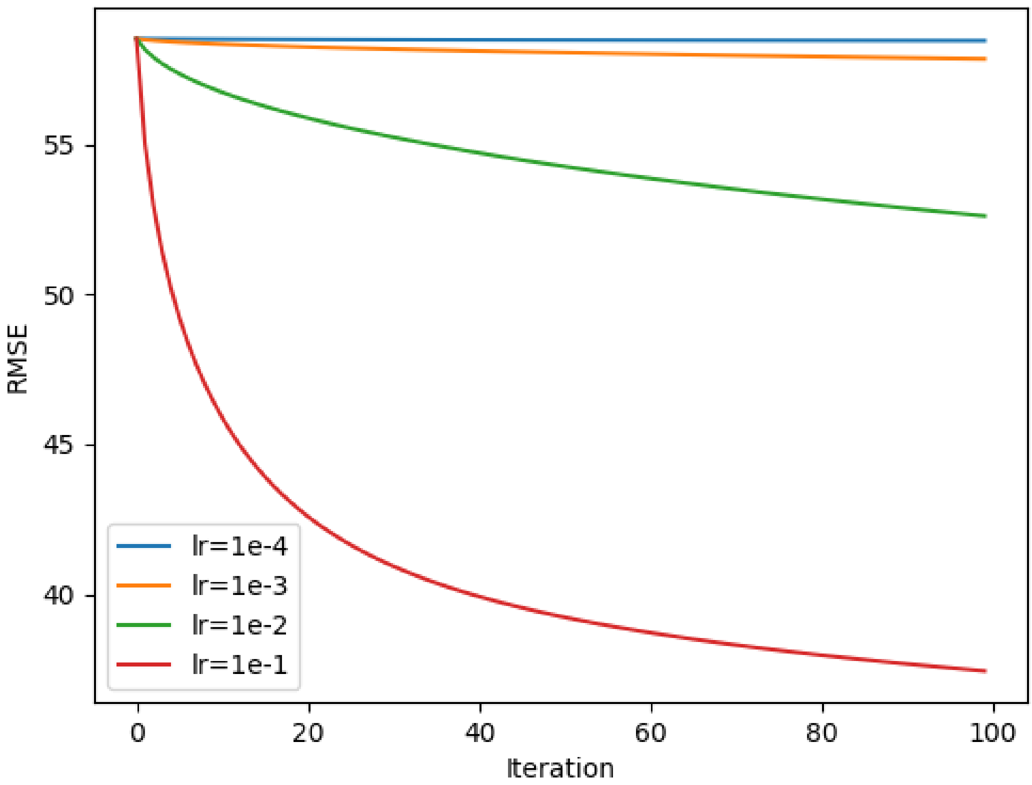
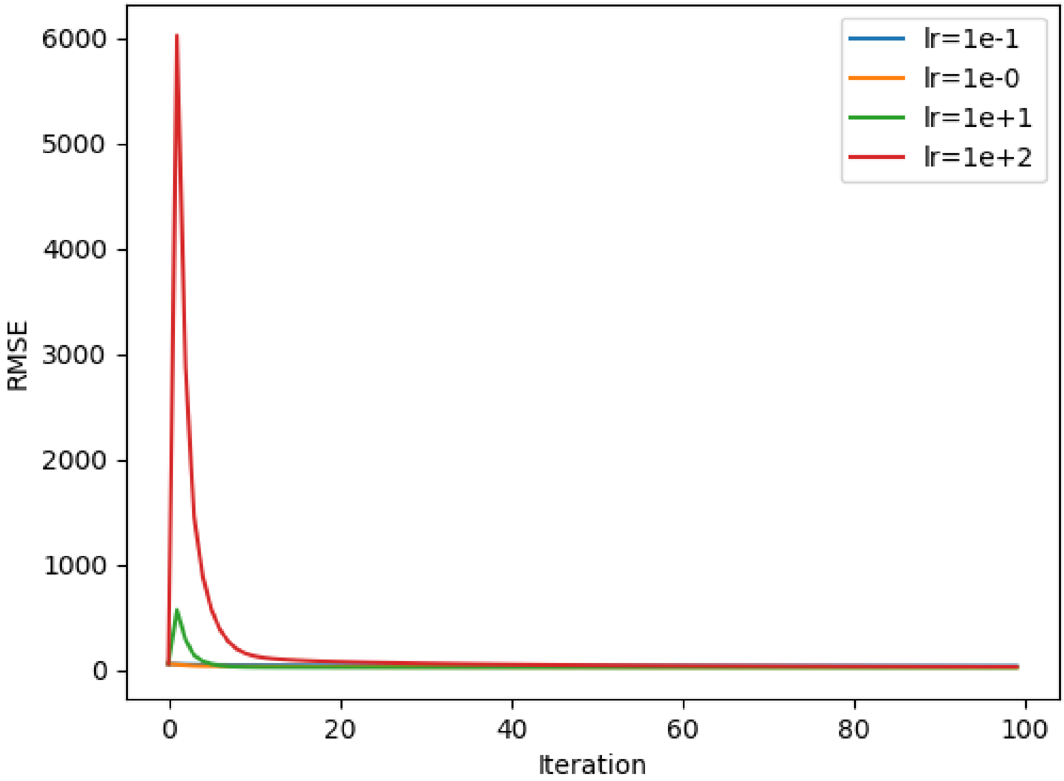
Homework 1 Report - PM2.5 Prediction

