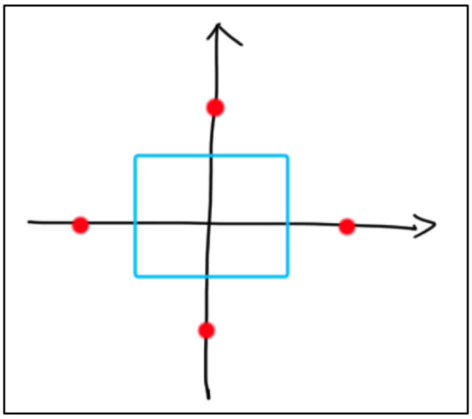
$$\frac{2^5}{4^5} = \frac{1}{2^5} = \frac{1}{32}$$

可以發現，$\frac{1}{32} \approx 0.03$,$\frac{31}{256} \approx 0.12$，BAD Data 發生的機率有所不同。

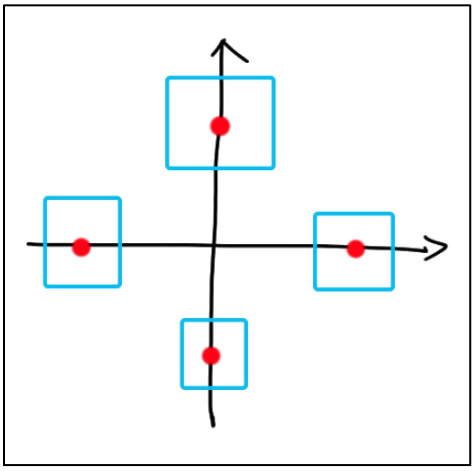
● VC Dimension

5. negative rectangle

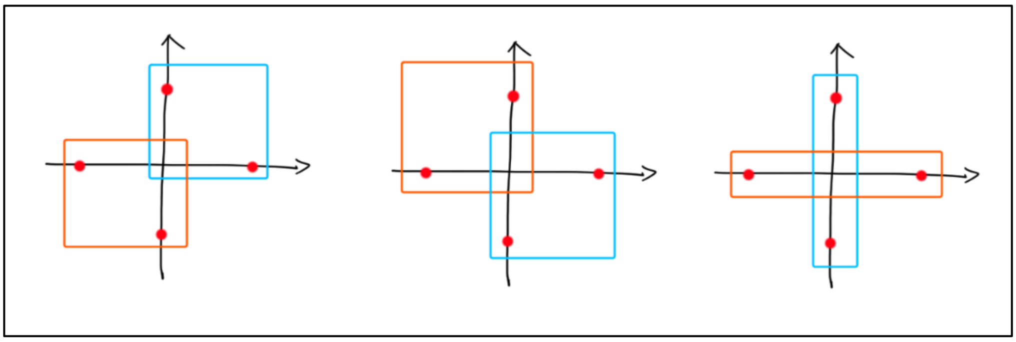
下面的圖中，紅色點是 4 個 input vectors，而每一個藍色框框或橘色框框都是一個 hypothesis，他們都屬於 negative rectangle 這個 hypothesis set。



