1.

項目	狀態	截止時間	權重	成績
作業三 測驗	已通過	1月20日 15:59 CST	50%	**100%**

2.

此敘述為 True

proof:

分成兩種狀況討論:

①當 y 與 $w^T x$ 同號，$err(w)= \max(0, -yw^T x)=0$:

PLA:此時分類正確，PLA 演算法並不會改動 weight，即 $w_{t+1}= w_t+0$

SGD:此時梯度為 0，$w_{t+1}= w_t-0= w_t+0$，結果與 PLA 相同

②當 y 與 $w^T x$ 異號，$err(w)= \max(0, -yw^T x)= -yw^T x$:

PLA:此時分類錯誤，PLA 演算法改動 weight，即 $w_{t+1}= w_t+yx$

SGD:此時梯度為-yx，$w_{t+1}= w_t-(-yx)= w_t+yx$，結果與 PLA 相同

可知在使用 SGD 的情況下，PLA 的 error function 可以是 $err(w)= \max(0, -yw^T x)$

3.

把 $E(u+\Delta u, v+\Delta v)$ 泰勒展開至二階形式

$$E(u+\Delta u, v+\Delta v) = E(u,v)+\frac{\partial E(u,v)}{\partial u}\Delta u+\frac{\partial E(u,v)}{\partial v}\Delta v+\frac{1}{2}\left[\frac{\partial^2 E(u,v)}{\partial u^2}\Delta u^2+2\frac{\partial^2 E(u,v)}{\partial u\partial v}\Delta u\Delta v+\frac{\partial^2 E(u,v)}{\partial v^2}\Delta v^2\right]$$

寫成矩陣的表達形式:

$$E(u+\Delta u, v+\Delta v) = E(u,v)+\left[\frac{\partial E}{\partial u}, \frac{\partial E}{\partial v}\right]\begin{bmatrix}\Delta u\\ \Delta v\end{bmatrix}+\frac{1}{2}\begin{bmatrix}\Delta u & \Delta v\end{bmatrix}\begin{bmatrix}\frac{\partial^2 E}{\partial u^2} & \frac{\partial^2 E}{\partial u\partial v}\\ \frac{\partial^2 E}{\partial v\partial u} & \frac{\partial^2 E}{\partial v^2}\end{bmatrix}\begin{bmatrix}\Delta u\\ \Delta v\end{bmatrix}$$

$$= E(u,v)+\nabla E(u,v)\cdot\Delta X+\frac{1}{2}\nabla^2 E(u,v)\Delta X^2 \quad (\text{其中}\Delta X=\begin{bmatrix}\Delta u\\ \Delta v\end{bmatrix})$$

當 $\Delta u, \Delta v$ 非常接近於 0 時，上面等式可以寫成

$$E(u,v) = E(u,v)+\nabla E(u,v)\cdot\Delta X+\nabla^2 E(u,v)\Delta X^2$$

$$\Rightarrow \nabla E(u,v)\cdot\Delta X+\nabla^2 E(u,v)\Delta X^2 = 0$$

$$\Rightarrow \nabla E(u,v)+\nabla^2 E(u,v)\cdot\Delta X = 0$$

$$\Rightarrow \Delta X = \frac{-\nabla E(u,v)}{\nabla^2 E(u,v)}$$

4.

$$\max_{\bar{i}} \frac{1}{N}\sum_{n=1}^{N} \frac{\exp(w_{y_n}^T x_n)}{\sum_{k=1}^{K}\exp(w_k^T x_n)} = \max_{\bar{i}} \frac{1}{N}\sum_{n=1}^{N} \ln \frac{\exp(w_{y_n}^T x_n)}{\sum_{k=1}^{K}\exp(w_k^T x_n)}$$

$$= \max_{\bar{i}} \frac{1}{N}\sum_{n=1}^{N} \left[\ln(\exp(w_{y_n}^T x_n) - \ln\sum_{k=1}^{K}\exp(w_k^T x_n)\right]$$

$$= \max_{\bar{i}} \frac{1}{N}\sum_{n=1}^{N} \left[w_{y_n}^T x_n - \ln\sum_{k=1}^{K}\exp(w_k^T x_n)\right]$$

$$= \min_{\bar{i}} \frac{1}{N}\sum_{n=1}^{N} \left[\ln\sum_{k=1}^{K}\exp(w_k^T x_n) - w_{y_n}^T x_n\right]$$

故 E_{in} 為 $\frac{1}{N}\sum_{n=1}^{N}\left[\ln\sum_{k=1}^{K}\exp(w_k^T x_n) - w_{y_n}^T x_n\right]$

5.

