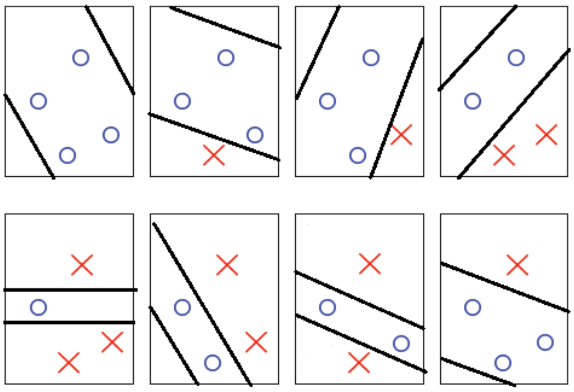
Machine Learning Foundation (NTU, Fall 2018) Homework #2

陳熙 R07922151

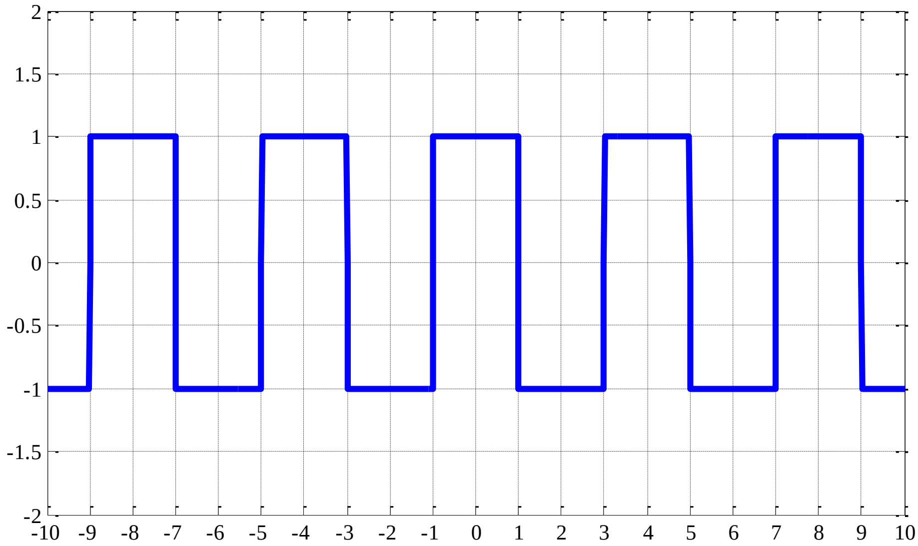
1.

2. $|w_0 + w_1x_1 + w_2x_2| \le \theta$ 可以寫成 $-\theta \le w_0 + w_1x_1 + w_2x_2 \le \theta$ 即 $-\theta - w_0 \le w_1x_1 + w_2x_2 \le \theta - w_0$。所以，"negative thick line" 表示的假說為兩條平行直線之間為+1，其他部分為-1。通過畫圖（剩下的 8 種情況通過翻轉樣本的 y，以及調整直線位置同樣可以實現）我們可以看到，



當 $N=4$，"negative thick line"這個假說可以 shatter。所以，"negative thick line" 的 VC-Dimension 至少為 4。

3. "triangle wave"的 VC-Dimension 為無窮大。當 α 為 1 時，圖形如下：



可以觀察到圖形為方波的形狀。而 α 可以任意改變，使得方波的週期可以為任意小。我們就可以使得每個+1 和-1 取值的區間無限小。這樣不管對於怎樣的樣本，我們只要讓 α 取適當值就可以讓樣本落入合適的區間，從而 shatter 所有的二分集合，因此”triangle wave”這個假設的 VC-Dimension 為無窮大。

4. 因為假說集合$\mathcal{H}_1$和假說集合$\mathcal{H}_2$的 intersection 指的是對於相同的 $\boldsymbol{x}$ 集合，$\mathcal{H}_1$和$\mathcal{H}_2$有相同的 y 集合的假說集合。這個部分只有在$\mathcal{H}_2$為$\mathcal{H}_1$子集或者$\mathcal{H}_2$和$\mathcal{H}_1$相同時，這個部分才和$\mathcal{H}_2$相同。其餘情況下，這個部分都僅僅是$\mathcal{H}_2$的一個子集。因此，假說$\mathcal{H}_1$和假說$\mathcal{H}_2$的 intersection 能夠 shatter 的最多就是$\mathcal{H}_2$能夠 shatter 的部分。假說$\mathcal{H}_2$不能 shatter 的部分，$\mathcal{H}_1$和$\mathcal{H}_2$的 intersection 一定也不能 shatter。所以，假說集合$\mathcal{H}_1$和假說集合$\mathcal{H}_2$的交集部分的 VC-Dimension 小於或等於假說集合$\mathcal{H}_2$的 VC-Dimension。

