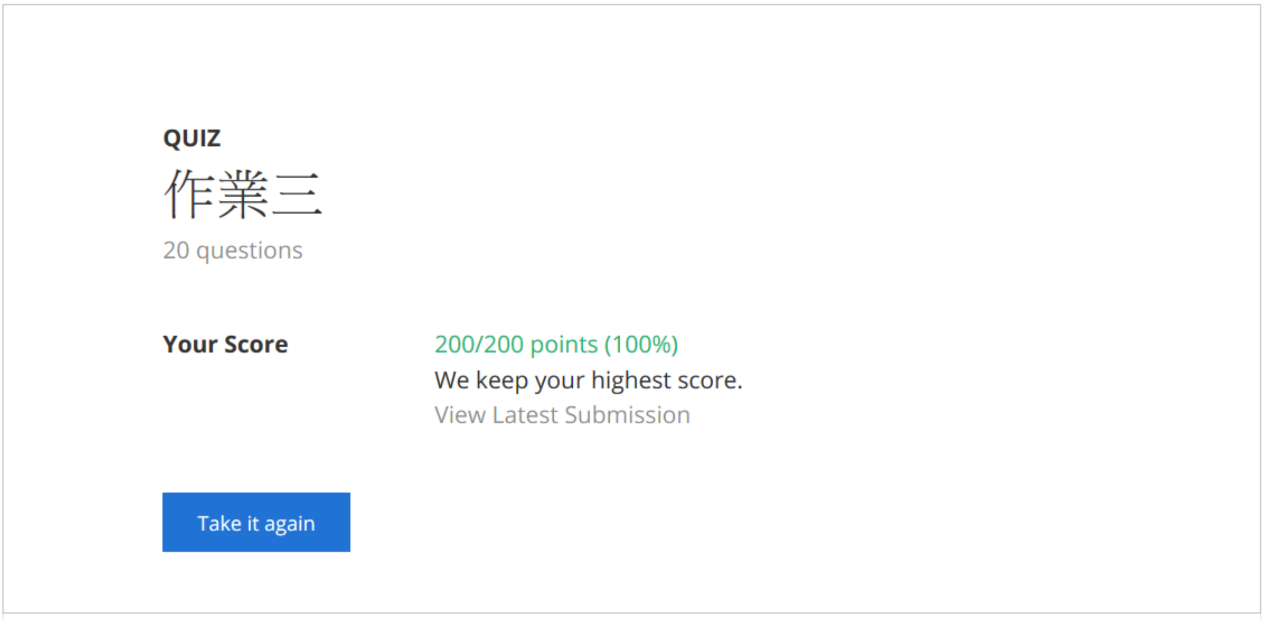
ML Foundation: HW3

b04902053 鄭淵仁

December 31, 2017

1



2

Claim: $H^2 = H$
Proof:

$$
\begin{aligned}
H^2 &= (X(X^TX)^{-1}X^T)^2 \\
&= (X(X^TX)^{-1}X^T)(X(X^TX)^{-1}X^T) \\
&= X(X^TX)^{-1}[(X^TX)(X^TX)^{-1}]X^T \\
&= X(X^TX)^{-1}X^T \\
&= H
\end{aligned}
$$

With the claim above, we can prove that:

$$
\begin{aligned}
(I-H)^2 &= I^2 - 2IH + H^2 \\
&= I - 2H + H \\
&= I - H
\end{aligned}
$$

3

TODO: 看不懂題目在寫什麼？QQ

4

$$\hat{E}_2(\Delta u, \Delta v) = E(u,v) + \nabla E(u,v)\cdot(\Delta u, \Delta v) + \frac{1}{2}(\Delta u, \Delta v)^T \nabla^2 E(u,v)(\Delta u, \Delta v)$$

Set the partial differences of $\hat{E}_2(\Delta u, \Delta v)$ be 0, we have :

$$\begin{cases} 0 = \dfrac{\partial \hat{E}_2(\Delta u, \Delta v)}{\partial \Delta u} = \dfrac{\partial E}{\partial u} + \dfrac{1}{2}\left(2\dfrac{\partial^2 E}{\partial u^2}\Delta u + 2\dfrac{\partial^2 E}{\partial u \partial v}\Delta v\right) \\ \qquad\qquad\qquad\quad = \dfrac{\partial E}{\partial u} + \dfrac{\partial^2 E}{\partial u^2}\Delta u + \dfrac{\partial^2 E}{\partial u \partial v}\Delta v \\ 0 = \dfrac{\partial \hat{E}_2(\Delta u, \Delta v)}{\partial \Delta v} = \dfrac{\partial E}{\partial v} + \dfrac{\partial^2 E}{\partial v^2}\Delta v + \dfrac{\partial^2 E}{\partial v \partial u}\Delta u \end{cases}$$

