- Learning Problems

1. 自我監督式學習舉例：

   如同老師上課提到的應用，在自然語言處理中自我監督學習幾乎是第一個階段；會先在目前所擁有的資料文本中挖空格，讓機器去根據上下文學會某個空格應該要填入甚麼單詞；或是去學會如果出現某個句子，之後應該要生成出甚麼詞或句子。經由這樣的「熱身動作」，機器好像對這些文本資料有了某種「背景知識」，從這樣的訓練中學會了詞與句子之間的某種關聯，讓機器在做下一階段的訓練時可以有更好的表現。
2. ML for shortest path of a maze
   我認為如果是對於那種有著固定牆壁和終點的迷宮，可以用已知的演算法找出最佳路徑，像是使用廣度優先搜尋可以找到最短的路徑；或者有可能路徑具有權重，可能可以使用 Dijkstra's algorithm 來找到抵達終點的最短路徑，不需要使用機器學習中的技巧，像是 ChatGPT 在題目中所提到的強化學習 Q-Learning 或是深度強化學習 Deep Q-Networks 來去學會找到最佳路徑。但如果是像移動迷宮或其他奇奇怪怪的迷宮，可能牆壁依照某種人類很難觀察出來的規則移動，或是地板以某種神奇難以推斷出的規則產生斷路陷阱無法前進，可能真的會如同 ChatGPT 所說的，機器學習可能可以帶給我們一些驚喜。
3. Machine Learning speed up any off-the-shelf algorithm
   我覺得 ChatGPT 會做出這樣的回答，除了因為他的資料是基於 2021 年 9 月之前，所以不會知道最近發生的事情之外，另一個原因是因為他所獲得的知識根據大量的文本而來的，也就是說如果某個知識的文本量越多，他就學的越精確；相對的，對於文本量少的知識，他對該知識的理解就可能非常少或不正確。Google DeepMind 所找到的新的更好的算法，除了時間是在近期這個因素之外，我覺得另一個原因是 ChatGPT 根據大文本量所學到的知識是在高階層語言的層面，而不是在組合語言的層面。

- Perceptron Learning Algorithm

4. 根據 PLA 的更新公式：

   $$\mathbf{w}_{t+1} = \mathbf{w}_t + y_{n(t)}\mathbf{x}_{n(t)}$$

   由於 w_0 一開始是 0，而經過了 T_+ 跟 T_- 次分別發生在 $y_{n(t)} = 1$ 跟 $y_{n(t)} = -1$ 的修正後，由於 x_0 都定為 1，所以可以根據更新公式得知：

   $$w_0 = T_+ \times \underbrace{1}_{y} \times \underbrace{1}_{x_0} + T_- \times \underbrace{(-1)}_{\mathrm{y}} \times \underbrace{1}_{x_0} = T_+ - T_-$$

5. 首先根據老師的講義中最後推導出的關係式：

$$1 \geq \frac{\mathbf{w}_f^T}{\|\mathbf{w}_f\|}\frac{\mathbf{w}_T}{\|\mathbf{w}_T\|} \geq \sqrt{T}\frac{\rho}{R}$$

$$\Rightarrow T \leq \left(\frac{R}{\rho}\right)^2 = \left(\frac{\underset{n}{max}\|\mathbf{x}_n\|}{\underset{n}{min}\frac{y_n\mathbf{w}_f^T\mathbf{x}_n}{\|\mathbf{w}_f\|}}\right)^2 = \frac{\left(\underset{n}{max}\|\mathbf{x}_n\|\right)^2}{\left(\underset{n}{min}\frac{\mathbf{w}_f^T\mathbf{x}_n}{\|\mathbf{w}_f\|}\right)^2}$$

其中 $\rho = \min\limits_{\mathrm{n}} \frac{y_n\mathbf{w}_f^T\mathbf{x}_n}{\|\mathbf{w}_f\|}$，$R = \max\limits_{\mathrm{n}}\|\mathbf{x}_n\|$， T 是修正次數。

