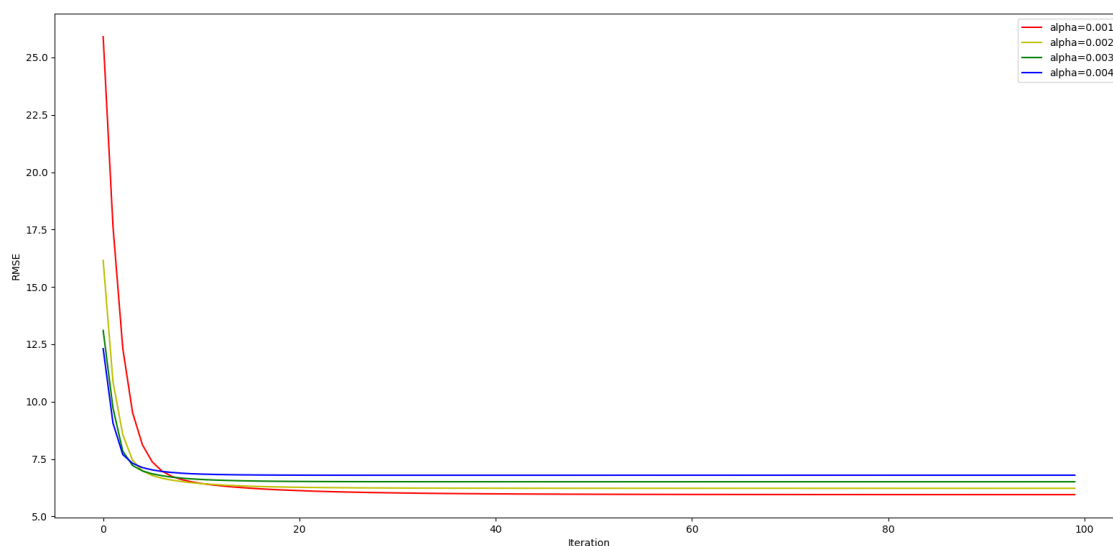


Homework 1 Report - PM2.5 Prediction

B05901022 電機三 許睿洋

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。

以下為示意圖：



在 learning rate 較小的時候，初始的 loss 相當大，但其收斂到的數值卻是比較小的；反之，在 learning rate 較大的時候，雖然有相當好的初始 loss，卻只能收斂到一個較大的數值。因此，慎選 learning rate 能夠確保訓練過程的 loss 能夠下降到較好的狀況。

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）。

	Public Score	Private Score
With all features	7.65725	7.44538
Only with PM2.5	8.54986	8.37499

由於 PM2.5 的濃度還可能受到其他因素影響(例如:下雨)，只針對 PM2.5 進行訓練可能會有較差的結果。

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

	Training	Public	Private	L2Norm
$\lambda = 0.0001$	6.426675	7.66664	7.45865	15.669713
$\lambda = 0.001$	6.537951	7.71292	7.36896	15.019553
$\lambda = 0.01$	7.280320	7.97906	7.56481	12.144796
$\lambda = 0.1$	11.062287	9.96809	8.06805	8.504792

加入 regularization 可以使 loss 在訓練過程中能夠有相對穩定的變化, 但是在這次的 case 中, 加入過大的 λ 可能會使訓練結果變差, 且可以從結果看出來可能是 L2Norm 被 λ 大幅壓縮所導致的惡化現象, 因此如若要考慮加入 regularization, 慎選參數是必要的。

4~6 數學題:

4.Collaborator: B05901009 高瑋聰、B05901034 劉奎元

4-a

$$w^* = \arg \min_w \frac{1}{2} \sum_{n=1}^N r_n (t_n - w^T x_n)^2 = \arg \min_w E_0(w)$$

$$w = [w_1, w_2, w_3, \dots, w_N]^T$$

$$\frac{\partial E_0(w)}{\partial w_i} = \frac{1}{2} \sum_{n=1}^N r_n \cdot 2(t_n - w^T x_n) \cdot (-x_n^i) = 0, \forall i \in [1, N]$$

$$\Rightarrow \sum_{n=1}^N r_n t_n x_n^i = \sum_{n=1}^N r_n w^T x_n x_n^i \quad \forall i \in [1, N]$$

$$\sum_{n=1}^N (r_n t_n) x_n = \sum_{n=1}^N (r_n w^T x_n) x_n$$

$$\Rightarrow \text{Let } R = \begin{bmatrix} r_1 & 0 & \dots & 0 \\ 0 & r_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & r_N \end{bmatrix}, T = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix}, X = [x_1 \dots x_N]$$

$$XRT = XRX^T w$$

$$\bar{T} = X^T w \Rightarrow w = (XX^T)^{-1} X \bar{T}$$

(Assume XX^T invertible)

4-2 ~~2b~~: (Assume X^T non-invertible) (using SVD)

(A) $T = X^T W$

SVD $\Rightarrow X^T = U \Sigma V^T$

$(X^T)^+ = V \Sigma^+ U^T$

$\Rightarrow W = (X^T)^+ T = V \Sigma^+ U^T T$ #

4-6 $W^* = \left(\begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 5 & 1 \\ 5 & 6 \end{bmatrix} \right)^{-1} \begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix} \begin{bmatrix} 0 \\ 10 \\ 5 \end{bmatrix}$ (invertible)