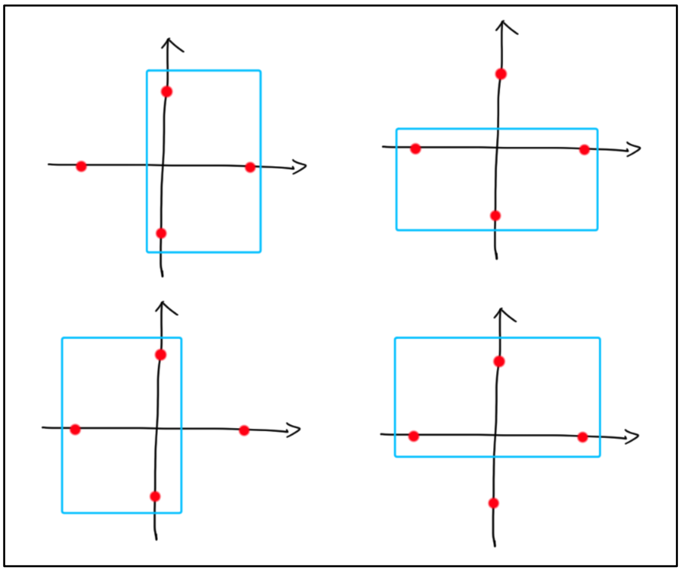
上圖是 0 個點回傳−1，框框沒有包含任何點，全部點回傳+1。



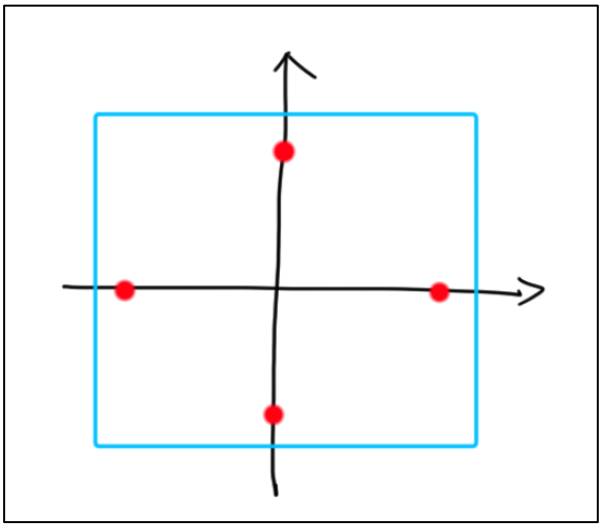
上圖是 1 個點回傳−1，圖中的四個藍色框框代表四個 hypothesis。對於一個藍色框框來說，框框內的點回傳−1，另外三個點回傳+1。



上圖是 2 個點回傳−1，圖中的每個藍色跟橘色框框各自代表 1 個 hypothesis。對於一個框框來說，框框內的兩個點回傳−1，另外兩個點回傳+1。



上圖是 3 個點回傳−1，圖中的每個藍色框框代表 1 個 hypothesis。框框內的三個點回傳−1，另外一個點回傳+1。



最後一張圖是 4 個點回傳−1，也就是每個點都回傳−1。

6. Multiple intervals

$2M+1$ 個參數，就是有 M 個 intervals 可以用，我們先討論 positive interval。首先，可以從課堂的 one positive interval 對 N 個點的公式：

$$C\binom{N+1}{0}+C\binom{N+1}{2}=1+\frac{N(N+1)}{2}$$

推廣到如果是 M 個 positive intervals 對 N 個點的公式：

$$C\binom{N+1}{0}+C\binom{N+1}{2}+C\binom{N+1}{4}+\cdots+C\binom{N+1}{2M}$$

這時候如果回憶高中所學：

$$(1+x)^n=C\binom{n}{0}\cdot 1^n\cdot x^0+C\binom{n}{1}\cdot 1^{n-1}\cdot x^1+\cdots+C\binom{n}{n}\cdot 1^0\cdot x^n$$

$$\Rightarrow (1+1)^n=2^n=C\binom{n}{0}+C\binom{n}{1}+\cdots+C\binom{n}{n}$$

$$(1+-1)^n = C\binom{n}{0}\cdot 1^n \cdot -1^0 + C\binom{n}{1}\cdot 1^{n-1}\cdot -1^1 + \cdots + C\binom{n}{n}\cdot 1^0 \cdot -1^n$$

$$\Rightarrow 0 = C\binom{n}{0} - C\binom{n}{1} + C\binom{n}{2} - C\binom{n}{3} \ldots + C\binom{n}{n}$$

$$\Rightarrow 2^n + 0 = 2\mathrm{C}\binom{n}{0} + 2\mathrm{C}\binom{n}{2} + \cdots + \begin{cases} 2\mathrm{C}\binom{n}{\mathrm{n}-1}, \text{n is odd} \\ 2\mathrm{C}\binom{n}{n}, \text{n is even} \end{cases}$$

$$\Rightarrow 2^{n-1} = C\binom{n}{0} + C\binom{n}{2} + \cdots + \begin{cases} \mathrm{C}\binom{n}{\mathrm{n}-1}, \text{n is odd} \\ \mathrm{C}\binom{n}{n}, \text{n is even} \end{cases}$$

