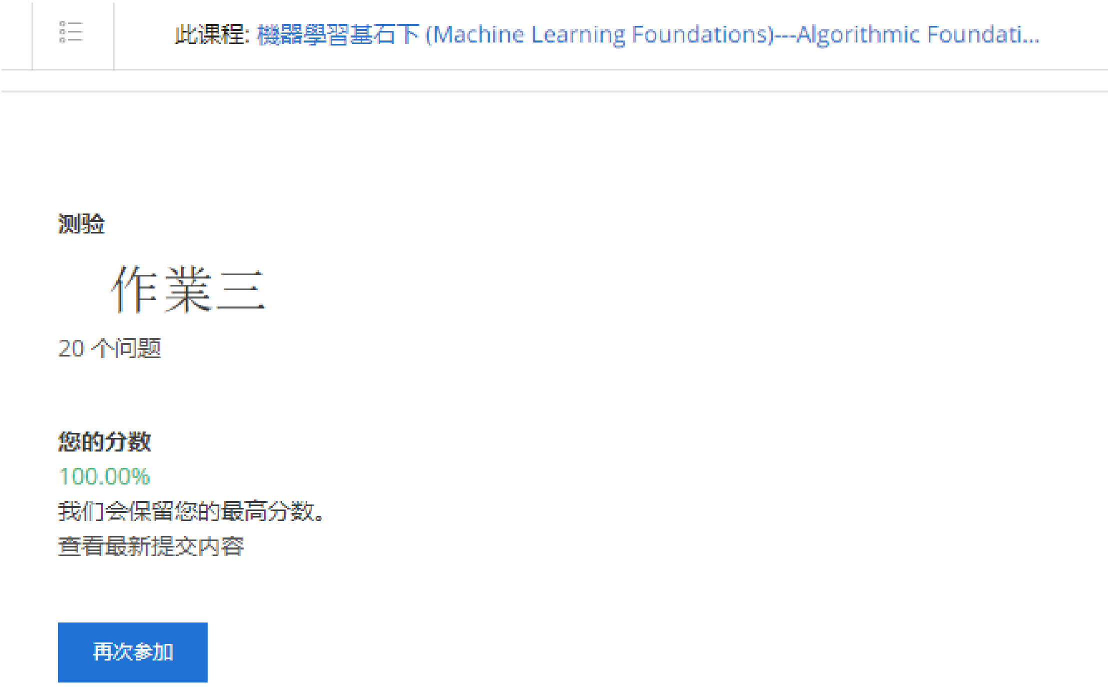
Machine Learning Foundation (NTU, Fall 2018) Homework #3

陳熙 R07922151

1.



2. 在 SGD 使用這個 error function 即，

$$
\begin{aligned}
\mathbf{w}_{t+1} &= \mathbf{w}_t - \nabla \operatorname{err}(\mathbf{w}_t) \\
&= \mathbf{w}_t - \nabla \max\left(0, -y\mathbf{w}_t^T\mathbf{x}\right) \\
&= \begin{cases} \mathbf{w}_t, y\mathbf{w}_t^T\mathbf{x} \geq 0 \\ \mathbf{w}_t - y\mathbf{x}, y\mathbf{w}_t^T\mathbf{x} < 0 \end{cases}
\end{aligned}
$$

其中， $y\mathbf{w}_t^T\mathbf{x} \geq 0$ 表示 y 和 $\mathbf{w}_t^T\mathbf{x}$ 同號，即 $\operatorname{sign}\left(\mathbf{w}_t^T\mathbf{x}\right) = y$ ，此時 w 不變。當 $\operatorname{sign}\left(\mathbf{w}_t^T\mathbf{x}\right) \neq y$ 時， $\mathbf{w}_{t+1} = \mathbf{w}_t - y\mathbf{x}$ 。因此，這一算法和 PLA 的結果相同。

3.

$$
\begin{aligned}
E_{in} &= \frac{1}{N}\sum_{n=1}^{N} -\ln\left(h_{y_n}(\mathbf{x}_n)\right) \\
&= \frac{1}{N}\sum_{n=1}^{N} -\ln\left(\frac{\exp\left(\mathbf{w}_{y_n}^T\mathbf{x}_n\right)}{\sum_{k=1}^{K}\exp\left(\mathbf{w}_k^T\mathbf{x}_n\right)}\right)
\end{aligned}
$$

$$= \frac{1}{N}\sum_{n=1}^{N}\ln\left(\frac{\sum_{k=1}^{K}\exp\left(\mathbf{w}_k^T\mathbf{x}_n\right)}{\exp\left(\mathbf{w}_{y_n}^T\mathbf{x}_n\right)}\right)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(\ln\left(\sum_{k=1}^{K}\exp\left(\mathbf{w}_k^T\mathbf{x}_n\right)\right)-\ln\left(\exp\left(\mathbf{w}_{y_n}^T\mathbf{x}_n\right)\right)\right)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(\ln\left(\sum_{k=1}^{K}\exp\left(\mathbf{w}_k^T\mathbf{x}_n\right)\right)-\mathbf{w}_{y_n}^T\mathbf{x}_n\right)$$

