

Homework 1 Report - PM2.5 Prediction

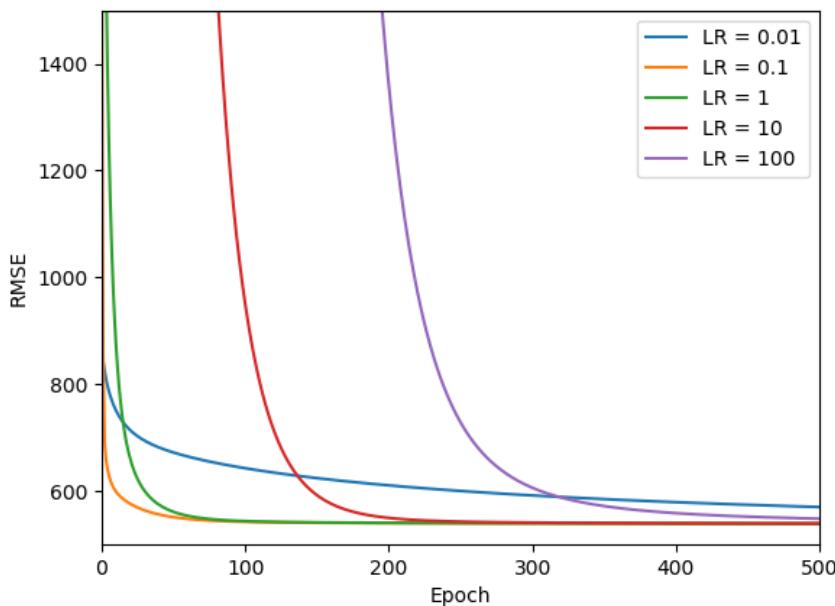
學號：R06521504 系級：土木所交通組碩二 姓名：陳譽仁

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。

使用下表中的參數，皆為一次項：

PM10 (2 hr 前)	PM10 (6 hr 前)	PM2.5 (1 hr 前)	PM2.5 (2 hr 前)	PM2.5 (5 hr 前)
PM2.5 (6 hr 前)	NO2 (2 hr 前)	NO2 (4 hr 前)	WIND_SPEED (3 hr 前)	O3 (2 hr 前)
O3 (4 hr 前)	RH (1 hr 前)	SO4 (1 hr 前)	NO (4 hr 前)	

分別使用以下的 learning rate 進行測試：0.01, 0.1, 1, 10, 100



結果如下：

可以發現在 learning rate 大於 0.1 時，RMSE 遞減的速率隨 learning rate 降低而升高，代表此時權重在每次迭代之間的差距過大，因此降低 learning rate 可改善遞減的速度，但是 learning rate 太小的時候，權重隨每次迭代所改善的量便會太小，造成圖中 LR=0.01 的結果。

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) 。

	RMSE (for training data)	Private Score	Public Score
引入所有 feature	523.764859	8.80455	8.87944
僅九小時 PM2.5	545.486409	9.71130	9.59175

從上表的結果可以發現只考慮 PM2.5 的模型與引入所有 feature 的結果相比，在表中所有指標下都比較差，其原因應該是預測的 PM2.5 不只與過去的 PM2.5 有關，還會與其他 feature 有關，因此只有考慮 PM2.5 的模型表現會比引入所有 feature 還差。

3. (1%) 請分別使用至少四種不同數值的 regularization parameter λ 進行 training (其他參數需一致)，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。

本題與第一題使用同樣的 feature，並測試以下幾種不同的 regularization parameter λ : 0, 1, 10, 100, 1000, 10000, 100000, 1000000，結果如下表：

λ	RMSE (training)	Private Score	Public Score	Average Score	L2 norm for weight
0	538.991768	7.91773	8.31483	8.11628	1.026404
1	538.991771	7.91773	8.31483	8.11628	1.026399
10	538.991797	7.91769	8.31492	8.11631	1.026363
100	538.992066	7.91729	8.31575	8.11652	1.026000
1000	538.995531	7.91345	8.32372	8.11859	1.022577
10000	539.076423	7.88196	8.37885	8.13041	1.002836
100000	540.662392	7.81134	8.54075	8.17604	0.993421
1000000	560.428195	8.50081	9.09563	8.79822	0.975616

從結果可以發現當 λ 在 100000 以下時，RMSE 的差距不管在 train, test 資料上都很小，但是從權重的 L2 norm 觀察，確實可以發現他隨著 λ 的增加而逐漸下降，代表 regularization parameter 確實有展現處罰過大權重的作用。但是當 λ 太大時，其影響就過大，使得模式的表現變得比較差。

4. 過程如以下掃描圖檔

4(a)

$$\text{Let } \hat{\mathbf{t}} = [\sqrt{r_1} t_1, \dots, \sqrt{r_n} t_n], \quad \hat{\mathbf{x}} = [\sqrt{r_1} x_1, \dots, \sqrt{r_n} x_n]$$

