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]

根據課程講義的定理 $R^2 = \max_n \|x_n\|^2$, $\rho = \min_n y_n \frac{\|w_f^T\|}{\|w_f\|} x_n$, PLA 需要疊代的次數 T 的上限是 R^2/ρ^2, 分子和分母的 scaling 的係數會消掉, 因此 scaling 上限是一樣的。

4. Answer: [c]

Stage 1:

$$
\begin{aligned}
w_f^T w_t &= w_f^T w_{t-1} + \frac{1}{\|x_{n(t-1)}\|} y_{n(t-1)} w_f^T x_{n(t-1)} \frac{\|w_f\|}{\|w_f\|}, \\
&\geq w_f^T w_{t-1} + y_{n(t-1)} \|w_f\| \min_n \frac{w_f^T x_n}{\|w_f\| \|x_n\|}, \\
&\geq w_f^T w_{t-1} + \|w_f\| \hat{\rho}, \\
&\geq w_f^T w_{t-2} + 2\|w_f\| \hat{\rho}, \text{, 由數學歸納法, 遞迴代入 } w_{t-3}, w_{t-4}, ..., \text{ 可得 } w_f^T w_t \geq t \|w_f\| \hat{\rho}
\end{aligned} \tag{1}
$$

Stage 2:

$$\begin{aligned}
\|w_t\|^2 &= \|w_{t-1}\|^2 + 2\frac{1}{\|x_{n(t-1)}\|} y_{n(t-1)} w_{t-1}^T x_{n(t-1)} + 1, \\
&\text{因為 } 2\frac{1}{\|x_{n(t-1)}\|} y_{n(t-1)} w_{t-1}^T x_{n(t-1)} \le 0, \text{ 可得到下面結果,} \\
&\le \|w_{t-1}\|^2 + 1, \\
&\le \|w_{t-2}\|^2 + 2, \text{ 由數學歸納法, 遞迴代入 } w_{t-3}, w_{t-4}, ..., \text{ 可得 } \|w_t\|^2 \le t
\end{aligned} \tag{2}$$

綜合 Stage 1 和 Stage 2:

$$1 \ge \frac{w_f^T w_T}{\|w_f\|\|w_T\|} \ge \sqrt{T}\hat{\rho}, \text{ 得 } T \quad \le \frac{1}{\hat{\rho}^2} \tag{3}$$

5. Answer: [d]

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

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] 參數不會往答對的方向更新, 錯的會持續答錯。 [d] 證明如下, 令 $M_t = \frac{-y_{n(t)} w_t^T x_{n(t)}}{\|x_{n(t)}\|^2}$:

Stage 1:

$$\begin{aligned}
w_f^T w_T &= w_f^T w_{T-1} + y_{n(T-1)} w_f^T x_{n(T-1)} \cdot \lfloor 1 + M_{T-1} \rfloor, \\
&\ge w_f^T w_{T-1} + (\min_n y_{n(T-1)} w_f^T x_{n(T-1)}) \cdot (1 + M_{T-1}), \text{ 令 } (\min_n y_{n(T-1)} w_f^T x_{n(T-1)}) \text{ 常數為 } c_1, \\
&\ge [w_f^T w_{T-2} + c_1(1 + M_{T-2})] + c_1(1 + M_{T-1}), \text{ 由數學歸納法可得到下面結果,} \\
&\ge c_1 \cdot \sum_{i=1}^{T} M_i + c_1 \cdot T, \text{ 因為 } M_i > 0 \text{ , 得 } w_f^T w_T \ge c_1 \cdot T
\end{aligned} \tag{5}$$

Stage 2:

$$\begin{aligned}
\|w_T\|^2 &\le \|w_{T-1}\|^2 + 2(1 + M_{T-1}) \cdot y_{n(T-1)} w_{T-1}^T x_{n(T-1)} + \|(1 + M_{T-1}) \cdot y_{n(T-1)} x_{n(T-1)}\|^2, \\
&= \|w_{T-1}\|^2 + 2y_{n(T-1)} w_{T-1}^T x_{n(T-1)} - 2 \cdot (\frac{w_{T-1}^T x_{n(T-1)}}{\|x_{n(T-1)}\|})^2 + \|x_{n(T-1)}\|^2 \\
&- 2y_{n(T-1)} w_{T-1}^T x_{n(T-1)} + (\frac{w_{T-1}^T x_{n(T-1)}}{\|x_{n(T-1)}\|})^2, \\
&= \|w_{T-1}\|^2 - (\frac{w_{T-1}^T x_{n(T-1)}}{\|x_{n(T-1)}\|})^2 + \|x_{n(T-1)}\|^2, \\
&\le \|w_{T-1}\|^2 + \|x_{n(T-1)}\|^2, \\
&\le \|w_{T-1}\|^2 + \max_n \|x_{n(T-1)}\|^2, \text{ 令常數 } \max_n \|x_{n(T-1)}\|^2 \text{ 為 } c_2, \\
&\le \|w_{T-2}\|^2 + 2 \cdot c_2, \text{ 由數學歸納法, 得 } \|w_T\|^2 \le c_2 \cdot T
\end{aligned} \tag{6}$$