5. 與假說$\mathcal{H}_1$相同（假說$\mathcal{H}_2$只是把假說$\mathcal{H}_1$的 y 翻轉），假說$\mathcal{H}_2$的成長函數 $m_{\mathcal{H}_2}(N)=N+1$。假說$\mathcal{H}_1$和假說$\mathcal{H}_2$重複的地方在於 y 全為+1 和全為-1 的情況。因此，有

$$m_{\mathcal{H}_1\cup\mathcal{H}_2}(N)=2(N+1)-2=2N$$

例如，當 N=3 時，$\mathcal{H}_1$ 有的 dichotomy 為

x_1	x_2	x_3
○	○	○
×	○	○
×	×	○
×	×	×

$\mathcal{H}_2$ 有的 dichotomy 為

x_1	x_2	x_3
×	×	×
○	○	×
○	×	×
○	○	○

而 $\mathcal{H}_1 \cup \mathcal{H}_2$ 有的 dichotomy 為：

x_1	x_2	x_3
○	○	○
×	○	○
×	×	○
×	×	×
○	○	×
○	×	×

因此，$\mathcal{H}_1 \cup \mathcal{H}_2$ 出現 break point 的地方為 $N=3$，所以 $d_{vc}\left(\mathcal{H}_1 \cup \mathcal{H}_2\right)=2$。

6. 因為 $f(x)=\overline{s}(x)+\text{noise}$，其中 noise 翻轉了 20% 的結果。令 $\lambda = P\left(f(x)=\overline{s}(x) \mid x\right)$，

$$P\left(f(x) \mid x\right)=\begin{cases}0.8, f(x)=\overline{s}(x)\\ 0.2, f(x)=-\overline{s}(x)\end{cases}$$

有 $\lambda=0.8$。我們令 $\mu=P\left(h_{s,\theta}(x)\neq \overline{s}(x) \mid x\right)$，有

$$\begin{aligned}
E_{out}\left(h_{s,\theta}\right) &= P\left(f(x)\neq h_{s,\theta}(x) \mid x\right)\\
&= P\left(\overline{s}(x)\neq h_{s,\theta}(x) \mid f(x)=\overline{s}(x)\right)P\left(f(x)=\overline{s}(x) \mid x\right)\\
&\quad +P\left(\overline{s}(x)= h_{s,\theta}(x) \mid f(x)\neq\overline{s}(x)\right)P\left(f(x)\neq\overline{s}(x) \mid x\right)\\
&= P\left(\overline{s}(x)\neq h_{s,\theta}(x)\right)P\left(f(x)=\overline{s}(x) \mid x\right)+P\left(\overline{s}(x)= h_{s,\theta}(x)\right)P\left(f(x)\neq\overline{s}(x) \mid x\right)\\
&= \mu\lambda+(1-\mu)(1-\lambda)
\end{aligned}$$

根據 s 和 θ 不同情況有：

(1) 當 $s=1$，$0\le\theta\le1$時，$h_{s,\theta}(x)=\text{sign}(x-\theta)$，此時因為 x 為在[-1,1]上的均勻分佈，$\overline{s}(x)$ 在 x 大於 0 時為+1，小於 0 時為-1。$h_{s,\theta}(x)$ 在 x 大於 θ 時（不考慮等於的邊界情況，對計算機率無影響）為+1，x 小於 θ

時為-1。此時，在$0 \le x \le \theta$這個區間$h_{s,\theta}(x) \ne \bar{s}(x)$，此時$\mu = \frac{\theta - 0}{2} = \frac{\theta}{2}$。

(2) 當$s = 1$，$-1 \le \theta \le 0$時，此時，在$\theta \le x \le 0$這個區間$h_{s,\theta}(x) \ne \bar{s}(x)$，此時$\mu = \frac{0-\theta}{2} = \frac{-\theta}{2}$。