白話來說就是偶數項加起來等於2^{n-1}。
所以我們可以知道，根據 M 個 positive intervals 的公式，如果我們令 $N = 2M$：

$$C\binom{2M+1}{0} + C\binom{2M+1}{2} + \cdots + C\binom{2M+1}{2M} = 2^{2M}$$

也就是說，M 個 positive intervals 可以 shatter 2M 個點。
但如果是$2M+1$個點：

$$C\binom{2M+2}{0} + C\binom{2M+2}{2} + \cdots + C\binom{2M+2}{2M}$$

會發現少了最後一項$C\binom{2M+2}{2M+2}$，如果我們幫他補上去：

$$C\binom{2M+2}{0} + \cdots + C\binom{2M+2}{2M} + C\binom{2M+2}{2M+2} - C\binom{2M+2}{2M+2}$$

$$= 2^{2M+1} - C\binom{2M+2}{2M+2} = 2^{2M+1} - 1$$

也就代表 M 個 positive intervals 無法 shatter 任何$2M+1$個點。

不過當我們舉一個實際的例子來看，例如$M = 2$的時候，上面的推論可以知道我們可以 shatter $N = 4$ 個點，無法 shatter 任何 $N = 5$ 個點，那為甚麼 positive intervals 無法 shatter $N = 5$ 個點，或者說剛好只差 1 個情形做不到？答案就是$+1, -1, +1, -1, +1$這樣的分布情形，因為我們只有 2 個 positive intervals，但是想要弄出$+1, -1, +1, -1, +1$需要 3 個。

這時候就是 negative interval 上場的時候了，上面的這種情形其實就是 2 個 negative interval，所以如果我們連同 negative interval 也考慮進來，那麼 M 個 intervals 就可以 shatter $2M+1$個點了。

那如果是$2M+2$個點呢？我們一樣先從 positive interval 開始討論：

$$C\binom{2M+3}{0}+C\binom{2M+3}{2}+\cdots+C\binom{2M+3}{2M}$$

會發現少了最後一項$C\binom{2M+3}{2M+2}$，如果我們幫他補上去：

$$C\binom{2M+3}{0}+\cdots+C\binom{2M+3}{2M}+C\binom{2M+3}{2M+2}-C\binom{2M+3}{2M+2}$$

$$=2^{2M+2}-C\binom{2M+3}{2M+2}=2^{2M+2}-\frac{(2M+3)!}{(2M+2)!}=2^{2M+2}-(2M+3)$$

從上面推導的過程可以知道，positive interval 想要 shatter $2M+2$個點還差$(2M+3)$種情形做不到，我們一樣以上面 2 個 positive interval 來做舉例，對於$N=6$，哪 7 種情形做不到？其實就跟剛剛很類似，需要 3 個 positive interval 的情形我們就無法辦到：

$$+1,-1,+1,-1,+1,-1$$

$$+1,-1,+1,-1,-1,+1$$

$$+1,-1,-1,+1,-1,+1\ \ldots$$

除了上面 3 種外還有另外 4 種。那這 7 種我們有辦法靠 negative interval 做到嗎？可以發現有一些可以，例如：

$$+1,-1,-1,+1,-1,+1\ \ldots$$

但是，有一些不行：

$$+1,-1,+1,-1,+1,-1$$

這種的他要嘛需要 3 個 positive interval，要嘛 3 個 negative interval。

如果推廣到 $2M+2$ 個點，我們無法透過 M 個 interval，去擊敗那種要嘛需要 $M+1$ 個 positive interval，要嘛 $M+1$ 個 negative interval 的情形；因此我們無法 shatter 任何 $2M+2$ 個點。

因此我們最終可以知道，M個 interval 的 Hypothesis Set，我們可以 shatter $2M+1$個點，但無法 shatter 任何 $2M+2$ 個點，也就是說 VC dimension 是$2M+1$。

