# Machine Learning Homework 1

資工碩一 r09922055 陳柏妤

1. **(d)**

   因為 (a) 沒有可供機器學習的 underlying pattern，(b)(c) 則是有 programmable definition 不需要用到機器學習。只有 (d) 具有某種 underlying pattern 且無法被我們準確定義，因此會需要機器學習。

2. **(e)**

   (a) 是 random 做分類，與 spam 的任何性質皆無關，也沒有讓機器學習如何辨認 spam。

   (b) 是以人工做分類，沒有讓機器學習。

   (c) 也是以人工制定規則，讓機器單純依照那些 rule 工作，並沒有學到辨認 spam 的 underlying pattern。

   (d) 設定 “10 words” 作為 threshold 並非機器學習，應該讓機器從不同的 threshold hypothesis 中，依照 data 選擇一個分類得最好的 threshhold 作為標準

3. **(d)**

   by 課後證明 $T \leq \frac{R^2}{\rho^2}$

   · assume $\min_n y_n w_f^T x_n = \|w_f\| \cdot \rho$，若 $x_n$ scale down 4 倍，$\min_n y_n w_f^T x_n' = \|w_f\| \cdot \frac{\rho}{4} = \|w_f\| \cdot \rho'$

   · assume $\max_n \|x_n\|^2 = R^2$，若 $x_n$ scale down 4 倍，$\max_n \|x_n'\|^2 = \frac{R^2}{16} = R'^2$

   · $T \leq \frac{R'^2}{\rho'^2} = \frac{\frac{R^2}{16}}{\frac{\rho^2}{4^2}} \leq \frac{R^2}{\rho^2}$ （unchanged）

4. **(c)**

延伸自上課的「Convergence of Perceptron Learning Algorithm」證明

(a) 證 $w_f^T w_{t+1} \geq w_f^T w_t + \hat{\rho} \cdot \|w_f\|$：

" for linear separable data with both class of examples, a separable $w_f$ means $\|w_f\| > 0$ and $\rho > 0$ "，代表 $y_n w_f^T x_n > 0$。又 $y_n = \pm 1$，因此 $y_n w_f^T x_n = |y_n w_f^T x_n| = |w_f^T x_n|$

By 題目， $w_{t+1} = w_t + \eta_t \cdot y_{n(t)} x_{n(t)}$，則：

$w_f^T w_{t+1} = w_f^T (w_t + \eta_t \cdot y_{n(t)} x_{n(t)}) \geq w_f^T w_t + \eta_t \cdot \min_n y_{n(t)} w_f^T x_{n(t)} = w_f^T w_t + \eta_t \cdot \min_n |w_f^T x_{n(t)}|$

$= w_f^T w_t + \eta_t \cdot \hat{\rho} \cdot \|w_f\| \|x_n\| = w_f^T w_t + \hat{\rho} \cdot \|w_f\|$

(b) 證 $w_f^T w_T \geq T \cdot \hat{\rho} \cdot \|w_f\|$：

By (a)，assume $w_0 = 0$，可以得到：

$w_f^T w_0 = 0$

$w_f^T w_1 \geq w_f^T w_0 + \hat{\rho} \cdot \|w_f\|$

$w_f^T w_2 \geq w_f^T w_1 + \hat{\rho} \cdot \|w_f\|$

$\quad \ldots \quad \ldots$

$w_f^T w_T \geq w_f^T w_{T-1} + \hat{\rho} \cdot \|w_f\|$

把這些不等式加總，就會得到 $w_f^T w_T \geq T \cdot \hat{\rho} \cdot \|w_f\|$

(c) 證 $\|w_{t+1}\|^2 \leq \|w_t\|^2 + 1$：

By 題目， $w_{t+1} = w_t + \eta_t \cdot y_{n(t)} x_{n(t)}$，則：

$\|w_{t+1}\|^2 = \|w_t + \eta_t \cdot y_{n(t)} x_{n(t)}\|^2 = \|w_t\|^2 + 2\eta_t \cdot y_{n(t)} w_t^T x_{n(t)} + \|\eta_t \cdot y_{n(t)} x_{n(t)}\|^2$

$\leq \|w_t\|^2 + 0 + \|\eta_t \cdot y_{n(t)} x_{n(t)}\|^2 = \|w_t\|^2 + (\frac{1}{\|x_{n(t)}\|} \|y_{n(t)} x_{n(t)}\|)^2 = \|w_t\|^2 + 1$

