HW1

1.

self-supervised learning 一個常見的應用是拿來做 pre-training。簡單來說就是希望用 self-supervised learning 預先學習到 data 中的某些特徵，這樣之後在其他任務作 fine-tune 時可以利用原先學習到的特徵來讓 fine-tune 的效果更好。像是之前的 bert 就是這樣的，用填空題的方法進行 unsupervised learning，讓模型學到文字的某些規律，像是文法、習慣等等。具體而言，就是拿一堆沒有 label 的句子，然後對於一個句子，把某個單字抽出來，把剩下的句子放進 model，讓 model 去 predict 抽出來的單字是什麼。訓練完後，會預期說 model 學到了這個語言的某些特性。這樣其他人把預先訓練好的 bert 拿去做其他任務時，就可以利用原本學到的特性去更好的學會其他任務。目前為止我們還沒有學到說這個應用是否有嚴謹的數學證明，但至少從 bert 論文中提出的數據而言，至少結果上來說是符合預期的。

2.

就像老師所說的，ML 基本上無法 100%絕對可以做到某些事。但或許可以透過學習達成一定的成功率。從老師講的機器學習 3 大要素來說。首先就是有沒有一個 performance 的標準。這點明顯

就是到終點的步數。一個是有沒有某種規律，直觀上很難說這算不算是某種意義上的規律，畢竟這不是直觀能說「規律就是這樣」的問題。但演算法角度上來說最短路徑就是用 bfs 到終點得出的答案，勉強還是能以這個角度去做判定，所以我覺得這可以算是有規律。一個是有沒有 data，這點並不難用程式生成。所以我覺得是可以用 ML 的。Chatgpt 的回答並沒有明確給出如何訓練，所以單就 ML 能不能找到最短路徑這個問題我是同意他的說法的。ML 或許能學到某些東西，或許真的能用學到的東西來找到最短路徑，但 performance 就無法保證。

3.

我覺得不能說 Chatgpt 錯誤。Chatgpt 回答說「ML 可以在某些特定的情況優化某些演算法，但不一定能加速演算法，還是要看問題還有演算法的狀況。」可以說 chatgpt 回答得很保守，並沒有很肯定地否認說 ML 絕對不能加速任何演算法。Chatgpt 覺得是有一種可能性說當問題和演算法的狀況可以的話，某些演算法或許是可以被 ML 加速。所以我們只能說 chatgpt 的看法偏悲觀，但沒有完全去否認。因此當 alphadev 出來後，我們也不能說 chatgpt 的這個回答完全錯誤，畢竟他這個回答還是有留點餘地的。

4.

至少就講義上的定義而言，初始 w 的w_0是 0。根據公式$h(x) = \text{sign}(w_0 + \sum_{i=1}^{n} w_i x_i)$，我們可以想說，對於每個點 x，會多增加一維$x_0$，其值是 1，乘上−threshold，為了能跟 w 維度一樣。每次當$y_{n(t)} = 1$或-1 時，根據公式$w_{t+1} = w_t + y_{n(t)} x_{n(t)}$，$w_t$中的$w_0$項會加上$y_{n(t)} * x_0 = y_{n(t)} * 1 = y_{n(t)}$。所以經過了$T_+$次$y_{n(t)} = 1$，$w_0$會增加$T_+$。經過了$T_-$次$y_{n(t)} = -1$，$w_0$會減少$T_-$。所以最後$w_0 = 0 + T_+ - T_- = T_+ - T_-$。

5.