學號：R06922132　系級：資工所碩二　姓名:何羿辰

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。





本題採用七種 learning rate 0.0001/0.001/0.01/0.1/1/10/100，從圖中可以看出 learning rate 為 0.1 時收斂的最快，如果 learning rate 再繼續提高到 1.0 或 10.0 時 RMSE 會呈現一個暴增的山峰然後才慢慢收斂，原因是因為本題採用 gradient descent，雖然過程中會修正每一步的位移量，但太極端的 learning rate 則需要經過多個回合的修正才能開始收斂。

不過本圖也可看出在 gradient descent 的方法裡，learning rate 過高會比過低的收斂速度來的快。

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）。

Score	Private	Public
All Feature	8.73094	8.35278
PM2.5	9.71723	9.61511

根據結果可以發現 feature 全取比只取 PM2.5 不管在 private 還是 public 的結果都還要好，因此可以判斷出除了 PM2.5 以外至少還有一個 feature 會影響未來 PM2.5 的數值。

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。

λ	Train RMSE	L2 norm	Private Score	Public Score
10.0	25.008061	353.875457	9.42694	8.79099
0.1	23.115629	692.218842	8.66211	8.29411
0.01	23.113938	702.882058	8.66601	8.30637
0.00001	23.113909	704.103165	8.66655	8.30788

根據結果可以發現 λ 越大則 L2 norm 越小，原因是因為 minimize error 的過程中若λ越大則 weight 必須越小才能最小化，但 training RMSE 則會上升，原因是因為阻礙了 weight 的更新(變化較平滑)，而比較 Private 與 Public 的結果可以發現λ從大至小的 score 表現先降後升，原因與老師所講的結論一樣，function 不夠平滑與太平滑都不好。

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:

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

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

$\mathrm{w}^* = \arg\min_w E_D(w)$ ➔ compute $\frac{dE_D}{dw}|_{w=w^0}$ ➔find $\frac{\partial E_D}{\partial w} = 0$

對其微分展開得$\frac{\partial E_D}{\partial w} = \sum_{n=1}^{N} r_n(t_n - w^T x_n)(-x_n) = 0$ **，故只需找到w* 使其滿足左式即可最小化 error function**

$$\frac{\partial E_D}{\partial w_j} = \sum_{n=1}^{N} \frac{\partial}{\partial w_j}\left(\frac{1}{2} r_n (t_n - w^T x_n)^2\right) = \sum_{n=1}^{N} r_n x_{n,j}(t_n - w^T x_n) = 0$$

