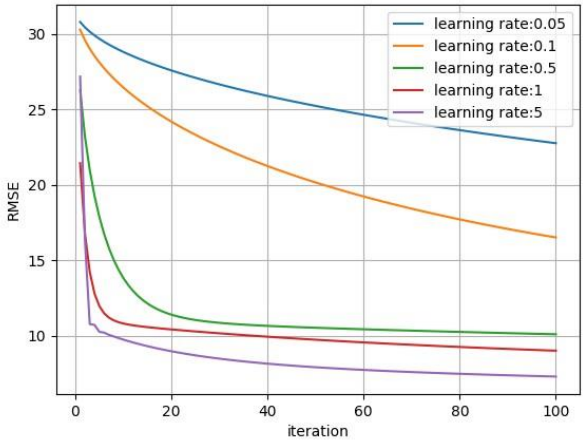
Homework 1 Report - PM2.5 Prediction

學號: B04505025　系級: 電機四　姓名: 陳在賢

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。

以下統一使用無整理過的 training/testing data 及 gradient descent 搭配 adagrad 演算法訓練模型，並以 validation 檢查其 RMSE。



根據上圖可看出，隨著 learning-rate 上升，RMSE 會更快得下降到穩定值。這部份很容易理解，畢竟較大的 learning-rate 允許權重向量有較大的變化量，因此很快就能離開初始權重向量。相反得，往後的 iteration 在較小的 learning rate 反而對應較大的 RMSE 下降幅度，這部分可歸因於 adagrad。因為前面過小的權重向量變化，反而會加快之後的變化，而不至於讓權重向量卡在過大的 RMSE。

另外，我也有比較更大 learning-rate 的趨勢，起初大致與 rate 為 5 的相同。但之後會因過大的權重變化而有「衝過頭」的問題（即便使用 adagrad 都無法抑制此現象），導致 RMSE 不斷抖動。因此，謹慎得透過 validation error 觀察適合的 learning-rate 參數也是相當重要的。

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）。

Training Set	*Training error*	*Testing error*
All data	22.65	9.39
PM 2.5	23.34	9.56

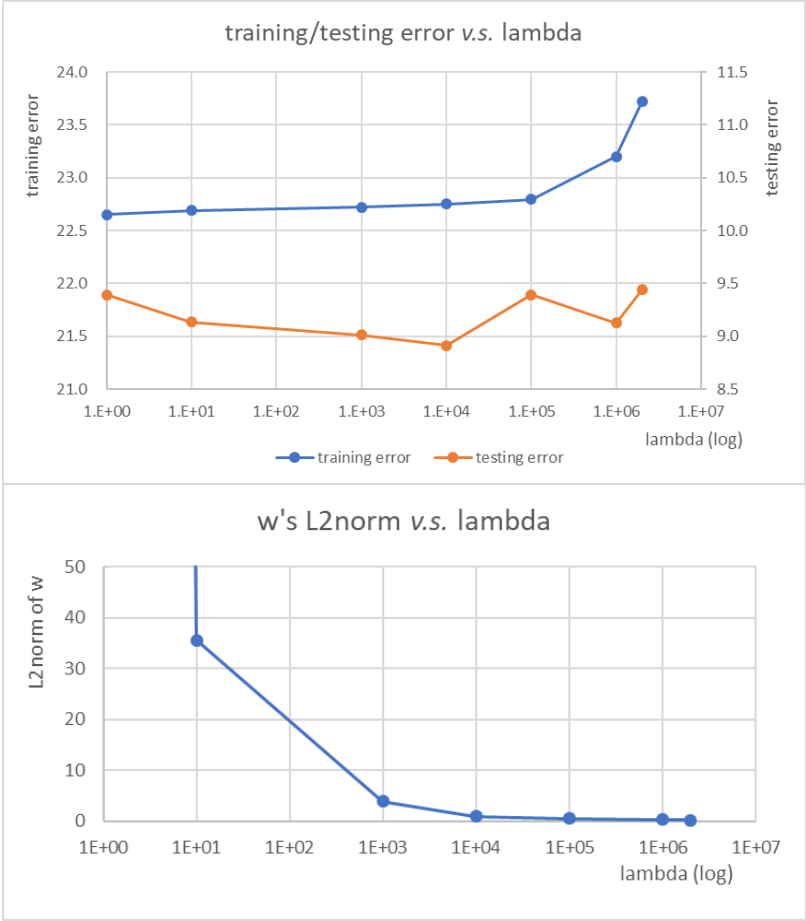
以下統一使用無整理過的 training/testing data 做運算，並以 closed-form solution（公式解）訓練模型。

