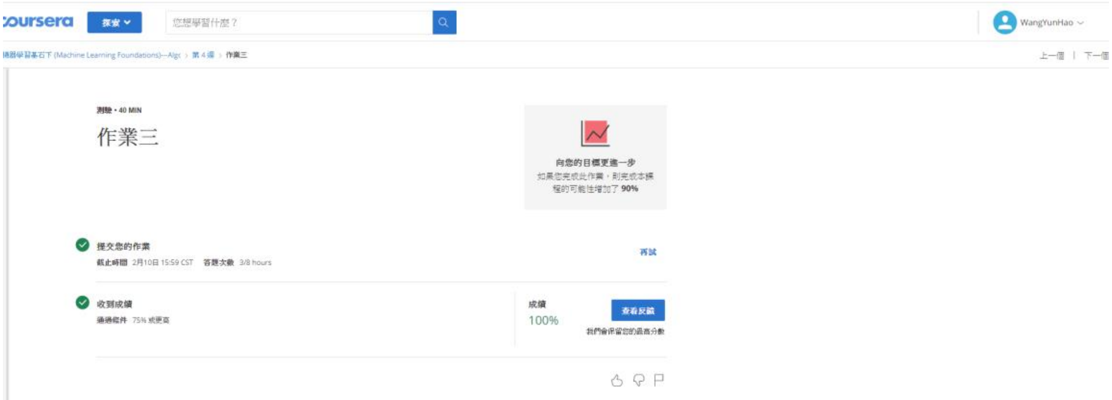
1.


oursera
探索
您想學習什麼？
WangYunHao
測驗 • 40 MIN
作業三
向您的目標更進一步
如果您完成此作業，則完成本課程的可能性增加了 90%
提交您的作業
截止時間 2月10日 15:59 CST 答題次數 3/8 hours
再試
收到成績
通過條件 75% 或更高
成績
100%
查看反饋
我們會保留您的最高分數

2.

2.

PLA: $W_{t+1} = W_t + [y \neq \mathrm{sign}(W_t^T x)]\, yx$

SGD: $W_{t+1} = W_t - \eta \nabla E_{in}(W_t)$. η: learning rate

$y=+1,\ w_t^T x > 0$ $\begin{cases} y \neq \mathrm{sign}(w_t^T x) \to \text{false} \Rightarrow W_{t+1} = W_t \\ \nabla E_{in} = 0 \Rightarrow W_{t+1} = W_t \end{cases}$

$y=-1,\ w_t^T x < 0$ $\begin{cases} y \neq \mathrm{sign}(w_t^T x) \to \text{true} \Rightarrow W_{t+1} = W_t + x \\ \nabla E_{in} = \nabla err(w,x,y) = \nabla(-w^T x) = -x \Rightarrow W_{t+1} = W_t + x. \end{cases}$

$y=-1,\ w_t^T x > 0$ $\begin{cases} y \neq \mathrm{sign}(w_t^T x) \to \text{true} \Rightarrow W_{t+1} = W_t - x \\ \nabla E_{in} = \nabla err(W,x,y) = \nabla(w^T x) = x \Rightarrow W_{t+1} = W_t - x \end{cases}$

$y=-1,\ w_t^T x < 0$ $\begin{cases} y \neq \mathrm{sign}(w_t^T x) \to \text{false} \Rightarrow W_{t+1} = W_t \\ \nabla E_{in} = 0 \Rightarrow W_{t+1} = W_t \end{cases}$

$\to$ 由上得知, SGD using error function $\max(0, -y w^T x)$ results in PLA.

3. $\hat{E}_2(\Delta u, \Delta v) = b_{uu}(\Delta u)^2 + b_{vv}(\Delta v)^2 + b_{uv}(\Delta u)(\Delta v) + b_u \Delta u + b_v \Delta v + b. \approx E(u+\Delta u, v+\Delta v)$

target: $\nabla \hat{E}_2(\Delta u, \Delta v) = 0 \longleftrightarrow \nabla E(u+\Delta u, v+\Delta v) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$\nabla E(u+\Delta u, v+\Delta v) \approx \begin{bmatrix} b_{uu}\Delta u + b_{uv}\Delta v + b_u \\ b_{vv}\Delta v + b_{uv}\Delta u + b_v \end{bmatrix} = \begin{bmatrix} b_u \\ b_v \end{bmatrix} + \begin{bmatrix} b_{uu} & b_{uv} \\ b_{uv} & b_{vv} \end{bmatrix} \times \begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix} = \nabla E(u,v) + \nabla^2 E(u,v) \begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix}$