$\Rightarrow \sum_{n=1}^{N} r_n x_{n,j}\, t_n = \left(\sum_{n=1}^{N} r_n x_{n,j}\, x_n\right) w$

令 A=$\left(\sum_{n=1}^{N} r_n x_{n,j}\, x_n\right)$，b=$\sum_{n=1}^{N} r_n x_{n,j}\, t_n$，則 Aw=b 解出 w=A^{-1}b 即可，若 A 不可逆則採用$\sum_{n=1}^{N} r_n (t_n - w^T x_n)(-x_n) = 0$ 解聯立方程式一樣可求出w*，底下 4-b 即是採用解聯立方法求出

(4-b)

Following the previous problem(2-a), if

$$\mathbf{t} = [t_1 t_2 t_3] = [0 \quad 10 \quad 5], \mathbf{X} = [\mathbf{x_1 x_2 x_3}] = \begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix}$$

$$r_1 = 2, r_2 = 1, r_3 = 3$$

Find the solution $\mathbf{w}^*$.

令w* = $[w_1\ w_2]^T$，代入 4-a 所得公式可得兩個聯立方程式，展開解聯立方程式得

$\mathbf{w}^* = \begin{bmatrix} 2.282752536391707 \\ -1.135862373180415 \end{bmatrix}$

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

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

將x_n加上 noise: $\widetilde{x_n}=x_n+\epsilon_n$，則 minimizing E averaged over the noise distribution 為

$$\min\{\frac{1}{2}\sum_{n=1}^{N}(w_0+\sum_{i=1}^{D}w_i(x_i+\epsilon_i)-t_n)^2\}=\min\{\frac{1}{2}\sum_{n=1}^{N}(w_0+\sum_{i=1}^{D}w_ix_i+\sum_{i=1}^{D}w_i\epsilon_i-t_n)^2\}$$

令$A_n=w_0+\sum_{i=1}^{D}w_ix_i-t_n$，並將常數項$\frac{1}{2}$省略不影響 min，繼續展開

原式$=\min\left\{\sum_{n=1}^{N}\left(A_n+\sum_{i=1}^{D}w_i\epsilon_i\right)^2\right\}=\min\{\sum_{n=1}^{N}({A_n}^2+2*A_n*\sum_{i=1}^{D}w_i\epsilon_i+(\sum_{i=1}^{D}w_i\epsilon_i)^2)\}$

其中期望值$\mathbb{E}[{A_n}^2]={A_n}^2$，$\mathbb{E}[2*A_n*\sum_{i=1}^{D}w_i\epsilon_i]=2A_n\sum_{i=1}^{D}w_i\mathbb{E}[\epsilon_i]=2A_n\sum_{i=1}^{D}w_i*0=0$，

$\mathbb{E}\left[\left(\sum_{i=1}^{D}w_i\epsilon_i\right)^2\right]=\mathbb{E}[({w_1}^2{\epsilon_1}^2+{w_2}^2{\epsilon_2}^2+\cdots+{w_D}^2{\epsilon_D}^2)]+2\mathbb{E}[(w_1\epsilon_1w_2\epsilon_2+w_1\epsilon_1w_3\epsilon_3+\cdots+w_{D-1}\epsilon_{D-1}w_D\epsilon_D)]=({w_1}^2\delta_{11}\sigma^2+{w_2}^2\delta_{22}\sigma^2+\cdots+{w_D}^2\delta_{DD}\sigma^2)+2(w_1w_2\delta_{12}\sigma^2+w_1w_3\delta_{13}\sigma^2+\cdots+w_{D-1}w_D\delta_{D-1D}\sigma^2)=\sum_{i=1}^{D}{w_i}^2\sigma^2+0=\sigma^2\sum_{i=1}^{D}{w_i}^2$ 帶回原式

原式$=\min\{\sum_{n=1}^{N}({A_n}^2+\sigma^2\sum_{i=1}^{D}{w_i}^2\}=\min\{\sum_{n=1}^{N}((w_0+\sum_{i=1}^{D}w_ix_i-t_n)^2+\sigma^2\sum_{i=1}^{D}{w_i}^2\}$

前面項$(w_0+\sum_{i=1}^{D}w_ix_i-t_n)^2$是 minimizing the sum-of-squares error for noise-free input variables，而後面項 $\sigma^2\sum_{i=1}^{D}{w_i}^2$是 addition of a weight–decay regulation term，故兩者結果相等得證#

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.