首先討論 training error，All data 訓練出的模型的 training error 較 PM2.5 低，這是因為 All data 包含所有 PM2.5 數據，因此可訓練出的模型也就完全包含另一者。且因為公式解保證能得最佳解，所以 All data 對應的 training data 保證小於等於另一者（等號通常不成立，畢竟多了非常多參數值可調整）。

至於 testing error 就沒有一定的大小關係。但就本次作業而言，我猜這結果是因為 PM2.5 數據本身為含有較多 noise 的數據（包含多個明顯不合理的測驗數據），因此單就 PM2.5 數據訓練出的模型也就容易有大誤差。

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。

訓練方式同上題。



從上二圖可以看出隨著 lambda 增加，training error 持續增加、w 的 L2 norm 持續下降，而 testing error 則有不規則跳動，但大致呈先降後升的趨勢。

首先討論 regularization 的意義，在原公式加入 lambda 參數是希望能夠減小 weight vector 的長度，以避免參數過大造成 overfit；而增加 lambda 代表著對權重長度的限制更嚴苛。因此 w 的 L2 norm 也理所當然得隨著 lambda 增加有明顯得下降趨勢。

再來討論 training error，由於公式解保證能找到使 training error 最小的解，因此當 lambda 為 0（意謂著直接對原 training data 訓練）時，training error 也會有極小值。然而隨著 lambda 增加，對原數據的汙染也就更重，使得訓練出來的模型在 training data 上的表現越來越偏差，造成 training error 持續增加。

最終討論 testing error，當 lambda 過小或過大時會有較大的 error。這是因為較小的 lambda 會使模型過度得 overfit 於 training data（尤其訓練數據中又有幾個六百以上的暴增數據）；而當太大時，又會過度得限制 weight vector。兩者皆會造成模型無法近似於 testing data 而得較大的誤差。因此適當得選擇 lambda 值，才能在 testing data 中有最佳的表現。

4~6 (3%) 請參考數學題目（連結：），將作答過程以各種形式（latex 尤佳）清楚地呈現在 pdf 檔中（手寫再拍照也可以，但請注意解析度）。

4-a

$$\vec{w}^* = \arg\min_{\vec{w}} E_D(\vec{w}) = \arg\min_{\vec{w}} \frac{1}{2}\sum_{n=1}^{N} r_n (t_n - \vec{w}^T \vec{x}_n)^2$$

$$= \arg\min_{\vec{w}} \sum_{n=1}^{N} (r_n^{\frac{1}{2}} t_n - \vec{w}^T r_n^{\frac{1}{2}} \cdot \vec{x}_n)^2 = \arg\min_{\vec{w}} \sum_{n=1}^{N} (\vec{w}^T \vec{x}_n' - y_n')^2 ;\ \vec{x}_n' = r_n^{\frac{1}{2}} \vec{x}_n,\ y_n' = r_n^{\frac{1}{2}} t_n$$

$$\Rightarrow \vec{w}^* = (X'^T X')^{-1} X'^T \vec{y}',\ \text{where } X' = \begin{bmatrix} \sqrt{r_1}\,\vec{x}_1^T \\ \vdots \\ \sqrt{r_n}\,\vec{x}_n^T \end{bmatrix},\ \vec{y}' = \begin{bmatrix} \sqrt{r_1}\,t_1 \\ \vdots \\ \sqrt{r_n}\,t_n \end{bmatrix}$$