$\because \nabla^2 E(u,v)$ is positive define $\therefore \lambda > 0,\ \lambda \in \lambda(\nabla^2 E(u,v))$. $\because \lambda > 0, \lambda \in \lambda(\nabla^2 E(u,v)) \therefore \nabla^2 E(u,v)^{-1}$ exits

$\therefore \begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix} = -(\nabla^2 E(u,v))^{-1} \nabla E(u,v)$

4. $y = \{1, 2, \dots, k\}$. $h_y(x) = (\exp(w_y^T x)) / (\sum_{k=1}^{k} \exp(w_k^T x))$

$\to$ maximum likelihood of $h_y(x)$

$\to \max_h \prod_{n=1}^{N} h_y(x_n) \to -\min \frac{1}{N} \ln \prod_{n=1}^{N} \exp(w_y^T x_n) - \ln \prod_{n=1}^{N} (\sum_{k=1}^{k} \exp(w_k^T x_n))$

$\longrightarrow -\min \frac{1}{N} \sum_{n=1}^{N} w_y^T x_n - \sum_{n=1}^{N} \ln(\sum_{k=1}^{k} \exp(w_k^T x_n)) \Rightarrow \min \frac{1}{N} \sum_{n=1}^{N} (\ln(\sum_{k=1}^{k} \exp(w_k^T x_n) - w_y^T x_n)$

$\therefore E_{in}(w_1, \dots, w_k) = \frac{1}{N} \sum_{n=1}^{N} \ln(\sum_{k=1}^{k} \exp(w_k^T x_n) - w_y^T x_n)$.

5. $x = \{x_1, \dots, x_n\}^T$. $y = \{y_1, \dots, y_n\}^T$. $\tilde{x} = \{\hat{x}_1 \dots \hat{x}_k\}^T$. $\tilde{y} = \{\hat{y}_1 \dots \tilde{y}_k\}^T$.

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

$\Rightarrow \min_w \frac{1}{N+k} [(w^T X^T X w + 2 w^T X^T y + y^T y) + (w^T \tilde{x}^T \hat{x} w + 2 w^T \tilde{x}^T \hat{y} + \tilde{y}^T y)] = E_{in}(w)$

$\nabla E_{in}(w) = \frac{2}{N+k} (X^T X w - X^T y + \tilde{x}^T \tilde{x} w - \hat{x}^T \tilde{y}) = 0 \Rightarrow (x^T x + \hat{x}^T \hat{x}) w = x^T y + \tilde{x}^T \hat{y}$

$w = (x^T x + \tilde{x}^T \hat{x})^{-1} (x^T y + \hat{x}^T \hat{y})$.

6.

6. $w_{reg} = \arg\min_w \frac{\lambda}{N}\|w\|^2 + \frac{1}{N}\|Xw - y\|^2$

$\longrightarrow \frac{2\lambda}{N} w_{reg} + \frac{2}{N} X^T(Xw_{reg} - y) = 0$

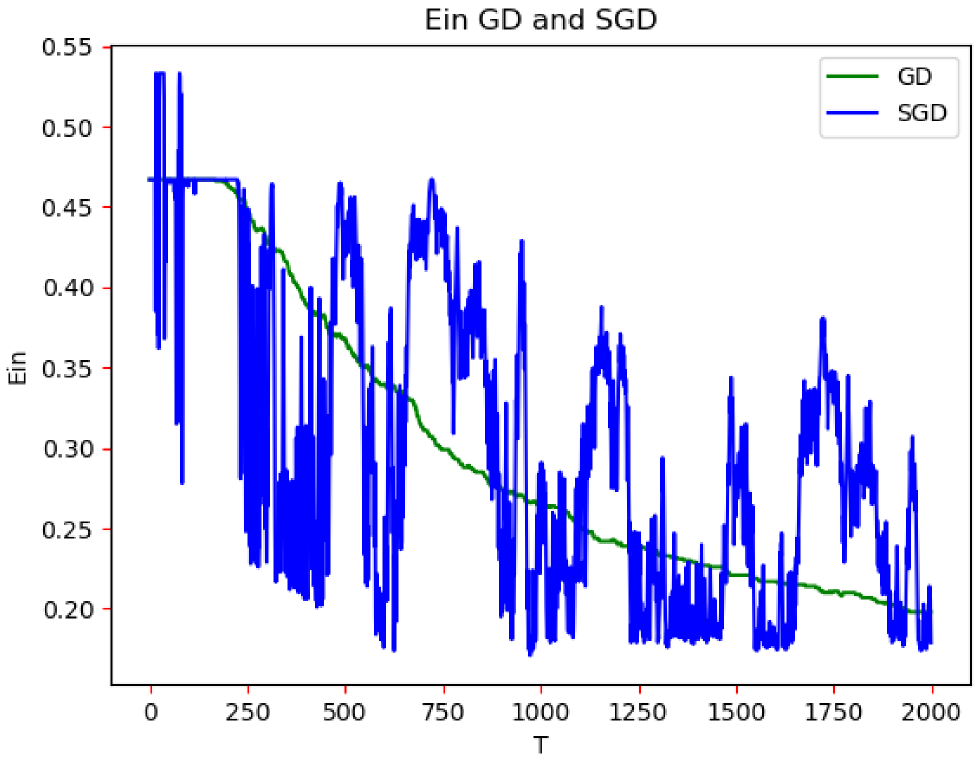
$\longrightarrow \frac{2\lambda}{N} w_{reg} + \frac{2}{N}\left(X^TX w_{reg} - X^Ty\right) = 0$

