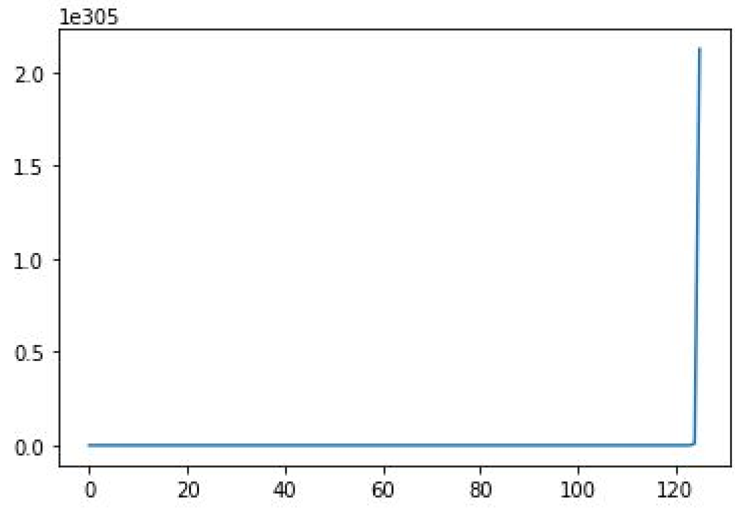
Homework 1 Report - PM2.5 Prediction

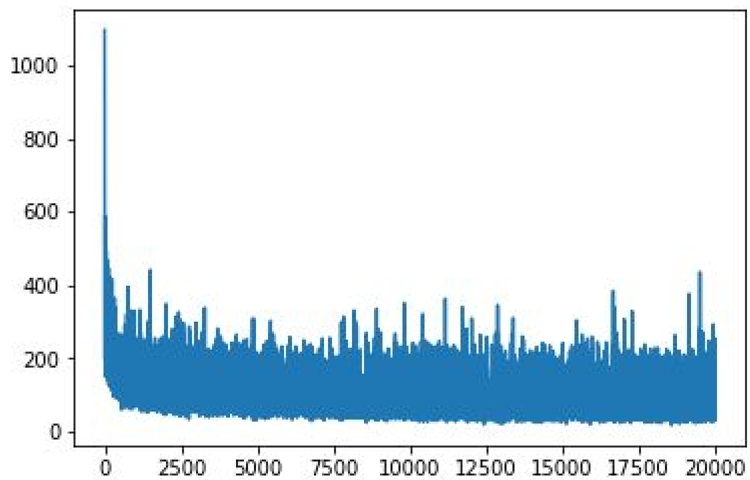
學號：r07942086　系級：電信碩二　姓名: 顏宏宇

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。

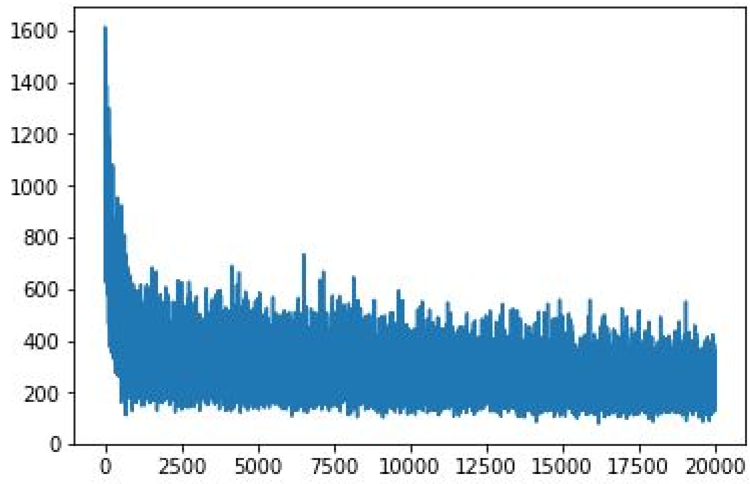
learning rate 10 ^ -5



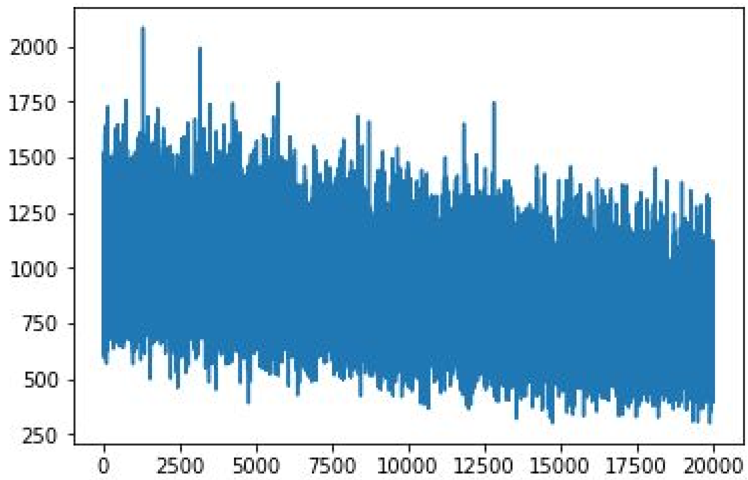
learning rate 10 ^ -7



learning rate 10 ^ -9



learning rate 10 ^ -11



可以看出當learning rate太大(10^-5)的時候會導致loss無法收斂，反而會爆炸。
而當learning rate太小(10^-11)的時候會讓模型很難收斂，loss降不下去。
對於我的模型而言10^-6 是最合適的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）。

output.csv 4 days ago by r07942086_hyyen only 9 pm2.5 model	9.86511	9.68834
ans_[77.035192].csv 5 days ago by r07942086_hyyen whole data model	7.78278	8.19495

可以明顯看出只使用PM2.5之9小時資料訓練出來的模型，其RMSE比使用所有feature的模型大了許多，代表其他資料也是有助於訓練更精確的模型。

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。

ans_[486.5153657].csv 4 days ago by r07942086_hyyen λ = 1000000, L2 norm = 0.07685345124616812	15.48262	15.00746
ans_[228.84403582].csv 4 days ago by r07942086_hyyen λ = 10000, L2 norm = 0.38489263070550417	9.01399	10.08392
ans_[296.42139113].csv 4 days ago by r07942086_hyyen λ = 100, L2 norm = 0.6432965947997006	9.69615	10.53389
ans_[157.56833104].csv 4 days ago by r07942086_hyyen λ = 0 , L2 norm = 0.658115519967589	8.42384	9.08432

以我的模型來說，沒有regulization(λ=0)的時候狀況最好，基本上越大的λ導致越大的loss。
這可能是因為我在調learning rate的時候就已經找到最佳解了，而regulization反而會讓weight無法收斂到最佳解。
而L2 norm的確是當 λ 越大的時候L2 norm越小。

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）。

4-6題 collaborator : d07946003 王嘉澤