題目有說明只有當垃圾字(spam-like words)的出現次數大於正常字(less spam-like words)的時候，才會判斷是垃圾郵件，讓輸出 $y = 1$，而且透過 -0.5 這個閾值項，當垃圾字跟正常字的數量一樣時，sign 函數裡面的值不會是 0 而是-0.5 ；並且從那個上天知道的神祕 f 當中，我們可以發現他給垃圾字的權重都是 1 ，正常字的權重都是-1 。

$$f(x) = \mathrm{sign}(z_+(x) - z_-(x) - 0.5)$$

下面我們來嘗試建構出可以完美分類的權重：

1. 首先我們可以發現 w_0 不可以為 0。如果為 0 的話對於完全沒有字出現的郵件，$\mathbf{w}\mathbf{x}_n$ 算出來會是 0，代表該點落在該超平面上；但是題目有說我們可以完美的分開這些資料，所以 w_0 不可以為 0。
2. 令所有垃圾字的權重都是 k ，所有正常字的權重都是 $-k$，k 是一個大於 1 的正整數。
   - 也就是說權重的絕對值要一樣，因為我們要計算的是字出現的「個數」，每種字的影響力是一樣的。
   - $k > 1$ 是因為 w_0 這個閾值項的權重只能是負整數，跟神祕 f 的 -0.5 不一樣。
3. w_0 是一個大於 $-k$ 的負整數
   - 因為要讓沒有任何字、或是垃圾字數量等於正常字的時候 sign 函數裡面要小於 0，不會判定是垃圾郵件。
   - 同時若郵件中就只有一個字，且是垃圾字的時候、或者垃圾字比正常字數量多一個的時候，sign 函數要大於 0，因此 w_0 不可以等於 $-k$ 。
   - 下面令 $w_0 = p$。

統整後可以發現，PLA 得到的最佳 $\mathbf{w}$ 形式如下：

$$(p, k/{-k}, k/{-k}, \ldots, k/{-k})$$

其中 $k/{-k}$ 代表可能是兩者中的其中一個： k 總共會有d_+ 個，$-k$ 總共會有d_-個；p 是我們 w_0 的值；只要最後的權重 $\mathbf{w}$ 符合這個形式，他就跟 f 一樣可以完美正確的分出各種郵件到底是垃圾郵件還是一般郵件。

有了這些代號，根據開頭 T 的關係式還有題目的定義，可以推得：

$$T \le \frac{\left(\max_n \|\mathbf{x}_n\|\right)^2}{\left(\min_n \frac{\mathbf{w}_f^T \mathbf{x}_n}{\|\mathbf{w}_f\|}\right)^2} = \frac{1+m}{\frac{p^2}{p^2+dk^2}} = \left(1+d\left(\frac{k}{p}\right)^2\right)(1+m)$$

其中最大的 $\mathbf{x}_n$ 因為裡面最多就 m 個字，所以會有 $1+m$ 個 1，長度平方就是 $1+m$；最小的 $\mathbf{x}_n$ 就是信裡面最多 0 個字，所以 $\mathbf{w}_f^T\mathbf{x}_n$ 的值就只剩下 $w_0x_0 = w_0 = p$，而長度平方根據上面的結論可以知道有 d 個 k^2 和一個 p^2，整合起來後就變成了上面的公式。因此如果是以上面的方式建構，得到的 Bound 會是：

$$T \le \left(1+d\left(\frac{k}{p}\right)^2\right)(1+m)$$

至於題目的 Bound 仔細一看可以發現，其實就是代入 $k=2p$。

所以可以知道 $(4d+1)(m+1)$ 不是真正的 Bound；然而上面的建構方式也未必是最 general 的，因為權重或許可以有點花樣，建構方式也會不同，得到的 Bound 要考慮的因素就更多了。

6. 一樣先列出那個厲害的關係式：

$$1 \ge \frac{\mathbf{w}_f^T}{\|\mathbf{w}_f\|}\frac{\mathbf{w}_T}{\|\mathbf{w}_T\|} \ge \sqrt{T}\frac{\rho}{R}$$

$$T \le \frac{\left(\max_n \|\mathbf{x}_n\|\right)^2}{\left(\min_n \frac{\mathbf{w}_f^T \mathbf{x}_n}{\|\mathbf{w}_f\|}\right)^2}$$

令 $\mathbf{w}_{PLA}$ 跟 $\mathbf{w}'_{PLA}$ 得到的關係式如下：

$$T_{pos} \le \frac{\left(\max_n \|\mathbf{x}_n\|\right)^2}{\left(\min_n \frac{\mathbf{w}_{PLA}^T \mathbf{x}_n}{\|\mathbf{w}_{PLA}\|}\right)^2}$$

$$T_{neg} \le \frac{\left(\max_n \|\mathbf{x}'_n\|\right)^2}{\left(\min_n \frac{\mathbf{w}'^T_{PLA} \mathbf{x}'_n}{\|\mathbf{w}'_{PLA}\|}\right)^2}$$

