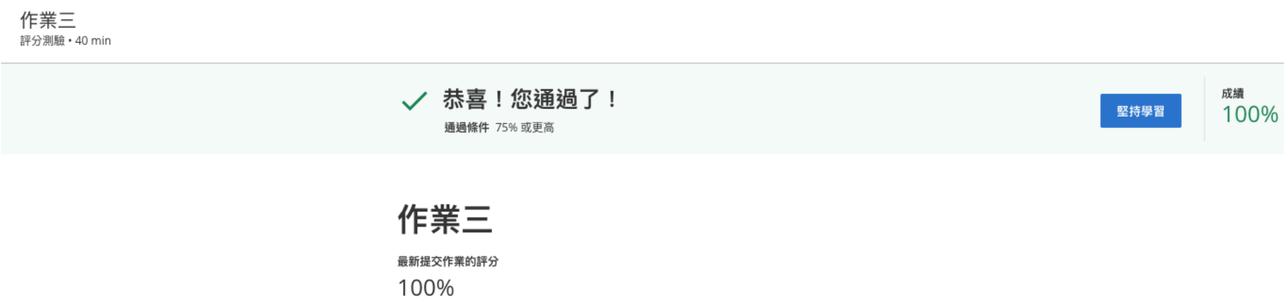
r08944035 呂翊愷

Machine Learning Foundations
Homework #3

1.



2.

當$w^T x$與y同號時，代表y點分類正確，故不修正，梯度為 0。

當$w^T x$與y異號時，則代表y點分類錯誤，需作修正，故對$-yw^T x$作微分。

即$\dfrac{\partial - yw^T x}{\partial w} = -yx \Rightarrow w_{t+1} = w_t + \eta(-yx)$

$\therefore err(w) = max(0, -yw^T x)$ results in PLA

3.

一元的二階泰勒展開式：

$$f(x+\Delta x) = f(x) + f'(x)\Delta x + \frac{1}{2} f''(x)\Delta x^2$$

當Δx趨近於0，上式等價於：

$$f'(x) + f''(x)\Delta x = 0$$

$$\Rightarrow \Delta x = -\frac{f'(x_n)}{f''(x_n)}$$

推廣到二元，

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

4.

根據課程投影片10中的第11頁，

cross entropy error $= \min\limits_{w} \left\{ \dfrac{1}{N} \sum_{n=1}^{N} -\ln\left(\theta\left(y_n w_y^T x_n\right)\right) \right\}$

又根據課程投影片10中的第10頁，$h_y(x_n) = \theta(y_n w_y^T x_n)$

已知本題的$h_y(x_n) = \dfrac{\exp(w_y^T x)}{\sum_{i=1}^{K} \exp(w_i^T x)}$，將此式代入cross entropy error的$\theta(y_n w_y^T x_n)$

$\Rightarrow$ cross entropy error = $\dfrac{1}{N} \sum_{n=1}^{N} \left(-\ln\left(\dfrac{\exp(w_{y_n}^T x)}{\sum_{i=1}^{K} \exp(w_i^T x)} \right) \right)$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(-\ln\left(\exp(w_{y_n}^T x_n)\right) + \ln\left(\sum_{i=1}^{K}\exp(w_i^T x_n)\right)\right)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(\left(\ln\sum_{i=1}^{K}\exp(w_i^T x_n)\right) - w_{y_n}^T x_n\right)$$

5.

根據課程投影片9中的第7頁，

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

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

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

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

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

$$\nabla E_{in} = 0$$

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

$$\Rightarrow \left(X^T X + \tilde{X}^T\tilde{X}\right) w = X^T y + \tilde{X}^T\tilde{y}$$

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

6.

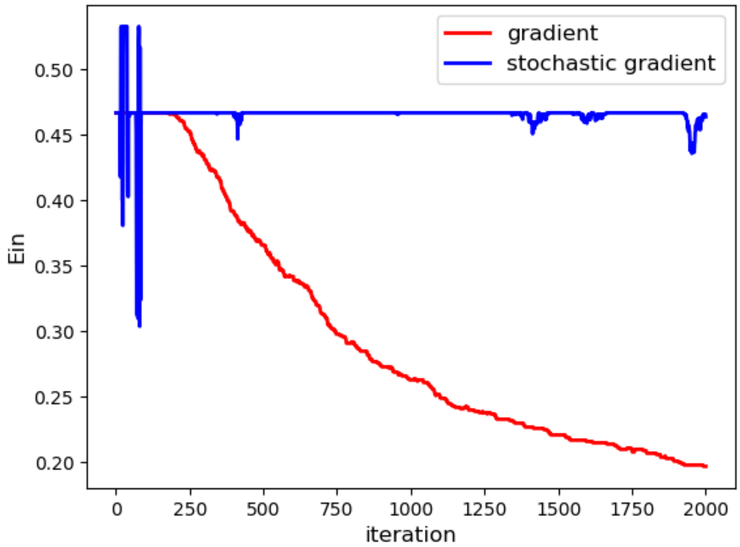
本題的w和上一題的w等價，

所以將$\tilde{X} = \sqrt{\lambda} I, \tilde{y} = 0$代入上一題的$w = \left(X^T X + \tilde{X}^T\tilde{X}\right)^{-1}\left(X^T y + \tilde{X}^T\tilde{y}\right)$

