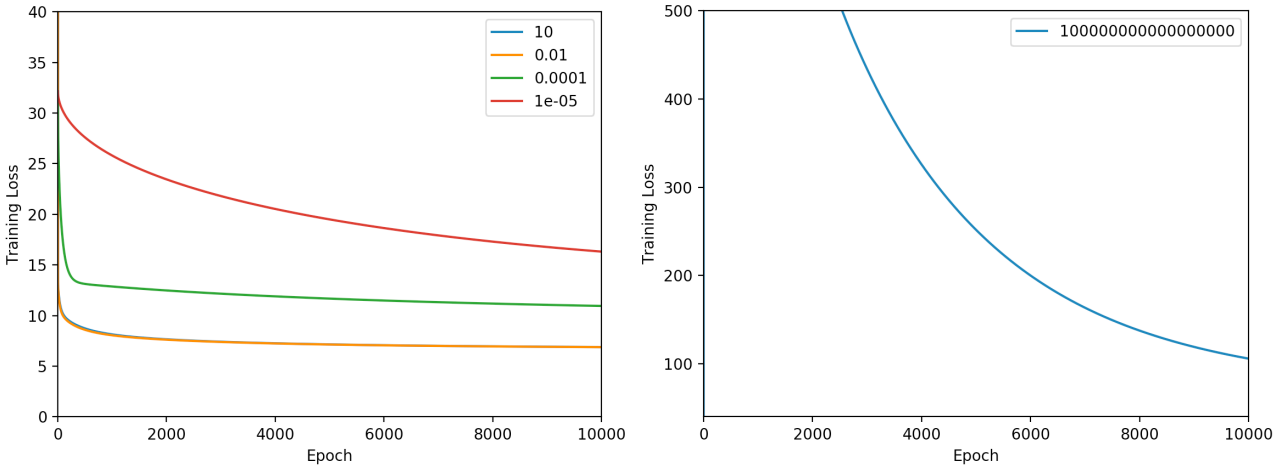
Homework 1 Report - PM2.5 Prediction

學號：r07921052　系級：電機所碩一　姓名: 劉兆鵬

1. (1%) 請分別使用至少4種不同數值的learning rate進行training（其他參數需一致），對其作圖，並且討論其收斂過程差異。



圖一、左圖為針對不同的Learning rate所訓練出的 Training Loss，紅色為 0.00001，綠色為 0.0001，橘色為 0.1，藍色為 10，右圖為極大 Learning rate 的 Loss 變化圖。

如圖一之左圖所示，隨著 Learning rate 逐漸下降，我們可以發現 Training Loss 的下降速率逐漸減緩，可以讓模型學習的過程中較為平滑，但由於模型參數每次調整的範圍過小，導致訓練的時間拉長，且由於我使用的 Optimizer 為 Adagrad，使得 Learning rate 會隨著 epoch 的增加而逐漸下降，導致後來參數調整的幅度趨近於 0 而使得 Training Loss 不再明顯下降，但隨著 Learning rate 的增加，可以觀察到 Training Loss 的收斂速度變快，在第 2000 次的 epoch 時，紅色 Learning rate 的 Training Loss 還在 25 左右，但是橘色及藍色 Learning rate 的 Training Loss 卻已經下降至 10，這代表著當 Learning rate 越大時，其模型參數調整越快，更快到達 Loss 的低點，但儘管使用大於橘色 100 倍的藍色 Learning rate ，其收斂速度竟是相差不遠，兩者的 Training Loss 變化量幾乎一致，代表著 Learning rate 是有其飽和程度的

