Machine Learning Foundation (NTU, Fall 2018) Homework #1

陳熙 R07922151

1.

2. 因爲對於所有 $\mathbf{x}$， $f(\mathbf{x})=+1$，而 $g(\mathbf{x})$在 k 爲偶數且 $N+1\le k\le N+L$ 時爲-1。此時兩者不相等。因此，

$$E_{OTS}(g,f)=\frac{1}{L}\sum_{N+1\le k\le N+1} I(k\text{為偶數})$$

根據 N 和 $N+L$ 的奇偶性可以分爲以下情況討論：

（1） 當 N 和 $N+L$ 都為偶數時，此時 $N+1\le k\le N+L$ 中 k 為偶數的數目為 $\frac{N+L}{2}-\frac{N}{2}$。

（2） 當 N 為奇數，$N+L$ 為偶數時，此時 $N+1\le k\le N+L$ 中 k 為偶數的數目為 $\frac{N+L}{2}-\frac{N-1}{2}=\frac{N+L}{2}-\left\lfloor\frac{N}{2}\right\rfloor$。

（3） 當 N 為奇數，$N+L$ 也為奇數時，此時 $N+1\le k\le N+L$ 中 k 為偶數的數目為 $\frac{N+L-1}{2}-\frac{N-1}{2}=\left\lfloor\frac{N+L}{2}\right\rfloor-\left\lfloor\frac{N}{2}\right\rfloor$

（4） 當 N 為偶數，$N+L$ 為奇數時，此時 $N+1 \le k \le N+L$ 中 k 為偶數的數目為 $\frac{N+L-1}{2} - \frac{N}{2} = \left\lfloor \frac{N+L}{2} \right\rfloor - \frac{N}{2}$ 。

綜上，$E_{OTS}(g,f) = \frac{1}{L}\left(\left\lfloor \frac{N+L}{2} \right\rfloor - \left\lfloor \frac{N}{2} \right\rfloor \right)$ 。

3. 令演算法$\mathcal{A}_1$、$\mathcal{A}_2$ 產生的假說分別爲 g_1、g_2。由題目可以知道

$$\mathbb{E}_f\left\{E_{OTS}\left(\mathcal{A}_1(\mathcal{D}), f\right)\right\} = \mathbb{E}_f\left\{\frac{1}{L}\sum_{\ell=1}^{L}\left[\!\left[g_1\left(\mathbf{x}_{N+\ell}\right) \ne f\left(\mathbf{x}_{N+\ell}\right)\right]\!\right]\right\}$$

因爲對於所有的 f，每個 f 出現的機率相等。而對於一個固定的 g，其中的一個 $g\left(\mathbf{x}_{N+\ell}\right)$，某個 f 對應的 $f\left(\mathbf{x}_{N+\ell}\right)$ 正確或錯誤的機率都爲 $\frac{1}{2}$ ，即

$$\mathbb{E}_f\left[\!\left[g\left(\mathbf{x}_{N+\ell}\right) \ne f\left(\mathbf{x}_{N+\ell}\right)\right]\!\right] = \frac{1}{2}$$

所以有，

$$\begin{aligned}\mathbb{E}_f\left\{E_{OTS}\left(\mathcal{A}_1(\mathcal{D}), f\right)\right\} &= \mathbb{E}_f\left\{\frac{1}{L}\sum_{\ell=1}^{L}\left[\!\left[g_1\left(\mathbf{x}_{N+\ell}\right) \ne f\left(\mathbf{x}_{N+\ell}\right)\right]\!\right]\right\} \\ &= \frac{1}{L}\sum_{\ell=1}^{L}\mathbb{E}_f\left[\!\left[g_1\left(\mathbf{x}_{N+\ell}\right) \ne f\left(\mathbf{x}_{N+\ell}\right)\right]\!\right] \\ &= \frac{1}{L} \times L \times \frac{1}{2} \\ &= \frac{1}{2}\end{aligned}$$