(d) 證 $\|w_T\|^2 \leq T$：

By (c)，assume $w_0 = 0$，可以得到：

$\|w_0\|^2 = 0$

$\|w_1\|^2 \leq \|w_0\|^2 + 1$

$\|w_2\|^2 \leq \|w_1\|^2 + 1$

$\quad \ldots \quad \ldots$

$\|w_T\|^2 \leq \|w_{T-1}\|^2 + 1$

把這些不等式加總，就會得到 $\|w_T\|^2 \leq T$

(e) 證 $T \leq \hat{\rho}^{-2}$：

Divide (b) by (d)，$\dfrac{w_f^T w_T}{\|w_T\|} \geq \dfrac{T \cdot \hat{\rho} \cdot \|w_f\|}{\|w_T\|} \geq \dfrac{T \cdot \hat{\rho} \cdot \|w_f\|}{\sqrt{T}} = \sqrt{T} \cdot \hat{\rho} \cdot \|w_f\|$

又 $1 \geq cos\theta_T = \dfrac{w_f^T w_T}{\|w_f^T\| \|w_T\|} \geq \sqrt{T} \cdot \hat{\rho} \Rightarrow 1 \geq T \cdot \hat{\rho}^2$

所以 $T \leq \hat{\rho}^{-2}$

5. **(d)**

(a)(b) by 原本的 PAL，不一定會讓 $y_{n(t)}w_{t+1}^{T}x_{n(t)} > 0$

(c)

$$w_{t+1} \leftarrow w_t + y_{n(t)}x_{n(t)} \cdot (\frac{-y_{n(t)}w_t^T x_{n(t)}}{\|x_{n(t)}\|^2})$$

$$y_{n(t)}w_{t+1}^T x_{n(t)} = y_{n(t)}w_t^T x_{n(t)} + (y_{n(t)})^2\|x_{n(t)}\|^2 \cdot (\frac{-y_{n(t)}w_t^T x_{n(t)}}{\|x_{n(t)}\|^2})$$

$$= y_{n(t)}w_t^T x_{n(t)} + \|x_{n(t)}\|^2 \cdot (\frac{-y_{n(t)}w_t^T x_{n(t)}}{\|x_{n(t)}\|^2})$$

$$= 0$$

(d)

$$w_{t+1} \leftarrow w_t + y_{n(t)}x_{n(t)} \cdot \lfloor \frac{-y_{n(t)}w_t^T x_{n(t)}}{\|x_{n(t)}\|^2} + 1 \rfloor$$

$$y_{n(t)}w_{t+1}^T x_{n(t)} = y_{n(t)}w_t^T x_{n(t)} + (y_{n(t)})^2\|x_{n(t)}\|^2 \cdot \lfloor \frac{-y_{n(t)}w_t^T x_{n(t)}}{\|x_{n(t)}\|^2} + 1 \rfloor$$

$$= y_{n(t)}w_t^T x_{n(t)} + \|x_{n(t)}\|^2 \cdot \lfloor \frac{-y_{n(t)}w_t^T x_{n(t)}}{\|x_{n(t)}\|^2} + 1 \rfloor \text{（因為 } y_{n(t)} = \pm 1\text{）}$$

$$> y_{n(t)}w_t^T x_{n(t)} + \|x_{n(t)}\|^2 \cdot \frac{-y_{n(t)}w_t^T x_{n(t)}}{\|x_{n(t)}\|^2}$$

（因為 $\frac{-y_{n(t)}w_t^T x_{n(t)}}{\|x_{n(t)}\|^2} > 0$，所以高斯符號去掉後減一就會小於原來的數）

$$= y_{n(t)}w_t^T x_{n(t)} + (-y_{n(t)}w_t^T x_{n(t)}) = 0$$

(e)

$$w_{t+1} \leftarrow w_t - y_{n(t)}x_{n(t)} \cdot \lfloor \frac{-y_{n(t)}w_t^T x_{n(t)}}{\|x_{n(t)}\|^2} + 1 \rfloor$$

$$y_{n(t)}w_{t+1}^T x_{n(t)} = y_{n(t)}w_t^T x_{n(t)} - (y_{n(t)})^2\|x_{n(t)}\|^2 \cdot \lfloor \frac{-y_{n(t)}w_t^T x_{n(t)}}{\|x_{n(t)}\|^2} + 1 \rfloor$$

$$= y_{n(t)}w_t^T x_{n(t)} - \|x_{n(t)}\|^2 \cdot \lfloor \frac{-y_{n(t)}w_t^T x_{n(t)}}{\|x_{n(t)}\|^2} + 1 \rfloor \text{（因為 } y_{n(t)} = \pm 1\text{）}$$

$$= y_{n(t)}w_t^T x_{n(t)} - \|x_{n(t)}\|^2 \cdot \lfloor \frac{-y_{n(t)}w_t^T x_{n(t)}}{\|x_{n(t)}\|^2} \rfloor - \|x_{n(t)}\|^2 < 0$$

