

# ML HW3

1.

← 作業三  
評分測驗 • 40 min

截止時間 2月9日 23:59 PST

✓ 恭喜！您通過了！

通過條件 75% 或更高

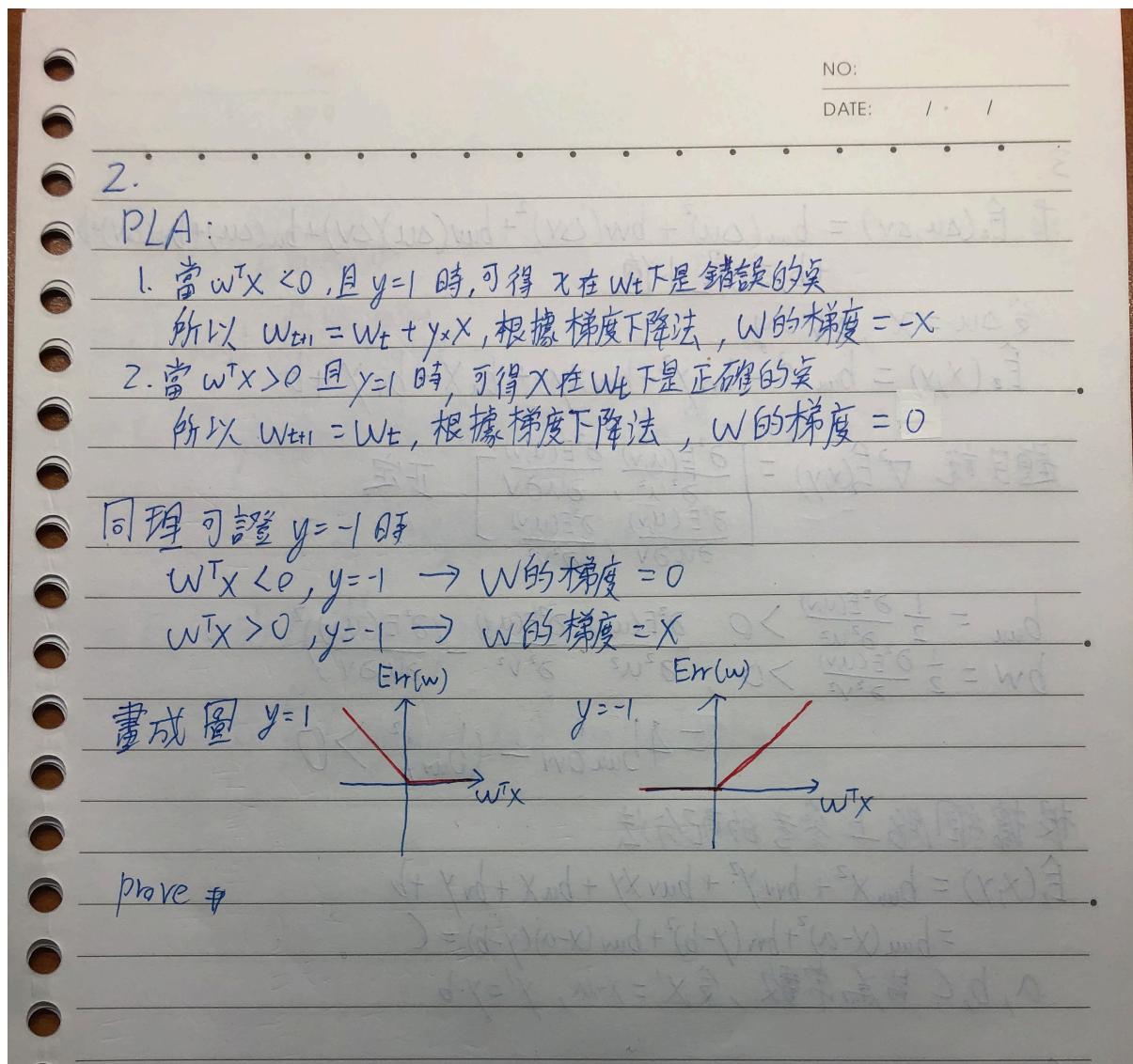
堅持學習

成績  
100%

## 作業三

最新提交作業的評分  
100%

2.



