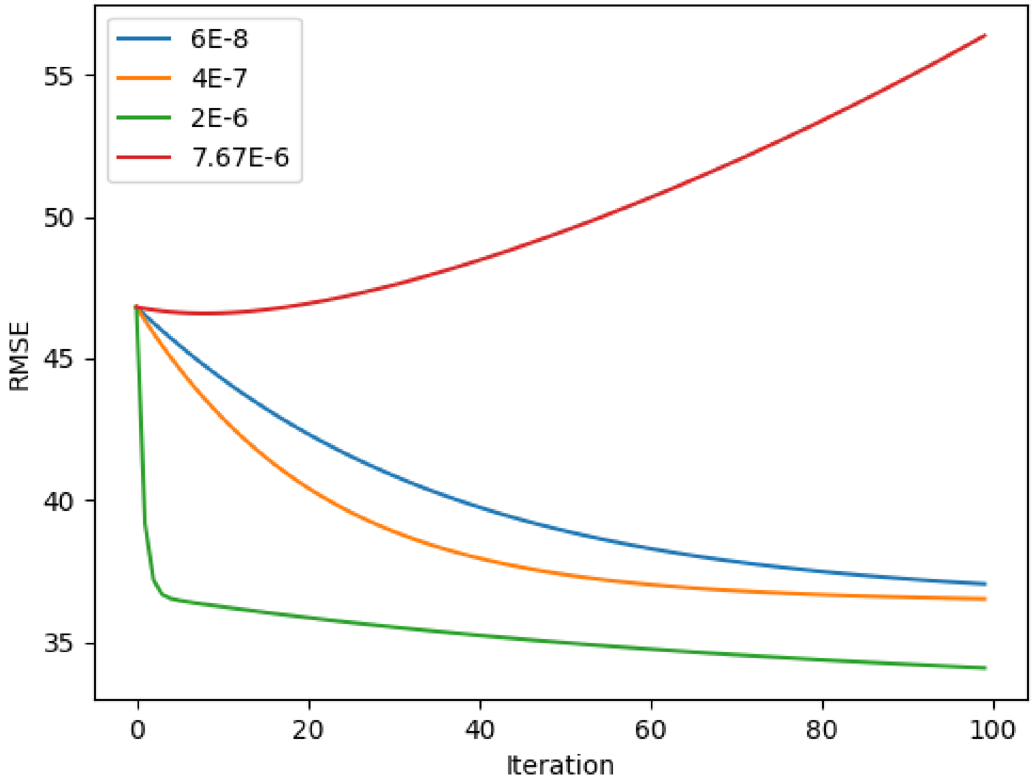
Homework 1 Report - PM2.5 Prediction

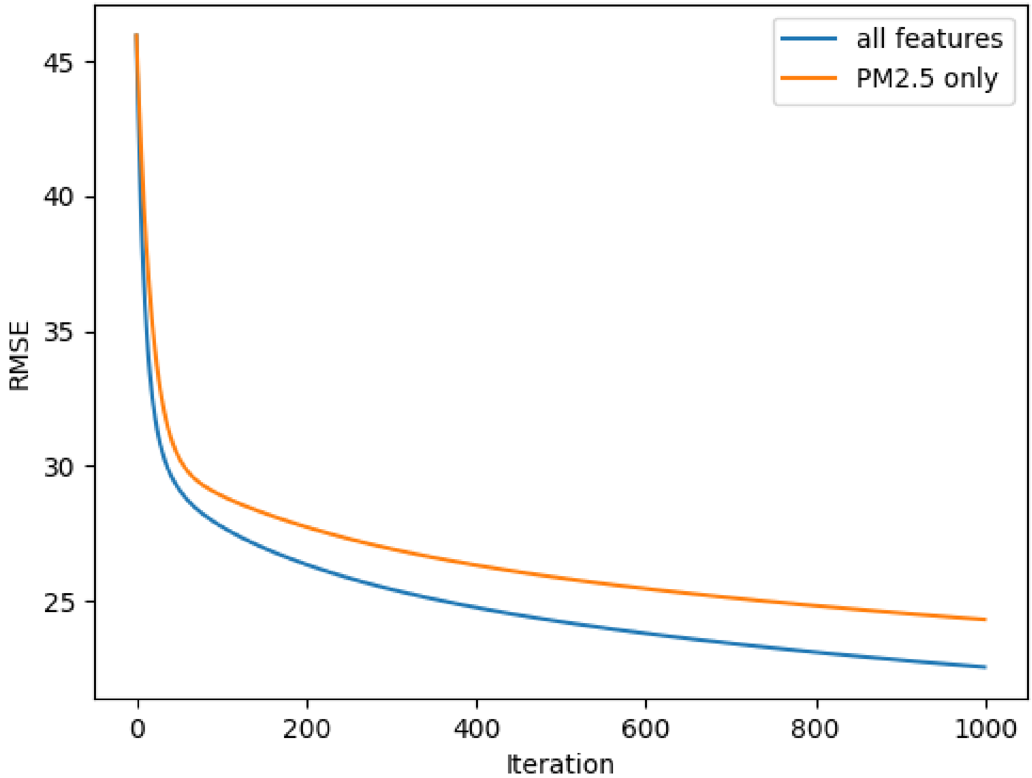
學號：B06901031　系級：電機二　姓名: 楊宗賢

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。



$\eta=6\times10^{8}$ 和 $\eta=4\times10^{7}$ 都太小，以致於雖然藍線與黃線皆嚴格遞減，斜率仍然太小； $\eta=7.67\times10^{6}$ 太大了，因此很快就向外發散； $\eta=2\times10^{6}$ 一下就把 RMSE 從初值 46 減至 36，且在 iteration 100 處亦有不錯的斜率，是較剛好的 learning rate。

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： training RMSE=22.537， public score=10.32572， private score=10.37691；
只用 PM2.5： training RMSE=24.302， public score=12.30281， private score=12.51373。

兩種 model 的學習曲線形狀類似，但只用 PM2.5 的 model 總是略遜一籌。我認為使用較多的 feature 才能得到更佳的預測結果。

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。

4~6 (3%) 請參考數學題目，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-b

take $w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$ $\Rightarrow$ $\dfrac{\partial E_D(w)}{\partial w_k} = \dfrac{1}{2} \sum_{n=1}^{3} 2 r_n (t_n - w^T x_n)(-x_{nk})$