6. **(c)**

(a)(b) 承第三題的結果，可以知道加上 learning rate 後更新次數一樣會被 $\frac{R^2}{\rho^2}$ bound 住

(c) 承第五題的 (c) 選項，已證明 $y_{n(t)}w_{t+1}^T x_{n(t)} = 0$，則：

$w_{t+1} \leftarrow w_t + y_{n(t)}x_{n(t)} \cdot (\frac{-y_{n(t)}w_t^T x_{n(t)}}{\|x_{n(t)}\|^2}) = w_t + y_{n(t)}x_{n(t)} \cdot 0 = w_t$，代表 $w$ 不會更新且和 $x_n$ 的內積永遠是零。若定義 sign(0) 為 -1，那 $y_n = 1$的點永遠會錯；反之若定義 sign(0) 為 1，那 $y_n = -1$的點永遠會錯，因此 PLA 不會停止。

(d)(e) 已知 (d)(e) 的更新方向相反，但 (e) 會讓 $y_{n(t)}w_{t+1}^T x_{n(t)} < 0$，也就是說 (e) 更新的方向永遠不是正確的（會遠離正確的方向），因此 PLA 不會停止。而 (d) 會讓 $y_{n(t)}w_{t+1}^T x_{n(t)} > 0$，i.e. 修正後的 $w$ 就能正確預測 $x_n$，若 data 是 linear separable 的，那可以預期 PLA 最終會停止。

7. **(e)**
因為他是透過 judge system 來「懲罰不好的反應」並「獎勵好的反應」，藉此來學習正確的下棋方法（因為我們人類也不知道怎樣的下法才是正確的，只能透過結果的好壞來給予反饋）。此種反饋機制即為 reinforcement learning 的主要概念。
8. **(b)**
・semi-supervised learning：只有一筆資料有給 human record（資料標記），剩下的都是單純的路況影片（沒有標記）
・batch learning：他是一次餵給機器一整批的影片資料。而非一次喂一筆、逐步更新算法（online learning），也沒有讓機器問問題並給予反饋（active learning）
・raw features ：影片資料大多是 raw features，因為是輸入每個 frame 的 pixel 值，後續還要經過 detection、segmentation、tracking 等處理，而非有具體 physical meaning 的資料。
9. **(e)**
假設抽取前三筆資料（即：(1, 0)、(3, 2)、(0, 2)）作為 data，則：
$sign(0*x_0-1*x_1+2.5)$ 與 $sign(0*x_0-1*x_1+6)$ 兩個 perceptron 皆可使得 $E_{in}(g)$ 為 0，然而前者的 $E_{ots}(g)=0$ 而後者的 $E_{ots}(g)=1$，因此答案為 (e)。
10. **(b)**
・by 題目敘述，假設骰到正面的機率是 $\mu=\frac{1}{2}+\epsilon$。當 sample data 中的反面比正面多時代表壞事發生（即：$\nu<\frac{1}{2}\Rightarrow-\nu>\frac{-1}{2}$）。套用 Hoeffding's Inequality，計算 $|\mu-\nu|$：
$|\mu-\nu|=|\frac{1}{2}+\epsilon-\nu|>|\frac{1}{2}+\epsilon-\frac{1}{2}|=\epsilon$ （Hoeffding's Inequality 中的 $\epsilon$ 就是題目的 $\epsilon$）
・題目說好事發生的機率至少要 $1-\delta$，代表壞事發生的機率要 $<\delta$，即：
$P[|\mu-\nu|>\epsilon]\leq 2exp(-2\epsilon^2N)\leq\delta$
$\Rightarrow -2\epsilon^2N\leq log\frac{\delta}{2}$
$\Rightarrow N\geq\frac{1}{2\epsilon^2}log\frac{2}{\delta}$
11. **(c)**
因為 $f(x)=sign(x_1)$、$h_2(x)=sign(x_2)$，表示當 $x_1$ 和 $x_2$ 同號時，$f(x_n)=h_2(x_n)$
如果從 $[-1,+1]\times[-1,+1]$ 的實數區間內取值，$x_1$ 和 $x_2$ 有 1/2 的機率同號，sample 5 次皆同號才能使得 $E_{in}(h_2)=0$，機率為 $(1/2)^5=1/32$