4. (a)

4. (a)

$$w^* = \arg\min \sum_{n=1}^{N} r_n (t_n - w^T x_n)^2$$

$$SSE = \frac{1}{2}\left(W^T Y R X^T W - 2W^T X R Y^T + Y R Y^T\right)$$

$$= \frac{1}{2}[W^T X - Y] R [W^T X - Y]^T$$

$$= \frac{1}{2}[W^T X R - Y R][W^T X - Y]^T$$

$$= \frac{1}{2}[W^T X R W^T X - Y R X^T W - W^T X R Y^T + Y R Y^T]$$

$$= \frac{1}{2}[W^T X R W^T X - 2W^T X R Y^T + Y R Y^T]$$

$$\frac{d}{dW} \cdot \frac{1}{2}[W^T X R W^T X - 2W^T X R Y^T + Y R Y^T]$$

$$= (X R X^T) W - X R Y^T$$

$$\Rightarrow X R X^T W - X R Y^T = 0 \qquad X R X^T W = X R Y^T$$

$$W = (X R X^T)^{-1} X R Y^T \quad \#$$

4. (b)

(b)

$$\left(\begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix}\begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix}\begin{bmatrix} 2 & 3 \\ 5 & 1 \\ 5 & 6 \end{bmatrix}\right)^{-1}\begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix}\begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix}\begin{bmatrix} 0 \\ 10 \\ 5 \end{bmatrix}$$

$$= \left(\begin{bmatrix} 4 & 5 & 15 \\ 6 & 1 & 18 \end{bmatrix}\begin{bmatrix} 2 & 3 \\ 5 & 1 \\ 5 & 6 \end{bmatrix}\right)^{-1}\begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix}\begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix}\begin{bmatrix} 0 \\ 10 \\ 5 \end{bmatrix}$$

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

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

$$= \frac{1}{2267}\begin{bmatrix} 127 & -107 \\ -107 & 108 \end{bmatrix}\begin{bmatrix} 125 \\ 100 \end{bmatrix} = \frac{1}{2267}\begin{bmatrix} 5175 \\ -2575 \end{bmatrix}$$

5.