$x_n = \begin{bmatrix} x_{n1} \\ x_{n2} \end{bmatrix}$ $\qquad = \sum_{n=1}^{3} r_n (w^T x_n - t_n)(x_{nk})$

the minimum $E_D(w)$ locates at where $\dfrac{\partial E_D(w)}{\partial w_1} = \dfrac{\partial E_D(w)}{\partial w_2} = 0$

$$\Rightarrow \begin{cases} \begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix} \begin{bmatrix} 2 \times 2 \\ 1 \times 5 \\ 3 \times 5 \end{bmatrix} = \begin{bmatrix} 0 & 10 & 5 \end{bmatrix} \begin{bmatrix} 2 \times 2 \\ 1 \times 5 \\ 3 \times 5 \end{bmatrix} \\ \begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix} \begin{bmatrix} 2 \times 3 \\ 1 \times 1 \\ 3 \times 6 \end{bmatrix} = \begin{bmatrix} 0 & 10 & 5 \end{bmatrix} \begin{bmatrix} 2 \times 3 \\ 1 \times 1 \\ 3 \times 6 \end{bmatrix} \end{cases} \Rightarrow \begin{cases} 108 w_1 + 107 w_2 = 125 \\ 107 w_1 + 127 w_2 = 100 \end{cases}$$

$$\Rightarrow w_1 = \frac{5175}{2267}, \quad w_2 = \frac{-2575}{2267} \Rightarrow w^* = \begin{bmatrix} \frac{5175}{2267} \\ \frac{-2575}{2267} \end{bmatrix}$$

