CSIE 5432 — Machine Learning Foundations

Name: 李吉昌 **Homework 1**
Student Number: r08922a27 **Due Date:** October 16 2020, 13:30

The Learning Problem

1. Answer: [d]

[a] 純粹隨機問題不存在 pattern, 也就沒辦法學習。

[b] 已經確定可以找到正確答案也就沒使用學習的需求。

[c] 已經確定可以找到正確答案也就沒使用學習的需求。

[d] 可以使用regression去預測芒果的分數(品質)。

[e] none of the other choices

2. Answer: [e]

[a] 判斷的方式並沒有基於任何資料的性質, 純粹抽籤不能算學習。

[b] 這是工人智慧, 不是學習。

[c] 已經確定找到明確 rule 去達成這個目的, 不需要使用學習。

[d] 已經確定找到明確 rule 去達成這個目的, 不需要使用學習。

[e] 這段敘述可以 formulate 成一個 regression 或是 classification 的問題, 可以投入學習方法。

Perceptron Learning Algorithm

3. Answer: [d]

令 scaling factor 為 α, 根據講義定理 $R^2 = \alpha^2 \cdot \max_n \|x_n\|^2$, $\rho = \alpha \cdot \min_n y_n \frac{w_f^T x_n}{\|w_f\|}$, PLA 需要疊代的次數 T 的上限是 $(\cancel{\alpha^2} \cdot \max_n \|x_n\|^2)/(\cancel{\alpha^2} \cdot (\min_n y_n \frac{w_f^T x_n}{\|w_f\|})^2)$, 分子和分母的係數會消掉, 得到的上限是一樣的。

4. Answer: [c]

Stage 1:

$$
\begin{aligned}
w_f^T w_t &= w_f^T w_{t-1} + \frac{1}{\|x_{n(t-1)}\|} y_{n(t-1)} w_f^T x_{n(t-1)}, \text{ 乘上 } \frac{\|w_f\|}{\|w_f\|}, \\
&= w_f^T w_{t-1} + y_{n(t-1)} \|w_f\| \cdot \frac{w_f^T x_n}{\|w_f\| \|x_n\|}, \\
&\geq w_f^T w_{t-1} + y_{n(t-1)} \|w_f\| \cdot \min_n \frac{w_f^T x_n}{\|w_f\| \|x_n\|}, \text{ 代入 } \hat{\rho}, \text{ 並由數學歸納法可得下面結果}, \\
&\geq w_f^T w_{t-1} + \|w_f\| \hat{\rho}, \geq w_f^T w_{t-2} + 2\|w_f\| \hat{\rho}, \geq w_f^T w_{t-3} + 3\|w_f\| \hat{\rho}, \ldots \geq w_f^T w_0 + t\|w_f\| \hat{\rho}, \\
&\text{因設初始化 } w_0 \text{ 為零向量, 可得 } w_f^T w_t \geq t\|w_f\| \hat{\rho}
\end{aligned} \tag{1}
$$

Stage 2:

$$\begin{aligned}
\|w_t\|^2 &= \|w_{t-1}\|^2 + 2\frac{1}{\|x_{n(t-1)}\|} y_{n(t-1)} w_{t-1}^T x_{n(t-1)} + 1, \\
&\text{因為 } 2\frac{1}{\|x_{n(t-1)}\|} y_{n(t-1)} w_{t-1}^T x_{n(t-1)} \leq 0, \text{ ,並由數學歸納法可得到下面結果,} \\
&\leq \|w_{t-1}\|^2 + 1, \leq \|w_{t-2}\|^2 + 2, \leq \|w_{t-3}\|^2 + 3, ... \leq \|w_0\|^2 + t \\
&\text{因設初始化 } w_0 \text{ 為零向量, 可得 } \|w_t\|^2 \leq t
\end{aligned} \tag{2}$$

綜合 Stage 1 和 Stage 2:

$$1 \geq \frac{w_f^T w_T}{\|w_f\|\|w_T\|} \geq \sqrt{T}\hat{\rho}, \text{ 移項後得 } T \quad \leq \frac{1}{\hat{\rho}^2} \tag{3}$$

