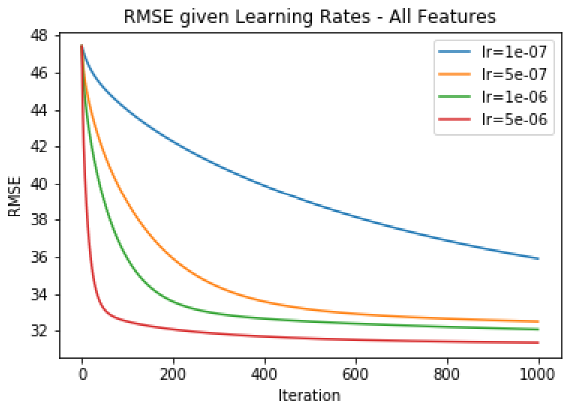
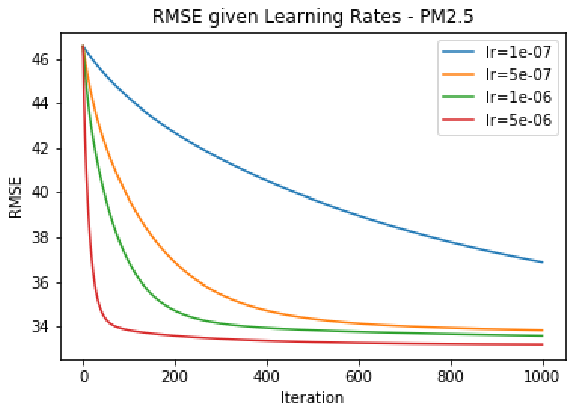
Homework 1 Report - PM2.5 Prediction

學號：R07946007　系級： 資料科學學程碩一　姓名: 陳庭安

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。





左右圖分別為考慮所有 features 與僅考慮 PM2.5 資料，在 Learning rate 為 10^{-7}、$5x10^{-7}$、10^{-6} 及 $5x10^{-6}$ 時，1000 次 iterations RMSE 的值。兩圖 RMSE 均隨更新參數次數增加而越小，Learning rate 設太小，如 10^{-7}、$5x10^{-7}$，RMSE 就越慢才收斂，Learning rate 適當取大一些的值如 10^{-6}，收斂較快，約 40 次 iterations 後收斂。

2. (1%) 請分別使用每筆 data 9 小時內所有 feature 的一次項（含 bias 項）以及每筆 data 9 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）。

Features	Training_RMSE	Testing_RMSE(Public/Private)
所有 features	31.5359	17.72361 / 13.99641
PM2.5	32.3827	9.63926 / 10.25875

用 PM2.5 每 9 小時 data 作為 features 去 train，得到的 Training_RMSE 略高於用所有 features 的 model 去 train 得到的 loss；然而在 Testing 時，無論是 public 還是 private 的結果，僅以 PM2.5 為 model features 的 RMSE 反而下降不少。使用了所有 features 的 model 雖然在 training set 表現得不錯，但在 testing 卻表現得很差，很可能是因模型過於複雜而產生 overfitting 的問題。

3. (1%) 請分別使用至少四種不同數值的 regularization parameter λ 進行 training（其他參數需一致），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。

Features / Lambdas	Training_RMSE	Testing_RMSE(Public / Private)	weight 的 L2 norm
所有 features / 0.0	31.53599123359141	17.74519 / 14.01537	17.7440
所有 features / 0.3	31.53596761374975	17.73865 / 14.00962	17.7430
所有 features / 0.6	31.535949543595223	17.73217 / 14.00393	17.7420
所有 features / 1.0	31.535933927239096	17.72361 / 13.99641	17.7407
PM2.5 / 0.0	32.3827276201007	9.63926 / 10.25875	17.7407
PM2.5 / 0.3	32.38273340914199	9.63905 / 10.25808	17.7407
PM2.5 / 0.6	32.38273935556728	9.63884 / 10.25742	17.7407
PM2.5 / 1.0	32.382747528789125	9.63857 / 10.25654	17.7407

同 2，用 PM2.5 每 9 小時 data 作為 features 去 train，得到的 Training_RMSE 略高於用所有 features 的 model 去 train 得到的 loss；然而在 Testing 時，僅以 PM2.5 為 model features 的 RMSE 反而下降不少。使用所有 features 的 model 在 training set 表現得不錯，但在 testing 卻表現得很差，很可能是因模型過於複雜而產生 overfitting 的問題。

