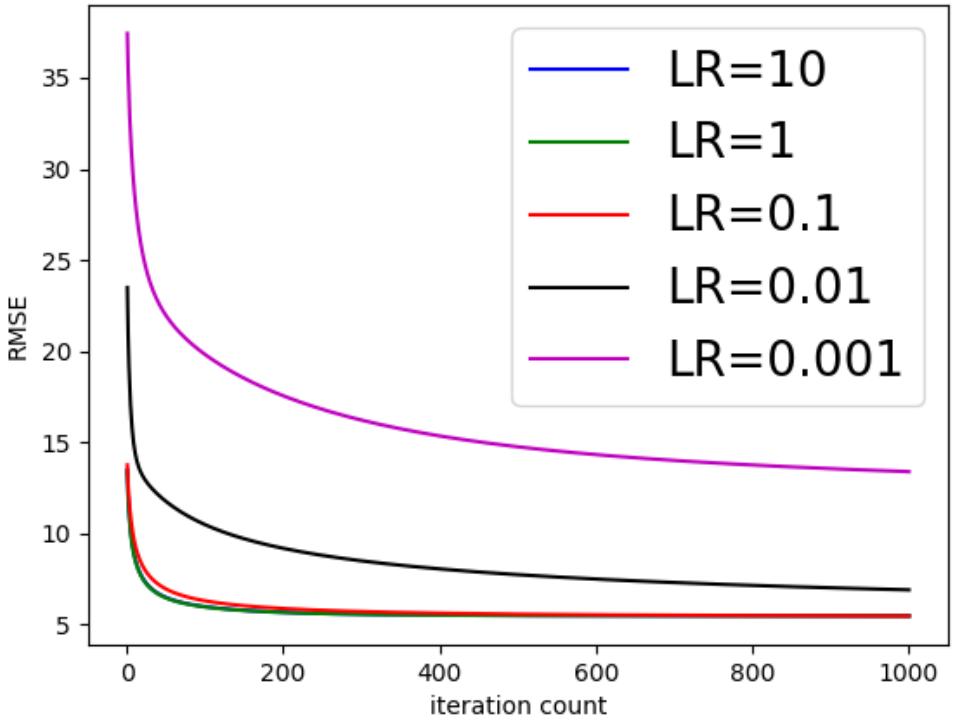
Homework 1 Report - PM2.5 Prediction

學號：R07943123　系級：電子所碩一　姓名：馬咏治

- Report.pdf 檔名錯誤（-1%）
- 學號系級姓名錯誤（-0.5%）

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。

以下為示意圖：



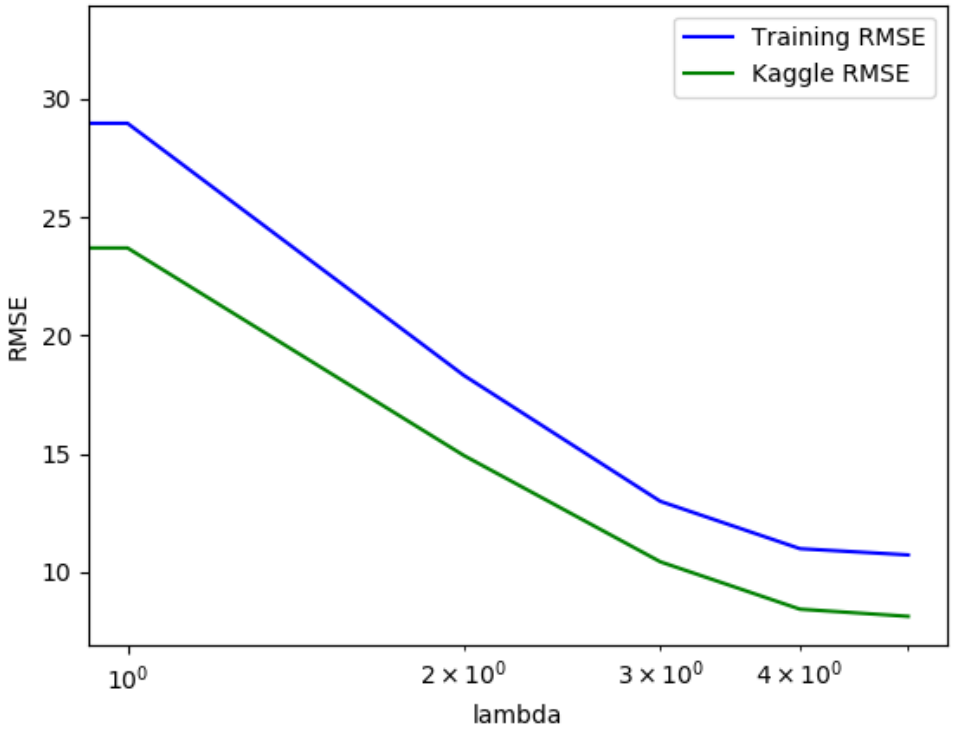
由上圖來看，我們的 learning rate 越高，出來的 performance 會越好。我想這是因為我在實作 Gradient Descent 時是使用 Ada Grad 這個方法來調整 learning rate。Ada Grad 會讓 learning rate decay 的非常快，如果一開始 learning rate 設太低，那麼就會被第一筆資料所 dominant，第二甚至後面幾筆資料都無法把它從 suboptimal tune 出來。基於這些原因 LR 在較低的情況，會困在 suboptimal，所以收斂速度較慢，反之 LR 較高的情況，收斂速度較快。