3.

3.

求  $\hat{E}_2(\Delta u, \Delta v) = b_{uu}(\Delta u)^2 + b_{vv}(\Delta v)^2 + b_{uv}(\Delta u)(\Delta v) + b_u(\Delta u) + b_v(\Delta v) + b$   
的最小值。

令  $\Delta u = x, \Delta v = y$

$$\hat{E}_2(x, y) = b_{uu}x^2 + b_{vv}y^2 + b_{uv}xy + b_u x + b_v y + b$$

題目說  $\nabla^2 \hat{E}_2(x, y) = \begin{bmatrix} \frac{\partial^2 \hat{E}(u, v)}{\partial u^2}, & \frac{\partial^2 \hat{E}(u, v)}{\partial u \partial v} \\ \frac{\partial^2 \hat{E}(u, v)}{\partial v \partial u}, & \frac{\partial^2 \hat{E}(u, v)}{\partial v^2} \end{bmatrix}$  正定

$$\begin{aligned} b_{uu} &= \frac{1}{2} \frac{\partial^2 \hat{E}(u, v)}{\partial u^2} > 0, \quad \frac{\partial^2 \hat{E}(u, v)}{\partial u^2} \cdot \frac{\partial^2 \hat{E}(u, v)}{\partial v^2} - \left( \frac{\partial^2 \hat{E}(u, v)}{\partial u \partial v} \right)^2 \\ b_{vv} &= \frac{1}{2} \frac{\partial^2 \hat{E}(u, v)}{\partial v^2} > 0, \quad \frac{\partial^2 \hat{E}(u, v)}{\partial u^2} \cdot \frac{\partial^2 \hat{E}(u, v)}{\partial v^2} - \left( \frac{\partial^2 \hat{E}(u, v)}{\partial u \partial v} \right)^2 \\ &= 4b_{uu}b_{vv} - (b_{uv})^2 > 0 \end{aligned}$$

根據網路上參考的配分法

$$\begin{aligned} \hat{E}_2(x, y) &= b_{uu}x^2 + b_{vv}y^2 + b_{uv}xy + b_u x + b_v y + b \\ &= b_{uu}(x-a)^2 + b_{vv}(y-b)^2 + b_{uv}(x-a)(y-b) + C \end{aligned}$$

$a, b, C$  皆為常數，令  $x' = x-a, y' = y-b$

$$\hat{E}_2(x, y) = b_{uu}x'^2 + b_{vv}y'^2 + b_{uv}x'y' + C$$

$$\begin{aligned} &= b_{uu}(x' + \frac{b_{uv}}{2b_{uu}}y')^2 + \left(b_{vv} - \frac{b_{uv}^2}{4b_{uu}}\right)y'^2 + C, \text{ 因 } 4b_{uu}b_{vv} - (b_{uv})^2 > 0 \\ \Rightarrow \hat{E}_2(\Delta u, \Delta v) &\geq C, \text{ 當等號成立時 } \Delta u = x = a, \Delta v = y = b \quad b_{vv} - \frac{b_{uv}^2}{4b_{uu}} > 0 \end{aligned}$$

