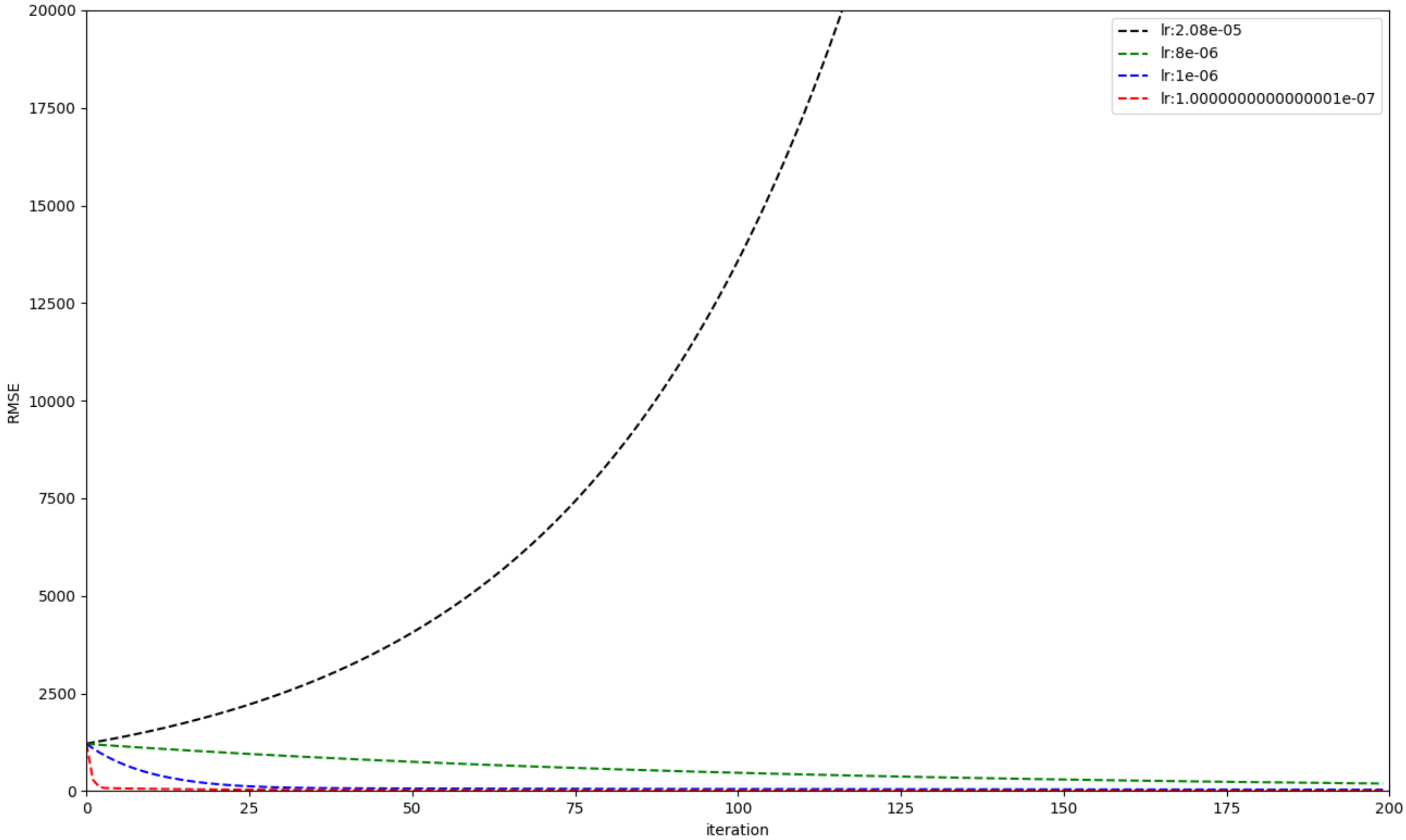
Homework 1 Report - PM2.5 Prediction

學號：R07922135　系級：資工研一　姓名:顏百謙

助教您好：我是有和您說我們要出國參加研討會的同學，所以 1-3 題的內容都只有用 public 來表示。

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。



在 train 每小時氣溫，臭氧，pm10，pm2.5 的狀況下以

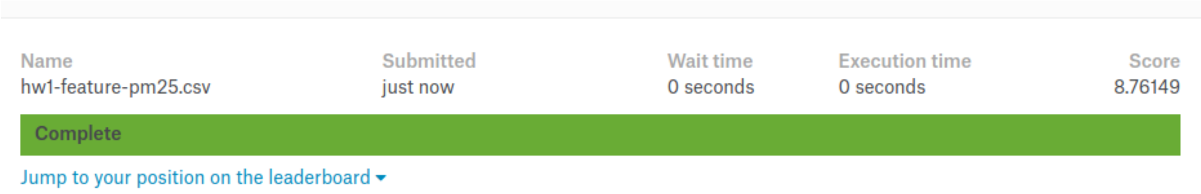
1. 綠線 learning rate = 0.0000001
2. 藍線 learning rate = 0.000001
3. 紅線 learning rate = 0.000008
4. 黑線 learning rate = 0.0000208

Train 200 個 iteration，並對其 RMSE 變化作圖，由圖可知當 learning rate 適量增加時可有效提昇收斂速度，但若 learning rate 過高會反而造成結果變差。

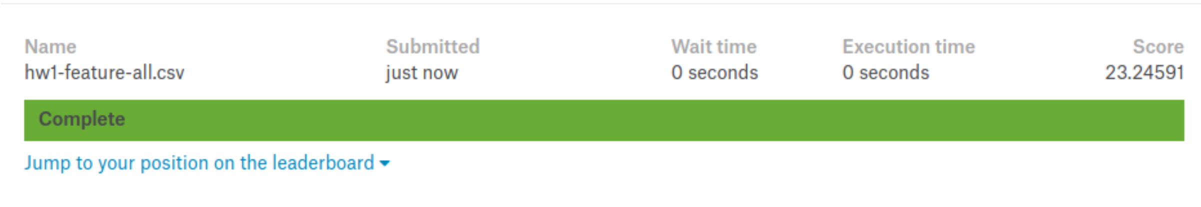
2. (1%) 請分別使用每筆 data9 小時內所有 feature 的一次項（含 bias 項）以及每筆 data9 小時內 PM2.5 的一次項（含 bias 項）進行 training，比較並討論這兩種模型的 root mean-square error（根據 kaggle 上的 public/private score）。

使用處理過的training data / learning rate = 0.000001 / iteration = 20000 進行training 得到

(1.) 僅使用pm2.5的model得到了8.76149 Score



(2.) 使用了全部feature 的model 得到了23.24591 Score



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。

使用處理過的training data / learning rate = 0.000001 / iteration = 20000 / feature = 9 小時內每小時的「氣溫」，「臭氧」，「pm10」，「pm2.5」進行training
得到

λ = 0 時得到 9.08804 Score 且weight 的 L2 norm = 0.9350074747035682

Name	Submitted	Wait time	Execution time	Score