由題目，

$$\begin{aligned} E_D(w) &= \frac{1}{2} \sum_{n=1}^N r_n (t_n - w^\top x_n)^2 \\ &= \frac{1}{2} \sum_{n=1}^N (\sqrt{r_n} t_n - w^\top (\sqrt{r_n} x_n))^2 \\ &= \frac{1}{2} (\hat{\mathbf{t}} - w^\top \hat{\mathbf{x}})^\top (\hat{\mathbf{t}} - w^\top \hat{\mathbf{x}}) \\ &= \frac{1}{2} (\hat{\mathbf{t}}^\top - \hat{\mathbf{x}}^\top w) (\hat{\mathbf{t}} - w^\top \hat{\mathbf{x}}) \\ &= \frac{1}{2} (\hat{\mathbf{t}}^\top \hat{\mathbf{t}} - \hat{\mathbf{t}}^\top w^\top \hat{\mathbf{x}} - \hat{\mathbf{x}}^\top w \hat{\mathbf{t}} + \hat{\mathbf{x}}^\top w w^\top \hat{\mathbf{x}}) \\ &= \frac{1}{2} (\hat{\mathbf{t}}^\top \hat{\mathbf{t}} - \hat{\mathbf{t}}^\top w^\top \hat{\mathbf{x}} - \hat{\mathbf{x}}^\top w \hat{\mathbf{t}} + w^\top \hat{\mathbf{x}} \hat{\mathbf{x}}^\top w) \end{aligned}$$

$$\begin{aligned} \Rightarrow \frac{\partial E_D(w)}{\partial w} &= \frac{1}{2} (-\hat{\mathbf{x}} \hat{\mathbf{x}}^\top - \hat{\mathbf{x}} \hat{\mathbf{x}}^\top + 2 \hat{\mathbf{x}} \hat{\mathbf{x}}^\top w) \\ &= -\hat{\mathbf{x}} \hat{\mathbf{x}}^\top + \hat{\mathbf{x}} \hat{\mathbf{x}}^\top w = 0, \quad \text{當 } E_D(w) \text{ 最小值發生時, } \frac{\partial E_D(w)}{\partial w} = 0 \end{aligned}$$

$$\Rightarrow \hat{\mathbf{x}} \hat{\mathbf{x}}^\top w^* = \hat{\mathbf{x}} \hat{\mathbf{x}}^\top \Rightarrow \boxed{w^* = (\hat{\mathbf{x}}^\top)^{-1} \hat{\mathbf{x}}^\top}$$

4(b) 根據(a)小題的定義，

$$\hat{\mathbf{t}} = [\sqrt{r_1} t_1, \sqrt{r_2} t_2, \sqrt{r_3} t_3] = [0 \ 10 \ 5\sqrt{3}]$$

$$\hat{\mathbf{x}} = [\sqrt{r_1} x_1, \sqrt{r_2} x_2, \sqrt{r_3} x_3] = \begin{bmatrix} 2\sqrt{2} & 5 & 5\sqrt{3} \\ 3\sqrt{2} & 1 & 6\sqrt{3} \end{bmatrix}$$

$$\therefore w^* = (\hat{\mathbf{x}}^\top)^{-1} \hat{\mathbf{t}}^\top = \begin{bmatrix} 2\sqrt{2} & 3\sqrt{2} \\ 5 & 1 \\ 5\sqrt{3} & 6\sqrt{3} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 10 \\ 5\sqrt{3} \end{bmatrix}$$

$$= \begin{bmatrix} -0.0698 & 0.2148 & 0.0142 \\ 0.0945 & -0.1659 & 0.0649 \end{bmatrix} \begin{bmatrix} 0 \\ 10 \\ 5\sqrt{3} \end{bmatrix}$$

$$= \begin{bmatrix} 2.2717 \\ -1.0970 \end{bmatrix}$$

5. 過程如以下掃描圖檔

$$5. \quad \hat{w} = \begin{pmatrix} w_0 \\ \vdots \\ w_D \end{pmatrix}, \quad \hat{x} = \begin{pmatrix} 1 \\ x \end{pmatrix}, \quad x = (x_1, x_2, \dots, x_N), \quad \hat{\varepsilon} = \begin{pmatrix} 0 \\ \varepsilon \end{pmatrix}, \quad \varepsilon = (\varepsilon_1, \dots, \varepsilon_N),$$

$$\text{由題目, } E(w) = \frac{1}{2} \sum_{n=1}^N (y(x_n, w) - t_n)^2 \quad t = \begin{pmatrix} t_1 \\ \vdots \\ t_N \end{pmatrix}$$

$$= \frac{1}{2} (\hat{x}^\top w + \hat{\varepsilon}^\top w - t)^\top (\hat{x}^\top w + \hat{\varepsilon}^\top w - t)$$

