2.

分以下兩種情形來討論：

(i) err = 0: 不用做更新

(ii) err≠0:

其梯度為$\nabla \mathrm{err}(\mathrm{w}) = -\mathrm{yx}$。故若是今天有一筆資料$(x_n, y_n)$要拿其去做 SGD 進行修正，結果會是如下：

$$w_{t+1} \leftarrow w_t - \eta(-y_n x_n)$$

而若是做 PLA 的修正，結果會是：

$$w_{t+1} \leftarrow w_t + y_n x_n$$

可以看出兩者修正的方向都是相同的。

3.

首先，由 16 題可知：

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

對 sigma 內前項做微分，可得：(由鏈鎖律)

$$\frac{1}{\sum_{k=1}^{K} e^{w_k^T x_n}}(x_n e^{w_k^T x_n}) = h_i(x_n)$$

對 sigma 內後項做微分：

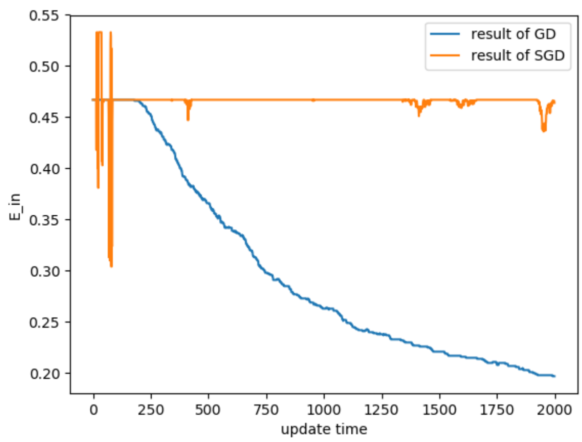
(i) 若$\mathrm{w}_{\mathrm{y_n}} = w_i$，對$\mathrm{w_i}$微分則為$x_n$

(ii) 若$\mathrm{w}_{\mathrm{y_n}} \neq w_i$，對$\mathrm{w_i}$微分則為 0

結合以上可得整個$\mathrm{E_{in}}$對$\mathrm{w_i}$的微分即為：

$$\frac{1}{N}\sum_{n=1}^{N}(h_i(x_n) - [[\, y_n = i\,]]x_n)$$

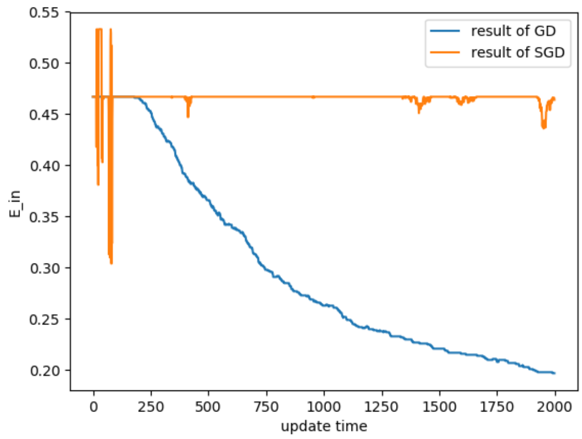
4.



(i) 使用 stochastic gradient descent(SGD)在一開始做更新時，變動幅度很大。而

且在做 2000 次以內的更新對於降低 E_in 的效果相當有限，E_in 大約都落在 0.46 附近。
(ii) 使用 gradient descent(GD)能夠穩定使 E_in 下降，可以可降低製 0.2 附近。

5.



(i) E_out 的趨勢和 E_in 相同，故不再贅述。
(ii) 大體而言，E_out 會大於 E_in。
*註：兩題程式我用同一個 code(45.py)來跑，並製出兩張圖。一張是 E_{in} 對更新次數，另一張是 E_{out} 對更新次數。

6.
Consulting: 林首志、陳義榮
令$Y = [y_1, y_2, \dots, y_n]^T$，$\bar{h}_i = [h_i(x_1), h_i(x_2), \dots, h_i(x_n)]^T$
故原本的函式可以表成：$(Y - h_i)^2 = Ne_i^2$

展開可得$Y^TY - 2\bar{h}_i^TY + \bar{h}_i^2 = Ne_i^2$。又因$\left(Y - \bar{h}_0\right)^2 = Y^TY = Ne_0^2$，故可得出：

$$\bar{h}_i^TY = N(e_0^2 - e_i^2) + \bar{h}_i^2$$

再令：(便於表示)

$$H_j' = \begin{bmatrix} h_1(x_j) \\ h_2(x_j) \\ \dots \\ h_k(x_j) \end{bmatrix}, H' = \begin{bmatrix} h_1(x_1) & h_1(x_2) & \dots & h_1(x_n) \\ h_2(x_1) & h_2(x_2) & \cdots & h_2(x_n) \\ \dots & \dots & & \dots \\ h_k(x_j) & h_k(x_2) & \dots & h_k(x_n) \end{bmatrix}, W = \begin{bmatrix} w_1 \\ w_2 \\ \dots \\ w_k \end{bmatrix}$$

而題目所求為 RMSE(H)的最小值，其極值發生點會和$(\mathrm{RMSE(H)})^2$相同，即要找以下式子：

$$\min_w \frac{1}{N} \sum_{j=1}^{N} (y_j - W^T H_j')^2$$

將其對w_i做偏微可得：

$$\frac{1}{N}\sum_{j=1}^{N} -2h_i(x_j)\,(y_j - w_i h_i(x_j)) = \frac{2}{N}\left(\bar{h}_i{}^T \bar{h}_i w_i - \bar{h}_i^T Y\right)$$

把全部的$\mathrm{w_i}$和在一起，可得其梯度：

$$\nabla(\mathrm{RMSE(H)})^2 = \frac{2}{\mathrm{N}}\left(H'H'^T W - H'Y\right)令 = 0$$

於是可得：

$$\mathrm{W} = \left(\mathrm{H'H'}^T\right)^{-1} H'Y$$

又因為前面曾算出：

$$\bar{\mathrm{h}}_\mathrm{i}^T Y = N(e_0^2 - e_i^2) + \bar{\mathrm{h}}\,_\mathrm{i}^2$$

故可得：

$$\mathrm{H'Y} = \begin{bmatrix} N(e_0^2 - e_1^2) + \bar{\mathrm{h}}\,_1^2 \\ N(e_0^2 - e_2^2) + \bar{\mathrm{h}}\,_2^2 \\ \cdots \\ N(e_0^2 - e_n^2) + \bar{\mathrm{h}}\,_\mathrm{n}^2 \end{bmatrix}$$

結合以上可得：

$$\mathrm{W} = \left(\mathrm{H'H'}^T\right)^{-1} \begin{bmatrix} N(e_0^2 - e_1^2) + \bar{\mathrm{h}}\,_1^2 \\ N(e_0^2 - e_2^2) + \bar{\mathrm{h}}\,_2^2 \\ \cdots \\ N(e_0^2 - e_n^2) + \bar{\mathrm{h}}\,_\mathrm{n}^2 \end{bmatrix}, 其中\mathrm{H'} = \begin{bmatrix} h_1(x_1) & h_1(x_2) & \dots & h_1(x_n) \\ h_2(x_1) & h_2(x_2) & \cdots & h_2(x_n) \\ \dots & \dots & & \dots \\ h_k(x_j) & h_k(x_2) & \dots & h_k(x_n) \end{bmatrix}$$