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）。

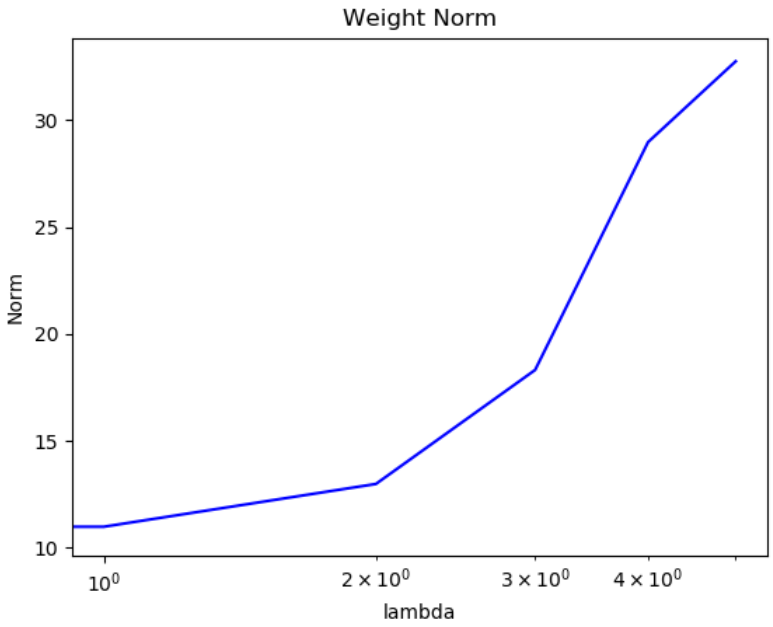
	RMSE error（public）	RMSE error（private）
ALL Feature	8.584	8.19754
One Feature	8.63806	8.50318

從上圖上看，取所有 feature 和單取 PM2.5，在 Public 並沒有太大的差異，但在 Private 就有顯著的差異。取所有 feature 的 performance 略勝過一個 feature，我想原因在於多 feature 可以更好的去 representPM2.5 的預測值。

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。



這裡分別使用 regularization lambda = 1000,100000,1000000,100000000 進行實驗，由於我 training 出來的 weight 值都較小，因此 regularization 強度要跳到這麼高才看得出差別。從實驗結果發現，regularization term 並不會影響我們的結果，因此當 regularization 值越低時，performance 越好，原因在於原本我 train 出來的 weight 值以及很小，並不需要再壓。



當我們的 regularization term 越大時，我們的 weight norm 越來越小，這個是符合我們的直覺的，regularization 會抑制 weight 的生長。

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 (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:

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} r_n (t_n - \mathbf{w}^{\mathrm{T}}\mathbf{x}_n)^2$$

Find the solution $\mathbf{w}^*$ that minimizes the error function.

Solution:

Let $\mathbf{R} = \begin{bmatrix} r_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & r_n \end{bmatrix}$, $\mathbf{W} = \begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix}$, $\mathbf{T} = \begin{bmatrix} t_1 \\ \vdots \\ t_n \end{bmatrix}$, $\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}$

$$\text{So that SSE} = E_D(w) = (\mathbf{T} - \mathbf{XW})^{\mathbf{T}}\mathbf{R}(\mathbf{T} - \mathbf{XW})$$

$$= \mathbf{T}^{\mathbf{T}}\mathbf{RT} - \mathbf{w}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{RT} + \mathbf{w}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{RXw} - \mathbf{T}^{\mathbf{T}}\mathbf{RXw}$$

From Eq. above, we derive its derivation

$$\frac{\text{dSSE}}{\text{d}\mathbf{w}} = -\mathbf{X}^{\mathbf{T}}\mathbf{RT} - \mathbf{X}^{\mathbf{T}}\mathbf{RT} + 2\mathbf{X}^{\mathbf{T}}\mathbf{RXw}$$

$$\text{Hence } \mathbf{w} = (\mathbf{X}^{\mathbf{T}}\mathbf{RX})^{-1}(\mathbf{X}^{\mathbf{T}}\mathbf{RT})_{\#}$$

(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 $\mathbf{w}^*$.