由於 $\mathbf{x}_n$ 跟 $\mathbf{x}'_n$ 只差在 x_0 一個是 1 一個是 -1，可以額外發現：

$$\left(\max_n \|\mathbf{x}_n\|\right)^2 = \left(\max_n \|\mathbf{x}'_n\|\right)^2$$

但重點是，根據公式，對這兩種資料跑的 PLA，只要資料是線性可分的，最終都一定會停下來；也就是說最終停下來後得到的 $\mathbf{w}_{PLA}$ 跟 $\mathbf{w}'_{PLA}$ 都是可以完美正確的分開「訓練資料」的權重，他們兩個在「訓練資料」內做的事情是

一樣的，他們是等價的 equivalent；但是 $\mathbf{w}_{PLA}$ 跟 $\mathbf{w'}_{PLA}$ 很有可能終究還是兩個不一樣的權重，所以到了測試資料的環境，這兩個權重非常有可能會對一個資料的意見產生分歧，而做出不一樣的判斷，此時他們就不是等價的 not equivalent。

7. 一樣先列出那個超好用的關係式：

$$1 \geq \frac{\mathbf{w}_f^T}{\|\mathbf{w}_f\|}\frac{\mathbf{w}_T}{\|\mathbf{w}_T\|} \geq \sqrt{T}\frac{\rho}{R}$$

$$T \leq \frac{\left(\max_{\mathrm{n}}\|\mathbf{x}_{\mathrm{n}}\|\right)^2}{\left(\min_{\mathrm{n}}\frac{\mathbf{w}_{\mathbf{f}}^{\mathbf{T}}\mathbf{x}_{\mathbf{n}}}{\|\mathbf{w}_{\mathbf{f}}\|}\right)^2}$$

這是尚未 normalizing 時的長相，此時$\rho = \min_{\mathrm{n}}\frac{y_n\mathbf{w}_f^T\mathbf{x}_n}{\|\mathbf{w}_f\|}$，$R = \max_{\mathrm{n}}\|\mathbf{x}_n\|$。

然後現在將資料做 normalizing，全部 $\mathbf{z}_n$ 的長度都是 1，根據推導的過程，上面的關係式會修改成：

$$T \leq \frac{\left(\max_n\|\mathbf{z}_n\|\right)^2}{\left(\min_n\frac{\mathbf{w}_{\mathbf{f}}^{\mathbf{T}}\mathbf{z}_{\mathbf{n}}}{\|\mathbf{w}_{\mathbf{f}}\|}\right)^2} = \frac{1}{\left(\min_n\frac{\mathbf{w}_{\mathbf{f}}^{\mathbf{T}}\mathbf{z}_{\mathbf{n}}}{\|\mathbf{w}_{\mathbf{f}}\|}\right)^2}$$

所以如果令 $\rho_{\mathbf{z}} = \min_n\frac{y_n w_f^T \mathbf{z}_n}{\|w_f\|}$ ，可知 $\rho_{\mathbf{z}}^2 = \left(\min_n\frac{y_n w_f^T \mathbf{z}_n}{\|w_f\|}\right)^2 = \left(\min_n\frac{w_f^T \mathbf{z}_n}{\|w_f\|}\right)^2$

因此上面的式子就可以改寫成：

$$T \leq \frac{1}{\rho_{\mathbf{z}}^2}$$

這就是經過 normalizing 後的 Bound。

8. PAM 的特色是將更新的時候改成了當下面發生的時候：

$$y_n\mathbf{w}_{\mathbf{t}}^{\mathbf{T}}\mathbf{x}_{\mathbf{n}} \leq \tau$$

也就是說，對於一個完美可以分開全部點並滿足「邊界寬度要求」的 $\mathbf{w}_{\mathbf{f}}$ 可以知道：

$$y_{n(t)}\mathbf{w}_{\mathbf{f}}^{\mathbf{T}}\mathbf{x}_{\mathbf{n(t)}} \geq \min_n y_n\mathbf{w}_{\mathbf{f}}^{\mathbf{T}}\mathbf{x}_{\mathbf{n}} > \tau$$

