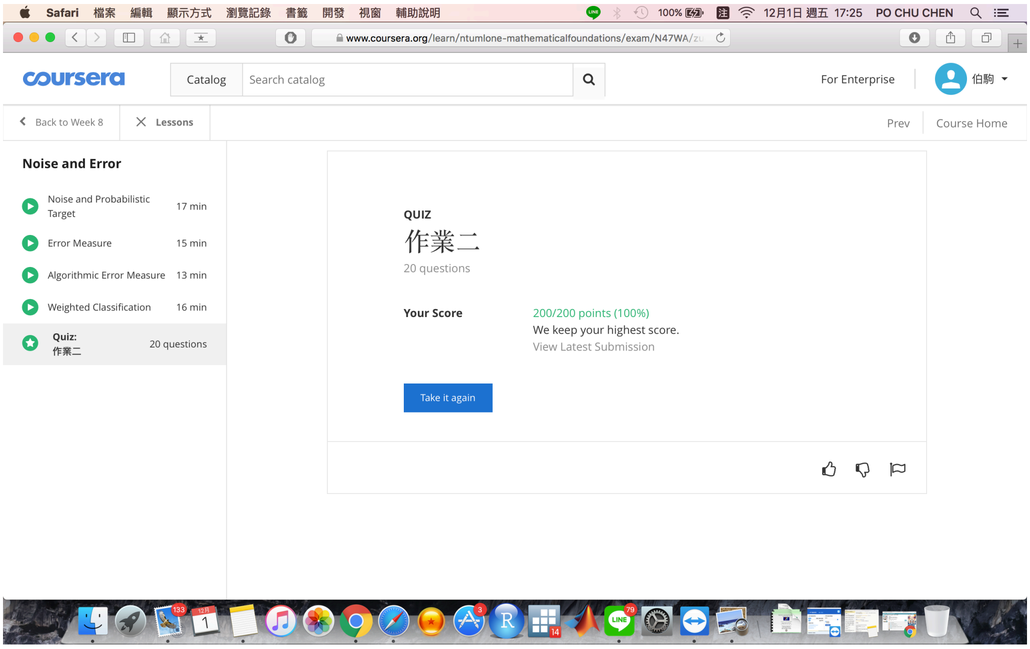
Home Work 2

Machine Learning Foundations

R04323050

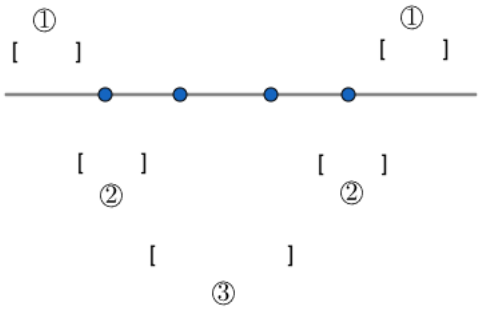
經濟碩三　陳伯駒

1.



2.

Positive & Negative interval on $\mathbb{R}$. 可能的區間有以下情形：



①: 全為"−" 或全為"+"

②: +− 或 −+

③: + − + 或 − + −

N 個點中間有 N-1 個間隔

$\therefore m_{\mathcal{H}}(N) = 2 \times (1 + \binom{N-1}{1} + \binom{N-2}{2}) = N^2 - N + 2$

3.

<u>**Claim:**</u> $d_{vc}(\mathcal{H}) = D + 1$

We can observe the Hypothesis Set $\mathcal{H}$ is a D-dim PLA from the slide in lecture 2.

$\therefore$ By the slide in lecture 7, we've shown $d_{vc}(\mathcal{H}) = D + 1$ during the class.

4.

"Triangle Waves" Hypothesis Set in $\mathbb{R}$:

$$\mathcal{H} = \{h_\alpha | \quad h_\alpha(x) = sgn(|(\alpha x)\, mod\, 4 - 2| - 1), \alpha \in \mathbb{R}\}$$

是個週期為 $\frac{4}{|\alpha|}$ 的三角波函數.[1]

$\because \alpha \in \mathbb{R} \therefore$ 週期可以任意小。i.e x 軸 ($\mathbb{R}^1$) 可以被該曲線切成無限多個區域，故 $d_{vc}(\mathcal{H}) = \infty$

5.