這裡直接引用 W2 講義中 fun time 的解答，PLA 最大錯誤次數是$\frac{R^2}{\rho^2}$。其中$R^2 = \max_n \left||x_n|\right|^2$，$\rho = \min_n y_n \frac{w_f^T}{\left||w_f|\right|} x_n$。因為 x 的 d+1 維中最多只有 m+1 個是 1 其他都是 0，所以$R^2 = \max_n \left||x_n|\right|^2 \le m + 1$。題目給出的 f 其實直接對應一個 w_f: for $i = 1{\sim}d$，$w_i = 2$ if x_i $is\ spam-like\ else$ $w_i = 2 = -2$。$w_0 = -1$。因為這樣$\text{sign}\left(w_f^T x\right) = sign(2z_+(x) - 2z_-(x) - 1) = sign(z_+(x) - z_-(x) - 0.5)$。我們不知道真的求出來的$w_f$長什麼樣子，但我們可以知道至少$w_f$是有可能長得跟 f 一樣的。所以這時$\rho^2 = \min_n \left(y_n \frac{w_f^T}{\left||w_f|\right|} x_n\right)^2 = \min_n \left(y_n \frac{w_f^T}{\sqrt{1+4d}} x_n\right)^2$。這時$y_n(2z_+(x) - 2z_-(x) - 1)$因為大於 0，所以最

低是 1，所以$\min_n\left(y_n\frac{w_f^T}{\sqrt{1+4d}}\mathrm{x_n}\right)^2 \geq \frac{1}{1+4d}$。因此$\frac{\mathrm{R}^2}{\rho^2} \leq (4d+1)(m+1)$。當$w_f$跟上面給出的 f 一樣時等號有可能會成立，所以可以說$(4d+1)(m+1)$是 PLA 錯誤數的一個 upper bound。

6.

根據公式$\mathrm{w_{t+1}} = w_t + y_{n(t)}x_{n(t)}$，一開始，對於原始 PLA 而言，$\mathrm{w_1} = w_0 + y_{n(0)}x_{n(0)} = y_{n(0)}x_{n(0)}$。對於改版 PLA 而言，$\mathrm{w'_1} = w'_0 + y_{n(0)}x'_{n(0)} = y_{n(0)}x'_{n(0)}$。這時$x'_{n(0)}$跟$x_{n(0)}$只有$x_0$這一項不一樣。當 PLA 選到第 t 個點$x_{n(t)}$時(t>=1)，代表$y_{n(t)}w_{t-1}^T x_{n(t)} < 0$。這時因為$y_{n(t)}w_{t-1}^T x_{n(t)} = y_{n(t)}\sum_{i=0}^n w_{(t-1)i}x_{n(t)i} = y_{n(t)}\sum_{i=1}^n w_{(t-1)i}x_{n(t)i} + w_{(t-1)0}x_{n(t)0} = y_{n(1)}\sum_{i=1}^n w_{0i}x_{n(1)i} + \left(-w_{(t-1)0}\right)\left(-x_{n(t)0}\right) = y_{n(t)}\sum_{i=1}^n w'_{(t-1)i}x'_{n(t)i} = y_{n(t)}w'^T_{(t-1)}x'_{n(t)} < 0$。所以第 t 個點的狀況，當$x_{n(t)}$對原始 PLA 是錯誤點，則$x_{n(1)}$對改版 PLA 也是錯誤點。所以第 t 個點的狀況原始跟改版的 PLA 都可以選到相同的點。根據數學歸納法，當初始 w 相同(都是 0)，且每次選到的錯誤點都是相同的情況時，$\mathrm{w_{PLA}}$跟$\mathrm{w'_{PLA}}$是一樣的。

7.

講義中推導：$\mathrm{w_f^T} w_{t+1} \geq \mathrm{w_f^T} w_t + \min_n y_n w_f^T \mathrm{x_n}$，在這題會變成$\mathrm{w_f^T} w_{t+1} \geq \mathrm{w_f^T} w_t + \min_n y_n w_f^T \mathrm{z_n}$，因此$\mathrm{w_f^T} w_t \geq \mathrm{w_f^T} w_0 + t *$