Simplify the equations :

$$\begin{cases} 0 = \dfrac{\partial E}{\partial u} + \dfrac{\partial^2 E}{\partial u^2}\Delta u + \dfrac{\partial^2 E}{\partial u \partial v}\Delta v \\ 0 = \dfrac{\partial E}{\partial v} + \dfrac{\partial^2 E}{\partial v^2}\Delta v + \dfrac{\partial^2 E}{\partial v \partial u}\Delta u \end{cases}$$

Now combine the two equations to one equation by vector (u, v) :

$$\begin{aligned} 0 &= \nabla E(u,v) + \nabla^2 E(u,v)\cdot(\Delta u, \Delta v) \\ -\nabla^2 E(u,v)\cdot(\Delta u, \Delta v) &= \nabla E(u,v) \\ (\Delta u, \Delta v) &= -\left(\nabla^2 E(u,v)\right)^{-1}\nabla E(u,v) \end{aligned}$$

Q.E.D.

5

$$\max_h \prod_{n=1}^{N} h_y(x_n) = \max_w \prod_{n=1}^{N} \frac{\exp\left(w_{y_n}^T x_n\right)}{\sum_{k=1}^{K}\exp(w_k^T x_n)}$$

Take natural log on the it :

$$\begin{aligned} &\max_w \ln \prod_{n=1}^{N} \frac{\exp\left(w_{y_n}^T x_n\right)}{\sum_{k=1}^{K}\exp(w_k^T x_n)} \\ &= \max_w \sum_{n=1}^{N} \ln \frac{\exp\left(w_{y_n}^T x_n\right)}{\sum_{k=1}^{K}\exp(w_k^T x_n)} \\ &= \max_w \sum_{n=1}^{N} \left(\ln\left(\exp\left(w_{y_n}^T x_n\right)\right) - \ln \sum_{k=1}^{K} \exp\left(w_k^T x_n\right)\right) \\ &= \min_w \sum_{n=1}^{N} \left(\ln \sum_{k=1}^{K} \exp\left(w_k^T x_n\right) - w_{y_n}^T x_n\right) \end{aligned}$$

Therefore the E_{in} is :

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

6

First we compute :

$$
\begin{aligned}
&\frac{\partial\left(\sum_{n=1}^{N}\left(\ln\sum_{k=1}^{K}\exp\left(w_k^T x_n\right)\right)\right)}{\partial w_i} \\
&= \sum_{n=1}^{N}\left(\frac{\exp\left(w_i^T x_n\right)}{\sum_{k=1}^{K}\exp(w_k^T x_n)} x_n\right) \\
&= \sum_{n=1}^{N}\left(h_i(x_n)x_n\right)
\end{aligned}
$$

Therefore the answer is :

$$
\begin{aligned}
\frac{\partial E_{in}}{\partial w_i} &= \frac{\partial\left(\frac{1}{N}\sum_{n=1}^{N}\left(\ln\sum_{k=1}^{K}\exp\left(w_k^T x_n\right) - w_{y_n}^T x_n\right)\right)}{\partial w_i} \\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\left(h_i(x_n)x_n\right) - [[y_n = i]]x_n\right) \\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\left(\left(h_i(x_n)\right) - [[y_n = i]]\right) x_n\right)
\end{aligned}
$$