現在求  $a, b$ . : 當  $x=a, y=b$  時,  $\hat{E}_2(\Delta u, \Delta v)$  有最小值。

$$\Rightarrow \hat{E}_2(x, y) = b_{uu}x^2 + b_{vv}y^2 + b_{uv}xy + b_u x + b_v y + b$$

$$\begin{aligned} \frac{\partial \hat{E}(x, y)}{\partial x} &= 0 = 2b_{uu}x + yb_{uv} + b_u \Rightarrow \begin{bmatrix} 2b_{uu} & b_{uv} \\ b_{uv} & 2b_{vv} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} -b_u \\ -b_v \end{bmatrix} \\ \frac{\partial \hat{E}(x, y)}{\partial y} &= 0 = 2b_{vv}y + xb_{uv} + b_v \end{aligned}$$

$$\Rightarrow \nabla^2 \hat{E}(u, v) \begin{bmatrix} a \\ b \end{bmatrix} = -\nabla \hat{E}(u, v) \Rightarrow \begin{bmatrix} a \\ b \end{bmatrix} = -(\nabla^2 \hat{E}(u, v))^{-1} \nabla \hat{E}(u, v)$$

4.

NO: \_\_\_\_\_  
DATE: / /

4.

$$h_y(x) = \frac{\exp(w_y^T x)}{\sum_{i=1}^k \exp(w_i^T x)}$$

$$\Rightarrow \prod_{j=1}^N h_y(x_j) = \prod_{j=1}^N \frac{\exp(w_y^T x_j)}{\sum_{i=1}^k \exp(w_i^T x_j)}$$

$$\Rightarrow \ln \left( \prod_{j=1}^N h_y(x_j) \right) = \sum_{j=1}^N \left( \ln(\exp(w_y^T x_j)) - \ln \left( \sum_{i=1}^k \exp(w_i^T x_j) \right) \right)$$

$$= \sum_{j=1}^N \left( w_y^T x_j - \ln \sum_{i=1}^k \exp(w_i^T x_j) \right)$$

$$\Rightarrow E_{in} = \frac{1}{N} \ln \left( \prod_{j=1}^N h_y(x_j) \right)$$

$$= \frac{1}{N} \cdot \sum_{j=1}^N \left( w_y^T x_j - \ln \sum_{i=1}^k \exp(w_i^T x_j) \right) \#$$

NO: \_\_\_\_\_  
DATE: / /

5.

根據題目，可發現  
只要將  $x$  和  $\tilde{x}$  合併成一個大  $X$ ,  $y$  和  $\tilde{y}$  合併成一個大  $Y$  即可  
原式  $\rightarrow \min_w \frac{1}{N+k} \|Xw - Y\|^2$  找  $w$  的最佳解  
最後，套入課程公式

$$\begin{aligned} w &= (X^T X)^{-1} X^T Y \\ &= ([x] [x])^{-1} [x]^T [y] \\ &= \frac{x^T y + \tilde{x}^T \tilde{y}}{x^T x + \tilde{x}^T \tilde{x}} \end{aligned}$$

6.

根據公式可得  
 $W_{reg} = (X^T X + \lambda I)^{-1} X^T y$   
 By 上一題的公式  $w = \frac{x^T y + \tilde{x}^T \tilde{y}}{x^T x + \tilde{x}^T \tilde{x}}$

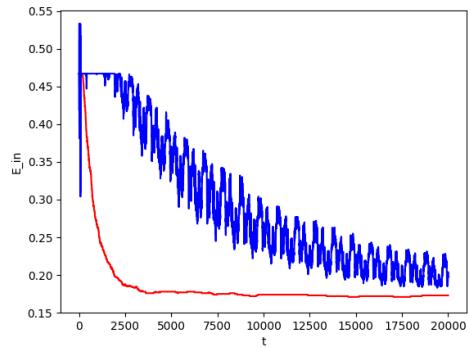
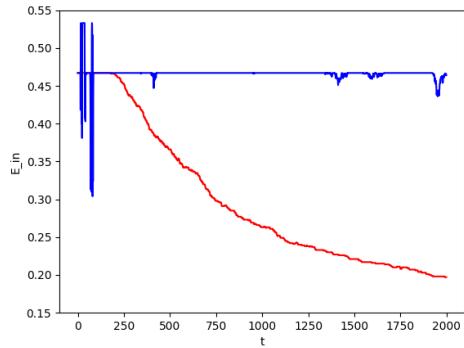
$$\frac{x^T y}{x^T x + \lambda I} = \frac{x^T y + \tilde{x}^T \tilde{y}}{x^T x + \tilde{x}^T \tilde{x}}$$

