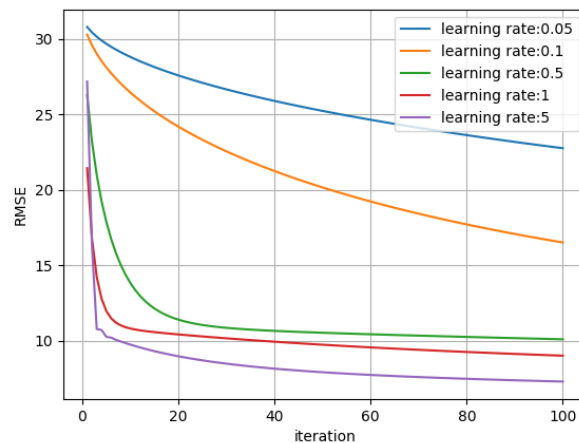


Homework 1 Report - PM2.5 Prediction

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，慎重得透過 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 |

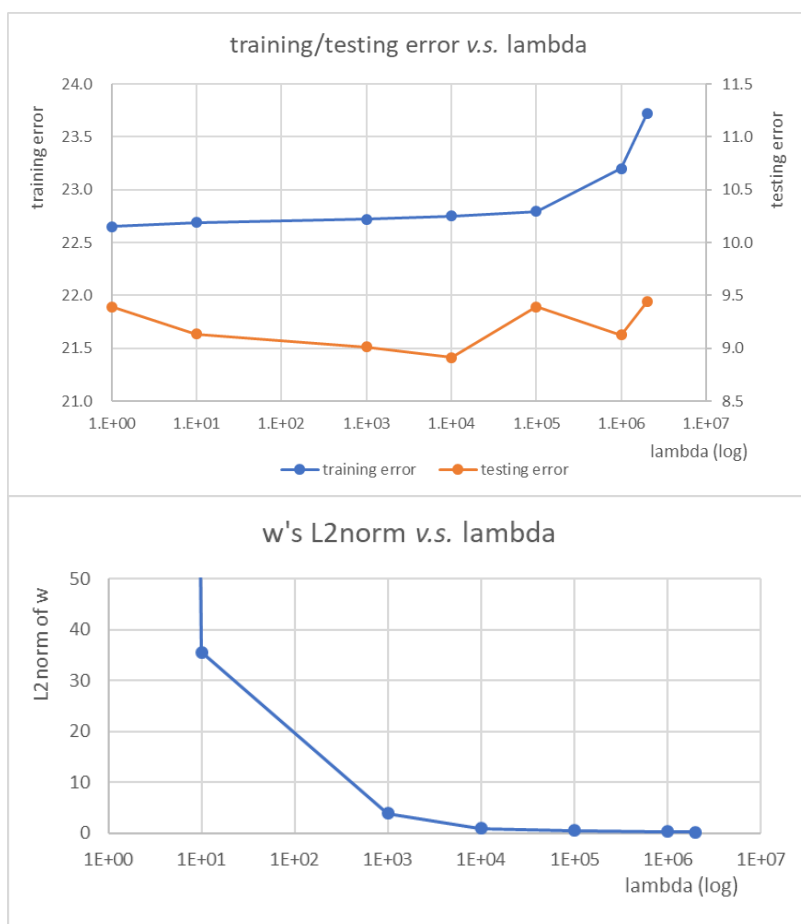
以下統一使用無整理過的 training/testing data 做運算，並以 closed-form solution（公式解）訓練模型。

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 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。

訓練方式同上題。



從上二圖可以看出隨著 λ 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 λ 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 λ 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 λ 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 λ 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 λ 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 λ 過小或過大時會有較大的 error。這是因為較小的 λ 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 λ 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

$$\begin{aligned}\vec{w}^* &= \underset{\vec{w}}{\operatorname{argmin}} E_0(\vec{w}) = \underset{\vec{w}}{\operatorname{argmin}} \frac{1}{2} \sum_{n=1}^N r_n (t_n - \vec{w}^T \vec{x}_n)^2 \\ &= \underset{\vec{w}}{\operatorname{argmin}} \sum_{n=1}^N (r_n^2 t_n - \vec{w}^T r_n^2 \vec{x}_n)^2 = \underset{\vec{w}}{\operatorname{argmin}} \sum_{n=1}^N (\vec{w}^T \vec{x}_n' - y_n')^2; \quad \vec{x}_n' = r_n^2 \vec{x}_n, \quad y_n' = r_n^2 t_n \\ \Rightarrow \vec{w}^* &= (X'^T X')^{-1} X'^T \vec{y}', \quad \text{where } X' = \begin{bmatrix} r_1^2 \vec{x}_1^T \\ \vdots \\ r_N^2 \vec{x}_N^T \end{bmatrix}, \quad \vec{y}' = \begin{bmatrix} r_1^2 t_1 \\ \vdots \\ r_N^2 t_N \end{bmatrix}\end{aligned}$$