以上可以知道，$E_{in} = \frac{1}{N}\sum_{n=1}^{N}\left(\ln\left(\sum_{k=1}^{K}\exp\left(\mathbf{w}_k^T\mathbf{x}_n\right)\right)-\mathbf{w}_{y_n}^T\mathbf{x}_n\right)$。

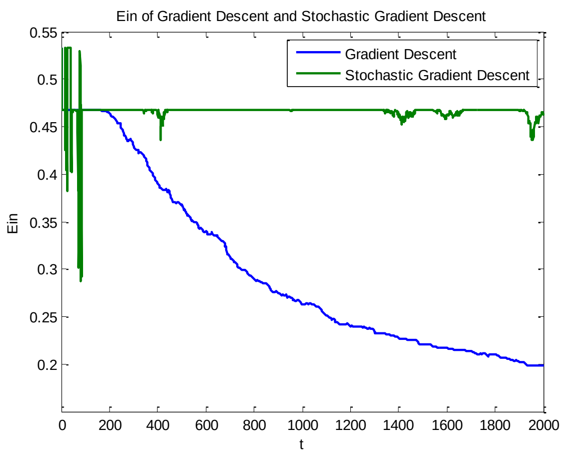
$$\frac{\partial E_{in}}{\partial w_i} = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{\sum_{k=1}^{K}\exp\left(\mathbf{w}_k^T\mathbf{x}_n\right)}\frac{\partial\sum_{k=1}^{K}\exp\left(\mathbf{w}_k^T\mathbf{x}_n\right)}{\partial w_i} - \;\; y_n = i \;\; \mathbf{x}_n\right)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(\frac{\exp\left(\mathbf{w}_i^T\mathbf{x}_n\right)}{\sum_{k=1}^{K}\exp\left(\mathbf{w}_k^T\mathbf{x}_n\right)}\frac{\partial\left(\mathbf{w}_i^T\mathbf{x}_n\right)}{\partial w_i} - \;\; y_n = i \;\; \mathbf{x}_n\right)$$

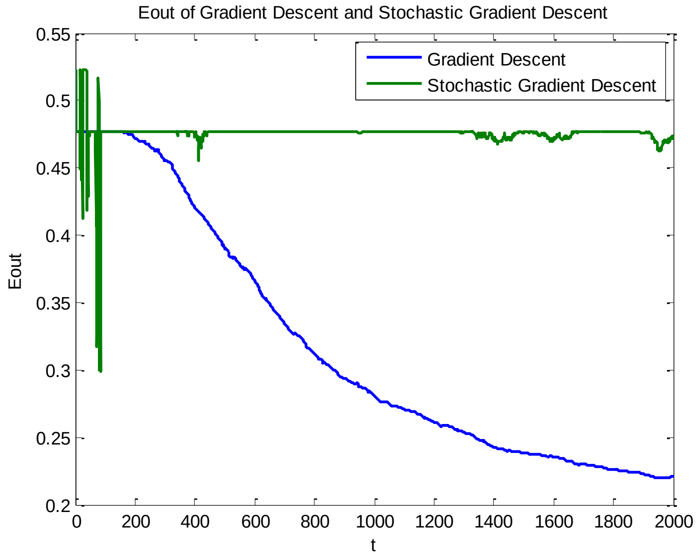
$$= \frac{1}{N}\sum_{n=1}^{N}\left(\frac{\exp\left(\mathbf{w}_i^T\mathbf{x}_n\right)}{\sum_{k=1}^{K}\exp\left(\mathbf{w}_k^T\mathbf{x}_n\right)}\mathbf{x}_n - \;\; y_n = i \;\; \mathbf{x}_n\right)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(\left(h_i\left(\mathbf{x}_n\right) - \;\; y_n = i \;\;\right)\mathbf{x}_n\right)$$