7

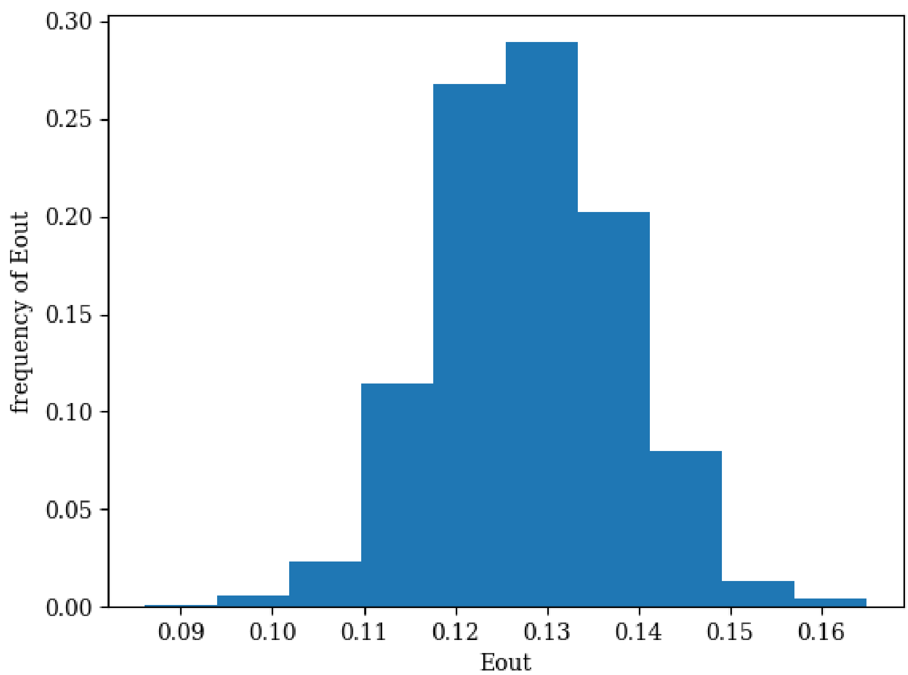
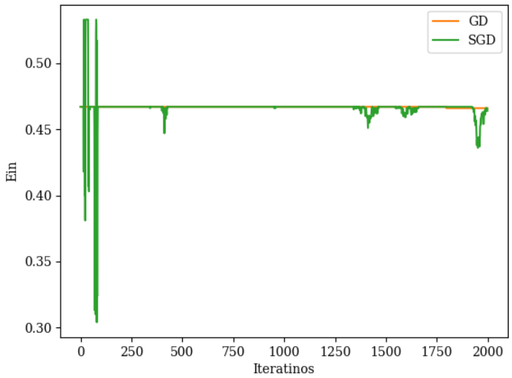


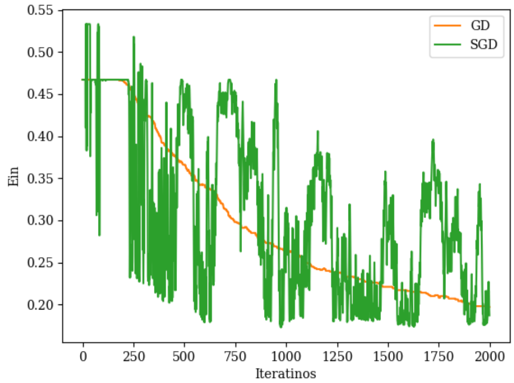
Figure 1: Histogram of E_{out}

Figure 1 shows the histogram of E_{out}.

8



(a) $lr = 0.001$



(b) $lr = 0.01$

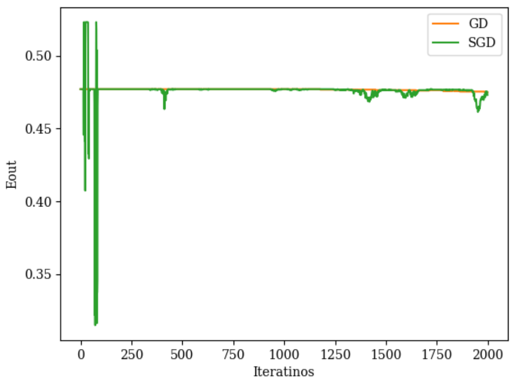
Figure 2: Comparison between GD and SGD in E_{in}.

My findings:

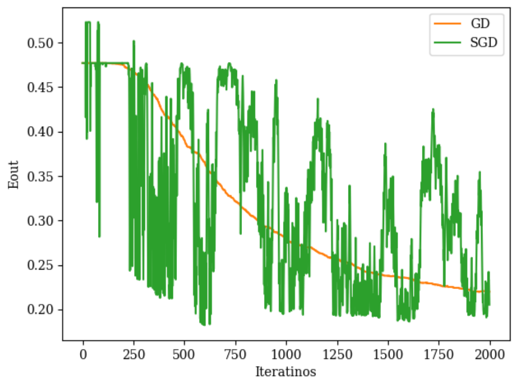
- GD 和 SGD 的差異：
  我發現 GD 的 E_{in} 很快就會穩定保持相同，或是穩定下降，而不會上下亂跳；相較之下，SGD 的 E_{in} 則是很容易上下浮動。
  我想這是因為 SGD 一次只會取一筆資料來計算 gradient ，如果這一筆資料有 noise 的話，算出來的 gradient 很容易會被 noise 影響；相較之下，GD 一次會用所有資料來計算 gradient ，依照 Hoeffding's Inequality 可以知道：取愈多資料，noise 有越高的機率會愈小，所以才會有這樣的結果。

- $lr = 0.001$ 和 $lr = 0.01$ 的差異：
  我發現 $lr = 0.001$ 的時候，除了剛開始 E_{in} 不穩定亂跳以外，

9



(a) $lr = 0.001$



(b) $lr = 0.01$

Figure 3: Comparison between GD and SGD in E_{out}.