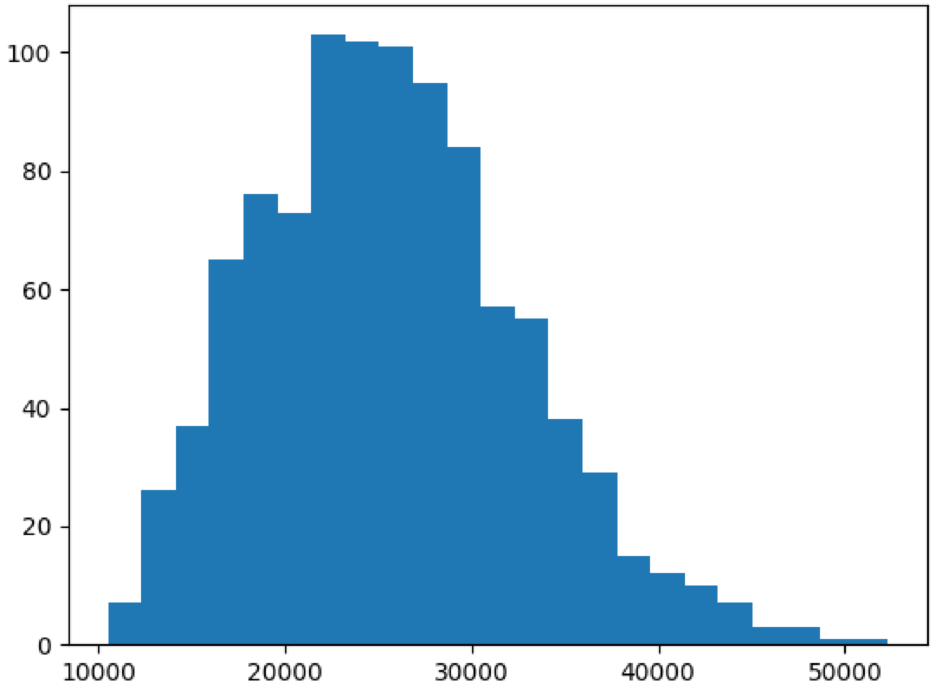
$\min_n y_n w_f^T \mathrm{z_n} = t * \min_n y_n w_f^T \mathrm{z_n}$ 。同時講義中推導：$\left|\left|w_{t+1}\right|\right|^2 \leq \left|\left|w_t\right|\right|^2 + \max_n \left|\left|\mathrm{x_n}\right|\right|^2$，這裡會變成$\left|\left|w_{t+1}\right|\right|^2 \leq \left|\left|w_t\right|\right|^2 + \max_n \left|\left|\mathrm{z_n}\right|\right|^2$。因此 $1 \geq \cos\theta = \frac{\mathrm{w_f}^T w_t}{\left|\left|w_f\right|\right|\left|\left|w_t\right|\right|} \geq \frac{t * \min_n y_n w_f^T \mathrm{z_n}}{\left|\left|w_f\right|\right|\left(\sqrt{\left|\left|w_0\right|\right|^2 + t * \max_n \left|\left|\mathrm{z_n}\right|\right|^2}\right)} \rightarrow \left(\frac{\max_n \left|\left|\mathrm{z_n}\right|\right|}{\min_n y_n \frac{w_f^T}{\left|\left|w_f\right|\right|} \mathrm{z_n}}\right)^2 \geq t$。其中$\left|\left|\mathrm{z_n}\right|\right| = 1$，所以$\left(\frac{\max_n \left|\left|\mathrm{z_n}\right|\right|}{\min_n y_n \frac{w_f^T}{\left|\left|w_f\right|\right|} \mathrm{z_n}}\right)^2 = \frac{1}{\rho_z^2} \geq t$

8.

