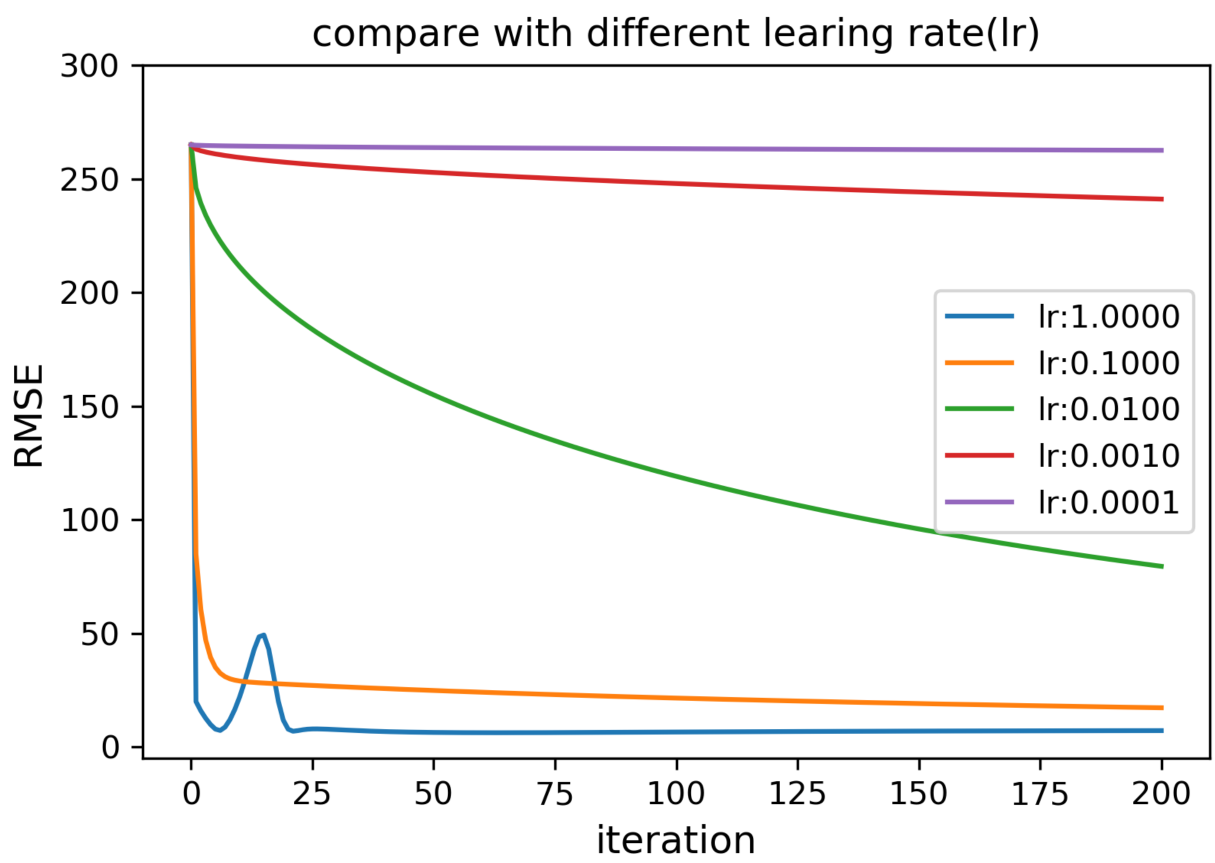
# Homework 1 Report - PM2.5 Prediction

學號：B06209027　系級：大氣二　姓名:李冠勳

## 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。




1. 資料整理:
    - 將PM2.5數值0以下或200以上的數值調整成與前一小時的數值相同
    - 將每個月20天的資料串聯
2. 使用PM2.5前九小時的資料的一次項加bias分析第十小時的PM2.5數值，並使用adagrad進行測試，然後比較五種不同的learning rate(分別是1,0.1,0.01,0.001,0.0001)下收斂的過程，分析結果如上圖
    - 紫色及紅色線:learning rate太小因此RMSE無法在短時間內找到最佳解
    - 綠線:可以看出收斂過程但速度還是太慢
    - 橘線:是這五個過程中最佳的，快速地達到最小值
    - 藍線:觀察圖中可以發現藍線有一個小峰然後又降到更低點，是因為learning rate太大導致到局部最佳解時，參數改變量依然很大，然後又掉入另一個局部最佳解，在本題卻也因此得到更好的結果但並不是我們所期待的

# 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）。

| Features | Training | public score | private score |
|---|---|---|---|
| All | 5.618217963847283 | 39.90107 | 39.41352 |
| pm2.5 | 5.831618085848395 | 7.01526 | 7.28936 |

可以發現維度較高的測試(All)在train時比維度較低的測試(PM2.5)較好一點點，然而在public和private的成績上就有很大的落差，All的成績明顯差了許多，因為維度較高時，可以找到一個較符合train的model但這個model卻僅僅只適用於train，因此換成test的資料時結果就非常糟糕，反觀PM2.5的測試test的結果並沒油比train差很多，算是一個可行的model。