綜合 Stage 1 和 Stage 2:

$$1 \geq \frac{w_f^T w_T}{\|w_f\|\|w_T\|} \geq \frac{c_1 \cdot T}{\|w_f\|\sqrt{c_2 \cdot T}} = (\frac{c_1}{\|w_f\|\sqrt{c_2}})\sqrt{T}\ ,\ 1 \geq (\frac{c_1}{\|w_f\|\sqrt{c_2}})^2 \cdot T\ ,\ 得\ T \leq (\frac{\|w_f\|^2 c_2}{{c_1}^2}) \tag{7}$$

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]

敘述中有提到 judge environment, 明確是一個能夠給予 reward 的環境。

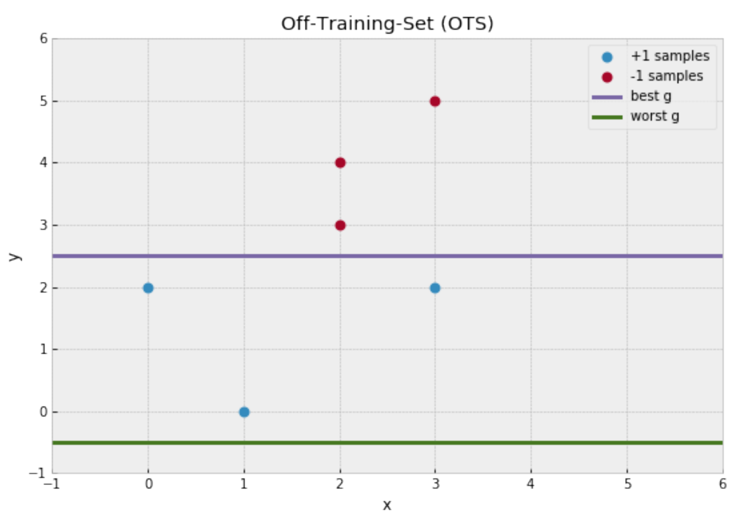
8. Answer: [b]

敘述中的 view 為 raw feature, sequence to sequence 的輸入輸出方式為 structure learning, 訓練資料中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, 我一定能找到最好跟最壞的 g, 能夠全部答對跟答錯。



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 $\mathbb{P}[|\mu-\nu|>\epsilon] \leq 2exp(-2\epsilon^2 N)$, 發現 probable side 的事件為 $\nu > \frac{1}{2}$, 而 $|\frac{1}{2}+\epsilon-\nu|>\epsilon$ 為

$\frac{1}{2} > \nu$ 和 $\frac{1}{2} + 2\epsilon < \nu$ 兩互斥事件的聯集, 如果發生 $\frac{1}{2} + 2\epsilon < \nu$ 則必發生, $\frac{1}{2} < \nu$, 因此, 找出 N 發生 $\frac{1}{2} + 2\epsilon < \nu$ 的條件必為 $\frac{1}{2} < \nu$ 的條件。

$$\begin{aligned} 2exp(-2\epsilon^2 N) &\geq \mathbb{P}[|\mu - \nu| > \epsilon] \\ &= \mathbb{P}[\frac{1}{2} + 2\epsilon] + \mathbb{P}[\frac{1}{2} + 2\epsilon < \nu] \\ &\geq \mathbb{P}[\frac{1}{2} + 2\epsilon < \nu] \end{aligned} \tag{8}$$