用所有 features 的 model，隨著 lambda 值越大，loss 變動並不是很大，很小幅度的下滑；weight 的 L2 norm 越小，表示對參數值有一定的懲罰，減少有時因參數值過大而造成預測結果波動大、不穩定的效果。

另外只用 PM2.5 的 features 的 model，在此隨著 lambda 值越大，loss、weight 的 L2 norm 變動都不是很大，預測結果波動較使用所有 features 的 model 穩定，即較簡單的模型，其 Variance 較小。

4 (a) $\underline{w}^* = \arg\min_{\underline{w}} \sum_{n=1}^{N} r_n (t_n - \underline{w}^T \underline{x}_n)^2$

$\stackrel{\wedge}{=} E = \sum_{n=1}^{N} r_n (t_n - \underline{w}^T \underline{x}_n)^2$, $\underline{t} = [t_1 t_2 \cdots t_N]$, $\underline{x} = [\underline{x}_1\ \underline{x}_2 \cdots \underline{x}_N]$,

$$\underline{r} = \begin{bmatrix} r_1 & 0 & \cdots & 0 \\ 0 & r_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_N \end{bmatrix}_{N \times N}$$

$$\Rightarrow E = (\underline{t} - \underline{w}^T \underline{x})\, \underline{r}\, (\underline{t} - \underline{w}^T \underline{x})^T$$

$$= (\underline{t}\,\underline{r} - \underline{w}^T \underline{x}\,\underline{r})(\underline{t}^T - \underline{x}^T \underline{w})$$

$$= \underline{t}\,\underline{r}\,\underline{t}^T - \underline{w}^T \underline{x}\,\underline{r}\,\underline{t}^T - \underline{t}\,\underline{r}\,\underline{x}^T \underline{w} + \underline{w}^T \underline{x}\,\underline{r}\,\underline{x}^T \underline{w}$$

$$\Rightarrow E(\underline{w} + \Delta \underline{w}) - E(\underline{w})$$

$$= \Big[\cancel{\underline{t}\,\underline{r}\,\underline{t}^T} - (\underline{w} + \Delta \underline{w})^T \underline{x}\,\underline{r}\,\underline{t}^T - \underline{t}\,\underline{r}\,\underline{x}^T (\underline{w} + \Delta \underline{w}) + (\underline{w} + \Delta \underline{w})^T \underline{x}\,\underline{r}\,\underline{x}^T (\underline{w} + \Delta \underline{w})\Big]$$

$$- \Big[\cancel{\underline{t}\,\underline{r}\,\underline{t}^T} - \underline{w}^T \underline{x}\,\underline{r}\,\underline{t}^T - \underline{t}\,\underline{r}\,\underline{x}^T \underline{w} + \underline{w}^T \underline{x}\,\underline{r}\,\underline{x}^T \underline{w}\Big]$$

$$= -\Delta \underline{w}^T \underline{x}\,\underline{r}\,\underline{t}^T - \underbrace{\underline{t}\,\underline{r}\,\underline{x}^T \Delta \underline{w}}_{= \Delta \underline{w}^T (\underline{t}\,\underline{r}\,\underline{x}^T)^T} + \underbrace{\underline{w}^T \underline{x}\,\underline{r}\,\underline{x}^T \Delta \underline{w}}_{= \Delta \underline{w}^T (\underline{w}^T \underline{x}\,\underline{r}\,\underline{x}^T)^T} + \Delta \underline{w}^T \underline{x}\,\underline{r}\,\underline{x}^T \underline{w} + \Delta \underline{w}^T \underline{x}\,\underline{r}\,\underline{x}^T \Delta \underline{w}$$

$$= 2 \Delta \underline{w}^T \underline{x}\,\underline{r}\,\underline{x}^T \underline{w} - 2 \Delta \underline{w}^T \underline{x}\,\underline{r}\,\underline{t}^T + \Delta \underline{w}^T \underline{x}\,\underline{r}\,\underline{x}^T \Delta \underline{w}$$

$$\doteq \Delta \underline{w}^T \left[2 \underline{x}\,\underline{r}\,\underline{x}^T \underline{w} - 2 \underline{x}\,\underline{r}\,\underline{t}^T\right]$$

$$\Rightarrow \stackrel{\wedge}{=} \nabla_{\underline{w}} E = 2 \underline{x}\,\underline{r}\,\underline{x}^T \underline{w} - 2 \underline{x}\,\underline{r}\,\underline{t}^T = 0$$

$$\Rightarrow \underline{w}^* = (\underline{x}\,\underline{r}\,\underline{x}^T)^{-1} \underline{x}\,\underline{r}\,\underline{t}^T$$