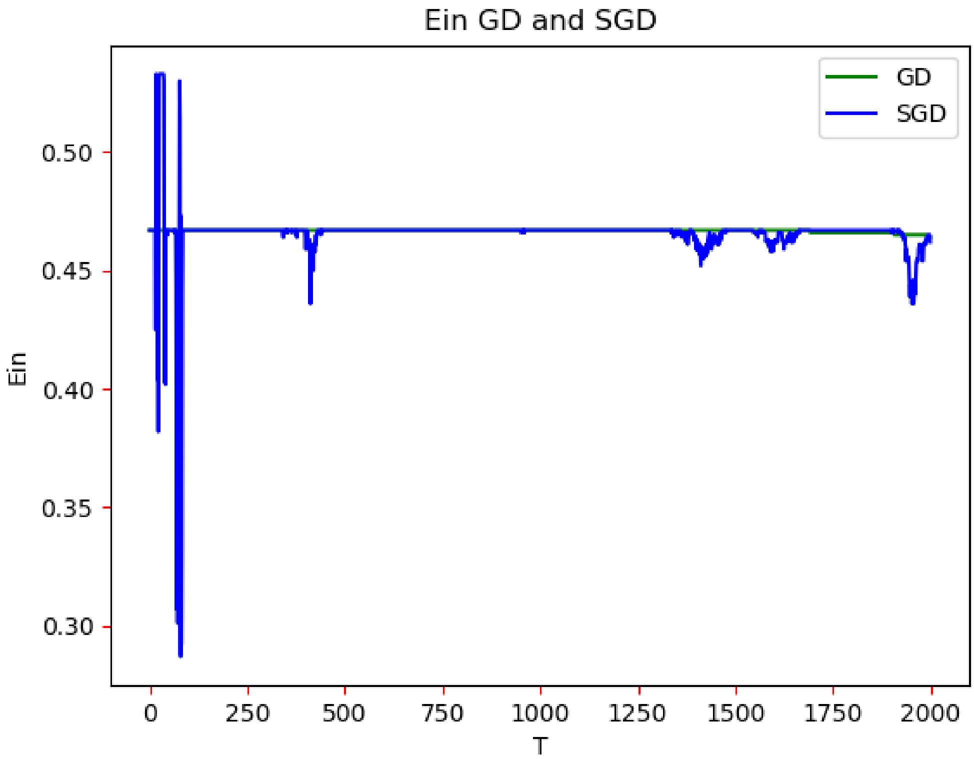
$\longrightarrow \lambda w_{reg} + X^TX w_{reg} \; X^Ty = 0$

$\longrightarrow (\lambda I + X^TX) w_{reg} - X^Ty = 0$

$\longrightarrow w_{reg} = (\lambda I + X^TX)^{-1} X^Ty \Rightarrow w_{reg} = w \Rightarrow \hat{X} = \sqrt{\lambda} I, \; \hat{y} = 0$