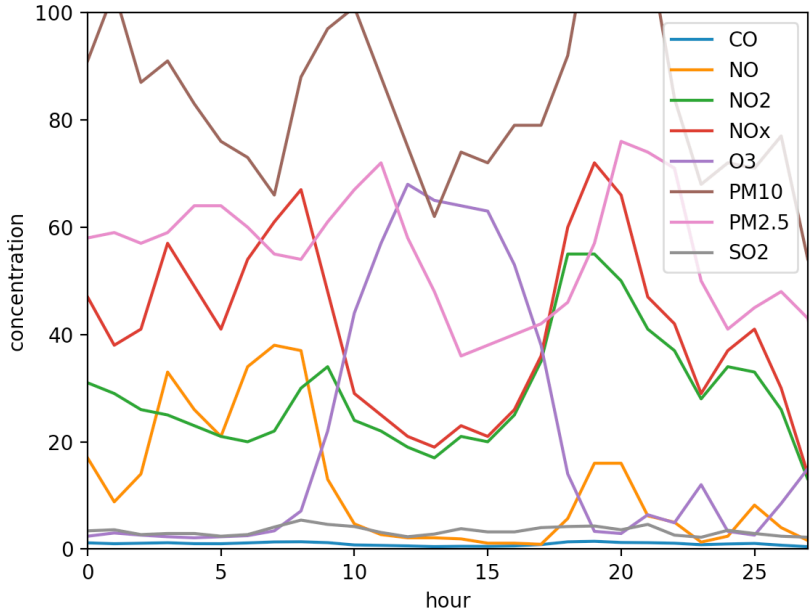
然而，這並不代表著我們應該使用更大的 Learning rate 來進行訓練，如圖一之右圖所示，此 Learning rate 的收斂速度非常緩慢，至第 10000 epoch 時，其 Training Loss 才至 100 左右，因為 Learning rate 太大時，會導致每次參數調整的幅度都過大，使得參數越過 Loss 的最低點，所以若將 Learning rate 設定過大時也會導致訓練成果不如預期。

2. (1%) 請分別使用每筆data9小時內所有feature的一次項（含bias項）以及每筆data9小時內PM2.5的一次項（含bias項）進行training，比較並討論這兩種模型的root mean-square error（根據kaggle上的public/private score）。

	所有 feature	PM2.5	CO, NO, NO2, NOx,O3,PM10, PM2.5, SO2
Public Score	8.38	7.35	6.95
Private Score	7.66	7.67	6.77
Training Loss	7.31	7.15	6.88

只用 PM2.5 來進行模型的訓練時，雖然能夠在 Training Loss 上獲得不錯的結果，但依然沒有辦法在 Testing set 中獲得最好的成績，我認為是因為影響 PM2.5 濃度的因素並不是只有 PM2.5 自己本身而已，所以當只使用過去九小時的 PM2.5 作為訓練參數時，則會因為參考的依據過少而無法獲得較好的預測結果．

而當使用所有的 Feature 來進行模型的訓練時，雖然在 Training Loss 上可以獲得與只使用 PM2.5 的結果差距不遠，但在 Testing set 中卻獲得了非常差的結果，這表示使用所有的 Feature 進行訓練會導致模型有 Overfitting 的問題．而 Overfitting 的原因可能為資料特性的問題．由於 Training data 中的各項 Feature 不一定全部都與 PM2.5 的變化有相關性，故使用所有的 Feature 進行模型的訓練時，則會使模型過度得 fit 於 Training data，且 Training data 中有許多資料呈現不合理的現象，如觀測值一大片為零的狀況，都會導致模型的預測結果與真實的 PM2.5 無法準確相符．



圖二、相關係數較高的 Feature 與 PM2.5 之作圖．

結合上述兩項資料的特性，若使用太少或太多的 Feature 都會無法準確的預測 PM2.5 的變化，故為了得知哪些 Feature 與 PM2.5 較有相關性，可將所有的 Feature 與 PM2.5 進行 Correlation 的運算，來取得各項 Feature 與 PM2.5 的相關係數，若相關係數的絕對值較高者，則代表此 Feature 與 PM2.5 的變化量比較有一定相關性，且透過將各項 Feature 與 PM2.5 進行作圖，如圖二所示．也能夠察覺到哪些 Feature 與 PM2.5 有相關的趨勢．故我在實作此作業便選取了 CO, NO, NO2, NOx, O3, PM10, PM2.5 及 SO2，並透過 Validation data 將 Training data 中不合理的資料進行丟棄，避免模型過度 fit 於錯誤的資料中，最後將此些 Feature 對各自欄位進行 min-max normalization 並進行訓練，則能夠獲得不錯的預測結果．

