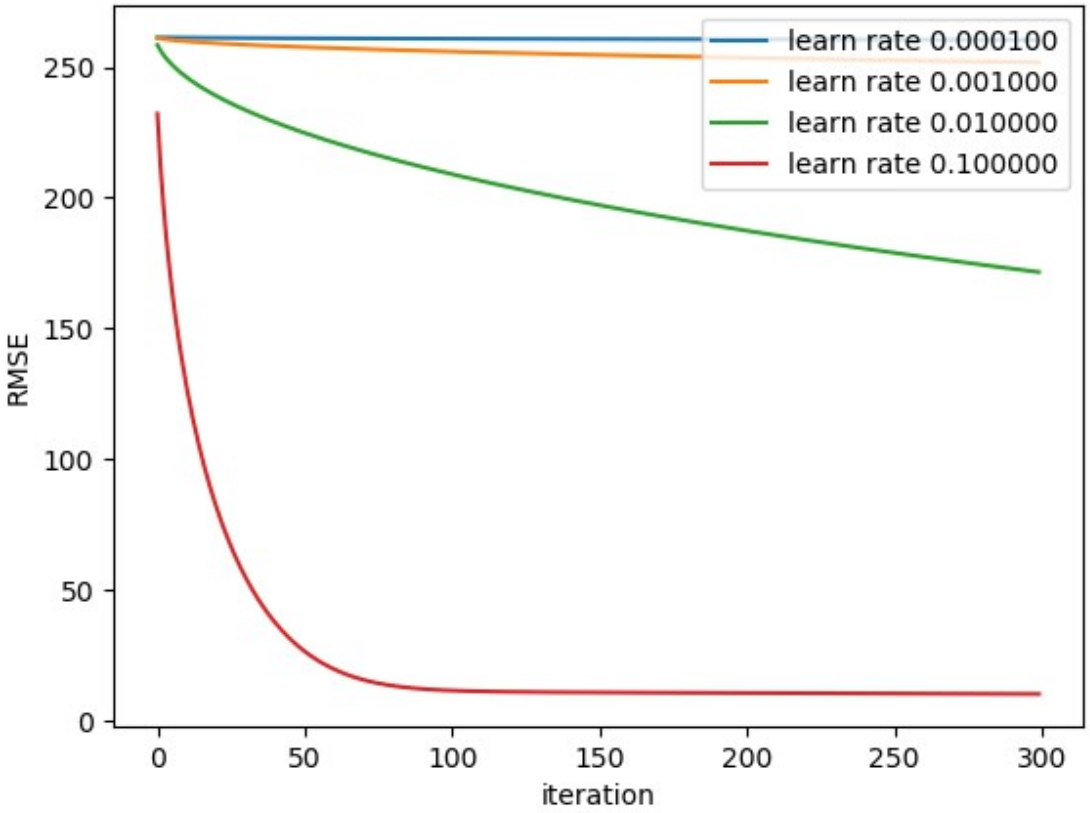
Homework 1 Report - PM2.5 Prediction

學號：B05902109 系級： 資工三 姓名: 柯上優

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。



以上只使用特徵pm2.5進行訓練，無進行任何資料整理。使用AdaGradient。可以發現，Learn rate越大，收斂速度越快，紅線最快抵達區域極值所以RMSE不再變動，綠色下降速度其次，橘線在後。反觀Learn rate最小的藍線學習太慢，一直收斂不了。

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	testing public	testing private
All	22.6044837272	9.02901	8.70808
pm2.5	5.5338208085	8.96531	8.70587

以上只將資料接起來，無進行特殊篩選，剔除無法以數學計算的RAINFALL值後，其他參數都一起丟入AdaGradient。觀察發現，「pm2.5」的Training Loss偏低，private testing loss卻比較高，可見pm2.5嚴重overfit。再來，「所有特徵」的Training Loss非常高，而「pm2.5」的Training Loss極低，兩者的private testing loss卻極為接近，可以發現，直接訓練的狀況下，「所有特徵」並沒有因此讓模型更加強大，或許需要資料的整理與篩選。此外，Training Loss遠大於testing表示，training data一定有很大的noise，若不進行資料清理，在優秀的模型也是枉然。

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。

λ	Training	testing public	testing private
1e0	5.533821	8.96531	8.70587
1e3	5.541666	8.94814	8.72264
1e6	8.376644	9.65600	10.09798
1e9	32.849194	36.57342	36.01847

和前一題一樣無進行資料處理，純粹資料接上後進行訓練。觀察發現，隨著λ的加大，理論上是為了限制住模型的複雜度，但是這個模型本身並不複雜(以特徵的一次項進行線性組合)，限制住$\mathbf{w}$的長度並沒有增加模型的強度，反而造成模型的能力下降。或許在更複雜的模型時，此方法能夠用來防止overfit，但在此狀況下，並沒有特殊成效。

4. (1%)

(4-a)

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

where d is the dimension of x_i.
R is a $N\mathbf{x}N$ diagonal matrix,

T is a Nx1 matrix,
X is a Nxd matrix,
and W is a dx1 matrix.

$$\frac{\mathrm{d}}{\mathrm{d}w}E_D(\mathbf{w}) = 2X^TRXW - 2X^TRT = O_{d\mathrm{x}1}$$

所求W為

$$W = (X^TRX)^{-1}X^TRT$$

(4-b)

方便套入上式，將題目的矩陣改為

$$T = \begin{bmatrix} 0 \\ 10 \\ 5 \end{bmatrix} X = \begin{bmatrix} 2 & 3 \\ 5 & 1 \\ 5 & 6 \end{bmatrix} R = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

由python計算得到

$$W = \begin{bmatrix} 2.28275254 \\ -1.13586237 \end{bmatrix}$$

5. (1%)

將雜訊加進 x，linear model變成

$$y((x_n + \epsilon_i), \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i(x_i + \epsilon_i) = \sum_{i=0}^{D} w_i(x_i + \epsilon_i)$$

where $x_0 = 1$. 而 $\epsilon_0 = 0$. 新的 sum-of-squares error function:

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

現在，我們取期望值，又根據期望值的線性特性

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

分別討論三項。第一項已經變成了沒有雜訊的error function。第二項由於$E[\epsilon_i]=0$所以消去。至於第三項

$$\mathbb{E}[(\sum_{d=1}^{D} w_d\epsilon_{nd})^2]$$
$$=\mathbb{E}[\sum_{i=1}^{D}\sum_{j=1}^{D} w_i w_j \epsilon_i\epsilon_j]$$
$$=\sum_{i=1}^{D}\sum_{j=1}^{D} w_i w_j \mathbb{E}[\epsilon_i\epsilon_j]$$
$$=\sum_{d=1}^{D} w_d w_d \sigma^2 = w^2\sigma^2$$

代入原式

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

即為所求，在有雜訊分布的E等同於沒有雜訊的E配上一個weight-decay regularization term，其中bias(parameter w_0)並不在regularization term。

6. (1%)

Collaborator: b05902074 魏佑珊

首先我們先證明

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

where A is a matrix.

proof>

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

現在我們設一個新的方陣 $B = lnA$。我們有其特性:

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

接著兩邊取 ln

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

最後我們微分他，並使用微分的連鎖率

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