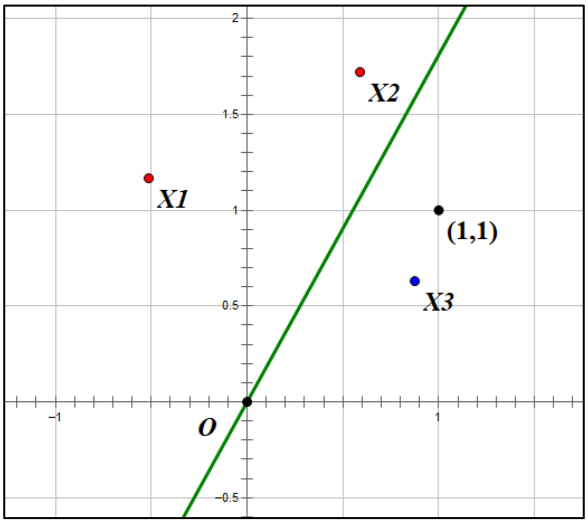
7. origin-passing perceptrons

由於 perceptron 得通過原點，就很像他被釘在了一個地方不可以任意移動只能旋轉。而在二維平面中，一條只能在原點旋轉的線，對於 N 個散佈在平面上的點，就等同把這些點壓到剩下一個維度，並在這僅剩的一個維度做「切一刀」的功能，也就是在線的一邊回傳$+1$，另一邊回傳-1。那麼其效果就跟課堂上提到的 positive and negative ray 效果是一樣的，growth function 是$2N$。

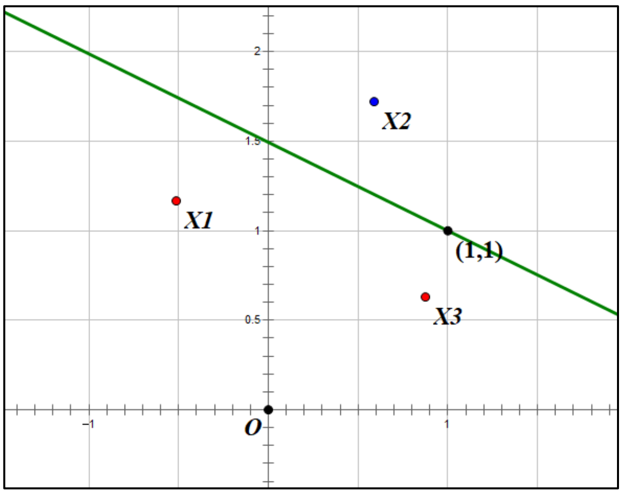
8. Union hypothesis set

上一題的 origin-passing perceptrons 可以知道 VC dimension 是 2，同理我們也可以知道(1,1)-passing perceptrons 的 VC dimension 也是 2，當我們把這兩個 Hypothesis set 聯集起來，可以分別對$N=3$跟$N=4$的資料做討論。

當 $N=3$ 時，令某組資料可以被 origin-passing perceptrons 分出 6 種$+1,-1$的情形，如果這時候有(1,1)-passing perceptrons 成員的幫助，其效果就好像是把原本的 origin-passing perceptrons 平移了一些距離，因此可以「換到另一個位置」去「切一刀」，也就 shatter 了該組資料。



例如在上面的圖中，資料中的三個點為$X1,X2,X3$，綠色的線是 origin-passing perceptrons 的其中一個成員，可以看到他在圖片中將$X1,X2,X3$分別分成了「紅，紅，藍」的情形，在這裡將紅色代表$+1$藍色代表-1；我們可以很明顯的知道，origin-passing perceptrons 沒有一個成員可以把 $X1,X2,X3$ 分成「藍，紅，藍」或「紅，藍，紅」，但是(1,1)-passing perceptrons 的某些成員可以，例如：