同理，對 g_2 也有同樣的結論。因此，

$$\mathbb{E}_f\left\{E_{OTS}\left(\mathcal{A}_1(\mathcal{D}), f\right)\right\} = \mathbb{E}_f\left\{E_{OTS}\left(\mathcal{A}_2(\mathcal{D}), f\right)\right\} = \frac{1}{2}\text{。}$$

4. $\nu \le 0.1$的機率為

$$\begin{aligned}P(\nu \le 0.1) &= P(10\text{個彈珠中有}1\text{個或}0\text{個橘色}) \\ &= (1-\mu)^{10} + \binom{10}{1}\mu(1-\mu)^9 \\ &= 0.2^{10} + 10 \times 0.8 \times 0.2^9 \\ &\approx 4.20 \times 10^{-6}\end{aligned}$$

$\nu \geq 0.9$的機率為

$$\begin{aligned} P(\nu \geq 0.9) &= P(10\text{個彈珠中有}9\text{個或}10\text{個橘色}) \\ &= \mu^{10} + \binom{10}{9}\mu^{9}(1-\mu) \\ &= 0.8^{10} + 10 \times 0.8^{9} \times 0.2 \\ &\approx 0.38 \end{aligned}$$

5. 從題目可以知道，A 中 1、3、5 為綠色，B 中 2、4、6 為綠色，C 中 4、5、6 為綠色，D 中 1、2、3 為綠色。取五個骰子，要得到五個綠色的 1，則這些骰子必須為 A 種或者 D 種，而 A，D 的數目在全部骰子中占一半。且每次取骰子的結果互相獨立。因此，

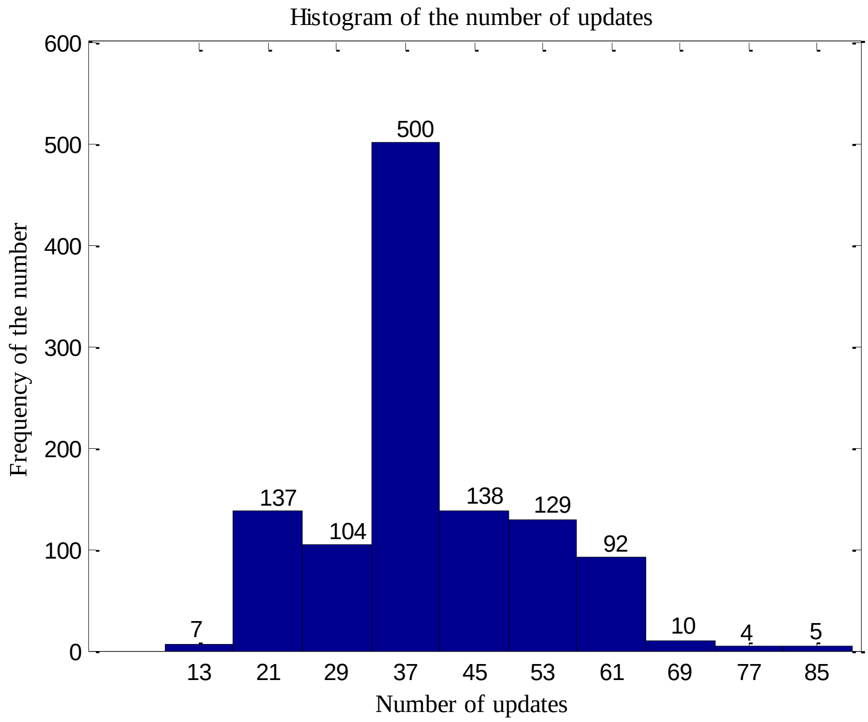
$$P(\text{擲}5\text{個骰子，得到}5\text{個綠色的}1) = \left(\frac{2}{4}\right)^{5} = \frac{1}{32}$$