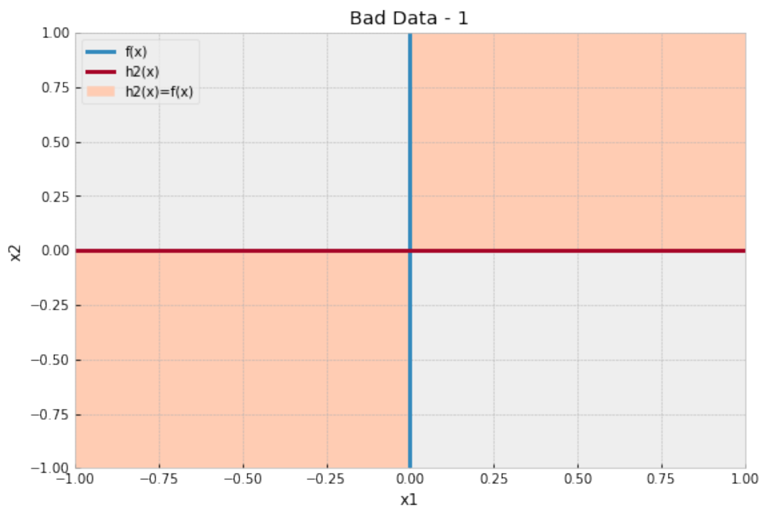
令 $\mathbb{P}[\frac{1}{2} + 2\epsilon < \nu]$ 為 δ, 則:

$$\delta \leq 2exp(-2\epsilon^2 N)\ ,\ \log(\frac{\delta}{2}) \leq -2\epsilon^2 N,\ 得\ N \leq \frac{1}{2\epsilon^2}\log(\frac{2}{\delta}) \tag{9}$$

Bad Data

11. Answer: [c]

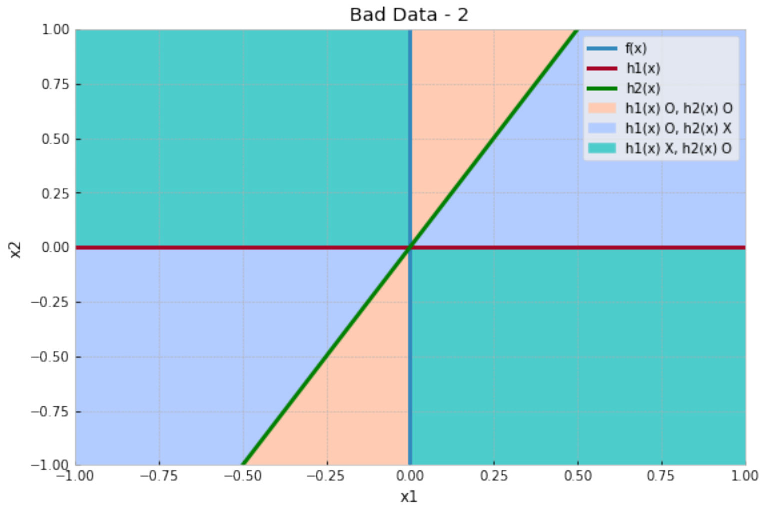
從下圖可以看到, $E_{in}(h_2) = 0$ 的部分占總面積的 $\frac{1}{2}$, 共獨立 sample 5 次, 機率是 $(\frac{1}{2})^5 = \frac{1}{32}$。



12. Answer: [d]

$E_{in}(h1) = E_{in}(h2)$ 會發生三種可能, 分別是 $E_{in}(h1) = E_{in}(h2) = 0$, $E_{in}(h1) = E_{in}(h2) = 1$, $E_{in}(h1) = E_{in}(h2) = 2$。 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, 各色塊占總面積比例即為各情況的機率。

$$\begin{aligned} &\mathbb{P}[E_{in}(h1) = E_{in}(h2)] \\ &= \mathbb{P}[E_{in}(h1) = E_{in}(h2) = 0] + \mathbb{P}[E_{in}(h1) = E_{in}(h2) = 1] + \mathbb{P}[E_{in}(h1) = E_{in}(h2) = 2] \\ &= (\frac{3}{8})^5 + (\frac{3}{8})^3(\frac{1}{8})(\frac{4}{8})C^5_2C^2_1 + (\frac{3}{8})(\frac{1}{8})^2(\frac{4}{8})^2C^5_4C^4_2 = \frac{3843}{32768} \end{aligned} \tag{10}$$



13. Answer: [b]

因為 $d+1$ 至 $2d$ 能分到的群體和 1 至 d 是一樣的, 差別只在標籤正負號, 且 1 至 d 的 hypothesis 至多能得出 d 種分法, 得 $C=d$。

Multiple-Bin Sampling

14. Answer: [d]

抽到綠色 3 號有 B 和 D 共 2 種選擇, [a] 的選擇有 0 種, [b] 的選擇有 C 共 1 種, [c] 的選擇有 A, B 和 D 共 3 種, [d] 的選擇有 A 和 B 共 2 種, [e] 的選擇有 D 共 1 種。 [d] 的選擇和綠色 3 號一樣, 根據題目敘述, 骰子選擇一樣則機率也會一樣。

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 $\frac{274}{1024}$。