5. Answer: [d]

$$\begin{aligned}
w_{t+1} &= w_t + y_{n(t)} x_{n(t)} \lfloor \frac{-y_{n(t)} w_t^T x_{n(t)}}{\|x_{n(t)}\|^2} + 1 \rfloor, \\
y_{n(t)} w_{t+1}^T x_{n(t)} &= y_{n(t)} w_t^T x_{n(t)} + \|x_{n(t)}\|^2 \cdot \lfloor \frac{-y_{n(t)} w_t^T x_{n(t)}}{\|x_{n(t)}\|^2} + 1 \rfloor, \\
&> y_{n(t)} w_t^T x_{n(t)} + \|x_{n(t)}\|^2 \cdot \frac{-y_{n(t)} w_t^T x_{n(t)}}{\|x_{n(t)}\|^2} = 0, \\
&\text{得 } y_{n(t)} w_{t+1}^T x_{n(t)} > 0
\end{aligned} \tag{4}$$

6. Answer: [c]

[a], [b] 可以看成對所有 x_n 做 scaling, 對疊代次數上限不影響, [c] $y_{n(t)} w_{t+1}^T x_{n(t)} = 0$ 不會停, [e] 在初始化為零向量的情狀下, PLA 會停下來的條件為找到一組 W_T 能夠全部答對, 但答錯的部分只會持續是錯的, 所以 PLA 不會停下來。 [d] 證明如下, 令 $M_t = \frac{-y_{n(t)} w_t^T x_{n(t)}}{\|x_{n(t)}\|^2}$:

Stage 1:

$$\begin{aligned}
w_f^T w_T &= w_f^T w_{T-1} + y_{n(T-1)} w_f^T x_{n(T-1)} \cdot \lfloor 1 + M_{T-1} \rfloor, \\
&\geq w_f^T w_{T-1} + (\min_n y_n w_f^T x_n) \cdot (1 + M_{T-1}), \text{ 後項皆為正, 得 } w_f^T w_T > w_f^T w_{T-1}, \\
&\text{令 } (\min_n y_n w_f^T x_n) \text{ 正號常數為 } c_1, \text{ 由數學歸納法可得到下面結果,} \\
&\geq w_f^T w_{T-1} + c_1(1 + M_{T-1}) \geq w_f^T w_{T-2} + c_1(2 + M_{T-1} + M_{T-2}), ... \\
&\geq w_f^T w_0 + c_1 \cdot \sum_{i=1}^{T} M_{T-i} + c_1 \cdot T, \\
&\text{因設初始化 } w_0 \text{ 為零向量, 且 } M_{T-i} \text{ 恆大於 0 , 得 } w_f^T w_T \geq c_1 \cdot T,
\end{aligned} \tag{5}$$

Stage 2:

$$
\begin{aligned}
\|w_T\|^2 &\leq \|w_{T-1}\|^2 + 2(1+M_{T-1})\cdot y_{n(T-1)} w_{T-1}^T x_{n(T-1)} + \|(1+M_{T-1})\cdot y_{n(T-1)} x_{n(T-1)}\|^2, \\
&= \|w_{T-1}\|^2 + 2y_{n(T-1)} w_{T-1}^T x_{n(T-1)} - 2\cdot(\frac{w_{T-1}^T x_{n(T-1)}}{\|x_{n(T-1)}\|})^2 + \|x_{n(T-1)}\|^2 \\
&- 2y_{n(T-1)} w_{T-1}^T x_{n(T-1)} + (\frac{w_{T-1}^T x_{n(T-1)}}{\|x_{n(T-1)}\|})^2, \text{ 經化減可得下面結果,} \\
&= \|w_{T-1}\|^2 - (\frac{w_{T-1}^T x_{n(T-1)}}{\|x_{n(T-1)}\|})^2 + \|x_{n(T-1)}\|^2, \text{ 因為 } (\frac{w_{T-1}^T x_{n(T-1)}}{\|x_{n(T-1)}\|})^2 > 0, \text{ 可得下面結果,} \\
&\leq \|w_{T-1}\|^2 + \|x_{n(T-1)}\|^2 \leq \|w_{T-1}\|^2 + \max_n \|x_n\|^2, \text{ 得 } \|w_T\|^2 \text{ 成長被 } \max_n \|x_n\|^2 \text{ 限制,} \\
&\text{令正號常數 } \max_n \|x_n\|^2 \text{ 為 } c_2, \text{ 由數學歸納法, 可得下面結果,} \\
&\leq \|w_{T-1}\|^2 + c_2 \leq \|w_{T-2}\|^2 + 2c_2 \leq \|w_{T-3}\|^2 + 3c_2, ... \leq \|w_0\|^2 + c_2\cdot T, \\
&\text{因設初始化 } w_0 \text{ 為零向量, 可得 } \|w_T\|^2 \leq c_2 \cdot T
\end{aligned} \tag{6}
$$