$$\frac{1}{N+K}\left(\sum_{n=1}^{N}(y_n - w^T x_n)^2 + \sum_{k=1}^{K}(\tilde{y}_k - w^T \tilde{x}_k)^2\right)$$

$$= \frac{1}{N+K}\left(\sum_{n=1}^{N}(w^T x_n - y_n)^2 + \sum_{k=1}^{K}(w^T \tilde{x}_k - \tilde{y}_k)^2\right)$$

$$= \frac{1}{N+K}\left(\sum_{n=1}^{N}(x_n^T w - y_n)^2 + \sum_{k=1}^{K}(\tilde{x}_k^T w - \tilde{y}_k)^2\right)$$

$$= \frac{1}{N+K}\left(\left\| \begin{bmatrix} x_1^T w - y_1 \\ x_2^T w - y_2 \\ \vdots \\ x_N^T w - y_N \end{bmatrix} \right\|^2 + \left\| \begin{bmatrix} \tilde{x}_1^T w - \tilde{y}_1 \\ \tilde{x}_2^T w - \tilde{y}_2 \\ \vdots \\ \tilde{x}_K^T w - \tilde{y}_K \end{bmatrix} \right\|^2\right)$$

$$= \frac{1}{N+K}\left(\left\| \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix} w - \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \right\|^2 + \left\| \begin{bmatrix} \tilde{x}_1^T \\ \tilde{x}_2^T \\ \vdots \\ \tilde{x}_K^T \end{bmatrix} w - \begin{bmatrix} \tilde{y}_1 \\ \tilde{y}_2 \\ \vdots \\ \tilde{y}_K \end{bmatrix} \right\|^2\right)$$

$$= \frac{1}{N+K}\left(\|X^T w - y\|^2 + \|\tilde{X}^T w - \tilde{y}\|^2\right)$$

$$= \frac{1}{N+K}\left[(w^T X^T X w - 2w^T X^T y + y^T y) + (w^T \tilde{X}^T \tilde{X} w - 2w^T \tilde{X}^T \tilde{y} + \tilde{y}^T \tilde{y})\right]$$

對上式作偏微分可得：

$$\frac{1}{N+K}\left[(2X^T X w - 2X^T y) + (2\tilde{X}^T \tilde{X} w - 2\tilde{X}^T \tilde{y})\right]$$

$$= \frac{2}{N+K}\left[(X^T X w - X^T y) + (\tilde{X}^T \tilde{X} w - \tilde{X}^T \tilde{y})\right]$$

令上式為 0

$$\frac{2}{N+K}\left[(X^T X w - X^T y) + (\tilde{X}^T \tilde{X} w - \tilde{X}^T \tilde{y})\right] = 0$$

$$\Rightarrow (X^T X + \tilde{X}^T \tilde{X}) w - (X^T y + \tilde{X}^T \tilde{y}) = 0$$

$$\Rightarrow w = (X^T X + \tilde{X}^T \tilde{X})^{-1}(X^T y + \tilde{X}^T \tilde{y})$$

6.

$$\frac{\lambda}{n}\|w\|^2+\frac{1}{n}\|Xw-y\|^2$$

$$=\frac{\lambda}{n}(w^Tw)+\frac{1}{n}(w^TX^TXw-2w^TX^Ty+y^Ty)$$

對上式作偏微分可得:

$$=\frac{2\lambda}{n}(w)+\frac{2}{n}(X^TXw-X^Ty)$$

令上式為0可得:

$$\frac{2\lambda}{n}(w)+\frac{2}{n}(X^TXw-X^Ty)=0$$

$$\Rightarrow \lambda\cdot w+X^TXw-X^Ty=0$$

$$\Rightarrow (\lambda\cdot I+X^TX)w=X^Ty$$

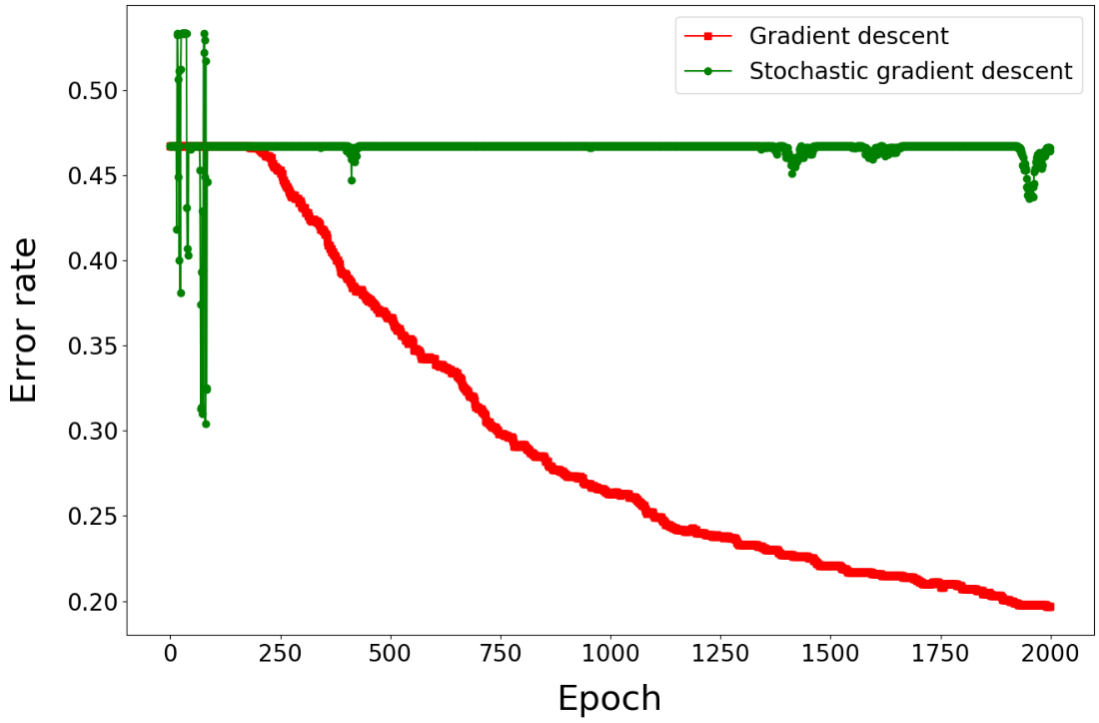
$$\Rightarrow w=(X^TX+\lambda I)^{-1}(X^Ty)$$