因此我們可以知道，兩個 Hypothesis Set 聯集之後，某種角度來說就好

像是讓 origin-passing perceptrons 有了「局部移動」的能力，可以 shatter 一組 $N = 3$ 的資料。

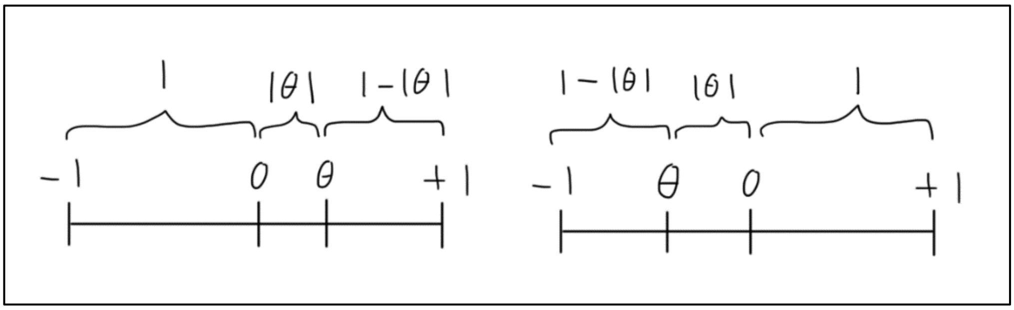
但是當$N = 4$時，就算 origin-passing perceptrons 獲得了一點點平移的能力，依舊不能改變他們「身為一個 2D perceptrons」的極限： 2 維平面的 perceptrons 的 VC dimension 是 3。因此兩個 Hypothesis Set 聯集之後無法 shatter 任何一組 $N = 4$ 的資料。以另一個直覺的角度來說，想像有四個點排成一直線，想要透過 perceptrons 切出「$+1, -1, +1, -1$」是不可能的。

因此我們可以知道聯集之後的 VC dimension 是 3。

● Decision Stumps

9. E_{out}

首先我們圖像化：



上圖畫出了兩種θ的情形。

可以知道不管θ是正的或負地，都可以劃分出三個區域：1、$|\theta|$跟 $1 - |\theta|$。接著分成 $s = 1$ 跟 -1 兩種情況探討，並且p是 noise 的機率，此題中為 10%：

$s = 1$：

可以知道在這三個區間內，「1」跟「$1 - |\theta|$」這兩個區域有p的比例會犯錯，「$|\theta|$」這個區域則是 $1 - p$ 的比例會犯錯，所以可以列出犯錯的式子：

$$1 \times p + (1 - |\theta|) \times p + |\theta|(1 - p) = p + p - |\theta|p + |\theta| - |\theta|p$$
$$= 2\mathrm{p} - 2|\theta|p + |\theta| = 2p + (1 - 2p)|\theta| = (1 - 2p)|\theta| - (1 - 2p) + 1$$
$$= (1 - 2p)(|\theta| - 1) + 1$$

但是不要忘記，我們的區是從-1到$+1$，所以要記得除 2 才會是錯誤率：

$$\Rightarrow (0.5 - p)(|\theta| - 1) + 0.5$$

$s = -1$：

流程跟上面一樣。「1」跟「$1 - |\theta|$」這兩個區域有$1 - p$ 的比例會犯錯，「$|\theta|$」這個區域則是 p 的比例會犯錯，所以可以列出犯錯的式子：

$1 \times (1 - p) + (1 - |\theta|) \times (1 - p) + |\theta|p$

$= 1 - p + 1 - p - |\theta| + |\theta|p + |\theta|p = 2 - 2p - |\theta| + 2|\theta|p$

