MLFoundation HW3

r07922100 楊力權

1.

2.

SGD : $w_{t+1} \leftarrow w_t - \eta \nabla E_{in}$ PLA : $w_{t+1} \leftarrow w_t + (y_n x_n)[[y_n \neq sign(w_t^T x_n)]]$

$\because err(w) = \max(0, -yw^Tx)$ $\therefore \nabla E_{in} = [[yw^Tx < 0]](-yx) = [[sign(w^Tx) \neq y]](-yx)$

SGD每次只使用一筆資料(x_n,y_n)更新w。

SGD : $w_{t+1} \leftarrow w_t - \eta[[sign(w^Tx_n) \neq y_n]](-y_nx_n) = w_t + \eta[[sign(w^Tx_n) \neq y_n]](y_nx_n)$

若learning rate $\eta = 1$則SGD在使用$err(w) = \max(0, -yw^Tx)$下等同於PLA。

3.

面對N筆資料，最大化likelihood，也就是最小化negative likelihood

$$\arg\max_w likelihood(w) \propto \prod_{n=1}^{N} h_{y_n}(y_n w^T x_n) \Rightarrow \arg\max_w \ln \prod_{n=1}^{N} h_{y_n}(y_n w^T x_n) = \arg\min_w \sum_{n=1}^{N} -\ln \frac{e^{(w_{y_n}^T x_n)}}{\sum_{k=1}^{K} e^{(w_k^T x_n)}}$$

$$= \arg\min_w \sum_{n=1}^{N} \ln \frac{\sum_{k=1}^{K} e^{(w_k^T x_n)}}{e^{(w_{y_n}^T x_n)}} = \arg\min_w \sum_{n=1}^{N} (\ln(\sum_{k=1}^{K} e^{(w_k^T x_n)}) - (w_{y_n}^T x_n))$$

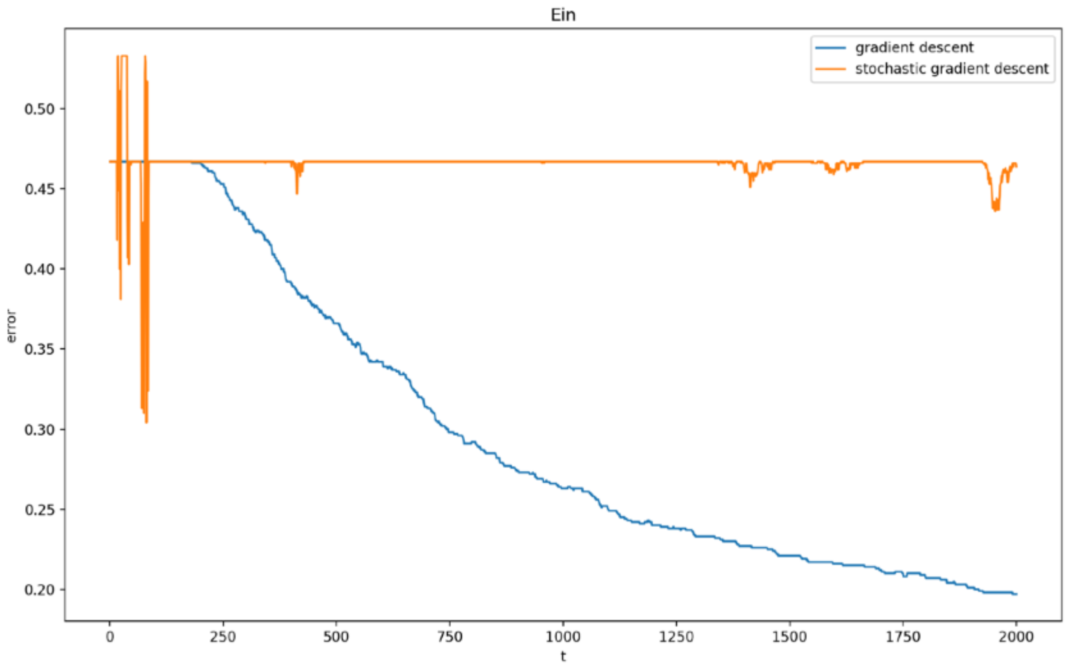
$$\text{error} = \sum_{n=1}^{N} (\ln(\sum_{k=1}^{K} e^{(w_k^T x_n)}) - (w_{y_n}^T x_n)) \Rightarrow E_{in} = \frac{\text{error}}{N} = \frac{1}{N} \sum_{n=1}^{N} (\ln(\sum_{k=1}^{K} e^{(w_k^T x_n)}) - (w_{y_n}^T x_n))$$