Derivation of Jacobi's formula by Laplace:

$d|A|=\sum_j A_{ij}adj^T(A)_{ij}$ 又 $|A|=F(A_{11},A_{12},\dots,A_{21},A_{22},\dots,A_{nn})$

➔ $d|A|=\sum_i\sum_j\frac{\partial F}{\partial A_{ij}}dA_{ij}$➔ $\frac{\partial|A|}{\partial A_{ij}}=\frac{\partial\sum_k A_{ik}adj^T(A)_{ik}}{\partial A_{ij}}=\sum_k\frac{\partial A_{ik}adj^T(A)_{ik}}{\partial A_{ij}}=\sum_k\frac{\partial A_{ik}}{\partial A_{ij}}adj^T(A)_{ik}+\sum_k\frac{\partial adj^T(A)_{ik}}{\partial A_{ij}}A_{ik}=\sum_k\frac{\partial A_{ik}}{\partial A_{ij}}adj^T(A)_{ik}+0=\sum_k\frac{\partial A_{ik}}{\partial A_{ij}}adj^T(A)_{ik}=adj^T(A)_{ij}$➔ $\mathbf{d|A|}=\boldsymbol{Tr(adj(A)dA)}$

因為 A 可逆➔ $\frac{\mathbf{d|A|}}{\boldsymbol{d\alpha}}=\boldsymbol{|A|Tr(A^{-1}\frac{dA}{d\alpha})}$

$$\frac{d}{d\alpha}\ln|A| = Tr\left(adj(\ln A)\frac{d\ln A}{d\alpha}\right) = Tr\left(\ln|A| * \frac{1}{\ln A} * \frac{d\ln A}{d\alpha}\right) = \ln|A|\,\mathrm{Tr}\left(\ln A^{-1}\frac{d\ln A}{d\alpha}\right)$$

$$= \ln|A|\,\mathrm{Tr}\left(\ln A^{-1}\frac{d\ln A}{dA} * \frac{dA}{d\alpha}\right) = \ln|A|\,\mathrm{Tr}\left(\frac{1}{\ln A} * A^{-1} * \frac{dA}{d\alpha}\right)$$

設 A 的 eigenvalues 為$\lambda_1, \lambda_2, \lambda_3 \dots, \lambda_n$，則$|\mathrm{A}| = \lambda_1\lambda_2\lambda_3 \dots \lambda_n$，$\mathrm{Tr(A)} = \lambda_1 + \lambda_2 + \lambda_3 + \cdots + \lambda_n$繼續展開

$$\frac{d}{d\alpha}\ln|A| = \ln(\lambda_1\lambda_2\lambda_3 \dots \lambda_n) * \frac{1}{\ln(\lambda_1) + \ln(\lambda_2) + \ln(\lambda_3) + \cdots + \ln(\lambda_n)} * \mathrm{Tr}\left(A^{-1} * \frac{dA}{d\alpha}\right)$$

又$\ln(\lambda_1\lambda_2\lambda_3 \dots \lambda_n) = \ln(\lambda_1) + \ln(\lambda_2) + \ln(\lambda_3) + \cdots + \ln(\lambda_n)$

故$\frac{d}{d\alpha}\ln|A| = \frac{1}{1}\mathrm{Tr}\left(A^{-1} * \frac{d}{d\alpha}A\right) = \mathrm{Tr}\left(A^{-1} * \frac{d}{d\alpha}A\right)$ 得證#