所以可以知道：

$$\mathbf{w}_f^T\mathbf{w}_{t+1} = \mathbf{w}_f^T\left(\mathbf{w}_t + y_{n(t)}\mathbf{x}_{n(t)}\right)$$

$$\geq \mathbf{w}_f^T\mathbf{w}_t + \min_n y_n\boldsymbol{w}_{\mathbf{f}}^{\boldsymbol{T}}\boldsymbol{x}_{\boldsymbol{n}}$$

$$\geq \mathbf{w}_f^T\mathbf{w}_t + \tau$$

也就是跟 PLA 一樣，每次的更新都好像讓 $\mathbf{w}_{t+1}$ 更靠近 $\mathbf{w}_f^T$ 。

所以接著是嘗試弄出單位向量的形式，先推導出下列：

$$\|\mathbf{w}_{t+1}\|^2 = \left\|\mathbf{w}_t + y_{n(t)}\mathbf{x}_{n(t)}\right\|^2 = \|\mathbf{w}_t\|^2 + 2y_{n(t)}\mathbf{w}_t^T\mathbf{x}_{n(t)} + \left\|y_{n(t)}\mathbf{x}_{n(t)}\right\|^2$$

$$\leq \|\mathbf{w}_t\|^2 + 2\tau + \left\|y_{n(t)}\mathbf{x}_{n(t)}\right\|^2 \leq \|\mathbf{w}_t\|^2 + 2\tau + \left(\max_n\|\mathbf{x}_n\|\right)^2$$

所以這部分也跟 PLA 一樣，每次的更新，長度的成長是有上限的。

所以只要把這兩個結論合併起來：

$$\mathbf{w}_f^T\mathbf{w}_T \geq T\min_n y_n\mathbf{w}_f^T\mathbf{x}_n$$

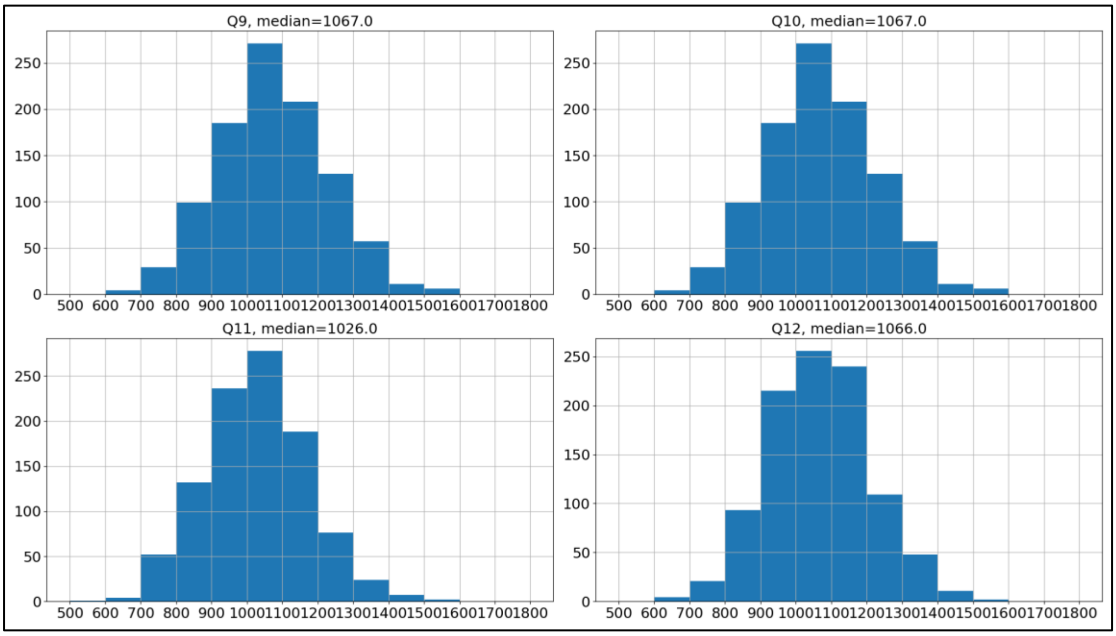
$$\|\mathbf{w}_T\|^2 \leq T\left(2\tau + \left(\max_n\|\mathbf{x}_n\|\right)^2\right) \Rightarrow \frac{1}{\|\mathbf{w}_T\|} \geq \frac{1}{\sqrt{T}\sqrt{2\tau + \left(\max_n\|\mathbf{x}_n\|\right)^2}}$$