Let $\bar{t} = [t_1, \dots, t_N]^T \in R^{N\times 1}$ $\quad \bar{x}_n = [1, x_1^n, \dots, x_D^n]^T \in R^{(D+1)\times 1}$

$\bar{w} = [w_0, w_1, \dots w_D]^T \in R^{(D+1)\times 1}$, $\quad \bar{\varepsilon}_n = [0, \varepsilon_1^n, \dots, \varepsilon_D^n]^T \in R^{(D+1)\times 1}$ $\quad n \in [1, N]$

$\tilde{X} = [\bar{x}_1, \dots, \bar{x}_n] \in R^{(D+1)\times N}$ $\quad \hat{E} = [\bar{\varepsilon}_1, \dots, \bar{\varepsilon}_N] \in R^{(D+1)\times N}$

$y(\bar{x}_n, \bar{w}) = w_0 + \sum_{i=1}^{D} w_i x_i = \bar{w}^T \bar{x}_n$ $\quad y(\bar{x}_n + \bar{\varepsilon}_n, \bar{w}) = \bar{w}^T(\bar{x}_n + \bar{\varepsilon}_n)$

$$E(\bar{w}) = \frac{1}{2}\sum_{n=1}^{N}[y(\bar{x}_n + \bar{\varepsilon}_n, \bar{w}) - t_n]^2 = \frac{1}{2}[(\tilde{X}+\tilde{E})^T\bar{w} - \bar{t}]^T[(\tilde{X}+\tilde{E})^T\bar{w} - \bar{t}]$$

$$= \frac{1}{2}[\bar{w}^T(\tilde{X}+\tilde{E}) - \bar{t}^T][(\tilde{X}+\tilde{E})^T\bar{w} - \bar{t}]$$

$$= \frac{1}{2}[\bar{w}^T(\tilde{X}+\tilde{E})(\tilde{X}+\tilde{E})^T\bar{w} - \bar{w}^T(\tilde{X}+\hat{E})\bar{t} - \bar{t}^T(\tilde{X}+\tilde{E})^T\bar{w} - \bar{t}^T\bar{t}]$$

$$= \frac{1}{2}[\bar{w}^T(\tilde{X}+\tilde{E})(\tilde{X}+\tilde{E})^T\bar{w} - 2\bar{w}^T(\tilde{X}+\tilde{E})\bar{t} - \bar{t}^T\bar{t}]$$

$$= \frac{1}{2}[\bar{w}^T(\tilde{X}+\tilde{E})(\tilde{X}+\tilde{E})^T\bar{w} - 2\bar{w}^T(\tilde{X}+\tilde{E})\bar{t} - \bar{t}^T\bar{t}]$$

$$= \frac{1}{2}[\bar{w}^T\tilde{X}\tilde{X}^T\bar{w} + 2\bar{w}^T\tilde{X}\tilde{E}^T\bar{w} + \bar{w}^T\tilde{E}\tilde{E}^T\bar{w} - 2\bar{w}^T\tilde{X}\bar{t} - 2\bar{w}^T\tilde{E}\bar{t} - \bar{t}^T\bar{t}]$$

$$\mathbb{E}[E(\bar{w})] = \frac{1}{2}[\mathbb{E}(\bar{w}^T\tilde{X}\tilde{X}^T\bar{w}) + 2\mathbb{E}(\bar{w}^T\tilde{X}\tilde{E}^T\bar{w}) + \mathbb{E}(\bar{w}^T\tilde{E}\tilde{E}^T\bar{w}) - 2\mathbb{E}(\bar{w}^T\tilde{X}\bar{t})$$
$$- 2\mathbb{E}(\bar{w}^T\tilde{E}\bar{t}) - \mathbb{E}(\bar{t}^T\bar{t})]$$

$$= \frac{1}{2}[\bar{w}^T\tilde{X}\tilde{X}^T\bar{w} + 2\bar{w}^T\tilde{X}\,\mathbb{E}(\tilde{E}^T)\bar{w} + \bar{w}^T\mathbb{E}(\tilde{E}\tilde{E}^T)\bar{w} - 2\bar{w}^T\tilde{X}\bar{t}$$
$$- 2\bar{w}^T\mathbb{E}(\tilde{E})\bar{t} - \bar{t}^T\bar{t}]$$