綜合 Stage 1 和 Stage 2:

$$
\begin{aligned}
1 &\geq \frac{w_f^T w_T}{\|w_f\|\|w_T\|} \geq \frac{c_1\cdot T}{\|w_f\|\sqrt{c_2\cdot T}} = (\frac{c_1}{\|w_f\|\sqrt{c_2}})\sqrt{T}, \\
1 &\geq (\frac{c_1}{\|w_f\|\sqrt{c_2}})^2\cdot T \text{ , 移項後得 } T \leq (\frac{\|w_f\|^2 c_2}{{c_1}^2})
\end{aligned} \tag{7}
$$

在[a], [b] 和 [d] 三個選項的參數更新次數 T 的上限會被限制在一常數上, 所以訓練資料在 linear separable 的時候, 一定能夠在有限次數找到 perfect line。

Types of Learning

7. Answer: [e]

雖然和 unsupervised learning 一樣, 沒有使用明確的 label 資料, 但敘述中提到的 judge environment 是一個 reward 的提供者, 環境給予的回饋仍然能使模型得到和期望表現正相關的 gradient。

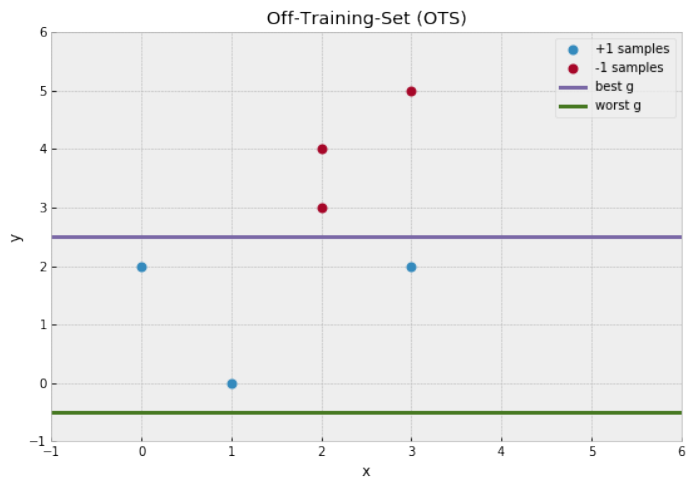
8. Answer: [b]

敘述中的 view 是一般 image, 為 raw feature; sequence to sequence 的輸入輸出方式為 structure learning, 訓練時共餵進兩次成批的資料, 為 batch learning; 第二次餵進 learner 的訓練資料沒有 human record, 為 semi-supervised learning。

Off-Training-Set Error

9. Answer: [e]

從下圖可以看到, sample 很明顯是線性可分的, 如果我抽到的 3 個是 $(0,2),(3,2),(2,3)$, 我一定能找到最好的 g 將 unseen sample 全部答對; 如果我抽到的是 $(0,2),(3,2),(1,0)$, 我一定能找到最壞的 g 將 unseen sample 全部答錯。