4-b

from Python…

$$\Rightarrow \vec{w} = X'^{+} \vec{y}' = \begin{bmatrix} 2.283 \\ -1.136 \end{bmatrix}$$

5.

$$E(\vec{w}) = \frac{1}{2}\sum_{n=1}^{N} \left[y(\vec{x}_n, \vec{w}) - t_n \right]^2 = \frac{1}{2}\sum_{n=1}^{N} \left[w_0 + \sum_{i=1}^{D} w_i x_{n,i} - t_n \right]^2 \quad (\text{令 } C_n = w_0 - t_n)$$

$$= \frac{1}{2}\sum_{n=1}^{N} \left[\sum_{i=1}^{D}\sum_{j=1}^{D} w_i \cdot w_j \cdot x_{n,i} \cdot x_{n,j} - 2C_n \sum_{i=1}^{D} x_{n,i} + C_n^2 \right]$$

考慮 input variable 受雜訊干擾，E 變為 $\tilde{E}$，且 $\tilde{x}_{n,i} = x_{n,i} + \varepsilon_{n,i}$，則

$$\Rightarrow \mathbb{E}[\tilde{E}(\vec{w})] = \frac{1}{2}\sum_{n=1}^{N} \left[\sum_{i=1}^{D}\sum_{j=1}^{D} \mathbb{E}\left[w_i \cdot w_j \cdot (x_{n,i} + \varepsilon_{n,i}) \cdot (x_{n,j} + \varepsilon_{n,j}) \right] - 2C_n \sum_{i=1}^{D} \mathbb{E}\left[x_{n,i} + \varepsilon_{n,i} \right] + C_n^2 \right]$$

$$\left(= w_i w_j \cdot \mathbb{E}\left[(x_{n,i} + \varepsilon_{n,i})(x_{n,j} + \varepsilon_{n,j}) \right] \right.$$

$$\left. = w_i w_j \left[x_{n,i} x_{n,j} + x_{n,j}\mathbb{E}(\varepsilon_{n,i}) + x_{n,i}\mathbb{E}(\varepsilon_{n,j}) + \mathbb{E}(\varepsilon_{n,i}\varepsilon_{n,j}) \right] \begin{cases} 0, & i \neq j \\ \sigma^2, & i = j \end{cases} \right)$$

$$= \frac{1}{2}\sum_{n=1}^{N} \left[\sum_{i=1}^{D} w_i^2 D\sigma^2 + \sum_{i=1}^{D}\sum_{j=1}^{D} w_i w_j x_{n,i} x_{n,j} - 2C_n \sum_{i=1}^{D} x_{n,i} + C_n^2 \right] = E(\vec{w}) + \frac{N \cdot D}{2} \cdot \sigma^2 \cdot \|\vec{w}\|^2 \ \#\ \text{得證}$$

$$\Rightarrow \mathbb{E}[\tilde{E}(\vec{w})] = \mathbb{E}(\vec{w}) + f(\|\vec{w}\|^2)$$，故極小化 $\tilde{E}(\vec{w})$ 可視為極小化原誤差 + regularization

6.

$$\frac{d}{d\alpha}\ln(|A|) = \frac{1}{|A|}\frac{d}{d\alpha}|A| \overset{\text{pf}}{=\!=\!\Downarrow} \frac{1}{|A|} \cdot |A|\,\mathrm{Tr}\left[A^{-1} \cdot \frac{d}{d\alpha} A \right] = \mathrm{Tr}\left(A^{-1} \frac{d}{d\alpha} A \right)$$ 得證

from Jacobi's formula: $\frac{d}{d\alpha}|A| = \mathrm{Tr}\left[\mathrm{adj}(A) \cdot \frac{dA}{d\alpha} \right]$（adj(A)：A 的伴隨矩陣）

且已知 $A^{-1} = \frac{1}{|A|}\mathrm{adj}(A)$，故將 $\mathrm{adj}(A)$ 以 $|A| \cdot A^{-1}$ 代回上式

得 $\frac{d}{d\alpha}|A| = |A| \cdot \mathrm{Tr}\left[A^{-1} \cdot \frac{dA}{d\alpha} \right]$