(3) 當 s = -1，$0 \le \theta \le 1$時，$h_{s,\theta}(x) = -\mathrm{sign}(x-\theta)$，此時$h_{s,\theta}(x)$在$x$大於$\theta$時（不考慮等於的邊界情況，對計算機率無影響）為-1，x小於θ時為+1。此時，只在$0 \le x \le \theta$這個區間$h_{s,\theta}(x) = \bar{s}(x)$，此時$\mu = 1 - \frac{\theta - 0}{2} = \frac{2-\theta}{2}$。

(4) 當s=-1，$-1 \le \theta \le 0$時，此時，只在$\theta \le x \le 0$這個區間$h_{s,\theta}(x) = \bar{s}(x)$，此時$\mu = 1 - \frac{0-\theta}{2} = \frac{2+\theta}{2}$。

θ在其他情況下，全為在[+1,-1]區間，$h_{s,\theta}(x)$全為+1 或-1，此時$\mu = \frac{1}{2}$。綜上，μ可以表示為

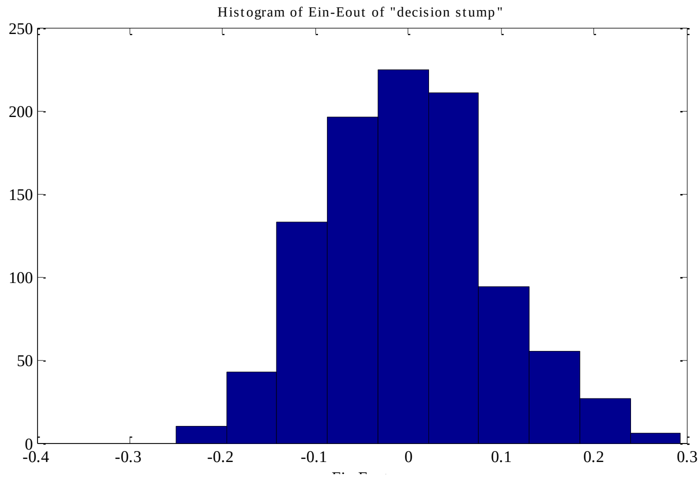
$$\mu = \frac{s+1}{2}\frac{|\theta|}{2} - \frac{s-1}{2}\frac{2-|\theta|}{2} = \frac{s|\theta| - s + 1}{2}\text{。}$$

因此，有

$$\begin{aligned} E_{out}(h_{s,\theta}) &= 0.8 \times \frac{s|\theta| - s + 1}{2} + 0.2 \times \left(1 - \frac{s|\theta| - s + 1}{2}\right) \\ &= 0.3s|\theta| - 0.3s + 0.5 \\ &= 0.5 + 0.3s(|\theta| - 1) \end{aligned}$$

7. 可以發現$E_{in} - E_{out}$的分佈接近於零均值的正態分佈。$E_{in} - E_{out}$集中分佈在零點周圍，E_{in}與E_{out}較為接近。因此，在這個樣本中”decision stump”可以通過 minimize E_{in} 實現 minimize E_{out}。

直方圖如下：



8. 我們通過限制每個 dichotomy 中-1 的數目小於 k（即小於或等於 k-1）從而構造一個 dichotomy 的集合，這樣這個 dichotomy 集合將因為缺少一 N 個點中 k 點全為-1 的 dichotomy，讓這個集合成為無法 shatter k 個點的最大集合。但是再加上任一不同的 dichotomy（這個 dichotomy 中-1 的數目一定會大於或等于 k）都將 shatter k 個點。這個集合的數目相當於每次從 N 個點中挑選可以為-1 的點的組合數目之和 $\sum_{i=0}^{k-1}\binom{N}{i}$。

即 $i=0$ 時，$\binom{N}{0}$表示全為+1，無-1 的 dichotomy 的數目，而 $i=1$，$\binom{N}{1}$表示有一個點為-1，其他 N-1 個點為+1 的 dichotomy 的數目，以此類推，當 $i=k-1$ 時，$\binom{N}{k-1}$表示有 k-1 個點為-1 的 dichotomy 數目，因為當集合中有 k 個及 k 個以上-1 的 dichotomy 時，這個集合將 shatter k 個點，所以這個集合的 dichotomy 數目為 $\sum_{i=0}^{k-1}\binom{N}{i}$。

綜上，$B(N,k)=\sum_{i=0}^{k-1}\binom{N}{i}$。