Hoeffding Inequality

10. Answer: [b]

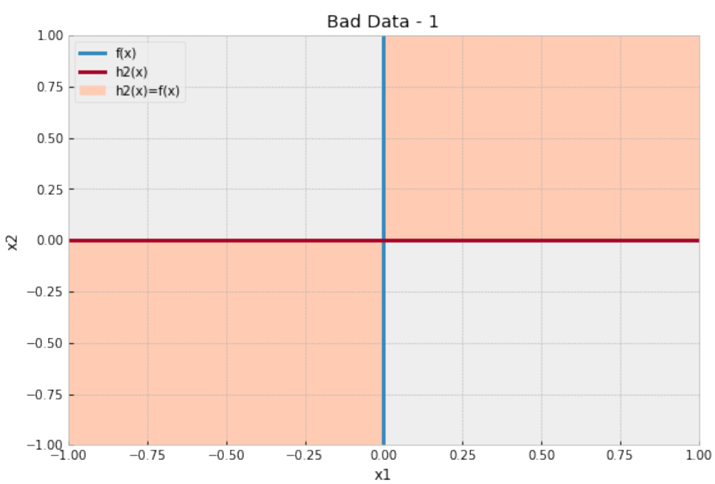
令 $\mu(=\frac{1}{2}+\epsilon)$ 為出現 probable side 的機率, 令 ν 為抽到 probable side 的 fraction, 由 Hoeffding Inequality 得 $\mathbb{P}[|\mu-\nu|>\epsilon] \le 2exp(-2\epsilon^2 N)$, 發現 probable side 的事件為 $\nu > \frac{1}{2}$, 依下列不等式可得, 當 N 達條件數量以上, $\mathbb{P}[|\mu-\nu|>\epsilon]$ 小於 δ 時, 則 $\mathbb{P}[\frac{1}{2}>\nu]$ 也必小於 δ, $\mathbb{P}[\frac{1}{2}<\nu]$ 必大於 $1-\delta$:

$$\begin{aligned}&1-(1-\delta) = 2exp(-2\epsilon^2 N) \ge \mathbb{P}[|\mu-\nu|>\epsilon] = \mathbb{P}[\frac{1}{2}>\nu] + \mathbb{P}[\frac{1}{2}+2\epsilon<\nu] \ge \mathbb{P}[\frac{1}{2}>\nu],\\ &1-\mathbb{P}[\frac{1}{2}>\nu] = \mathbb{P}[\nu \ge \frac{1}{2}] = \mathbb{P}[\nu > \frac{1}{2}] \ge 1-2exp(-2\epsilon^2 N) = 1-\delta \text{ , 移項整理得 } N = \frac{1}{2\epsilon^2}\log(\frac{2}{\delta})\end{aligned} \tag{8}$$

Bad Data

11. Answer: [c]

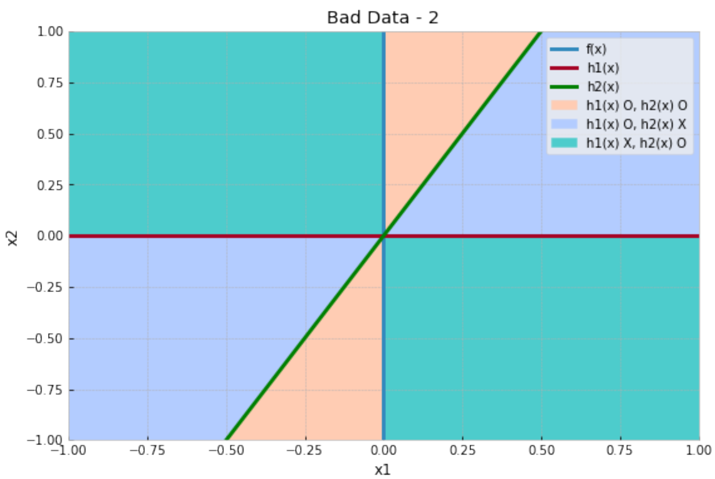
從下圖可以看到, 抽樣來自平均分布, 每次發生 $E_{in}(h_2)=0$ 的機率為 $\frac{1}{2}$, 即 $E_{in}(h_2)=0$ 發生的部分占總面積的比例, 獨立地抽樣 5 次的機率會是 $(\frac{1}{2})^5=\frac{1}{32}$ 。



