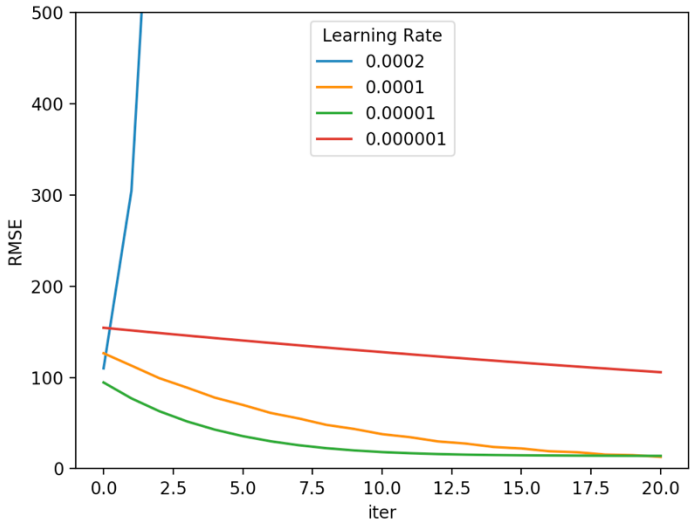
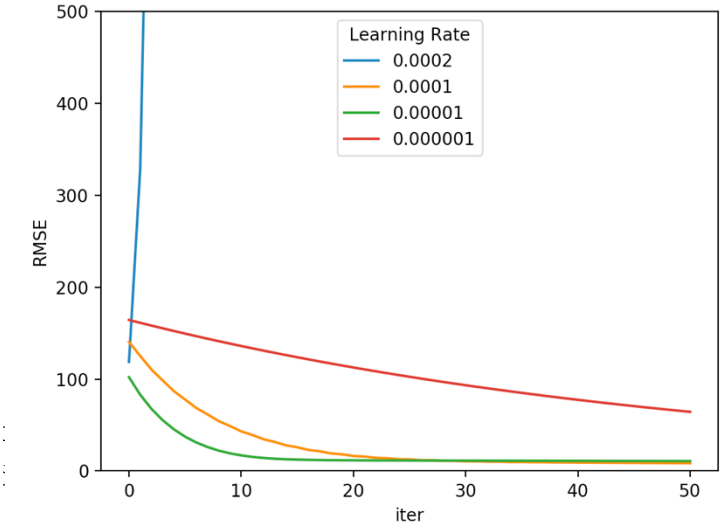
Homework 1 Report - PM2.5 Prediction

學號：B05902022　系級：資工三　姓名: 張雅信

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。



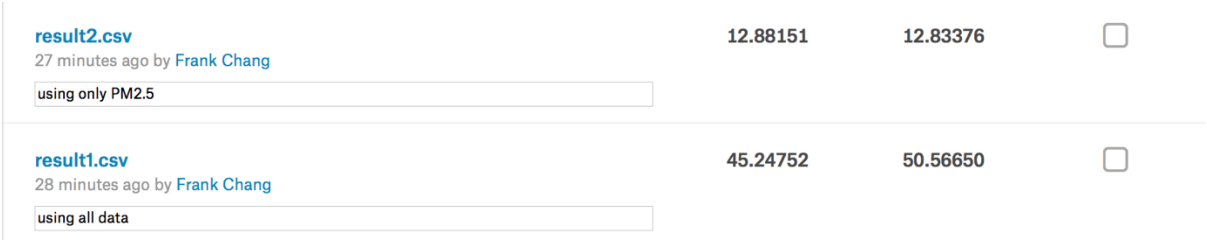


在learning rate大於 0.0002 的時候，一次gradient descent 後對參數的改變會直接跨越loss最小值，甚至使參數的位置距離loss最小值處更遠，以至於loss迅速增加，regression失敗。

在learning rate介於 0.0001-0.00001 的時候(橘線與綠線之間)，參數移動的距離適當，loss符合預期的逐漸下降，最後是好的結果。

在learning rate小於 0.000001 的時候，每次更新的距離太短，使得loss下降緩慢，需要更多step才能使loss逼近最小值。由於會浪費很多不必要的計算資源，這是不恰當的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）。



若使用所有一次項，因為考慮了很多不必要的因素，使得error 很大。相對的，僅使用PM2.5的資料，與要預測的項目相關性增加很多，表現較好。

但僅使用PM2.5 的資料，放棄了很多可能相關的資料，並不一定是最佳解。

3. (1%)請分別使用至少四種不同數值的regulization parameterλ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。

parameter	Training error	Public score	Private score	Weight L2 norm
0	6.70	8.75	9.49	0.66
0.01	6.79	9.10	9.72	0.63
1	6.85	9.02	10.04	0.72
100	7.64	9.97	10.67	0.25