$$\Rightarrow w = (X^T X + \sqrt{\lambda} I \sqrt{\lambda} I)^{-1}(X^T y + (\sqrt{\lambda} I)^T 0) = (X^T X + \lambda I)^{-1} X^T y$$

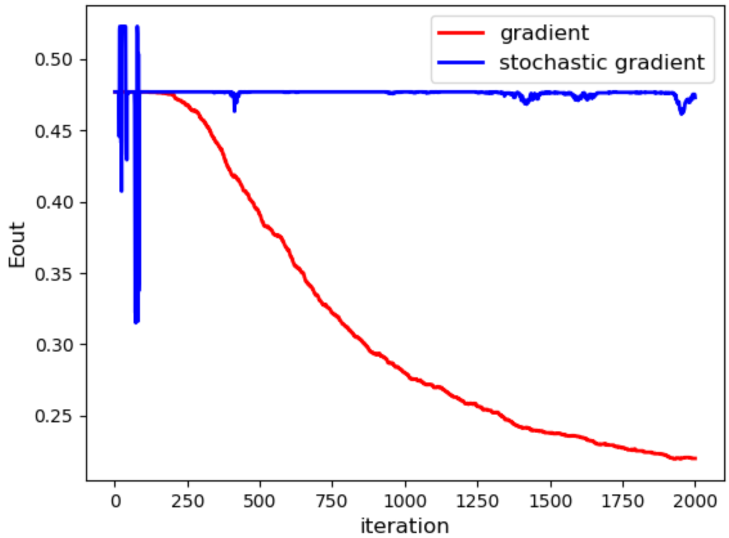
與課程投影片14的第10頁中Augmented Error的optimal solution相符

7.



根據上圖的結果，可以發現用gradient descent去更新w的話，Ein確實會隨著越來越小。然而，用stochastic gradient descent更新w在前幾次讓Ein十分不穩，一直到更新2000次的時候Ein都沒有下降。我認為stochastic gradient的效果會不如預期可能是因為learning rate較小，且iteration的次數只有2000次，或許對stochastic gradient來說不夠多次，可能更新次數要再多一點才會有較明顯的效果。

8.



根據上圖的結果，可以發現跟上一題的Ein相比，無論是gradient descent或是stochastic gradient descent所造成的Eout走向都和上一題差不多，唯獨Eout都會比Ein高一點點。其實這樣的效果蠻正常的，因為是使用training data去計算gradient，所以同樣用training data去算出Error的話本來就會比較低一點。

r08944035 呂翊愷

9.

(a)

已知 $X^T X w_{lin} = X^T y$，

又已知 $X = U\Gamma V^T, w_{lin} = V\Gamma^{-1}U^T y$，代入左式

左式 $= (U\Gamma V^T)^T(U\Gamma V^T)V\Gamma^{-1}U^T y = V\Gamma^T U^T U\Gamma V^T V\Gamma^{-1}U^T y$

又已知$U^T U = I, V^T V = I$，

$\therefore$ 左式$=V\Gamma^T\Gamma\Gamma^{-1}U^T y = V\Gamma^T U^T y$=右式

故得證

(b)

若$X^T Xw = X^T y$，則$\|Xw - y\|$會是最小值

$\Leftrightarrow Xw = Proj_{R(X)}y$

$\Leftrightarrow (y - Xw) \perp R(X)$

$\Leftrightarrow \langle y - Xw, Xb\rangle = 0,\ \forall b : (d+1)\times 1$

$\Leftrightarrow (Xb)^T(y - Xw) = 0,\ \forall b : (d+1)\times 1$

$\Leftrightarrow b^T X^T y - b^T X^T Xw = 0,\ \forall b : (d+1)\times 1$

$\Leftrightarrow b^T(X^T y - X^T Xw) = 0,\ \forall b : (d+1)\times 1$

$\Leftrightarrow \langle X^T y - X^T Xw, b\rangle = 0,\ \forall b : (d+1)\times 1$

$\Leftrightarrow X^T y - X^T Xw = 0$

$\Leftrightarrow X^T Xw = X^T y$