3. (1%)請分別使用至少四種不同數值的regulization parameter λ進行training（其他參數需一至），討論及討論其RMSE(traning, testing)（testing根據kaggle上的public/private score）以及參數weight的L2 norm。

	λ = 0	λ = 10	λ = 100	λ = 1000	λ = 10000
Public Score	6.95	6.96	6.98	7.25	8.66
Private Score	6.77	7.19	7.20	7.38	8.69
Training Loss	6.88	6.88	6.90	7.09	8.26

在這次作業中，regularization 並沒有對於 Testing 的結果有顯著的幫助．當 λ 介於 0 ~ 1000 時，Training Loss 並沒有明顯的變化，但在 Testing 上卻讓預測結果逐漸降低，而當 λ 超過 1000 後，便會使 Loss 變大幅度逐漸增加，故這次作業並不適合使用 regularization．

ML2018Fall Hw1 Report Problem 4~6 （於下頁開始）

4 . 數學題

4.

(4-a) $E_D(W)=\frac{1}{2}\sum r_n(t_n-W^TX_n)^2$, Let $X\in R^{M\times N}$

$\hat{X}\in R^{N\times M}$

$W\in R^{M\times 1}$

$T\in R^{N\times 1}$

$R\in R^{N\times N}$ 為 Diagonal 且主对角線為 r_n; $n=1\sim N$

$\Rightarrow \frac{1}{2}\sum r_n(t_n-\hat{x}_nW)^2$

$\Rightarrow \frac{1}{2}(T-\hat{X}W)^TR(T-\hat{X}W)$

$\Rightarrow \frac{1}{2}(T^T-W^T\hat{X}^T)R(T-\hat{X}W)$

$\Rightarrow \frac{1}{2}[T^TRT-T^TR\hat{X}W-W^T\hat{X}^TRT+W^T\hat{X}^TR\hat{X}W]$

Let $SSE(W+\Delta W)-SSE(W)=0$

$\Rightarrow (T^TRT-T^TR\hat{X}(W+\Delta W)-(W+\Delta W)^T\hat{X}^TRT+(W+\Delta W)^T\hat{X}^TR\hat{X}(W+\Delta W))$

$-(T^TRT-T^TR\hat{X}W-W^T\hat{X}^TRT+W^T\hat{X}^TR\hat{X}W)=0$

$\Rightarrow -T^TR\hat{X}\Delta W-\Delta W^T\hat{X}^TRT+W^T\hat{X}^TR\hat{X}\Delta W+\Delta W^T\hat{X}^TR\hat{X}W+\Delta W^T\hat{X}^TR\hat{X}\Delta W=0$

$\Rightarrow \Delta W^T\hat{X}^TRT-\Delta W^T\hat{X}^TRT+\Delta W^T\hat{X}^TR\hat{X}W+\Delta W^T\hat{X}^TR\hat{X}W+\underbrace{\Delta W^T\hat{X}^TR\hat{X}\Delta W}_{\text{省略}}=0$

$\Rightarrow -2\Delta W^T\hat{X}^TRT+2\Delta W^T\hat{X}^TR\hat{X}W=0$

$\Rightarrow \Delta W^T\hat{X}^TR\hat{X}W=\Delta W^T\hat{X}^TRT$

$\Rightarrow W=(\hat{X}^TR\hat{X})^{-1}\hat{X}^TRT$ ※

(4-b) $t=[t_1\ t_2\ t_3]=[0,10,5]$, $X=[x_1,x_2,x_3]=\begin{bmatrix}2&5&5\\3&1&6\end{bmatrix}_{M\times N}$