#

4.(b) $\quad w^* = (\underline{x}\,\underline{r}\,\underline{x}^T)^{-1}\,\underline{x}\,\underline{r}\,\underline{t}^T$

$$\underline{x}\,\underline{r}\,\underline{x}^T = \begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 5 & 1 \\ 5 & 6 \end{bmatrix} = \begin{bmatrix} 108 & 107 \\ 107 & 127 \end{bmatrix}$$

$$(\underline{x}\,\underline{r}\,\underline{x}^T)^{-1} = \frac{1}{2267} \begin{bmatrix} 127 & -107 \\ -107 & 108 \end{bmatrix}$$

$$\Rightarrow w^* = \frac{1}{2267} \begin{bmatrix} 127 & -107 \\ -107 & 108 \end{bmatrix} \begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 10 \\ 5 \end{bmatrix}$$

$$= \frac{1}{2267} \begin{bmatrix} -67 & 528 & -7 \\ 110 & -427 & 113 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 10 \\ 5 \end{bmatrix}$$

$$= \frac{1}{2267} \begin{bmatrix} -134 & 528 & -21 \\ 220 & -427 & 339 \end{bmatrix} \begin{bmatrix} 0 \\ 10 \\ 5 \end{bmatrix} = \begin{bmatrix} \frac{5175}{2267} \\ \frac{-2575}{2267} \end{bmatrix} = \begin{bmatrix} 2.2828 \\ -1.1359 \end{bmatrix}$$

#

5. 令 $\underline{W} = [w_1\ w_2 \cdots w_D]_{1\times D}$, $\underline{w_0} = [w_0\ w_0 \cdots w_0]_{1\times N}$

$\underline{x} = [\underline{x_1}\ \underline{x_2} \cdots \underline{x_N}]_{D\times N}$, $\underline{\varepsilon} = [\underline{\varepsilon_1}\ \underline{\varepsilon_2} \cdots \underline{\varepsilon_D}]_{N\times D}$

$\underline{t} = [t_1\ t_2 \cdots t_N]_{1\times N}$

$$E\left[\left(\underline{w}(\underline{x}^T+\underline{\varepsilon})^T+\underline{w_0}-\underline{t}\right)\left(\underline{w}(\underline{x}^T+\underline{\varepsilon})^T+\underline{w_0}-\underline{t}\right)^T\right]$$

$$= E\left[(\underline{w}\underline{x}+\underline{w}\underline{\varepsilon}^T+\underline{w_0}-\underline{t})(\underline{x}^T\underline{w}^T+\underline{\varepsilon}\underline{w}^T+\underline{w_0}^T-\underline{t}^T)\right]$$

$$= E\Big[(\underline{w}\underline{x}\underline{x}^T\underline{w}^T+\underline{w}\underline{x}\underline{\varepsilon}\underline{w}^T+\underline{w}\underline{x}\underline{w_0}^T-\underline{w}\underline{x}\underline{t}^T)+(\underline{w}\underline{\varepsilon}^T\underline{x}^T\underline{w}^T+\underline{w}\underline{\varepsilon}^T\underline{\varepsilon}\underline{w}^T+\underline{w}\underline{\varepsilon}^T\underline{w_0}^T-\underline{w}\underline{\varepsilon}^T\underline{t}^T)$$

$$+(\underline{w_0}\underline{x}^T\underline{w}^T+\underline{w_0}\underline{\varepsilon}\underline{w}^T+\underline{w_0}\underline{w_0}^T-\underline{w_0}\underline{t}^T)+(-\underline{t}\underline{x}^T\underline{w}^T-\underline{t}\underline{\varepsilon}\underline{w}^T-\underline{t}\underline{w_0}^T+\underline{t}\underline{t}^T)\Big]$$

$$= E\Big[(\underline{w}\underline{x}+\underline{w_0}-\underline{t})(\underline{w}\underline{x}+\underline{w_0}-\underline{t})^T+(\underline{w}\underline{x}\underline{\varepsilon}\underline{w}^T+\underline{w}\underline{\varepsilon}^T\underline{x}^T\underline{w}^T+\underline{w}\underline{\varepsilon}^T\underline{\varepsilon}\underline{w}^T+\underline{w}\underline{\varepsilon}\underline{w_0}^T-\underline{w}\underline{\varepsilon}^T\underline{t}^T$$

$$+\underline{w_0}\underline{\varepsilon}\underline{w}^T-\underline{t}\underline{\varepsilon}\underline{w}^T)\Big]$$