取期望值,

$$\begin{aligned} E(E(w)) &= E\left[\frac{1}{2} (\hat{x}^\top w + \hat{\varepsilon}^\top w - t)^\top (\hat{x}^\top w + \hat{\varepsilon}^\top w - t)\right] \\ &= E\left[\frac{1}{2} (w^\top \hat{x} + w^\top \hat{\varepsilon} - t)(\hat{x}^\top w + \hat{\varepsilon}^\top w - t)\right] \\ &= E\left[\frac{1}{2} (w^\top \hat{x} \hat{x}^\top w + w^\top \hat{x} \hat{\varepsilon}^\top w - w^\top \hat{x} t \right. \\ &\quad \left. + w^\top \hat{\varepsilon} \hat{x}^\top w + w^\top \hat{\varepsilon} \hat{\varepsilon}^\top w - w^\top \hat{\varepsilon} t \right. \\ &\quad \left. - t^\top \hat{x}^\top w - t^\top \hat{\varepsilon}^\top w + t^\top t\right)] \\ &= \frac{1}{2} \left\{ w^\top \hat{x} \hat{x}^\top w + w^\top \hat{x} E[\hat{\varepsilon}^\top] w - w^\top \hat{x} t \right. \\ &\quad \left. + w^\top E[\hat{\varepsilon}] \hat{x}^\top w + w^\top E[\hat{\varepsilon} \hat{\varepsilon}^\top] w - w^\top E[\hat{\varepsilon}] t \right. \\ &\quad \left. - t^\top \hat{x}^\top w - t^\top E[\hat{\varepsilon}^\top] w + t^\top t \right\} \rightarrow ① \end{aligned}$$

$$\because E[\varepsilon_i] = 0, \quad E[\varepsilon_i \varepsilon_j] = \delta_{ij} \sigma^2$$

$$\because E[\hat{\varepsilon}^\top] = 0, \quad E[\hat{\varepsilon} \hat{\varepsilon}^\top] = E\left(\begin{bmatrix} 0 & 0 & \cdots & 0 \\ \varepsilon_{1,1} & \varepsilon_{1,2} & \cdots & \varepsilon_{1,N} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{D,1} & \varepsilon_{D,2} & \cdots & \varepsilon_{D,N} \end{bmatrix} \begin{bmatrix} 0 & \varepsilon_{1,1} & \cdots & \varepsilon_{D,1} \\ 0 & \varepsilon_{1,2} & \cdots & \varepsilon_{D,2} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \varepsilon_{1,N} & \cdots & \varepsilon_{D,N} \end{bmatrix}\right)$$

$$= \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & N\sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & N\sigma^2 \end{bmatrix}$$

由前式，①可以寫為：

$$\begin{aligned} \mathbb{E}[E(w)] &= \frac{1}{2} [w^T \hat{x} \hat{x}^T w - w^T \hat{x} t - t^T \hat{x}^T w + t^T t] + \frac{1}{2} w^T \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & N\sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & N\sigma^2 \end{bmatrix} w \\ \Rightarrow \frac{\partial}{\partial w} \mathbb{E}[E(w)] &= \frac{1}{2} [2 \hat{x} \hat{x}^T w - \hat{x} t - t^T \hat{x}^T] + \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & N\sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & N\sigma^2 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{bmatrix} \\ &= \frac{1}{2} [2 \hat{x} \hat{x}^T w - \hat{x} t - \hat{x}^T t] + N\sigma^2 \begin{bmatrix} 0 \\ w_1 \\ w_2 \\ \vdots \\ w_D \end{bmatrix} \\ &= \hat{x} \hat{x}^T w - \hat{x} t + N\sigma^2 \begin{bmatrix} 0 \\ w_1 \\ w_2 \\ \vdots \\ w_D \end{bmatrix} \end{aligned}$$

當 $\mathbb{E}[E(w)]$ 最小值發生時， $\frac{\partial}{\partial w} \mathbb{E}[E(w)] = 0$

$$\Rightarrow \hat{x} \hat{x}^T w^* = \hat{x} t - N\sigma^2 \begin{bmatrix} 0 \\ w_1 \\ \vdots \\ w_D \end{bmatrix}$$

$$\Rightarrow w^* = (\hat{x}^T)^{-1} t - (\hat{x} \hat{x}^T)^{-1} N\sigma^2 \begin{bmatrix} 0 \\ w_1 \\ \vdots \\ w_D \end{bmatrix}, \text{ 確實由 noise-free 項與 weight-decay regularization term 組成，且 } w_0 \text{ 被消去}$$

6. 過程如以下掃描圖檔

b. 由題目， A 是 non-singular matrix，代表 A^{-1} 存在，

而且 Jacobi's formula 成立： $\frac{d}{d\alpha} |A| = \text{Tr} \left(\text{adj}(A) \frac{dA}{d\alpha} \right)$

再由於 $\text{adj}(A)$ 是 A 的餘因子方陣的轉置， $\text{adj}(A) = C^T$
(c)

而反矩陣 $A^{-1} = \frac{1}{|A|} (C^T)$ ，可以由行列式值與餘因子方陣求得

$$\Rightarrow A^{-1} = \frac{1}{|A|} \cdot \text{adj}(A)$$

$\Rightarrow \text{adj}(A) = |A| A^{-1}$ ，代入 Jacobi's formula

$$\begin{aligned} \frac{d}{d\alpha} |A| &= \text{Tr} \left(|A| A^{-1} \frac{dA}{d\alpha} \right) \\ &= |A| \cdot \text{Tr} \left(A^{-1} \frac{dA}{d\alpha} \right) \end{aligned}$$

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

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

#