$r_1,r_2,r_3=2,1,3$

$\Rightarrow \hat{X}=X^T=\begin{bmatrix}2&3\\5&1\\5&6\end{bmatrix}_{N\times M}$, $R=\begin{bmatrix}2&0&0\\0&1&0\\0&0&3\end{bmatrix}_{N\times N}$, $T=[0,10,5]$

$W=(\hat{X}^TR\hat{X})^{-1}\hat{X}^TRT$

$\Rightarrow \left(\begin{bmatrix}2&3\\5&1\\5&6\end{bmatrix}^T\begin{bmatrix}2&0&0\\0&1&0\\0&0&3\end{bmatrix}\begin{bmatrix}2&3\\5&1\\5&6\end{bmatrix}\right)^{-1}\begin{bmatrix}2&5&5\\3&1&6\end{bmatrix}\begin{bmatrix}2&0&0\\0&1&0\\0&0&3\end{bmatrix}[0,10,5]$

$=[2.28275254,-1.13586237]$ ※

5. 數學題

5, Given a linear model:

$$y(x,w) = w_0 + \sum_{i=1}^{M} w_i x_i$$

with a sum of square error function:

$$E(w) = \frac{1}{2}\sum (y(x,w) - t_n)^2$$

let $\hat{y}_n = w_0 + \sum_{i=1}^{M} w_i (x_i + \varepsilon_i)$

$= w_0 + \sum_{i=1}^{M} w_i x_i + \sum w_i \varepsilon_i$

$= y_n + \sum_{i=1}^{M} w_i \varepsilon_i$ * y_n 為 noise-free model 的第 n 個 output

$\Rightarrow \frac{1}{2}\sum_{n=1}^{N} (\hat{y} - t_n)^2$

$\Rightarrow \frac{1}{2}\sum_{n=1}^{N} \left(y_n^2 + 2y_n\sum w_i\varepsilon_i + \left(\sum_{i=1}^{M} w_i\varepsilon_i\right)^2 - 2y_n t_n - 2t_n\sum_{i=1}^{M} w_i\varepsilon_i + t_n^2 \right)$

$\Rightarrow \frac{1}{2}\sum_{n=1}^{N} \left(y_n^2 - 2y_n t_n + t_n^2 + \left(\sum_{i=1}^{M} w_i\varepsilon_i\right)^2 \right)$, $\because E[\varepsilon_i] = 0$, $\therefore 2y_n\sum w_i\varepsilon_i$ 與 $2t_n\sum w_i\varepsilon_i$ 為 0

$\Rightarrow \frac{1}{2}\sum (y_n - t_n)^2 + \frac{1}{2}\sum w_i^2\sigma^2$, * $E[(\sum w_i\varepsilon_i)^2] = \sum w_i^2\sigma^2$

$\because E[\varepsilon_i\varepsilon_j] = \sigma^2$, If $i=j$

$\Rightarrow \frac{1}{2}\sum (y_n - t_n)^2 + \frac{\sigma^2}{2}\sum w_i^2$

$\Rightarrow \frac{1}{2}\sum (y_n - t_n)^2 + \frac{\lambda}{2}\sum w_i^2$ ~~XX~~ * λ 為 σ^2, 故加上 noise 的 error 會等於 noise-free error 加上以 σ^2 為 λ 的 regularization 的 error 🙏

$\Rightarrow \frac{1}{2}[(X^Tw - T)^T(X^Tw - T) + \lambda w^Tw]$

$\Rightarrow \frac{1}{2}[(w^TX - T)(X^Tw - T) + \lambda w^Tw]$

$\Rightarrow \frac{1}{2}[w^TXX^Tw - w^TXT - T^TX^Tw + T^TT + \lambda w^Tw]$

$\because \delta E(w+\delta w) - \delta E(w) = 0$

$[((w+\delta w)^TXX^T(w+\delta w) - (w+\delta w)^TXT - T^TX^T(w+\delta w) + T^TT + \lambda(w+\delta w)^T(w+\delta w)) - (w^TXX^Tw - w^TXT - T^TX^Tw + T^TT + \lambda w^Tw)]$