<u>**Claim:**</u> If $\mathcal{H}_1 \subseteq \mathcal{H}_2$, then $d_{vc}(\mathcal{H}_1) \leq d_{vc}(\mathcal{H}_2)$

Suppose $d_{vc}(\mathcal{H}) > d_{vc}(\mathcal{H}_2)$，則代表 $\mathcal{H}_1$ 可以 shatter 的 inputs 個數超過 $\mathcal{H}_2$ 所可以 shatter 的 inputs。i.e 至少存在一個 inputs x 使得 $\mathcal{H}_1(x) \notin \mathcal{H}_2$, contradiction.($\because$

[1]Triangle Waves Function: http://bit.ly/2nneX45

$\mathcal{H}_1 \subseteq \mathcal{H}_2$, $\therefore$ $\mathcal{H}_1$ 所能夠產生的 dichotomies，$\mathcal{H}_2$ 也都要能夠產出)

Hence, $d_{vc}(\mathcal{H}) \le d_{vc}(\mathcal{H}_2)$

6.

By Q.15 on Couresa: **Claim:** $\max\{d_{vc}(\mathcal{H}_k)\}_{k=1}^{K} \le d_{vc}(\bigcup_{k=1}^{K}\mathcal{H}_k) \le K-1+\sum_{k=1}^{K} d_{vc}(\mathcal{H}_k)$

Left: 設 $\mathcal{H}_1, \mathcal{H}_2, ..., \mathcal{H}_K$ 所能夠 shatter 的最大 inputs 個數為 N，則 $\bigcup_{k=1}^{K}\mathcal{H}_k$ 也至少能夠 shatter N 個 inputs:

Suppose not, i.e suppose $d_{vc}(\bigcup_{k=1}^{K}\mathcal{H}_k) = N-1$，這些 $\mathcal{H}_1, \mathcal{H}_2, ..., \mathcal{H}_K$ 聯集起來所形成的 Hypothesis Set 最多只能 shatter $N-1$ 個 inputs，代表這之中所能 shatter 最多的 inputs 也只會到 $N-1$ 個, contradiction.

Hence, $\max\{d_{vc}(\mathcal{H}_k)\}_{k=1}^{K} \le d_{vc}(\bigcup_{k=1}^{K}\mathcal{H}_k)$

Right: 假設現在只有 $\mathcal{H}_1, \mathcal{H}_2$ 這兩種 Hypothesis Sets，$\mathcal{H}_1$ 是把平面上所有的點歸類為 $+1$; $\mathcal{H}_2$ 是把平面上所有點歸類為 -1，則我們知道 $d_{vc}(\mathcal{H}_1) = 0$ & $d_{vc}(\mathcal{H}_2) = 0$，$d_{vc}(\mathcal{H}_1 \cup \mathcal{H}_2) = 1$。

$\therefore$ 從 Coursera Q.15 的選項中，$d_{vc}(\mathcal{H}_1 \cup \mathcal{H}_2) = 1$, $\sum_{k=1}^{K} d_{vc}(\mathcal{H}_k) = 0$.

Hence, $d_{vc}(\mathcal{H}_1 \cup \mathcal{H}_2) = 1 \le 2-1+0 = K-1+\sum_{k=1}^{K} d_{vc}(\mathcal{H}_k) = 0$ 成立。

Therefore, $\max\{d_{vc}(\mathcal{H}_k)\}_{k=1}^{K} \le d_{vc}(\bigcup_{k=1}^{K}\mathcal{H}_k) \le K-1+\sum_{k=1}^{K} d_{vc}(\mathcal{H}_k)$.

Now let $\mathcal{H}_1$ be positive-ray hypothesis set and $\mathcal{H}_2$ be negative-ray hypothesis set. By the slides in lecture 5, we know: $m_{\mathcal{H}_1}(N) = N+1, \quad d_{vc}(\mathcal{H}_1) = 1$

$$m_{\mathcal{H}_2}(N) = N+1, \quad d_{vc}(\mathcal{H}_2) = 1$$

$\therefore \max\{d_{vc}(\mathcal{H}_k)\}_{k=1}^{2} = 1 \le d_{vc}(\mathcal{H}_1 \cup \mathcal{H}_2) \le K-1+\sum_{k=1}^{2} d_{vc}(\mathcal{H}_k) = 2-1+2 = 3$

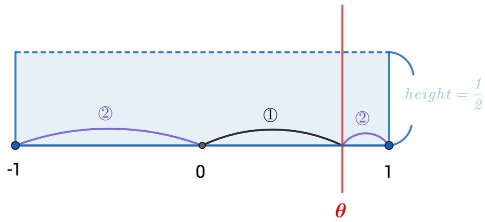
$\Rightarrow 1 \le d_{vc}(\mathcal{H}_1 \cup \mathcal{H}_2) \le 3$.