12. Answer: [d]

$E_{in}(h1) = E_{in}(h2)$ 會發生三種可能, 分別是 $E_{in}(h1) = E_{in}(h2) = 0$, $E_{in}(h1) = E_{in}(h2) = 0.2$, $E_{in}(h1) = E_{in}(h2) = 0.4$。 從下圖可以看到, $h1$ 和 $h2$ 同時答錯的情況並不會發生, 有可能發生的情況分別是全對或是兩種一對一錯的情況, 因為抽樣為平均分布, 各色塊占總面積比例即為各情況的機率。

$$\begin{aligned}&\mathbb{P}[E_{in}(h1) = E_{in}(h2)]\\&= \mathbb{P}[E_{in}(h1) = E_{in}(h2) = 0] + \mathbb{P}[E_{in}(h1) = E_{in}(h2) = 0.2] + \mathbb{P}[E_{in}(h1) = E_{in}(h2) = 0.4]\\&= (\frac{3}{8})^5 + (\frac{3}{8})^3(\frac{1}{8})(\frac{4}{8})C_2^5C_1^2 + (\frac{3}{8})(\frac{1}{8})^2(\frac{4}{8})^2C_4^5C_2^4 = \frac{3843}{32768}\end{aligned} \tag{9}$$



13. Answer: [b]

令 i 為 1 至 d 的任一 $index$, 由題意 $h_i = -h_{i+d}$, 可得 $E_{in}(h_i) = 1 - E_{in}(h_{i+d})$ 和 $E_{out}(h_i) = 1 - E_{out}(h_{i+d})$, 綜上兩式取絕對值可得 $|E_{in}(h_i) - E_{out}(h_i)| = |E_{in}(h_{i+d}) - E_{out}(h_{i+d})|$ 。 上述可知, 發生 BAD $\mathcal{D}$ for $h_i(|E_{in}(h_i) - E_{out}(h_i)| > \epsilon)$ 的事件必然會發生 BAD $\mathcal{D}$ for $h_{i+d}(|E_{in}(h_{i+d}) - E_{out}(h_{i+d})| > \epsilon)$, 因此 $2d$ 個事件的聯集機率 $\mathbb{P}$[(BAD $\mathcal{D}$ for h_1) or (BAD $\mathcal{D}$ for h_2) or ... or (BAD $\mathcal{D}$ for h_{2d})] 可以化減成 d 個事件的聯集機率 $\mathbb{P}$[(BAD $\mathcal{D}$ for h_1) or (BAD $\mathcal{D}$ for h_2) or ... or (BAD $\mathcal{D}$ for h_d)] , 上限為 $\sum_{i=1}^{d} \mathbb{P}$[(BAD $\mathcal{D}$ for h_i) $\leq 2d \cdot exp(-2\epsilon^2 N)$, 得 $C = d$。

Multiple-Bin Sampling

14. Answer: [d]

抽到綠色 3 號有 B 和 D 共 2 種選擇, 機率是 $(\frac{2}{4})^5$; [a] 的選擇有 0 種, 機率是 0 ; [b] 的選擇有 C 共 1 種, 機率是 $(\frac{1}{4})^5$; [c] 的選擇有 A, B 和 D 共 3 種, 機率是 $(\frac{3}{4})^5$; [d] 的選擇有 A 和 B 共 2 種, 機率是 $(\frac{2}{4})^5$; [e] 的選擇有 D 共 1 種, 機率是 $(\frac{1}{4})^5$。 [d] 的選擇和綠色 3 號的機率相同。

15. Answer: [c]

骰 5 次, 每次有 A, B, C, D 四種可能, 總共 $4^5 = 1024$ 種可能性。題目敘述的條件可分為全部綠色且號碼是 1 至 6 號六種可能的結果來討論, 全綠且 1 號的選擇為空集合, 全綠且 2 號的選擇為 A, B 和 D, 全綠且 3 號的選擇為 B 和 D, 全綠且 4 號的選擇為 B 和 D, 全綠且 5 號的選擇為 A 和 B, 全綠且 5 號的選擇為 D, 全綠且 6 號的選擇為 A, C。 可以注意到其實全綠且 3, 4 和 5 號的可能結果其實包含於全綠且 2 號當中, 所以可以直接不計, 我們只需要計算全綠 2 號和 6 號的結果即可。總共 $3^5 + (2^5 - 1) = 274$ 種可能, 機率為 $\frac{274}{1024}$。