hw1-randa-0.csv	just now	0 seconds	0 seconds	9.08804

Complete

Jump to your position on the leaderboard

λ = 1000 時得到 9.04907 Score 且weight 的 L2 norm = 0.9326304150157435

Name	Submitted	Wait time	Execution time	Score
hw1-randa-1000.csv	just now	0 seconds	0 seconds	9.04907

Complete

Jump to your position on the leaderboard

λ = 100000 時得到 8.77764 Score 且weight 的 L2 norm = 0.7476988271218137

Name	Submitted	Wait time	Execution time	Score
hw1-randa-100000.csv	just now	0 seconds	0 seconds	8.77764

Complete

Jump to your position on the leaderboard

λ = 1000000 時得到9.68059 Score 且weight 的 L2 norm = 0.3520363257057711

Name	Submitted	Wait time	Execution time	Score
hw1-randa-1000000.csv	just now	0 seconds	0 seconds	9.68059

Complete

Jump to your position on the leaderboard

由結果可知，在適度增加 λ 時可以使weight 趨向小值，且使結果變好，但當 λ 過大時會造成 x 的影響力過小(幾乎train 不到）反而使 RMSE 結果變差

4 (1%)

(4-a)

Given t_n is the data point of the data set $\mathcal{D} = \{t_1, \ldots, t_N\}$. Each data point t_n is associated with a weighting factor r_n > 0. The sum-of-squares error function becomes:
$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} r_n(t_n - \mathbf{w}^{\mathbf{T}}\mathbf{x}_n)^2$ Find the solution $\mathbf{w}^*$ that minimizes the error function.