6. 由上題分析可知，當 5 個骰子都為 A、D 種時，1 和 3 全為綠色；當 5 個骰子都為 B 種和 D 種時，2 全為綠色；當 5 個骰子全為 B 種和 C 種時，4 和 6 全為綠色；當 5 個骰子全為 A 種和 C 種時，5 全為綠色。而每種骰子組合的數目分別都占全部骰子數目的一半。由以上幾種情況的分析，可以知道，

$$\begin{aligned} P(\text{擲}5\text{個骰子，得到有些數字全為綠色}) &= P(5\text{個骰子全為AD或BD或BC或AC}) \\ &= 4 \times \left(\frac{1}{2}\right)^{5} - 4 \times \left(\frac{1}{4}\right)^{5} \\ &= \frac{31}{256} \end{aligned}$$

上式中第二行減去式子的原因在於在之前計算中重複了全為 A 種、全為 B 種、全為 C 種和全為 D 種的情況。我發現了取 5 個骰子，一些數字出現全為綠色的機率並不是數字 1 出現全為綠色機率的六倍，說明有些數字出現全為綠色的情況是相互關聯的，並不是獨立的。

7. 演算法停止前平均的停止次數為 40 次。直方圖如下。



8. 要使 $y_{n(t)}\mathbf{w}_{t+1}^T\mathbf{x}_{n(t)} > 0$，又有 $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + y_{n(t)}\mathbf{x}_{n(t)} \cdot M$ ，因此

$$y_{n(t)}\left(\mathbf{w}_t + y_{n(t)}\mathbf{x}_{n(t)} \cdot M\right)^T \mathbf{x}_{n(t)} > 0$$

$$y_{n(t)}\mathbf{w}_t^T\mathbf{x}_{n(t)} + M \cdot y_{n(t)}^2\mathbf{x}_{n(t)}^T\mathbf{x}_{n(t)} > 0$$

$$M > \frac{-\mathbf{w}_t^T\mathbf{x}_{n(t)}}{y_{n(t)}\mathbf{x}_{n(t)}^T\mathbf{x}_{n(t)}}$$

最後一個式子成立的原因在於，$y_{n(t)}^2\mathbf{x}_{n(t)}^T\mathbf{x}_{n(t)} > 0$。因此，$M_{n(t)} = \left\lceil \frac{-\mathbf{w}_t^T\mathbf{x}_{n(t)}}{y_{n(t)}\mathbf{x}_{n(t)}^T\mathbf{x}_{n(t)}} \right\rceil$。

令 $\rho = \min_{1\le n(t)\le N,\ \forall t>0} M_{n(t)} \cdot y_{n(t)}\left(\mathbf{w}^{*T}\mathbf{x}_{n(t)}\right)$，由於 $\mathbf{w}^*$ 正確預測了所有 $y_{n(t)}$。因此，有 $y_n=\text{sign}\left(\mathbf{w}^{*T}\mathbf{x}_n\right)$。所以對於所有的 n，有 $y_n\left(\mathbf{w}^{*T}\mathbf{x}_n\right) > 0$ ，$M_{n(t)}$也大於 0，因此 $\rho > 0$。

若 $y_{n(t)} \ne \text{sign}\left(\mathbf{w}_t^T\mathbf{x}_{n(t)}\right)$時，有

$$
\begin{aligned}
\mathbf{w}_{t+1}^T\mathbf{w}^* &= \left(\mathbf{w}_t + y_{n(t)}\mathbf{x}_{n(t)}\cdot M\right)^T \mathbf{w}^* \\
&= \mathbf{w}_t{}^T\mathbf{w}^* + M_{n(t)}\cdot y_{n(t)}\mathbf{w}^{*T}\mathbf{x}_{n(t)} \\
&\geq \mathbf{w}_t{}^T\mathbf{w}^* + \min_{1\leq n(t)\leq N,\ \forall t>0} M_{n(t)}\cdot y_n\left(\mathbf{w}^{*T}\mathbf{x}_n\right) \\
&= \mathbf{w}_t{}^T\mathbf{w}^* + \rho
\end{aligned}
$$
。