Let $\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{3}{2} \end{bmatrix}$, By obey the initial setting, we redefine the parameter t and X which is $\mathbf{t} = \begin{bmatrix} 0 \\ 10 \\ 5 \end{bmatrix}$ $\mathbf{X} = \begin{bmatrix} 2 & 3 \\ 5 & 1 \\ 5 & 6 \end{bmatrix}$

Substitute into the eq.

$$\mathbf{w} = \left(\begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{3}{2} \end{bmatrix}\begin{bmatrix} 2 & 3 \\ 5 & 1 \\ 5 & 6 \end{bmatrix}\right)^{-1}\begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{3}{2} \end{bmatrix}\begin{bmatrix} 0 \\ 10 \\ 5 \end{bmatrix}$$

$$=\left(\begin{bmatrix}54 & 53.5\\ 53.5 & 63.5\end{bmatrix}\right)^{-1}\begin{bmatrix}62.5\\ 50\end{bmatrix}$$

$$=\frac{1}{566.75}\begin{bmatrix}63.5 & -53.5\\ -53.5 & 54\end{bmatrix}\begin{bmatrix}62.5\\ 50\end{bmatrix}$$

$$=\begin{bmatrix}\frac{5175}{2267}\\ \frac{-2575}{2267}\end{bmatrix}\approx\begin{bmatrix}-2.283\\ -1.136\end{bmatrix}_{\#}$$

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

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

Noise case: $\mathrm{E}(w')$
Noise-free case: $\mathrm{E}(w)$

$$\mathrm{E}(w')=\frac{N}{2}\ E\ [(y(\mathbf{x}+\boldsymbol{\varepsilon},\mathbf{w})-\mathbf{t})^2]$$

$$=\mathrm{E}(\mathbf{w})+\frac{N}{2}\ E[2(\mathrm{y}(\mathbf{x},\mathbf{w})-\mathrm{t})\left(\textstyle\sum_{i=1}^{D}w_i\varepsilon_i\right)]+\ \frac{N}{2}\ E[\left(\textstyle\sum_{i=1}^{D}w_i\varepsilon_i\right)^2]$$

$$= \mathrm{E}(\mathbf{w}) + \frac{N}{2}\ E[w_1^2\varepsilon_1^2 + w_2^2\varepsilon_2^2 + \cdots + w_n^2\varepsilon_n^2]$$

$$= \mathrm{E}(\mathbf{w}) + \ \sigma^2\|\mathbf{w}\|^2$$

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.

Reference: https://en.wikipedia.org/wiki/Jacobi%27s_formula

From Jacobi Formula:

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}|\mathrm{A}| = \mathrm{Tr}(\mathrm{adj}(\mathrm{A})\frac{\mathrm{dA}(\alpha)}{\mathrm{d}\alpha})$$

Hence, by applying Jacobi Formula and Chain Rule

Notation:

$$\lambda_1, \lambda_2, \ldots, \lambda_n \ \ is \ \ the \ \ n \ \ eigenvalue \ \ of \ \ A$$

LHS:

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} ln|\mathrm{A}| = \frac{\mathrm{d}}{\mathrm{d}\alpha} \ln\lambda_1\lambda_2 \dots \lambda_n$$

$$= \frac{1}{\lambda_1\lambda_2 \dots \lambda_n} \frac{\mathrm{d}}{\mathrm{d}\alpha} \lambda_1\lambda_2 \dots \lambda_n$$

$$= \frac{1}{|\mathrm{A}|} \frac{\mathrm{dA}}{\mathrm{d}\alpha}$$

$$= \frac{1}{|\mathrm{A}|} \mathrm{Tr}(\mathrm{adj}(\mathrm{A}) \frac{\mathrm{dA}(\alpha)}{\mathrm{d}\alpha})$$

Since A is invertible

$$= \frac{1}{|\mathrm{A}|} |\mathrm{A}| \mathrm{Tr}(A^{-1} \frac{\mathrm{dA}(\alpha)}{\mathrm{d}\alpha})$$

$$= \mathrm{Tr}(A^{-1} \frac{\mathrm{dA}(\alpha)}{\mathrm{d}\alpha})_{\#} \quad \text{(RHS)}$$