$$Let(\sqrt{r_1}t_1, \sqrt{r_2}t_2, \cdots) = \hat{t} \quad , \begin{bmatrix} x_1^1\sqrt{r_1} & x_2^1\sqrt{r_2} & \ldots & x_n^1\sqrt{r_n} \\ x_1^2\sqrt{r_1} & x_2^2\sqrt{r_2} & \ldots & x_n^2\sqrt{r_n} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^m\sqrt{r_1} & x_2^m\sqrt{r_2} & \ldots & x_n^m\sqrt{r_n} \end{bmatrix} = \widehat{x}$$

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

$$Find\ w^*\ to\ minimize\ E_D(w) = Find\ \triangledown E_D(w^*) = 0$$

$$\triangledown E_D(w^*) = -\widehat{x}\hat{t}^T + \widehat{x}\widehat{x}^T w$$

$$So,\ when\ w^* = (\widehat{x}\widehat{x}^T)^{-1}(\widehat{x}\hat{t}^T)\ can\ minimize\ E_D(w)$$

(4-b)

Following the previous problem(2-a), if

$\mathbf{t} = [t_1 t_2 t_3] = [0 \quad 10 \quad 5], \mathbf{X} = [\mathbf{x_1}\mathbf{x_2}\mathbf{x_3}] = \begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix}$ $r_1 = 2, r_2 = 1, r_3 = 3$ Find the solution $\mathbf{w}^*$.

$$\widehat{x} = \begin{bmatrix} 2\sqrt{2} & 5 & 5\sqrt{3} \\ 3\sqrt{2} & 1 & 6\sqrt{3} \end{bmatrix} \quad , \quad \hat{t} = [0 \quad 10 \quad 5\sqrt{3}]$$

$$\begin{aligned} w^* &= (\begin{bmatrix} 2\sqrt{2} & 5 & 5\sqrt{3} \\ 3\sqrt{2} & 1 & 6\sqrt{3} \end{bmatrix} \begin{bmatrix} 2\sqrt{2} & 3\sqrt{2} \\ 5 & 1 \\ 5\sqrt{3} & 6\sqrt{3} \end{bmatrix})^{-1}(\begin{bmatrix} 125 \\ 100 \end{bmatrix}) \\ &= \begin{bmatrix} 108 & 107 \\ 107 & 127 \end{bmatrix}^{-1} \begin{bmatrix} 125 \\ 100 \end{bmatrix} \\ &= \begin{bmatrix} \frac{5175}{2267} \\ -\frac{2575}{2267} \end{bmatrix} \end{aligned}$$

5 (1%)

Given a linear model:

$y(x, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i$

with a sum-of-squares error function:

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

where t_n is the data point of the data set $\mathcal{D} = \{t_1, \ldots, t_N\}$

Suppose that Gaussian noise ϵ_i with zero mean and variance σ^2 is added independently to each of the input variables x_i. By making use of $\mathbb{E}[\epsilon_i \epsilon_j] = \delta_{ij} \sigma^2$ and $\mathbb{E}[\epsilon_i] = 0$, show that minimizing E averaged over the noise distribution is equivalent to minimizing the sum-of-squares error for noise-free input variables with the addition of a weight -decay regularization term, in which the bias parameter w_0 is omitted from the regularizer.

Hint

- $\delta_{ij} = \begin{cases} 1 (i = j), \\ 0 (i \neq j). \end{cases}$

$$\tilde{y}(x, w) = w_0 + \sum_{i=1}^{D} (w_i x_i)$$

$$\begin{aligned} \widetilde{E}(w) &= \frac{1}{2} \sum_{n=1}^{N} (w_0 + \sum_{i=1}^{D} (w_i x_{i,n}) - t_n)^2 + \lambda \sum_{i=1}^{D} w_i^2 \\ &= \frac{1}{2} \sum_{n=1}^{N} (\tilde{y}_n^2 - 2 \tilde{y}_n t_n + t_n^2) + \lambda \sum_{i=1}^{D} w_i^2 \end{aligned}$$