基本上就照著 PLA 講義裡面 PLA fact 來推就好了。首先對解答w_f，必有 $\min_n y_n w_f^T \mathrm{x_n} > \tau$（根據題目的定義），所以 $\mathrm{w_f}^T w_{t+1} \geq \mathrm{w_f}^T w_t + \min_n y_n w_f^T \mathrm{x_n} \geq \mathrm{w_f}^T w_t + \tau$，所以$\mathrm{w_f}^T w_{t+1} \geq \mathrm{w_f}^T w_0 + t\tau = t\tau$。然後 $\mathrm{w_{t+1}} = w_t + y_{n(t)} x_{n(t)} \rightarrow \left|\left|\mathrm{w_{t+1}}\right|\right|^2 = \left|\left|w_t + y_{n(t)} x_{n(t)}\right|\right|^2 = \left|\left|\mathrm{w_t}\right|\right|^2 + \left|\left|y_{n(t)} x_{n(t)}\right|\right|^2 + 2 y_n w_f^T \mathrm{x_n} \leq \left|\left|\mathrm{w_t}\right|\right|^2 + \max_n \left|\left|\mathrm{x_n}\right|\right|^2$，所以$\left|\left|\mathrm{w_{t+1}}\right|\right|^2 \leq \left|\left|\mathrm{w_0}\right|\right|^2 + t \max_n \left|\left|\mathrm{x_n}\right|\right|^2 = t \max_n \left|\left|\mathrm{x_n}\right|\right|^2$。最後，$1 \geq \cos\theta = \frac{\mathrm{w_f}^T w_t}{\left|\left|w_f\right|\right|\left|\left|w_t\right|\right|} \geq \frac{t\tau}{\left|\left|w_f\right|\right|\left(\sqrt{t \max_n \left|\left|\mathrm{x_n}\right|\right|^2}\right)} \rightarrow \frac{\left|\left|w_f\right|\right|^2 \max_n \left|\left|\mathrm{x_n}\right|\right|^2}{\tau^2} \geq \mathrm{t}$，因此，t 有上限，所以這種 PLA 也會在有限步內停止。

9.

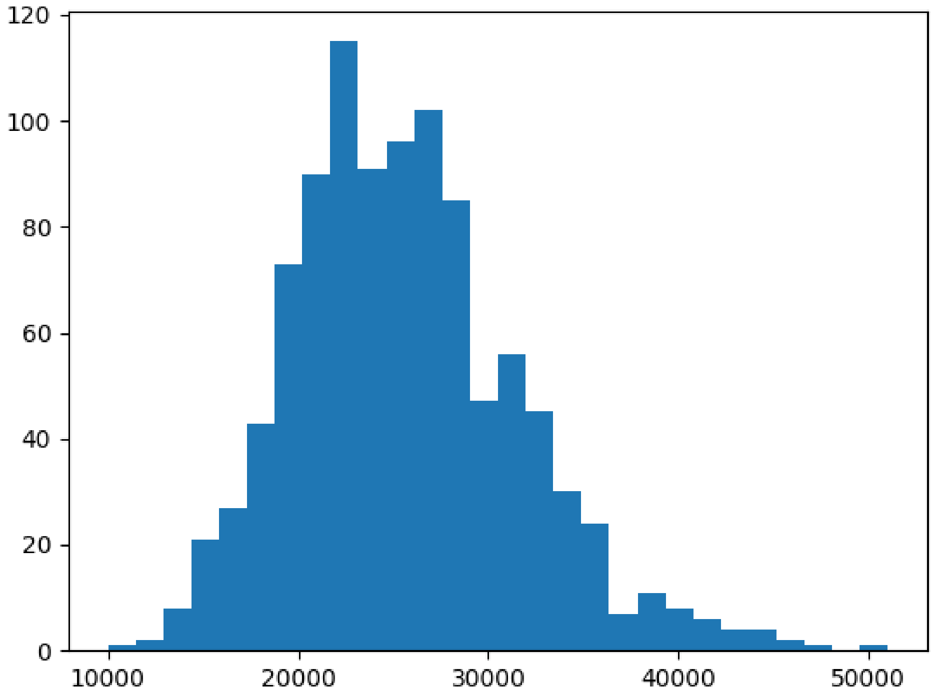
這個是我畫出來的圖:



其中位數為 25280

10.