$= \begin{bmatrix} 54 & 41 \\ 41 & 46 \end{bmatrix}^{-1} \begin{bmatrix} 95 \\ 40 \end{bmatrix} = \begin{bmatrix} \frac{1910}{803} \\ -\frac{915}{803} \end{bmatrix}$

5. Collaborator: (Self only)

5. $y(x, w) = w_0 + \underbrace{w^T}_{1 \times D} \underbrace{x}_{D \times 1}$

$E(w) = \frac{1}{2} \sum_{n=1}^N (w_0 + w^T x_n - t_n)^2$

$E^*(w) = \frac{1}{2} \sum_{n=1}^N (w_0 + w^T x_n + \underbrace{w^T \epsilon_n}_{1 \times 1} - t_n)^2$

$E[E^*(w)] = \frac{1}{2} \sum_{n=1}^N E[(w_0 + w^T x_n + w^T \epsilon_n - t_n)^2]$

$= \frac{1}{2} \sum_{n=1}^N E[w_0^2 + (w^T x_n)^2 + (w^T \epsilon_n)^2 + t_n^2$

$+ 2w_0 w^T x_n + 2w_0 w^T \epsilon_n - 2w_0 t_n$

$+ 2w^T x_n w^T \epsilon_n - 2w^T x_n t_n - 2w^T \epsilon_n t_n]$

$= \frac{1}{2} \sum_{n=1}^N [w_0^2 + (w^T x_n)^2 + t_n^2 + 2w_0 w^T x_n - 2w_0 t_n - 2w^T x_n t_n]$

$+ \frac{1}{2} \sum_{n=1}^N (w^T)^2 \sigma^2$ original (noise-free)

noise regularization

6. Collaborator:

B05901009 高瑋聰、B05901092 歐瀚墨、B05901011 許秉倫、

B05901034 劉奎元、B05901082 楊晟甫、B05901101 陳泓廷

6. $A \in \mathbb{R}^{n \times n}$ prove $\frac{d}{d\alpha} \ln|A| = \text{Tr}(A^{-1} \frac{d}{d\alpha} A)$

$$\frac{d}{d\alpha} \ln|A| = \frac{d}{d\alpha} \ln[\lambda_1 \lambda_2 \dots \lambda_n] = \frac{d}{d\alpha} \sum_{i=1}^n \ln \lambda_i$$

$$= \sum_{i=1}^n \frac{d \ln \lambda_i}{d \lambda_i} \frac{d \lambda_i}{d \alpha} = \sum_{i=1}^n \frac{1}{\lambda_i} \frac{d \lambda_i}{d \alpha}$$

$$A = P D P^{-1}, A^{-1} = P D^{-1} P^{-1}$$

$$\text{Tr}(A^{-1} \frac{d}{d\alpha} A) = \text{Tr}(P D^{-1} P^{-1} \frac{d}{d\alpha} (P D P^{-1}))$$

$$= \text{Tr}(P D^{-1} P^{-1} [\frac{dP}{d\alpha} D P^{-1} + P \frac{dD}{d\alpha} P^{-1}])$$

$$= \text{Tr}(P D^{-1} P^{-1} \frac{dP}{d\alpha} D P^{-1}) + \text{Tr}(P D^{-1} P^{-1} P [\frac{dD}{d\alpha} P^{-1} + D \frac{dP^{-1}}{d\alpha}])$$

$$= \text{Tr}(P D^{-1} P^{-1} \frac{dP}{d\alpha} D P^{-1}) + \text{Tr}(P D^{-1} \frac{dD}{d\alpha} P^{-1}) + \text{Tr}(P \frac{dP^{-1}}{d\alpha})$$

$$= \text{Tr}(\frac{dP}{d\alpha} P^{-1} + P \frac{dP^{-1}}{d\alpha}) + \text{Tr}(P D^{-1} \frac{dD}{d\alpha} P^{-1})$$

$$= \text{Tr}(\frac{d}{d\alpha} (P P^{-1})) + \text{Tr}(P D^{-1} \frac{dD}{d\alpha} P^{-1})$$

$$= \text{Tr}(\frac{d}{d\alpha} I) + \text{Tr}(P D^{-1} \frac{dD}{d\alpha} P^{-1})$$

$$= 0 + \text{Tr}(P D^{-1} \frac{dD}{d\alpha} P^{-1})$$

$$= \text{Tr}(P^{-1} \frac{dD}{d\alpha})$$

$$\left(D = \begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \lambda_n \end{bmatrix} \Rightarrow D^{-1} = \begin{bmatrix} \frac{1}{\lambda_1} & & \\ & \frac{1}{\lambda_2} & \\ & & \frac{1}{\lambda_n} \end{bmatrix} \right)$$

$$\frac{dD}{d\alpha} = \begin{bmatrix} \frac{d\lambda_1}{d\alpha} & & \\ & \frac{d\lambda_2}{d\alpha} & \\ & & \frac{d\lambda_n}{d\alpha} \end{bmatrix}$$

$$= \text{Tr} \left(\begin{bmatrix} \frac{1}{\lambda_1} \frac{d\lambda_1}{d\alpha} & & \\ & \frac{1}{\lambda_2} \frac{d\lambda_2}{d\alpha} & \\ & & \frac{1}{\lambda_n} \frac{d\lambda_n}{d\alpha} \end{bmatrix} \right)$$

$$= \sum_{i=1}^n \frac{1}{\lambda_i} \frac{d\lambda_i}{d\alpha} = \text{RHS} \quad \#$$