使用連鎖率求E_{in}對w_i的微分

$$\frac{\partial E_{in}}{\partial w_i} = \frac{1}{N} \sum_{n=1}^{N} (\frac{\partial}{\partial w_i} \ln(\sum_{k=1}^{K} e^{(w_k^T x_n)}) - \frac{\partial}{\partial w_i}(w_{y_n}^T x_n)) = \frac{1}{N} \sum_{n=1}^{N} (\frac{1}{\sum_{k=1}^{K} e^{(w_k^T x_n)}} \frac{\partial}{\partial w_i}(\sum_{k=1}^{K} e^{(w_k^T x_n)}) - [[y_n = i]]x_n)$$

$$= \frac{1}{N} \sum_{n=1}^{N} (\frac{e^{(w_i^T x_n)}}{\sum_{k=1}^{K} e^{(w_k^T x_n)}})x_n - [[y_n = i]]x_n) = \frac{1}{N} \sum_{n=1}^{N} (h_i(x_n)x_n - [[y_n = i]]x_n) = \frac{1}{N} \sum_{n=1}^{N} (h_i(x_n) - [[y_n = i]])x_n$$

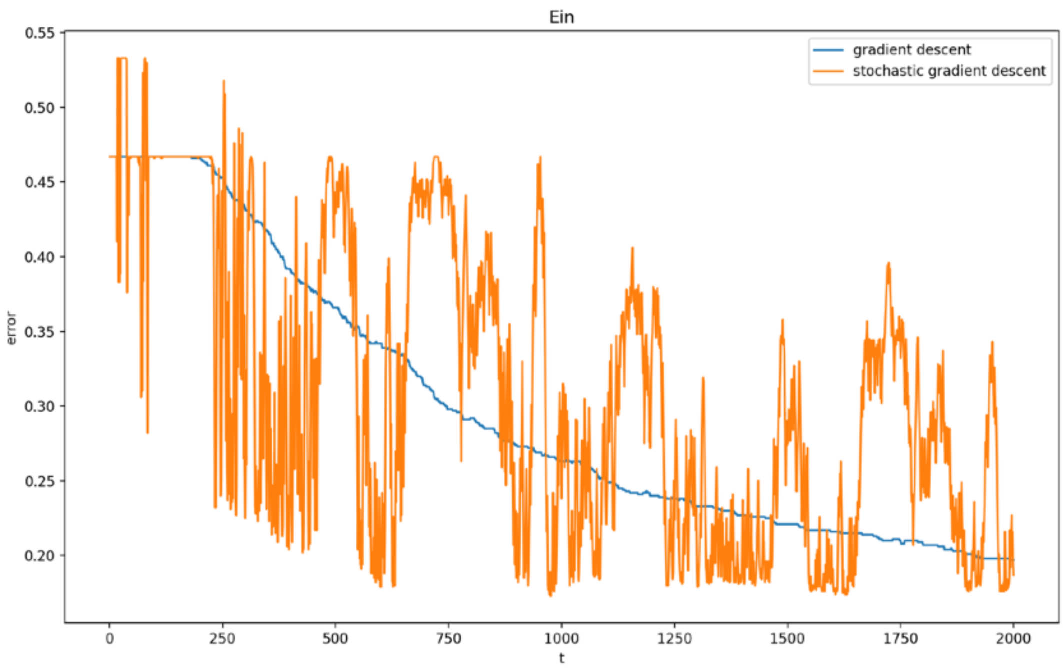
4.