利用數學歸納法，假設 $\mathbf{w}_t^T\mathbf{w}^* \geq t\rho$，則 $\mathbf{w}_{t+1}^T\mathbf{w}^* \geq \mathbf{w}_t{}^T\mathbf{w}^* + \rho \geq t\rho + \rho = (t+1)\rho$。

由於 $y_{n(t)} \neq \operatorname{sign}\left(\mathbf{w}_t{}^T\mathbf{x}_{n(t)}\right)$，所以有 $M\cdot y_{n(t)}\mathbf{w}_t{}^T\mathbf{x}_{n(t)} < 0$。

$$
\begin{aligned}
\left\|\mathbf{w}_{t+1}\right\|^2 &= \left\|\mathbf{w}_t + y_{n(t)}\mathbf{x}_{n(t)}\cdot M_{n(t)}\right\|^2 \\
&= \left\|\mathbf{w}_t\right\|^2 + M_{n(t)}{}^2\left\|\mathbf{x}_{n(t)}\right\|^2 + 2M_{n(t)}\cdot y_{n(t)}\mathbf{w}_t{}^T\mathbf{x}_{n(t)} \\
&\leq \left\|\mathbf{w}_t\right\|^2 + M_{n(t)}{}^2\left\|\mathbf{x}_{n(t)}\right\|^2
\end{aligned}
$$

令 $R = \max_{1\leq n(t)\leq N,\forall t>0} M_{n(t)}\left\|\mathbf{x}_{n(t)}\right\|$。利用數學歸納法，假設 $\left\|\mathbf{w}_t\right\|^2 \leq tR^2$。

則 $\left\|\mathbf{w}_{t+1}\right\|^2 \leq \left\|\mathbf{w}_t\right\|^2 + M_{n(t)}{}^2\left\|\mathbf{x}_{n(t)}\right\|^2 \leq tR^2 + \left(\max_{1\leq n(t)\leq N,\forall t>0} M_{n(t)}\left\|\mathbf{x}_{n(t)}\right\|\right)^2 = (t+1)R^2$。

由 $\mathbf{w}_{t+1}^T\mathbf{w}^* \geq (t+1)\rho$ 和 $\left\|\mathbf{w}_{t+1}\right\| \leq (t+1)R^2$，有 $\dfrac{\mathbf{w}_{t+1}^T\mathbf{w}^*}{\left\|\mathbf{w}_{t+1}\right\|} \geq \sqrt{t+1}\dfrac{\rho}{R}$。

由於 $1 \geq \cos\theta = \dfrac{\mathbf{w}_{t+1}^T\mathbf{w}^*}{\left\|\mathbf{w}_{t+1}^T\right\|\left\|\mathbf{w}^*\right\|} \geq \dfrac{\sqrt{t+1}}{\left\|\mathbf{w}^*\right\|}\dfrac{\rho}{R}$，所以有

$$
t+1 \leq \frac{R^2\left\|\mathbf{w}^*\right\|^2}{\rho^2} = \frac{\left(\max_{1\leq n(t)\leq N,\forall t>0} M_{n(t)}\left\|\mathbf{x}_{n(t)}\right\|\right)^2\left\|\mathbf{w}^*\right\|^2}{\left(\min_{1\leq n(t)\leq N,\ \forall t>0} y_{n(t)}M_{n(t)}\left(\mathbf{w}^{*T}\mathbf{x}_{n(t)}\right)\right)^2} = \left(\frac{\left(\max_{1\leq n(t)\leq N,\forall t>0} M_{n(t)}\left\|\mathbf{x}_{n(t)}\right\|\right)\left\|\mathbf{w}^*\right\|}{\min_{1\leq n(t)\leq N,\ \forall t>0} M_{n(t)}\mathbf{w}^{*T}\mathbf{x}_{n(t)}}\right)^2
$$

。

綜上，當資料集合線性可分時，新的更新規則將保證在有“完美的線”時停止。