4-a

take $w = \begin{bmatrix} w_1 \\ \vdots \\ w_k \end{bmatrix}$, $X = [x_1 \cdots x_n] = \begin{bmatrix} x_{11} & \cdots & x_{n1} \\ \vdots & & \\ x_{1k} & \cdots & x_{nk} \end{bmatrix}$, $r = \begin{bmatrix} r_1 \\ \vdots \\ r_n \end{bmatrix}$, $t = [t_1 \cdots t_n]$

then w^* is the solution of $w^T X \Big(\underbrace{[r\ r \cdots r]}_{k} \circ X^T \Big) = t \Big(\underbrace{[r\ r \cdots r]}_{k} \circ X^T \Big)$,

where $\circ$ is the entrywise product.

5.

after adding the noise, $y'(x_n,w) = w_0 + \sum_{i=1}^{D} w_i(x_i+\epsilon_i) = y(x_n,w) + \sum_{i=1}^{D} \epsilon_i w_i$,

$$\Rightarrow E'(w) = \frac{1}{2}\sum_{n=1}^{N}\left(y'(x_n,w)^2 - 2t_n y'(x_n,w) + t_n^2\right)$$

$$= \frac{1}{2}\sum_{n=1}^{N}\left[(y(x_n,w)-t_n)^2 + 2(y(x_n,w)-t_n)\left(\sum_{i=1}^{D}\epsilon_i w_i\right) + \left(\sum_{i=1}^{D}\epsilon_i w_i\right)^2\right]$$

$$= E(w) + \frac{1}{2}\sum_{n=1}^{N}\left[\sum_{i=1}^{D}\left(\epsilon_i \cdot 2w_i(y(x_n,w)-t_n)\right) + \sum_{i=1}^{D}\sum_{j=i+1}^{D}\left(\epsilon_i\epsilon_j \cdot 2w_iw_j\right) + \sum_{i=1}^{D}\left(\epsilon_i^2 w_i^2\right)\right],$$

because ϵ_i is stochastically independent to x, w and t,

$$\mathbb{E}(E'(w)) = \mathbb{E}(E(w)) + \frac{1}{2}\sum_{n=1}^{N}\Big[\overset{=0}{\mathbb{E}(\epsilon_i)}\,\mathbb{E}(2w_i(y(x_n,w)-t_n)$$
$$+ \overset{=0}{\mathbb{E}(\epsilon_i\epsilon_{j\,(i\neq j)})}\,\mathbb{E}(2w_iw_j)$$
$$+ \mathbb{E}(\epsilon_i^2)\,\mathbb{E}(w_i^2)\Big]$$

$$= \mathbb{E}\left(\frac{1}{2}\sum_{n=1}^{N}\left[(y(x_n,w)-t_n)^2 + \sigma^2\sum_{i=1}^{D} w_i^2\right]\right),$$

which is equivalent to adding

the L2 regularization term $\lambda\sum(w_i^2)$.

6.

the matrix A has n eigenvalues $\lambda_1 \sim \lambda_n$,

their corresponding eigenvectors are $X_1 \sim X_n$,

because A is real, symmetric and non-singular,

we can apply diagonalization on A:

$A = P \Lambda P^{-1}$, where $P = [X_1 \cdots X_n]$ and $\Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$.

$\Rightarrow A^{-1} = P^{-1} \begin{bmatrix} \lambda_1^{-1} & & \\ & \ddots & \\ & & \lambda_n^{-1} \end{bmatrix} P$, thus, A^{-1} has n eigenvalues $\lambda_1^{-1} \sim \lambda_n^{-1}$ ①;

$|A| = \prod_n \lambda_i \Rightarrow \ln|A| = \sum_n \ln(\lambda_i) \Rightarrow \frac{d}{d\alpha} \ln|A| = \sum_n \lambda_i^{-1} \left(\frac{d}{d\alpha} \lambda_i\right)$ ②;

$\frac{d}{d\alpha} A$ has n eigenvalues $\frac{d}{d\alpha}\lambda_1 \sim \frac{d}{d\alpha}\lambda_n$ ③;

……

This equation is equivalent to "Jacobi's formula".