$$\hat{y}(x,w) = w_0 + \sum_{i=1}^{D}(w_i x_i + \epsilon_i)$$

$$\begin{aligned}
\widehat{E}(w) &= \frac{1}{2}\sum_{n=1}^{N}(w_0 + \sum_{i=1}^{D}(w_i x_{i,n} + \epsilon_{i,n}) - t_n)^2 \\
&= \frac{1}{2}\sum_{n=1}^{N}(\hat{y}_n^2 - 2\hat{y}_n t_n + t_n^2) \\
&= \frac{1}{2}\sum_{n=1}^{N}(\tilde{y}_n^2 + (\epsilon_{1,n} + \epsilon_{2,n} + \cdots)(w_0 + w_1 x_{1,n} + \cdots) + (\epsilon_{1,n}^2 + \epsilon_{2,n}^2 + \cdots) \\
&\quad + (\sum_{i=1}^{D}\sum_{j=i+1}^{D}\epsilon_{i,n}\epsilon_{j,n}) + 2\tilde{y}_n t_n + (t_n\sum_{i=1}^{D}\epsilon_{i,n}) + t_n^2)
\end{aligned}$$

$$because \sum_{i=1}^{D}\sum_{j=i+1}^{D}\epsilon_{i,n}\epsilon_{j,n} = 0\ , \sum_{i=1}^{D}\epsilon_i = 0 \quad we\ can\ get$$

$$\begin{aligned}
&= \frac{1}{2}\sum_{n=1}^{N}(\tilde{y}_n^2 + (\epsilon_{1,n}^2 + \epsilon_{2,n}^2 + \cdots) + 2\tilde{y}_n t_n + t_n^2) \\
&= \frac{1}{2}\sum_{n=1}^{N}(\tilde{y}_n^2 - 2\tilde{y}_n t_n + t_n^2) + \frac{1}{2}\sum_{n=1}^{N}(\epsilon_{1,n}^2 + \epsilon_{2,n}^2 + \cdots) \\
&= \frac{1}{2}\sum_{n=1}^{N}(\tilde{y}_n^2 - 2\tilde{y}_n t_n + t_n^2) + \frac{N}{2}\sigma^2
\end{aligned}$$

因此當我們將 λ 令為 $\frac{N\sigma^2}{2\sum_{i=1}^{D} w_i^2}$ 可獲得相同的效果

6 (1%)

$\mathbf{A} \in \mathbb{R}^{n\times n}$, α is one of the elements of $\mathbf{A}$, prove that

$\frac{\mathrm{d}}{\mathrm{d}\alpha} ln|\mathbf{A}| = Tr\left(\mathbf{A}^{-1}\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathbf{A}\right)$ where the matrix $\mathbf{A}$ is a real, symmetric, non-sigular matrix.

Hint:

- The determinant and trace of $\mathbf{A}$ could be expressed in terms of its eigenvalues.

$$A\ is\ symmetric\ matrix$$

$$A \text{ 可分解為 } PDP^{-1}$$

$$\text{其中 } D = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

$$det(e^D) = e^{\lambda_1} e^{\lambda_2} e^{\lambda_3} \cdots\cdots e^{\lambda_n} = e^{Tr(D)}$$

$$Because\ A^k = PD^k P^{-1},\ for\ all\ k$$

$$we\ can\ get \quad A^{ln(e)} = PD^{ln(e)} P^{-1}$$

$$e^{ln(A)} = Pe^{ln(D)} P^{-1}$$

$$and \quad Tr(A) = Tr(PDP^{-1}) = Tr(D)$$

$$so \quad det(e^{ln(A)}) = det(e^{ln(D)}) = e^{Tr(ln(D))} = e^{Tr(ln(A))}$$

$$det(A) = e^{Tr(ln(A))}$$

$$ln(det(A)) = Tr(ln(A))$$

$$\begin{aligned} \frac{d}{d\alpha} ln|A| &= \frac{d}{d\alpha} Tr(ln(A)) \\ &= Tr(\frac{d}{d\alpha} ln(A)) \quad , By\ (\frac{\partial g(u)}{\partial x} = \frac{\partial g(u)}{\partial u} \frac{\partial u}{\partial x}) \\ &= Tr(\frac{\partial ln(A)}{\partial A} \frac{\partial A}{\partial \alpha}) \\ &= Tr(\frac{1}{A} \frac{dA}{d\alpha}) \end{aligned}$$