$$= (\underline{w}\underline{x}+\underline{w_0}-\underline{t})(\underline{w}\underline{x}+\underline{w_0}-\underline{t})^T+0+0+\underline{w}E(\underline{\varepsilon}^T\underline{\varepsilon})\underline{w}^T+0-0+0-0$$

$$= (\underline{w}\underline{x}+\underline{w_0}-\underline{t})(\underline{w}\underline{x}+\underline{w_0}-\underline{t})^T+\underline{w}\begin{bmatrix}\sigma^2 & & 0\\ & \sigma^2 & \\ 0 & \ddots & \sigma^2\end{bmatrix}_{D\times D}\underline{w}^T$$

$$= (\underline{w}\underline{x}+\underline{w_0}-\underline{t})(\underline{w}\underline{x}+\underline{w_0}-\underline{t})^T+\sigma^2\cdot\underline{w}\underline{w}^T$$

$$\therefore E\left[\frac{1}{2}\sum_{n=1}^{N}(y(x_n+\varepsilon_n,\underline{W})-t_n)^2\right]$$

$$= \frac{1}{2}E\left[(\underline{w}(\underline{x}^T+\underline{\varepsilon})^T+\underline{w_0}-\underline{t})(\underline{w}(\underline{x}^T+\underline{\varepsilon})^T+\underline{w_0}-\underline{t})^T\right]$$

$$= \frac{1}{2}(\underline{w}\underline{x}+\underline{w_0}-\underline{t})(\underline{w}\underline{x}+\underline{w_0}-\underline{t})^T+\frac{1}{2}\sigma^2\underline{w}\underline{w}^T$$

$$= \frac{1}{2}\sum_{n=1}^{N}(y(x_n-\underline{W})-t_n)^2+\frac{1}{2}\sigma^2\sum_{i=1}^{D}w_i^2$$

6. $A \in \mathbb{R}^{n\times n}$, symmetric, non-singular matrix.

Prove that $\frac{d}{d\alpha} \ln|A| = tr(A^{-1}\frac{d}{d\alpha}A)$.

<pf> $\because A \in \mathbb{R}^{n\times n}$ and is symmetric

$\therefore$ A is diagonalizable.

$\Rightarrow \exists$ invertible matrix $P \ni A = PDP^{-1}$, where D is diagonal.

左式 $= \frac{d}{d\alpha}\ln|A| = \frac{d}{d\alpha}\ln|PDP^{-1}| = \frac{d}{d\alpha}\ln(|P||D|\cdot\frac{1}{|P|})$

$= \frac{d}{d\alpha}\ln|D|$

右式 $= tr(A^{-1}\frac{d}{d\alpha}A) = tr((PDP^{-1})^{-1}\frac{d}{d\alpha}(PDP^{-1}))$

$= \underbrace{tr(PD^{-1}P^{-1}(\frac{d}{d\alpha}P)DP^{-1})}_{①} + \underbrace{tr(PD^{-1}P^{-1}P(\frac{d}{d\alpha}D)P^{-1})}_{②}$

$+ \underbrace{tr(PD^{-1}P^{-1}PD(\frac{d}{d\alpha}P^{-1}))}_{③}$

[$tr(AB)=tr(BA)$]

$① = tr((\frac{d}{d\alpha}P)DP^{-1}PD^{-1}P^{-1}) = tr((\frac{d}{d\alpha}P)P^{-1})$.

$② = tr(PD^{-1}(\frac{d}{d\alpha}D)P^{-1}) = tr((\frac{d}{d\alpha}D)P^{-1}PD^{-1}) = tr((\frac{d}{d\alpha}D)D^{-1})$

$③ = tr(P(\frac{d}{d\alpha}P^{-1}))$

$① + ③ = tr(\frac{d}{d\alpha}(PP^{-1})) = tr(\frac{d}{d\alpha}I) = 0$.

$\therefore$ 右式 $= ①+②+③ = ② = tr((\frac{d}{d\alpha}D)D^{-1})$

$= \sum_{i=1}^{n}\frac{\frac{d}{d\alpha}a_{ii}}{a_{ii}}$, where $a_{ii}, i=1,\cdots,n$ are the diagonal elements of D.

$= \sum_{i=1}^{n}\frac{d}{d\alpha}\ln a_{ii} = \frac{d}{d\alpha}\ln\prod_{i=1}^{n}a_{ii} = \frac{d}{d\alpha}\ln|D|$

故左式=右式，得證.