4-b

$$X' = \begin{bmatrix} r_1^2 \vec{x}_1^T & r_2^2 \vec{x}_2^T & r_3^2 \vec{x}_3^T \end{bmatrix}^T = \begin{bmatrix} 8 & 12 \\ 5 & 1 \\ 45 & 54 \end{bmatrix}; \quad \vec{y}' = \begin{bmatrix} r_1^2 t_1 \\ r_2^2 t_2 \\ r_3^2 t_3 \end{bmatrix} = \begin{bmatrix} 4 \cdot 0 \\ 1 \cdot 10 \\ 9 \cdot 5 \end{bmatrix} = \begin{bmatrix} 0 \\ 10 \\ 45 \end{bmatrix}$$

$$\text{Pseudo inverse of } X' = (X'^T X')^{-1} X'^T = \begin{bmatrix} 0.0036 & 0.0211 & -0.0052 \\ 0.0009 & -0.0006 & 0.0186 \end{bmatrix}$$

(by matlab)

$$\Rightarrow \vec{w} = X'^+ \vec{y}' = \begin{bmatrix} -0.0148 \\ 0.8296 \end{bmatrix}$$

5.

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

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

$$\begin{aligned}\Rightarrow E[\tilde{E}(\vec{w})] &= \frac{1}{2} \sum_{n=1}^N \left[\sum_{i=1}^D \sum_{j=1}^D E[w_i \cdot w_j \cdot (x_{n,i} + \varepsilon_{n,i}) \cdot (x_{n,j} + \varepsilon_{n,j})] - 2C_n \sum_{i=1}^D E[x_{n,i} + \varepsilon_{n,i}] + C_n^2 \right] \\ &= \frac{1}{2} \sum_{n=1}^N \left[\sum_{i=1}^D \sum_{j=1}^D E[w_i \cdot w_j \cdot (x_{n,i} x_{n,j} + x_{n,i} \varepsilon_{n,j} + \varepsilon_{n,i} x_{n,j} + \varepsilon_{n,i} \varepsilon_{n,j})] - 2C_n \sum_{i=1}^D E[x_{n,i} + \varepsilon_{n,i}] + C_n^2 \right] \\ &= \frac{1}{2} \sum_{n=1}^N \left[D \sigma^2 \sum_{i=1}^D \sum_{j=1}^D w_i w_j x_{n,i} x_{n,j} - 2C_n \sum_{i=1}^D x_{n,i} + C_n^2 \right] = E(\vec{w}) + \frac{ND}{2} \cdot \sigma^2 \# \\ &\Rightarrow \tilde{E} = E + \text{constant}, \text{ 故做對 } E \text{ 的極小化後, } \tilde{E}^* \text{ 仍為 } E^* + \text{constant} \text{ (兩者相差 } \frac{ND}{2} \sigma^2 \text{ 的誤差)}\end{aligned}$$

6.

$$\frac{d}{d\alpha} \ln(|A|) = \frac{1}{|A|} \frac{d}{d\alpha} |A| \stackrel{\text{PS}}{\downarrow} \frac{1}{|A|} |A| \operatorname{Tr} \left[A^{-1} \cdot \frac{d}{d\alpha} A \right] = \operatorname{Tr} \left(A^{-1} \frac{d}{d\alpha} A \right) \quad \text{得證}$$

$$\left(\begin{array}{l} \text{from Jacobi's formula: } \frac{d}{d\alpha} |A| = \operatorname{Tr} \left[\operatorname{adj}(A) \cdot \frac{dA}{d\alpha} \right] \quad A \text{ 的伴隨矩陣} \\ \text{且已知 } A^{-1} = \frac{1}{|A|} \operatorname{adj}(A), \text{ 故將 } \operatorname{adj}(A) \text{ 以 } |A| \cdot A^{-1} \text{ 代回上式} \\ \text{得 } \frac{d}{d\alpha} |A| = |A| \operatorname{Tr} \left[A^{-1} \cdot \frac{dA}{d\alpha} \right] \end{array} \right)$$