4. 通過圖形可以發現，GD 得到的 E_{in} 逐漸下降，而 SGD 不是十分穩定，也可能是 η 設置過小，或是 T 不夠大，使其在很小的區間內波動。



5. 通過觀察可以發現，E_{out} 圖形的趨勢和 E_{in} 圖形相似，說明這個參數下的 GD 此時表現得更好，更為穩定。



6. 根據題意，令 $\mathbf{y}=[y_1,\cdots,y_n]^T$，$\mathbf{w}=[w_1,\cdots,w_k]^T$，$\mathbf{h}(\mathbf{x}_n)^T=[h_1(\mathbf{x}_n),\cdots,h_k(\mathbf{x}_n)]$，$\mathbf{h}(X)=[\mathbf{h}(\mathbf{x}_1),\cdots,\mathbf{h}(\mathbf{x}_n)]^T$，有

$$\begin{aligned}\mathrm{RMSE}(H)&=\sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(y_n-H(\mathbf{x})\right)^2}\\&=\sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\sum_{k=1}^{K}w_k h_k(\mathbf{x}_n)-y_n\right)^2}\\&=\sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{h}^T(\mathbf{x}_n)\mathbf{w}-y_n\right)^2}\\&=\sqrt{\frac{1}{N}\left(\mathbf{h}(X)\mathbf{w}-\mathbf{y}\right)^T\left(\mathbf{h}(X)\mathbf{w}-\mathbf{y}\right)}\\&=\sqrt{\frac{1}{N}\left(\mathbf{w}^T\mathbf{h}^T(X)\mathbf{h}(X)\mathbf{w}-\mathbf{w}^T\mathbf{h}^T(X)\mathbf{y}-\mathbf{y}^T\mathbf{h}(X)\mathbf{w}+\mathbf{y}^T\mathbf{y}\right)}\end{aligned}$$

令 $f(\mathbf{w})=\mathbf{w}^T\mathbf{h}^T(X)\mathbf{h}(X)\mathbf{w}-\mathbf{w}^T\mathbf{h}^T(X)\mathbf{y}-\mathbf{y}^T\mathbf{h}(X)\mathbf{w}$，則

$\frac{df(\mathbf{w})}{d\mathbf{w}}=2\left(\mathbf{h}^T(X)\mathbf{h}(X)\mathbf{w}-\mathbf{h}^T(X)\mathbf{y}\right)$，令 $\frac{df(\mathbf{w})}{d\mathbf{w}}=0$，有

$$\mathbf{w}=\left(\mathbf{h}^T(X)\mathbf{h}(X)\right)^{-1}\mathbf{h}^T(X)\mathbf{y}$$

$$e_k^2=\frac{1}{N}\sum_{n=1}^{N}\left(y_n-h_k(\mathbf{x})\right)^2=\frac{1}{N}\left\|\mathbf{y}-\mathbf{h}_k(X)\right\|^2=\frac{1}{N}\left(\mathbf{y}^T\mathbf{y}+\mathbf{h}_k^T(X)\mathbf{h}_k(X)-2\mathbf{h}_k^T(X)\mathbf{y}\right)$$

$e_k^2 = \frac{1}{N}\left(e_0 + \mathbf{h}_k^T(X)\mathbf{h}_k(X) - 2\mathbf{h}_k^T(X)\mathbf{y}\right)$，則

$$\mathbf{h}_k^T(X)\mathbf{y} = \frac{1}{2}\left(e_0^2 + \mathbf{h}_k^T(X)\mathbf{h}_k(X) - Ne_k^2\right)$$

令 $\mathbf{e} = \left[e_1^2, \cdots, e_K^2\right]^T$，$\mathbf{e}_0 = \left[e_0^2, \cdots, e_0^2\right]^T, \tilde{\mathbf{h}}(X) = \left[\mathbf{h}_1^T(X)\mathbf{h}_1(X), \cdots \mathbf{h}_K^T(X)\mathbf{h}_K(X)\right]^T$

則，$\mathbf{h}^T(X)\mathbf{y} = \frac{1}{2}\left(\mathbf{e}_0^2 + \tilde{\mathbf{h}}(X) - N\mathbf{e}\right)$。

代入 $\mathbf{w} = \left(\mathbf{h}^T(X)\mathbf{h}(X)\right)^{-1}\mathbf{h}^T(X)\mathbf{y}$ 中，有

$$\mathbf{w} = \frac{1}{2}\left(\mathbf{h}^T(X)\mathbf{h}(X)\right)^{-1}\left(\mathbf{e}_0^2 - N\mathbf{e} + \tilde{\mathbf{h}}(X)\right)$$