本次模型使用Linear Regression，也就是線性的函數，不會有高維度的項讓模型overfit的機會，故較不需要regularization。就結果來看差異並不大，可能原本的參數就都沒有離 0 太遠。 weight的L2 norm 則會隨著 λ 增加而減少。

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)

$$Let \quad x_i' = \sqrt{r_i} x_i \quad and \quad t_i' = \sqrt{r_i} t_i$$

$$E_d(w) = \frac{1}{2}\sum_{n=1}^{N} r_n (t_n - w^T x_n)^2 = \frac{1}{2}\sum_{n=1}^{N} (\sqrt{r_n} t_n - \sqrt{r_i} x_n^T w)^2 = \frac{1}{2} \| \begin{bmatrix} -- & x_1' & -- \\ -- & x_2' & -- \\ \vdots & \vdots & \\ -- & x_n' & -- \end{bmatrix} w - \begin{bmatrix} t_1' \\ t_2' \\ \vdots \\ t_n' \end{bmatrix} \|^2$$

$$= \frac{1}{2}\|X'w - t'\|^2 = \frac{1}{2}(w^T X'^T X' w - 2w^T X'^T t' + t'^T t')$$

$$\nabla E_d(w) = X'^T X' w - X'^T t' \quad \text{(by definition)}$$

找 w* 使 $E_d(w)$最小：

$$\nabla E_d(w^*) = X'^T X' w^* - X'^T t' = 0$$

$$where \quad X' = \begin{bmatrix} -- & \sqrt{r_1} x_1 & -- \\ -- & \sqrt{r_2} x_2 & -- \\ \vdots & \vdots & \vdots \\ -- & \sqrt{r_n} x_n & -- \end{bmatrix}, \quad t' = \begin{bmatrix} \sqrt{r_1} t_1 \\ \sqrt{r_2} t_2 \\ \vdots \\ \sqrt{r_n} t_n \end{bmatrix}$$

$$Let \quad R = \begin{bmatrix} r_1 & 0 & 0 & \dots \\ 0 & r_2 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & r_n \end{bmatrix}$$

$$\nabla E_d(w^*) = X^T R X w^* - X^T R t = 0$$

$$\underline{\mathbf{w^* = (X^T R X)^{-1} X^T R t}}$$

(4-b)

$$w^* = (X^T R X)^{-1} X^T R t = (\begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 5 & 1 \\ 5 & 6 \end{bmatrix})^{-1} \begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 10 \\ 5 \end{bmatrix}$$

$$\underline{\mathbf{w}^* = \begin{bmatrix} 2.28275254 \\ -1.13586237 \end{bmatrix}}$$

5. (1%)
Collaborator: B05902109 柯上優

With weight decay:

$$E(w) = \frac{1}{2} \sum_{n=1}^{N} (t_n - w^T x_n)^2 + \frac{\lambda}{2} \|w\|_2 = \frac{1}{2} (w^T X^T X w - 2 w^T X^T t + t^T t) + \frac{\lambda}{2} w^T w$$

With Gaussian noise:

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

$$= \frac{1}{2} \sum_{n=1}^{N} ((y(x_n, w) - t_n)^2 + 2 * (y(x_n, w) - t_n) * \sum_{i=1}^{D} w_i \epsilon_{ni} + (\sum_{i=1}^{D} w_i \epsilon_{ni})^2)$$

取期望值，

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

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

上式的第一項相當於 對沒有 noise 的 x 取 E 的最小值，

第二項 $\mathbb{E}[2 * (y(x_n, w) - t_n) * \sum_{i=1}^{D} w_i \epsilon_{ni}]$

第三項 $\mathbb{E}[(\sum_{i=1}^{D} w_i \epsilon_{ni})^2] = \mathbb{E}[\sum_{i=1}^{D} \sum_{j=1}^{D} w_i w_j \epsilon_{ni} \epsilon_{nj}] = \sum_{i=1}^{D} w_i^2 \sigma^2$ 相當於 weight decay。

6. (1%)
Collaborator: B05902109 柯上優

若 λ_i 為 A 的第 i 個 eigenvalue，

$$|A| = \prod_{i=1}^{N} \lambda_i \quad and \quad Tr(A) = \sum_{i=1}^{N} \lambda_i$$

因此，

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

又因為 tr() 是 linear operator， $d(Tr(X)) = Tr(dX)$

故 $\frac{d}{d\alpha} ln|A| = \frac{d}{d\alpha} Tr(ln(A)) = Tr(\frac{d}{d\alpha} ln(A))$ ，

再透過連鎖律， $\frac{d}{d\alpha} ln|A| = Tr(\frac{d}{d\alpha} ln(A)) = Tr(A^{-1} \frac{d}{d\alpha} A)$