$$\Rightarrow 1 \geq \frac{\mathbf{w}_f^T}{\|\mathbf{w}_f\|}\frac{\mathbf{w}_T}{\|\mathbf{w}_T\|} \geq \frac{T\min_n y_n\mathbf{w}_f^T\mathbf{x}_n}{\|\mathbf{w}_f\|\sqrt{T}\sqrt{2\tau + \left(\max_n\|\mathbf{x}_n\|\right)^2}}$$

$$\Rightarrow T \leq \frac{2\tau + \left(\max_n\|\mathbf{x}_n\|\right)^2}{\left(\min_n \frac{\mathbf{w}_f^T\mathbf{x}_n}{\|\mathbf{w}_f\|}\right)^2}$$

所以我們就跟 PLA 的過程一樣，成功說明了 PMA 也會停下來。

- Experiments with Perceptron Learning Algorithm



9. 中位數是 1067。
10. 可以發現，中位數還有形狀分佈跟第 9 題的一模一樣。

11. 中位數是 1026。實驗結果中可以發現，修正次數超過 1100 的數量比第 9 題的低，低於 1100 的則均比第 9 題的高。
12. 中位數是 1066。跟第 9 題的圖形十分接近，但是可以發現變得更集中了，實驗結果中，位於 900 到 1200 區間內的實驗結果數量均高於第 9 題，反之以外的均低於第 9 題。

- Bonus

13. 在這裡我將 speed up 定義為，「最差的情形」能不能加速，所謂最差的情形是指，更新次數跟理論上限相差不會太多，上限的高與低會造成影響。
先列出原版 PLA 的最後推論結果：

$$\mathrm{T} \le \left(\frac{R}{\rho}\right)^2 = \frac{\left(\max\limits_n \|\mathbf{x}_\mathrm{n}\|\right)^2}{\left(\min\limits_n \frac{\mathbf{w}_\mathbf{f}^\mathbf{T}\mathbf{x}_\mathbf{n}}{\|\mathbf{w}_\mathbf{f}\|}\right)^2} = \frac{\left(\max\limits_n \|\mathbf{x}_\mathrm{n}\|\right)^2}{\left(\min\limits_n \|\mathbf{x}_\mathrm{n}\| cos\theta_{f,n}\right)^2}$$

接著是第 7 題的最後推論結果：

$$T \le \frac{1}{\rho_\mathbf{z}^2} = \frac{1}{\left(\min\limits_n \frac{\mathbf{w}_\mathbf{f}^\mathbf{T}\mathbf{z}_\mathbf{n}}{\|\mathbf{w}_\mathbf{f}\|}\right)^2} = \frac{1}{\left(\min\limits_\mathrm{n} cos\theta_{f,n}\right)^2}$$

上面兩行當中的 $\min\limits_\mathrm{n} cos\theta_{f,n}$ 是$\mathbf{w}_\mathbf{f}$ 跟內積最小值的點的夾角。

可以發現，兩者主要差別在於未經過 normalization 的上限會受最大和最小的 $\mathbf{x}_\mathrm{n}$ 的長度所影響。

但是要注意的是，兩者的 $\min\limits_n cos\theta_{f,n}$ 並沒有保證會一樣，也就是說，

就算透過 normalization 將資料點長度的影響消除掉了，但是資料跟 $\mathbf{w}_\mathbf{f}$ 的關係是另一個影響因素，如果 $cos\theta$ 的值太小，也就是說 $\mathbf{w}_\mathbf{f}$ 跟 $\mathbf{x}_\mathrm{n}$ 最小的夾角近乎 90 度，那麼更新次數的上限就會非常大；反之，如果 $\mathbf{w}_\mathbf{f}$ 跟 $\mathbf{x}_\mathbf{n}$ 最小的夾角近乎 0 度，那麼上限就會非常小。

如果換個角度來看，分母的 $\min\limits_n \frac{\mathbf{w}_\mathbf{f}^\mathbf{T}\mathbf{x}_\mathbf{n}}{\|\mathbf{w}_\mathbf{f}\|}$ 其實就是距離最佳超平面的最近

距離，所以一樣只能知道 normalization 將分子的最大資料點長度消除掉，但並不能知道 normalization 會把離超平面最近的點的距離造成何種影響。

所以我不認為第 7 題的 normalization 可以加速「最差的情形」。