$= 2 - 2p + (2p - 1)|\theta| = (2p - 1)|\theta| - (2p - 1) + 1$

$= -(1 - 2p)(|\theta| - 1) + 1$

一樣不要忘記除以 2：

$$\Rightarrow -(0.5 - p)(|\theta| - 1) + 0.5$$

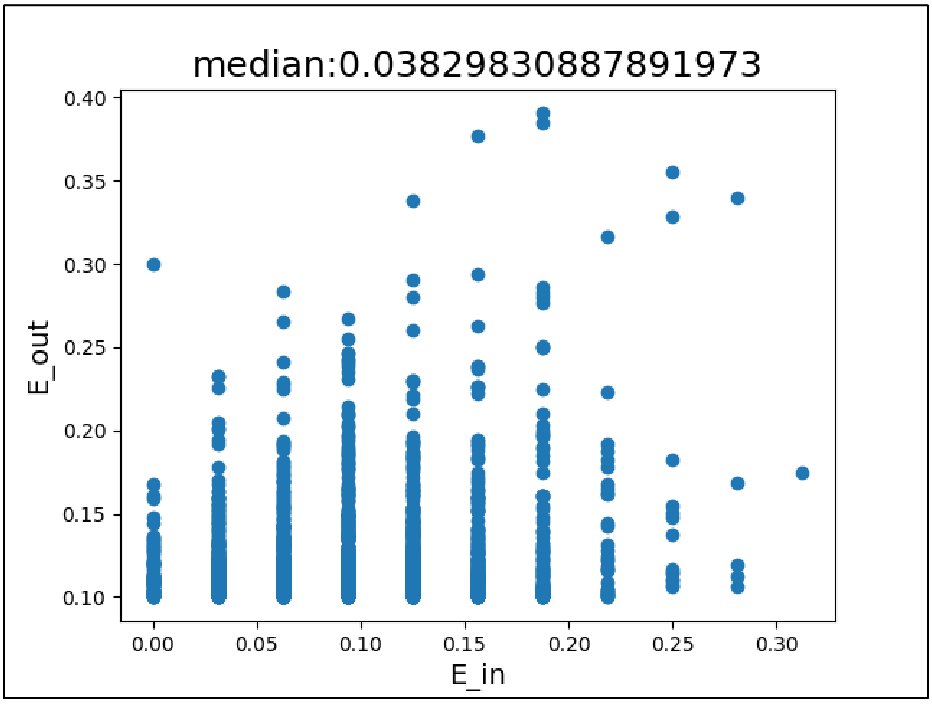
最後觀察$s = 1$ 跟 -1的結論，可以統整成：

$$s(0.5 - p)(|\theta| - 1) + 0.5$$

只要將$p = 0.1$代入就可以得到題目的公式了：

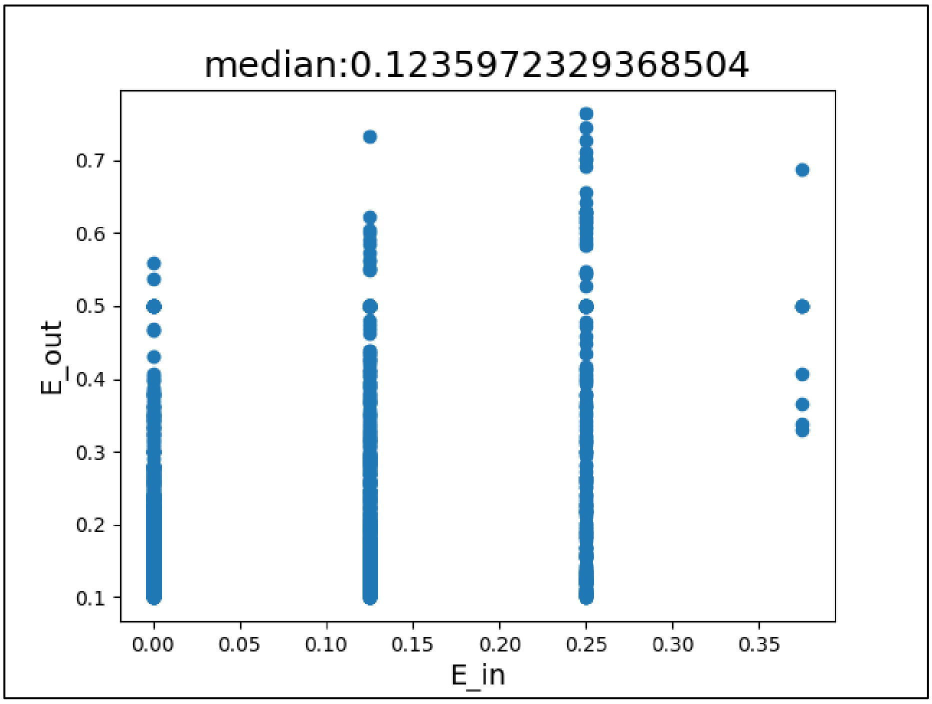
$$s(0.5 - 0.1)(|\theta| - 1) + 0.5 = 0.4s|\theta| - 0.4s + 0.5$$