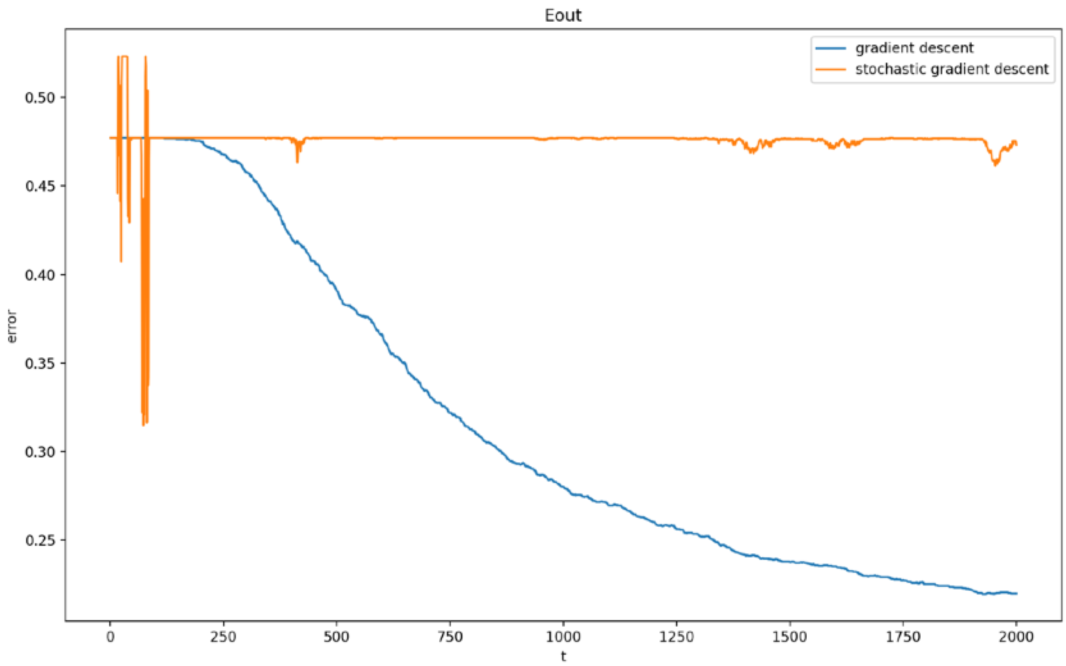
GD使用lr=0.01，而SGD使用lr=0.001。

gradient descent每個iteration使用所有的數據來計算平均梯度方向，因此訓練穩定，而SGD跳動劇烈且最後幾乎沒有降低Ein。

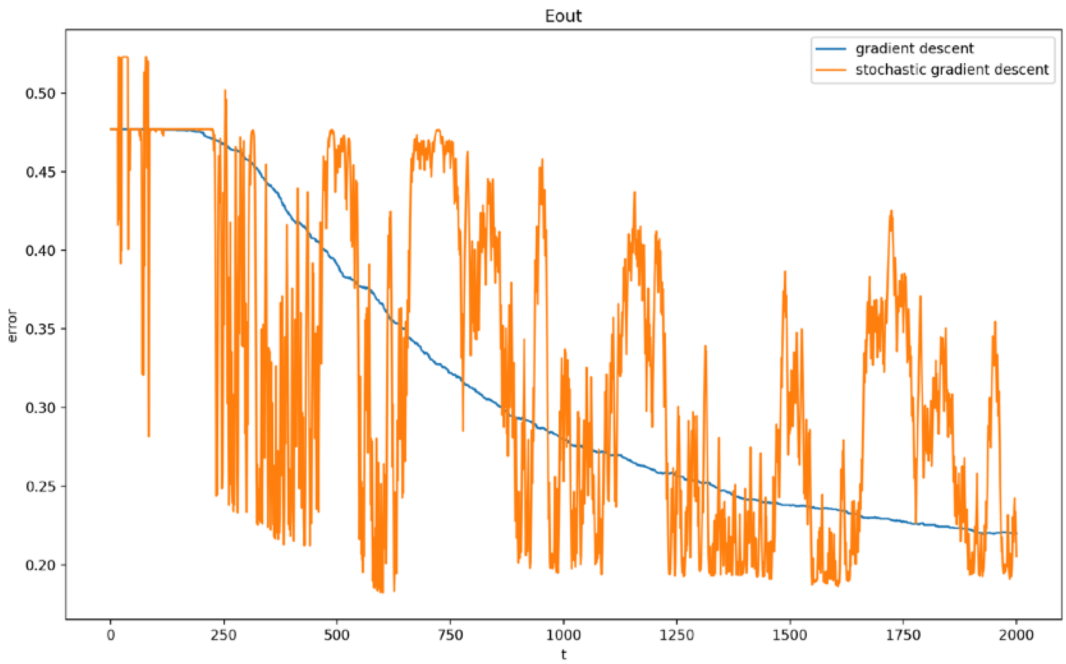
我認為這應該是出題時的問題，因為當SGD也使用lr=0.01時，產生下圖，GD與SGD的Ein皆有下降，而明顯不同之處是SGD非常不穩定的震盪，因為每筆數據有不同程度雜訊的緣故，GD能夠平均掉雜訊對梯度帶來的極端影響，而SGD只使用一筆數據對W的梯度。但是即便如此兩者都能成功降低Ein。



5.