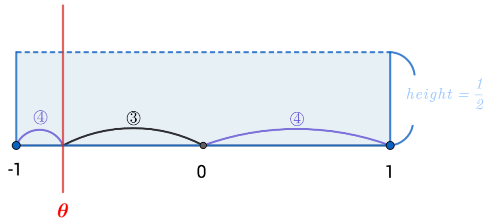
Also, we know the hypothesis set $\mathcal{H}_1 \cup \mathcal{H}_2$ is actually the 1-d perceptron. Hence, $m_{\mathcal{H}_1 \cup \mathcal{H}_2}(N) = 2N$ and $d_{vc}(\mathcal{H}_1 \cup \mathcal{H}_2) = 2$ by the slides in lecture 5 and 7, which holds in the above inequality.

7.

x is generated by a uniform distribution in [-1,1].

	θ	s	預測錯誤率 $\mu = P(h \neq f)$ $= P(s \cdot sgn(x-\theta) \neq sgn(x))$
①	>0	$+1$	$P(sgn(x-\theta) \neq sgn(x)) = \theta \times \frac{1}{2}$
②	>0	-1	$P(-sgn(x-\theta) \neq sgn(x))$ $= [1+(1-\theta)] \times \frac{1}{2} = 1 - \frac{\theta}{2}$
③	<0	$+1$	$P(sgn(x-\theta) \neq sgn(x)) = -\theta \times \frac{1}{2}$
④	<0	-1	$P(-sgn(x-\theta) \neq sgn(x))$ $= [1+(1+\theta)] \times \frac{1}{2} = 1 + \frac{\theta}{2}$





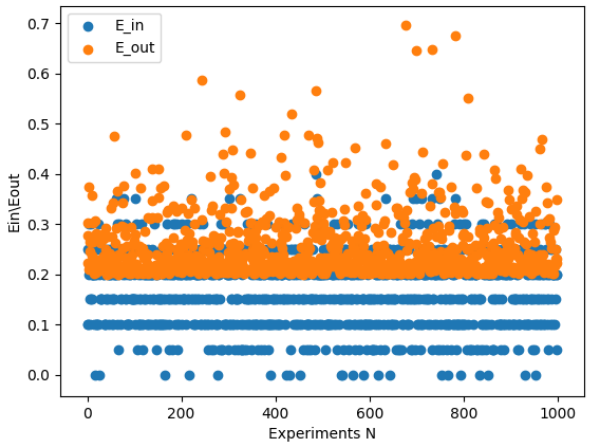
綜合①、②、③、④：$\mu = \begin{cases} |\theta| \times \frac{1}{2} & \text{if } s=+1 \\ 1-\frac{|\theta|}{2} & \text{if } s=-1 \end{cases} \xrightarrow{\text{兩點式}} \mu = \frac{1}{2} + (\frac{|\theta|-1}{2}) \times s$。[2]

By Q.1 on coursera, we know $E_{out}(h_{s,\theta}) = \lambda \cdot \mu + (1-\lambda) \cdot (1-\mu)$, where $\lambda = 1 - 0.2 = 0.8$

$$= 0.8 \times \mu + 0.2 \times (1-\mu)$$

$$= 0.5 + 0.3 \cdot (|\theta| - 1) \cdot s$$

8.

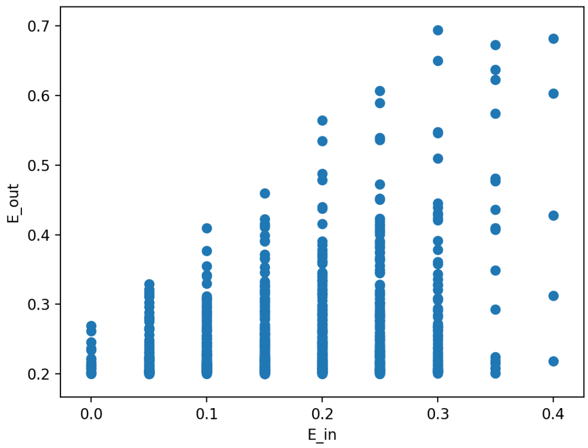


In the left figure, we can observe that the value of E_{in} is at least 0.2, which is exactly the probability of flipping noise. Intuitively, E_{out} is the expectation of $[\![g(x) \neq f(x)]\!]$ out of sample, now the flipping rate is 20%, then the above expectation term will be at least 20%.