經過 data 放大後，這個是我畫出來的圖:

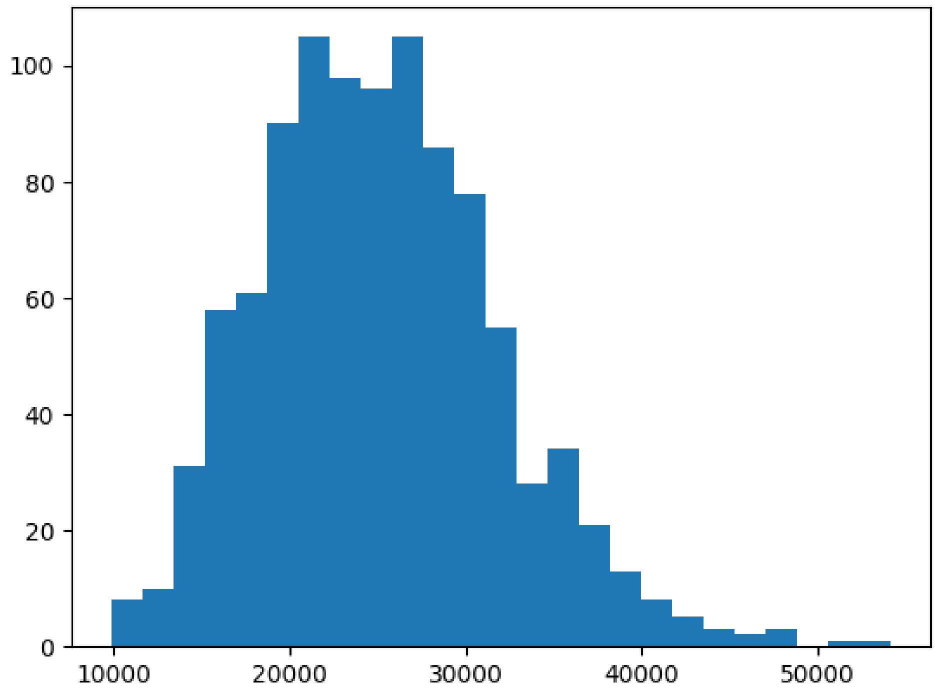


中位數是 25025.5。可以感覺出來跟前一個沒有統計意義上的差別。

其實可以想像，把資料放大 11.26 倍後，每次 w 做更新時就只是 w 變大 11.26 倍，數據之間的大小關係沒有變化，所以計算出來的 sign 是不會受到影響的。

11.

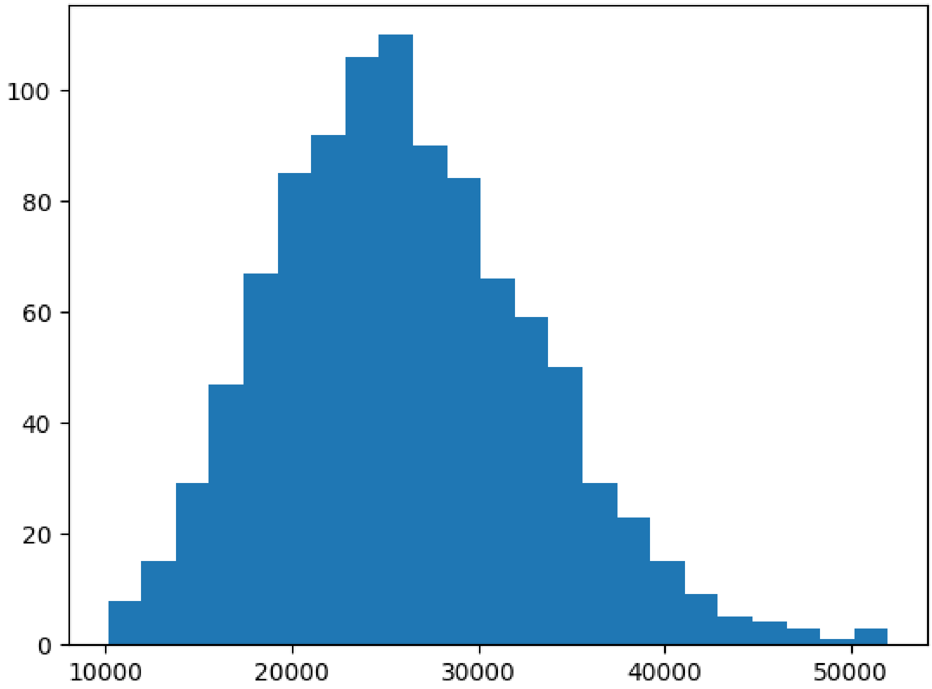
這是這題我畫出來的圖:



中位數是 24640.5。同樣跟前面沒有統計意義上的差別。畢竟 PLA 的計算基本上 w 的每個維度是獨立的。單個維度的放大不會影響到其他維度的計算，所以計算速度和 sign 的計算不太受影響。就有點類似第 6 題。

12.

這是我畫出來的圖:



中位數是 25519

一樣跟前面差距不大。感覺這題的算法像是前面算法的特例。而且就 PLA 的特性而言，至少就上課老師講義上 visualize 的 demo，同一個點不太會連續錯 2 次以上，所以事實上還是跟前面的算法很像。

13.

首先對解答w_f，必有$\forall n, y_n w_f^T \mathrm{x_n} > 0$(根據 PLA 的定義)且$\forall n, ||\mathrm{x_n}||$。所以$\forall n, y_n w_f^T \frac{\mathrm{x_n}}{||\mathrm{x_n}||} > 0$。所以說 normalize 前的$w_f$就是 normalize 後的$w_f$。接著，$\frac{\left(\frac{R}{\rho}\right)^2}{\frac{1}{\rho_z^2}} = \left(\frac{\rho_z R}{\rho}\right)^2$，而$\frac{\rho_z R}{\rho} = \frac{\min_n y_n \frac{w_f^T}{||w_f||}\mathrm{z_n} \ *\max_n ||x_n||}{\min_n y_n \frac{w_f^T}{||w_f||}\mathrm{x_n}}$。令$\min_n y_n \frac{w_f^T}{||w_f||}\mathrm{x_n}$

的 n 為 n”，$\min_n y_n \frac{w_f^T}{\|w_f\|} \mathrm{z_n}$ 的 n 為 n’，則 $\frac{\min_n y_n \frac{w_f^T}{\|w_f\|} \mathrm{z_n} \ * \ \max_n \|x_n\|}{\min_n y_n \frac{w_f^T}{\|w_f\|} \mathrm{x_n}} =$

$\frac{y_{n'} \frac{w_f^T}{\|w_f\|} \mathrm{z_{n'}} \ * \ \max_n \|x_n\|}{y_{n"} \frac{w_f^T}{\|w_f\|} \mathrm{x_{n"}}} = \frac{y_{n'} w_f^T \mathrm{z_{n'}} \ * \ \max_n \|x_n\|}{y_{n"} w_f^T \mathrm{x_{n"}}} \geq \frac{y_{n'} w_f^T \mathrm{z_{n'}} \ * \|\mathrm{x}_{n'}\|}{y_{n"} w_f^T \mathrm{x_{n"}}} = \frac{y_{n'} w_f^T \mathrm{x_{n'}}}{y_{n"} w_f^T \mathrm{x_{n"}}}$。因

$\min_n y_n w_f^T \mathrm{x_n}$ 的 n 為 n”，所以 $\frac{y_{n'} w_f^T \mathrm{x_{n'}}}{y_{n"} w_f^T \mathrm{x_{n"}}} \geq 1$。因此，$\frac{\rho_z R}{\rho} \geq 1 \rightarrow \frac{\left(\frac{R}{\rho}\right)^2}{\frac{1}{\rho_z^2}} \geq 1$，

可以說 normalize 可以加速 PLA。