上圖為GD lr=0.01，SGD lr=0.001；下圖為GD lr=0.01，SGD lr=0.01。
結果與第4題非常接近，因此觀察討論與第4.題一樣。而Eout比Ein高出一點點但差距不多，可以說明20維的資料在資料夠多時，可以做到Ein與Eout差不多的結果。



6.

設$x=[d_1,d_2,\cdots]\in R^d$，$X=\begin{bmatrix}x_1\\x_2\\\vdots\end{bmatrix}\in R^{N\times d}$，$h=\begin{bmatrix}u_1\\u_2\\\vdots\end{bmatrix}\in R^d$，$H=[h_1,h_2,\cdots]\in R^{d\times K}$

$W=\begin{bmatrix}w_1\\w_2\\\vdots\end{bmatrix}\in R^K$，$Y\in R^N$，$XHW\in R^N$

$$\text{RMSE}(H)=\sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n-\sum_{k=1}^{K}w_kh_k(x_n))^2}=\sqrt{\frac{1}{N}\|Y-XHW\|^2}$$

Optimal：$\nabla E_{in}=0\Rightarrow(XH)^T(XH)W=(XH)^TY\Rightarrow W=((XH)^T(XH))^{-1}(XH)^TY$

Y未知，但我們知道每個$e_k=RMSE(h_k)=\sqrt{\frac{1}{N}\|Y-Xh_k\|^2}=\sqrt{\frac{1}{N}(Y-Xh_k)^T(Y-Xh_k)}$

$$\Rightarrow Ne_k^2=Y^TY+(Xh_k)^T(Xh_k)-2(Xh_k)^TY$$

$$\Rightarrow N(e_k^2-e_m^2)=2(h_m^T-h_k^T)X^TY+(Xh_k)^T(Xh_k)-(Xh_m)^T(Xh_m)$$

且我們知道$h_0(x)=0$，因此把m用0代入

$$\Rightarrow N(e_k^2-e_0^2)=-2h_k^TX^TY+(Xh_k)^T(Xh_k)\Rightarrow(Xh_k)^TY=\frac{(Xh_k)^T(Xh_k)-N(e_k^2-e_0^2)}{2}$$

我們可以做一個矩陣如下(Y左方的矩陣必須Rank=N否則det=0無反矩陣，因此若Xh_k與$\{Xh_m|m<k\}$平行則必須跳過k)

$$\begin{bmatrix}(Xh_1)^T\\(Xh_2)^T\\\vdots\\(Xh_N)^T\end{bmatrix}Y=\begin{bmatrix}\frac{(Xh_1)^T(Xh_1)-N(e_1^2-e_0^2)}{2}\\\frac{(Xh_2)^T(Xh_2)-N(e_2^2-e_0^2)}{2}\\\vdots\\\frac{(Xh_N)^T(Xh_N)-N(e_N^2-e_0^2)}{2}\end{bmatrix}\Rightarrow Y=\begin{bmatrix}(Xh_1)^T\\(Xh_2)^T\\\vdots\\(Xh_N)^T\end{bmatrix}^{-1}\begin{bmatrix}\frac{(Xh_1)^T(Xh_1)-N(e_1^2-e_0^2)}{2}\\\frac{(Xh_2)^T(Xh_2)-N(e_2^2-e_0^2)}{2}\\\vdots\\\frac{(Xh_N)^T(Xh_N)-N(e_N^2-e_0^2)}{2}\end{bmatrix}$$

把Y代回

$$W=((XH)^T(XH))^{-1}(XH)^TY=((XH)^T(XH))^{-1}(XH)^T\begin{bmatrix}(Xh_1)^T\\(Xh_2)^T\\\vdots\\(Xh_N)^T\end{bmatrix}^{-1}\begin{bmatrix}\frac{(Xh_1)^T(Xh_1)-N(e_1^2-e_0^2)}{2}\\\frac{(Xh_2)^T(Xh_2)-N(e_2^2-e_0^2)}{2}\\\vdots\\\frac{(Xh_N)^T(Xh_N)-N(e_N^2-e_0^2)}{2}\end{bmatrix}$$

其中$X=\begin{bmatrix}x_1\\x_2\\\vdots\end{bmatrix}\in R^{N\times d}$，$H=[h_1,h_2,\cdots]\in R^{d\times K}$