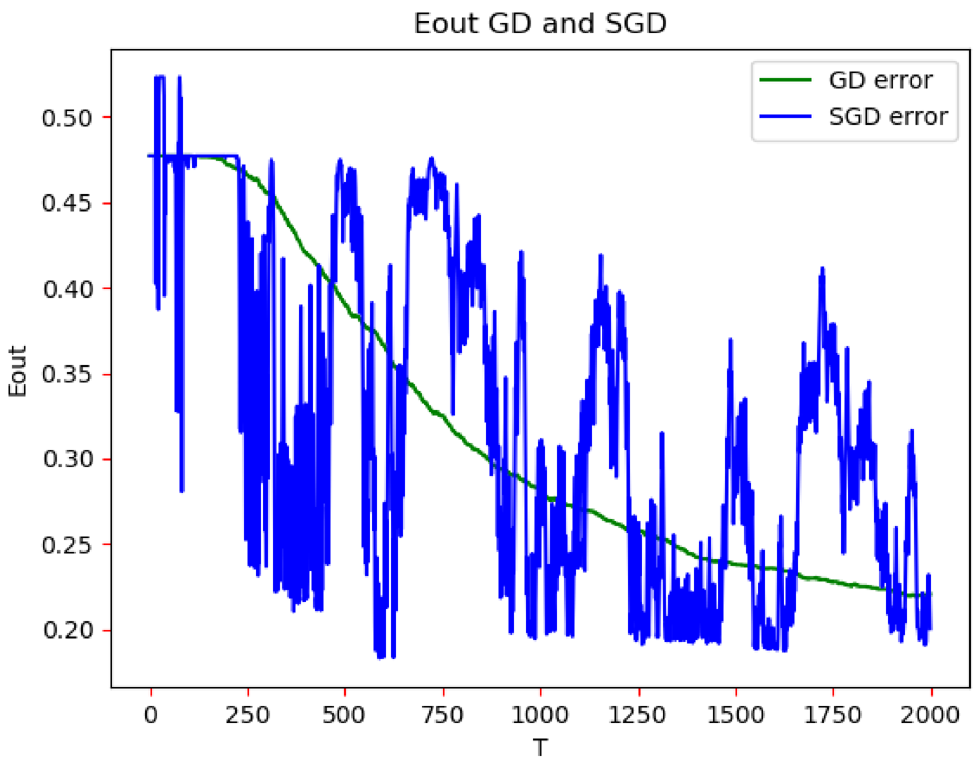
7.

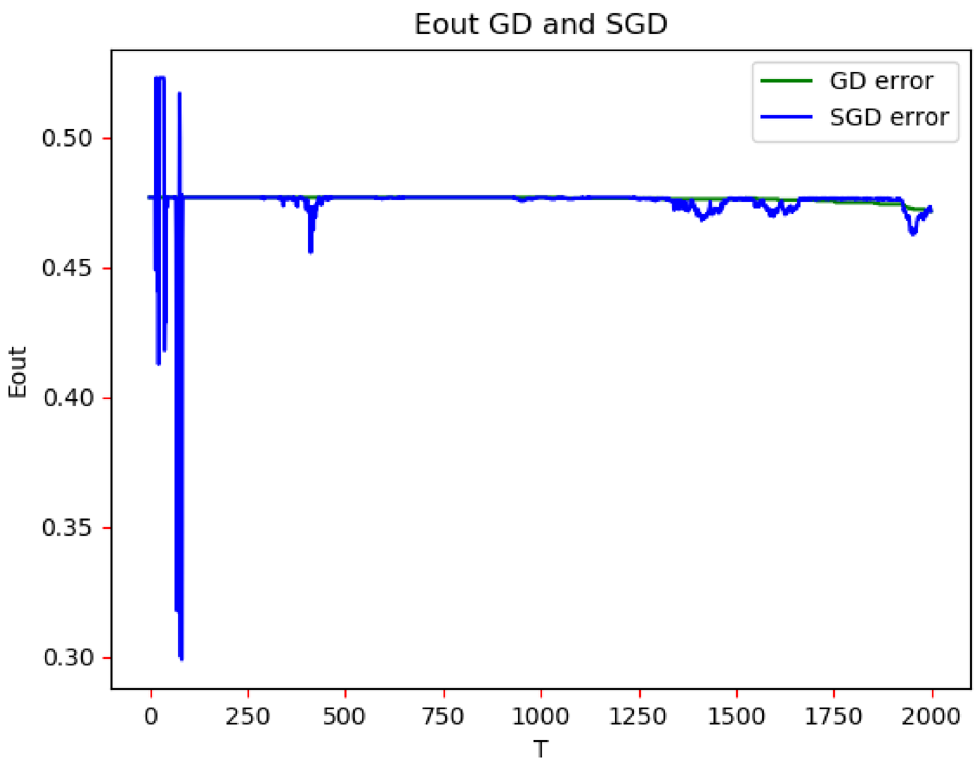

Ein GD and SGD
GD
SGD
Ein
T
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0
250
500
750
1000
1250
1500
1750
2000


Ein GD and SGD
GD
SGD
Ein
T
0.50
0.45
0.40
0.35
0.30
0
250
500
750
1000
1250
1500
1750
2000

Learning rate 0.01 時，GD 與 SGD 的 error rate 有明顯的下降，但 Learning rate 0.001 時，error rate 變化不明顯，但 SGD 還是有機會達到 error 較低的時候。

8.





Learning rate 0.01 時，SGD 有較明顯的震盪，Learning rate 0.001 時，SGD 前幾次更新有較劇烈的震盪，但最後 error rate 趨近 GD。