Though we also have noise in sample of E_{in}, we can choose s and θ to let E_{in} become smaller, so E_{in} could be less than 20%. However, the flipped y for out-of-sample is fol-

[2]Let $\mu = a \cdot s + b$. $|\theta| \times \frac{1}{2} = a + b - ①$, $1 - \frac{|\theta|}{2} = -a + b - ②$ By ①, ② $\Rightarrow a = \frac{|\theta|-1}{2}, b = \frac{1}{2}$

lowed a distribution(i.e our target funtcion) like Q.1 in coursera, which has a 20% filpped rate the optimal s and θ that I choose through E_{in}, so E_{out} will be at least 20%.



Moreover, if we put E_{in} and E_{out} in the same plot, we can observe that when E_{in} is smaller; the variation of E_{out} will also be smaller. This result corresponds to what we expect: we can let E_{out} be small enough as long as we choose optimal s and θ to minimize E_{in}. i.e Learning succeed: $E_{in} \approx E_{out}$ and E_{in}, E_{out} are small.

9.

Cover's Function Counting Theorem:

Let $\{x^1, x^2, ..., x^p\}$ be vectors in $\mathbb{R}^N$, then the number of distinct dichotomies applied to these points that can be realized by a plane through the origin is :

$$C(P,N) = 2 \times \sum_{k=0}^{N-1} \binom{P-1}{k}$$

在 d-維的 PLA 中，我們會對門檻值 w_0 再墊高一個向量 $x_0 = (1,1,1,...,1)$，用來突破分隔線只能通過原點的限制，而廣義上來說就是在 $\mathbb{R}^{d+1}$ 中的向量 $\{x_1, x_2, ..., x_N\}$ 做通過原點的 PLA。

$\therefore$ By Cover's theorem, $m_{\mathcal{H}}(N) = C(N, d+1) = 2 \times \sum_{i=0}^{d+1-1} \binom{N-1}{i} = 2 \times \sum_{i=0}^{d} \binom{N-1}{i}$

Proof of Cover's theorem:[3]

Denote the number of linearly separable partition by $C(P,N)$. We will find the expression for $C(P,N)$ by induction. Image first having p points and then adding one more point. Now, considering the linearly separable partitions of previous p points, there are two possibilities:

[3]Reference: http://bit.ly/2nnEtGC

Case 1: there is a separating hyperplane for the previous p points passing through the new point, in which case each such linearly separable partition of the previous p points gives rise to two distinct linearly separable partitions as the hyperplane can be shifted infinitesimally to place the new point in either class.
Case 2: there is no separating hyperplane for the previous p points passing through the new point, in which case each such linearly separable partition gives rise to only one linearly separable partition.
The number of linearly separable partition in Case 1 is precisely $C(P, N-1)$, because restricting the separating hyperplane to pass through a fixed point is the same as eliminating one degree of freedom and thus projecting the p points to a $N-1$-dim space. This can be understood if the new point is on the x-axis, for example - then the hyperplane has $N-1$ axes left to work with. If the point is not on the x-axis, then rotate the axes of space around to get the point on the x axis, and this of course has no effect on the geometry of the problem.
The recursive relation:
$C(P+1, N) = C(P, N) + C(P, N-1)$, where $C(P, N)$ is the number of separable hyperplanes in Case 2, and $C(P, N-1)$ is the number of separable hyperplanes in Case 1.

Iterating the recursion once, we have
$C(P+1, N) = C(P-1, N) + 2C(P-1, N-1) + C(P-1, N-2)$

Continue to iterate the recursion (twice)
$C(P+1, N) = C(P-2, N) + 3C(P-2, N-1) + 3C(P-2, N-2) + C(P-2, N-3)$

After $P-1$ iterations, we have
$C(P+1, N) = \binom{P}{0}C(1, N) + \binom{P}{1}C(1, N-1) + ... + \binom{P}{P}C(1, N-P)$, where $C(1, k) = 2$ for all $k \leq 1$.

So, finally we have $C(P+1, N) = 2 \times \sum_{i=0}^{N-1} \binom{P}{i}$, where $\binom{P}{i} = 0$ if $i > P$.