10. Artificial data with size 32



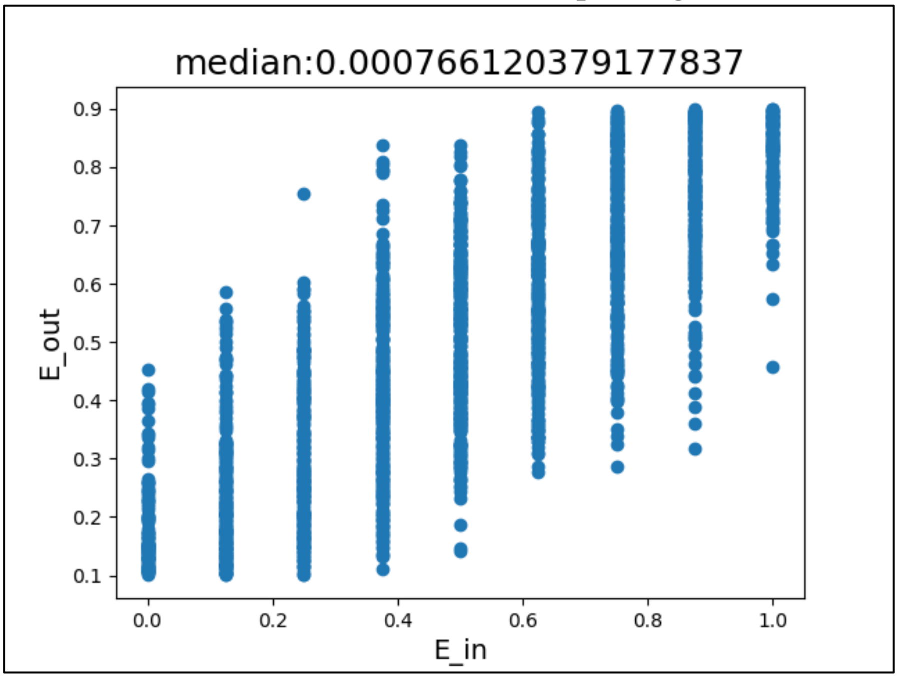
$E_{out} - E_{in}$ 的 median 大約是 0.038。

11. Artificial data with size 8



$E_{out} - E_{in}$ 的 median 大約是 0.12。可以發現 $E_{out} - E_{in}$ 的 median 變高了，並且 E_{out} 大於 0.5 的情形也變多了。回顧上面的公式，可以知道E_{out}要大於 0.5， s 必須是-1，也就是說資料量比較小時，noise 帶來的影響比較大，會讓 Decision Stump 選擇$s\ = -1$來達到低的E_{in}。

12. Artificial data with size 8 and random picking



$E_{out} - E_{in}$ 的 median 大約是 0.0007。可以發現 $E_{out} - E_{in}$ 的 median

變超低。首先因為均勻分布選取 θ 所產生的影響，圖形呈現斜線上升 E_{out} 會隨著 E_{in} 上升，而 E_{in} 的上升代表 s 的取值逐漸以 -1 為主，才可以逐漸拉高 E_{out}；可以發現 E_{in} 在 0 到 1 的範圍，雖然不好確定是否數量差不多，但是可以發現大致都有一定的分布。

● Bonus: Perceptrons that Pass Special Points

13. Cover's Theorem

就如同題目提供的 pdf 檔中，作者所提及的一段話：「Now, by forcing the hyperplane to pass through a certain fixed point, we are in fact moving the problem to one in $N-1$ dimensions, instead of N.」所以如果我們要求 perceptrons 通過 k 個「錨點 anchor points」，其效果就好像維度從 d 變成了 $d-k$，因此公式就會被改寫為：

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