12. **(d)**

如右圖，

藍色區域：$f(x_n) = h_1(x_n)$、$f(x_n) = h_2(x_n)$

綠色區域：$f(x_n) = h_1(x_n)$、$f(x_n) \neq h_2(x_n)$

紅色區域：$f(x_n) \neq h_1(x_n)$、$f(x_n) = h_2(x_n)$

sample 五次後 $E_{in}(h_2) = E_{in}(h_1)$ 出現的狀況有：

・五次都選在藍色：$(3/8)^5 = 243/32768$

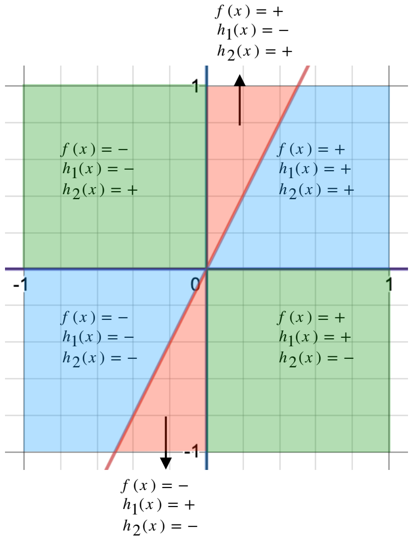
・一次紅色、一次綠色、三次藍色：

$(1/8) * (4/8) * (3/8)^3 * (5!/3!) = 2160/32768$

・兩次紅色、兩次綠色、一次藍色：

$(1/8)^2 * (4/8)^2 * (3/8) * (5!/(2! * 2!)) = 1440/32768$

三種情況加起來的機率即為：3843/32768




13. **(b)**

*以下的 i 範圍為 i=1, 2, ...d

by 題目定義，可得 $h_i(x) = sign(x_i) = -(h_{i+d}(x))$，也就是 $h_i(x)$ 和 $h_{i+d}(x)$ 預測的對錯會是相反的。即：

・$E_{out}(h_i) = 1 - E_{out}(h_{i+d})$

・$E_{in}(h_i) = 1 - E_{in}(h_{i+d})$

套入題目對 Bad D for H 的公式：

・若 $|E_{out}(h_i) - E_{in}(h_i)| > \epsilon$，則 $|E_{out}(h_{i+d}) - E_{in}(h_{i+d})| = |(1 - E_{out}(h_i)) - (1 - E_{in}(h_i))| > \epsilon$

・若 $|E_{out}(h_i) - E_{in}(h_i)| \leq \epsilon$，則 $|E_{out}(h_{i+d}) - E_{in}(h_{i+d})| = |(1 - E_{out}(h_i)) - (1 - E_{in}(h_i))| \leq \epsilon$

代表如果某筆 data 對 $h_i$ 是壞的，則對 $h_{i+d}$ 也是壞的；反之對 $h_i$ 是好的，則對 $h_{i+d}$ 也是好的。

因此產生 bad data 的機率一樣是被 $d \cdot 2exp(-2\epsilon^2 N)$ bound 住（by 課堂公式）

14. **(d)**

・five green 3's：B跟D有綠 3，$(\frac{2}{4})^5 = \frac{1}{32}$

(a) five green 1's：綠 1 不存在，機率為 0

(b) five orange 2's：C有橘 2，$(\frac{1}{4})^5 = \frac{1}{1024}$

(c) five green 2's：A 、B 、D有綠 2，$(\frac{3}{4})^5 = \frac{243}{1024}$

(d) five green 4's：A 、B有綠 4，$(\frac{2}{4})^5 = \frac{1}{32}$

(e) five green 5's：D有綠 5，$(\frac{1}{4})^5 = \frac{1}{1024}$

A：○○○○○○

B：○○○○○○

C：○○○○○○

D：○○○○○○

15. **(c)**

P（有些數字全綠）＝ P（至少 1 個數字全綠）- P（至少 2 個數字全綠）

＋P（至少 3 個數字全綠）- P（至少 4 個數字全綠）＋P（至少 5 個數字全綠）

・P（至少 1 個數字全綠）＝P（2 全綠）＋P（3 全綠）＋P（4 全綠）＋P（5 全綠）

＋P（6 全綠）＝ $(\frac{3}{4})^5+(\frac{2}{4})^5+(\frac{2}{4})^5+(\frac{1}{4})^5+(\frac{2}{4})^5+=(\frac{340}{1024})$

・P（至少 2 個數字全綠）＝P（23 全綠）＋P（24 全綠）＋P（25 全綠）＋P（26 全綠）＋P（34 全綠）＋P（35 全綠）＋P（46 全綠）$=(\frac{2}{4})^5+(\frac{2}{4})^5+(\frac{1}{4})^5+(\frac{1}{4})^5+(\frac{1}{4})^5+(\frac{1}{4})^5+(\frac{1}{4})^5=(\frac{69}{1024})$

・P（至少 3 個數字全綠）＝P（234 全綠）＋P（235 全綠）＋P（246 全綠）

$=(\frac{1}{4})^5+(\frac{1}{4})^5+(\frac{1}{4})^5=\frac{3}{1024}$

・P（至少 4 個數字全綠）＝0

・P（至少 5 個數字全綠）＝0

$\frac{340}{1024}-\frac{69}{1024}+\frac{3}{1024}=\frac{274}{1024}$

16. **(b)**
17. **(b)**
18. **(c)**
19. **(d)**
20. **(d)**