# 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。

| lambda | Training | public score | private score |
|---|---|---|---|
| 1 | 5.8422450751542065 | 7.05468 | 7.34491 |
| 10 | 6.039144334750001 | 7.42993 | 7.80754 |
| 100 | 7.467837591291416 | 9.32254 | 9.83354 |
| 1000 | 10.309557968227256 | 12.82488 | 13.15144 |

使用與前題PM2.5相同的model進行分析，發現此model加入lambda使取線更平滑並不能得到更好的結果，觀察以上4種不同的lambda，其中lambda越大可以使曲線更平滑但也使結果越來越差。

# 4 (1%)

## (4-a)

首先定義
$R$是nxn的矩陣
$T$是nx1的矩陣
$X$是mxn的矩陣
$W$是1xm的矩陣

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

$$\begin{aligned}\frac{d}{dw}E_D(\mathbf{w}) = \frac{d}{dw}(\frac{1}{2}R(T^{\mathbf{T}}T - 2TWX + (WX)^{\mathbf{T}}(WX))) \\ = \frac{1}{2}R(-2X^{\mathbf{T}}T + 2X^{\mathbf{T}}WX) \\ = X^{\mathbf{T}}RWX - X^{\mathbf{T}}RT \\ = 0_{1\text{x}m}\end{aligned}$$

$$W = (X^{\mathbf{T}}RX)^{-1}(X^{\mathbf{T}}RT)$$

### (4-b)

將題目給的矩陣帶入上式

$$T = \begin{bmatrix} 0 \\ 10 \\ 5 \end{bmatrix}$$

$$X^T = \begin{bmatrix} 2 & 3 \\ 5 & 1 \\ 5 & 6 \end{bmatrix}$$

$$R = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

$$W \approx \begin{bmatrix} 2.28 \\ -1.14 \end{bmatrix}$$

## 5 (1%)

Collaborator: b05902109 柯上優
將雜訊$\epsilon$加進 $x$，linear model變成

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

$$E_\epsilon(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\big(y((x_n+\epsilon_i),\mathbf{w})-t_n\big)^2$$
$$=\frac{1}{2}\sum_{n=1}^{N}\big(y(x_n,\mathbf{w})+\sum_{d=1}^{D}w_d\epsilon_{nd}-t_n\big)^2$$
$$=\frac{1}{2}\sum_{n=1}^{N}\left((y(x_n,\mathbf{w})-t_n)^2+2(y(x_n,\mathbf{w})-t_n)(\sum_{d=1}^{D}w_d\epsilon_{nd})+(\sum_{d=1}^{D}w_d\epsilon_{nd})^2\right)^2$$

得到三項後分別取期望值:

$$\mathbb{E}[E_\epsilon(\mathbf{w})]=\frac{1}{2}\sum_{n=1}^{N}\left((y(x_n,\mathbf{w})-t_n)^2+2(y(x_n,\mathbf{w})-t_n)(\sum_{d=1}^{D}w_d\mathbb{E}[\epsilon_{nd}])+\mathbb{E}[(\sum_{d=1}^{D}w_d\epsilon_{nd})^2]\right)^2$$

分別討論三項。第一項沒有雜訊，由於$E[\epsilon_i]=0$所以第二項等於0，至於第三項如下:

$$\mathbb{E}[(\sum_{d=1}^{D}w_d\epsilon_{nd})^2]$$
$$=\mathbb{E}[\sum_{i=1}^{D}\sum_{j=1}^{D}w_iw_j\epsilon_i\epsilon_j]$$
$$=\sum_{i=1}^{D}\sum_{j=1}^{D}w_iw_j\mathbb{E}[\epsilon_i\epsilon_j]$$
$$\text{因為}E[\epsilon_i\epsilon_j]=\delta_{ij}\sigma^2\text{且}i=j\text{時}\delta_{ij}=1\text{，}i\neq j\text{時}\delta_{ij}=0$$
$$=\sum_{d=1}^{D}w_dw_d\sigma^2=w^2\sigma^2$$

代入原式

$$\mathbb{E}[E_\epsilon(\mathbf{w})]$$
$$=\mathbb{E}[E(\mathbf{w})]+\frac{N}{2}w^2\sigma^2$$

得證，有雜訊E的最小值等於沒有雜訊E加上 weight -decay regularization term的最小值，此外後項的bias也被消去了，與題目要求符合。

## 6 (1%)

Collaborator: b05902109 柯上優 、 b05902074 魏佑珊
首先證明矩陣A有以下的性質

$$det(exp(A))=exp(Tr(A))$$

證明:

$$det(exp(A)) = \prod_{i=1}^{N} exp(\lambda_i) = exp(\sum_{i=1}^{N} \lambda_i) = exp(Tr(A))$$

$$\lambda_i\text{是矩陣}A\text{的對角線中第}i\text{個元素}$$

今假設矩陣 $B = ln(A)$ ，我們有以下特性:

$$det(A) = det(exp(lnA)) = det(exp(B)) = exp(Tr(B)) = exp(Tr(lnA))$$

接著兩邊取 $ln$:

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

最後以$\alpha$微分:

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} ln(det(\mathbf{A})) = \frac{\mathrm{d}}{\mathrm{d}\alpha} Tr(lnA) = Tr(A^{-1} \frac{\mathrm{d}}{\mathrm{d}\alpha} A)$$