$$(X^TX+\lambda I)^{-1}(X^Ty)=(X^TX+\tilde{X}^T\tilde{X})^{-1}(X^Ty+\tilde{X}^T\tilde{y})$$

$$\Rightarrow \text{①}\ \tilde{X}^T\tilde{X}=\lambda I \Rightarrow \tilde{X}=\sqrt{\lambda}\cdot I$$

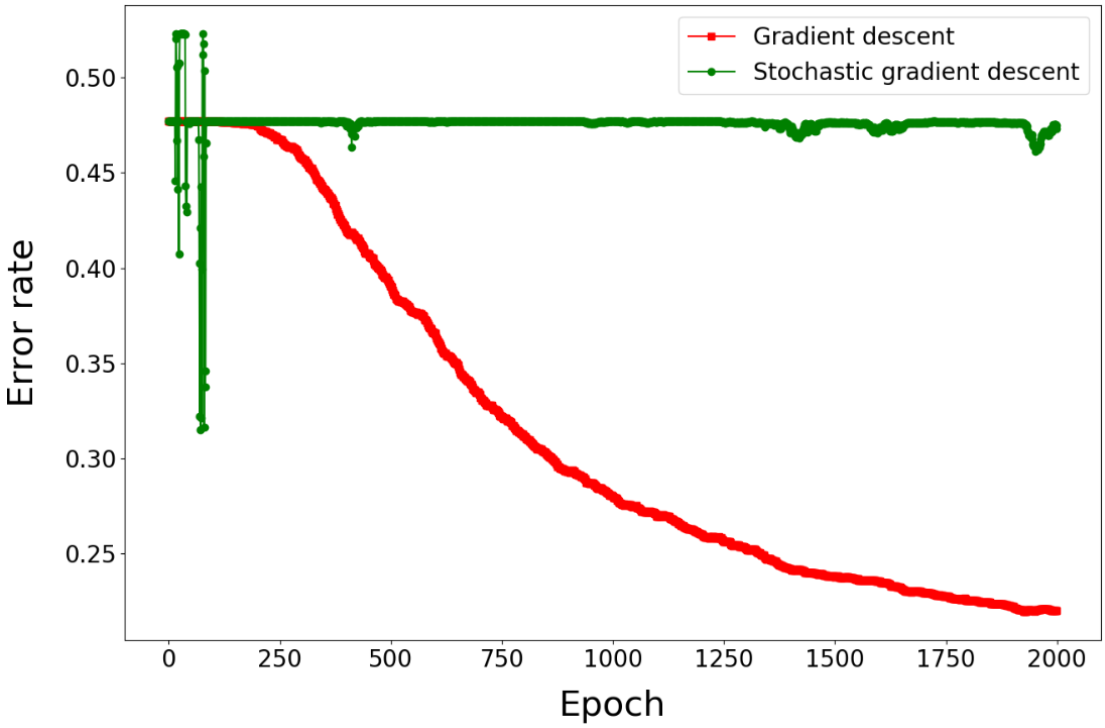
② $\tilde{X}^T\tilde{y}=0$ 且 $\tilde{X}=\sqrt{\lambda}\cdot I \Rightarrow \sqrt{\lambda}\cdot I\cdot\tilde{y}=0 \Rightarrow \tilde{y}=0$

7.



Stochastic gradient descent 的 error rate 基本上維持在 0.45 左右，相反的，Gradient descent 則一路下降直到 T=2000，且似乎還有再往下降的趨勢

8.



曲線基本上與第 7 題一樣，但在相同 T 的情況，不管是 Stochastic gradient descent 還是 Gradient descent，Ein 均比 Eout 來的低(下方是比較圖)