$\Rightarrow w^TXX^T\delta w + \delta w^TXX^Tw + \delta w^TXX^T\delta w - \delta w^TXT - T^TX^T\delta w + \lambda w^T\delta w + \lambda\delta w^Tw + \lambda\delta w^T\delta w$

$\Rightarrow 2\delta w^TXX^Tw - 2\delta w^TXT + 2\lambda\delta w^Tw = 0$

$\Rightarrow \delta w^TXX^Tw + \lambda\delta w^Tw = \delta w^TXT$

$\Rightarrow (XX^T + \lambda I)w = XT$

$\Rightarrow w = (XX^T + \lambda I)^{-1}XT$ 🙏 * λ 為 σ^2

6. 數學題

6. $A \in R^{n\times n}$, α is one of the elements of A, prove

$$\frac{d}{d\alpha}\ln|A| = Tr\left(A^{-1}\frac{d}{d\alpha}A\right)$$

證①

$$\frac{d}{d\alpha}\ln|A|$$
$$= \frac{d}{d\alpha}\ln(\lambda_1\times\lambda_2\times\cdots\times\lambda_n) \quad * \lambda_i, i=1\sim n \text{ 為 } A \text{ 的 eigenvalue}$$
$$= \frac{d}{d\alpha}(\ln\lambda_1+\ln\lambda_2+\cdots+\ln\lambda_n)$$
$$= \frac{d}{d\alpha}\ln\lambda_1 + \frac{d}{d\alpha}\ln\lambda_2+\cdots+\frac{d}{d\alpha}\ln\lambda_n$$
$$= \frac{1}{\lambda_1}\frac{d}{d\alpha}\lambda_1 + \frac{1}{\lambda_2}\frac{d}{d\alpha}\lambda_2+\cdots+\frac{1}{\lambda_n}\frac{d}{d\alpha}\lambda_n$$
$$= \sum_{i=1}^{n}\frac{1}{\lambda_i}\frac{d}{d\alpha}\lambda_i$$
$$= Tr\left(A^{-1}\frac{d}{d\alpha}A\right) \quad * \frac{1}{\lambda_i}, i=1\sim n \text{ 為 } A^{-1} \text{ 的 eigenvalue} \quad \#$$

證②

$$\frac{d}{d\alpha}\ln|A|$$
$$= \frac{1}{|A|}\frac{d}{d\alpha}|A|$$
$$= \frac{1}{|A|}\cdot|A|\cdot Tr\left(A^{-1}\cdot\frac{d}{d\alpha}A\right) \quad * \text{using Jacobi's formula}$$
$$= Tr\left(A^{-1}\cdot\frac{d}{d\alpha}A\right) \#$$

證③

$$\frac{d}{d\alpha}\ln|A|$$
$$= \frac{1}{\det(A)}\frac{d}{d\alpha}\det(A)$$
$$= \frac{1}{\det(A)}\sum_{i=1}^{n}\sum_{j=1}^{n}C_{ij}\cdot\frac{d}{d\alpha}A_{ij} \quad * C \text{ 為 } A \text{ 的 cofactor}$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{\det(A)}\cdot C_{ij}\frac{d}{d\alpha}A_{ij} \quad * \quad A^{-1} = \frac{1}{\det(A)}\cdot C^T$$
$$= \frac{1}{\det(A)}\cdot adj(A)$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{n}(A^{-1})_{ji}\frac{d}{d\alpha}A_{ij}$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{n}(A^{-1})_{ij}\frac{d}{d\alpha}A_{ij} \quad * A \text{ 為 symmetric}$$
故 A^{-1} 也為 symmetric
$$= Tr\left(A^{-1}\frac{d}{d\alpha}A\right) \#$$