$$\Rightarrow \tilde{x} = \sqrt{\lambda I}, \tilde{y} = 0 \#$$

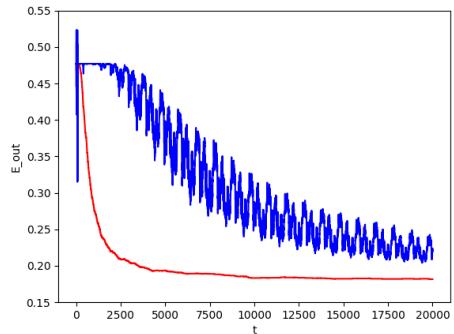
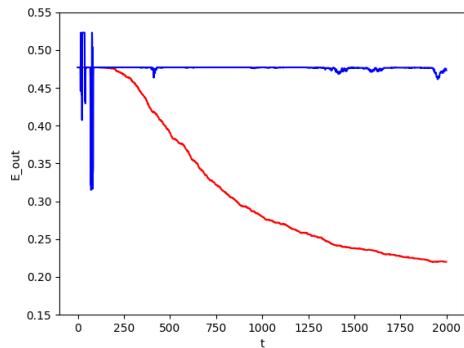
7.

藍色的線:Stochastic

紅色的線:Gradient



8.



## 7~8 Finding

觀察上面四圖即可發現，

當跑了 2000 次後，使用 GD 的  $E_{in}=0.195$ ， $E_{out}=0.22$

，使用 SGD 的  $E_{in}=0.466$ ， $E_{out}=0.475$

當跑了 20000 次後，使用 GD 的  $E_{in}=0.172$ ， $E_{out}=0.182$

，使用 SGD 的  $E_{in}=0.186$ ， $E_{out}=0.196$

這代表了使用 GD 的話，2000 次所達成的效果與跑 20000 次的效果差不多，代表 2000 次已夠，差不多已經接近最優解。

而使用 SGD 的話，2000 次所達成的效果與跑 20000 次的效果差頗多，代表 2000 次會離最優解比較遠，要跑到 20000 次才會差不多到最優解。

## Bonus

NO: \_\_\_\_\_  
DATE: / /

Bonus:

(a)

$$\begin{aligned} X^T X w_{\text{lin}} &= X^T y \\ \because X = U P V^T, X^T &= (U P V^T)^T \\ \therefore (U P V^T)^T (U P V^T) w_{\text{lin}} &= (U P V^T)^T y \\ \Rightarrow V P^T U^T U P V^T w_{\text{lin}} &= V P^T U^T y \\ \text{Also, } P &\text{ 是對角矩陣, 且 } U^T U = I \end{aligned}$$

$$\begin{aligned} \Rightarrow V P^T V^T w_{\text{lin}} &= V P^T U^T y \\ \Rightarrow P^T V^T w_{\text{lin}} &= P^T U^T y \\ \Rightarrow V^T w_{\text{lin}} &= P^T U^T y \\ \Rightarrow w_{\text{lin}} &= V P^T U^T y \end{aligned}$$

(b)

已知  $w_{\text{lin}} = X^T y$   
 設有一個  $w$  滿足  $X^T X w = X^T y$  (題目)  
 可推得

$$\begin{aligned} \Rightarrow X^T w &= y \\ \Rightarrow X^T X^T w &= X^T y \\ \Rightarrow X^T X^T w &= w_{\text{lin}} \end{aligned}$$

也就是說將  $w$  這個向量投影到  $X^T X$  這個維度 (線代)  
 而其長度必定不可能變短  
 將向量  $w$  投影到  $X^T X$  這個維度而有  $w_{\text{lin}}$   
 所以  $\|w\| \geq \|w_{\text{lin}}\|$  且