$\mathbb{E}(\hat{E}) = \tilde{O}_{(D+1),N}$. $\quad \mathbb{E}(\tilde{E}^T) = \tilde{O}_{N,(D+1)}$ $\quad$ where $\tilde{O}_{ij}$ is a zero matrix $\in R$

$$\mathbb{E}(\tilde{E}\tilde{E}^T) = \begin{bmatrix} 0 & & & \\ & \sigma^2 & & \\ & & \sigma^2 & \\ & & & \ddots \; \sigma^2 \end{bmatrix} \in \mathbb{R}^{(D+1)\times(D+1)} \Rightarrow W^T\mathbb{E}(\tilde{E}\tilde{E}^T)\bar{w} = \sigma^2\bar{w}^T\bar{w} - \sigma$$

D

$$\mathbb{E}[E(w)] = \frac{1}{2}(\underbrace{\bar{w}^T\tilde{X}\tilde{X}^T\bar{w} - 2\bar{w}^T\tilde{X}\bar{t} - \bar{t}^T\bar{t}}_{\text{SSE for no-noise input}} + \underbrace{\sigma^2\bar{w}^T\bar{w} - \sigma^2 w_0^2}_{\text{weight-decay regulariza}})$$

6.

6. $A \in R^{n \times n}$ α B one of the elements of A

prove. $\frac{d}{d\alpha} \ln|A| = Tr(A^{-1} \frac{d}{d\alpha} A)$.

LHS:

$$|A| = |QDQ^{-1}| = |Q||D||Q^{-1}| = |D| = \prod_i \lambda_i$$

$$\Rightarrow \ln|A| = \sum_i \ln \lambda_i \Rightarrow \frac{d}{d\alpha} \ln|A| = \frac{d}{d\alpha} \sum_i \ln \lambda_i = \sum_i \frac{1}{\lambda} \frac{d\lambda_i}{d\alpha}$$

RHS

$$\Rightarrow A^{-1} = QD^{-1}Q^{-1} \quad \frac{d}{d\alpha} A = \frac{d}{d\alpha}(QDQ^{-1}) = \frac{dQ}{d\alpha} DQ^{-1} + Q\frac{dD}{d\alpha}Q^{-1} + QD\frac{dQ^{-1}}{d\alpha}$$

Thus. $Tr(A^{-1}\frac{d}{d\alpha}A) = Tr(QD^{-1}Q^{-1}\frac{dQ}{d\alpha}DQ^{-1}) + Tr(QD^{-1}Q^{-1}Q\frac{dD}{d\alpha}Q^{-1}) + Tr($

$$\Rightarrow ① \; Tr(Q^{-1}D^{-1}Q\frac{dQ}{d\alpha}DQ^{-1}) = Tr(DQ^{-1}QD^{-1}Q^{-1}\frac{dQ}{d\alpha}) = Tr(Q^{-1}\frac{dQ}{d\alpha})$$

$$② \; Tr(QD^{-1}Q^{-1}Q\frac{dD}{d\alpha}Q^{-1}) = Tr(Q^{-1}QD^{-1}Q^{-1}Q\frac{dD}{d\alpha}) = Tr(D^{-1}\frac{dD}{d\alpha})$$

$$③ \; Tr(QD^{-1}Q^{-1}QD\frac{dQ^{-1}}{d\alpha}) = Tr(Q\frac{dQ^{-1}}{d\alpha})$$

$$①+③ = Tr(Q^{-1}\frac{dQ}{d\alpha}) + Tr(Q\frac{dQ^{-1}}{d\alpha}) = Tr(\frac{d}{d\alpha}(QQ^{-1})) = 0$$

$$\Rightarrow RHS = ①+②+③ = ② = Tr(D^{-1}\frac{dD}{d\alpha}) = \sum_i \frac{1}{\lambda}\frac{d\lambda_i}{d\alpha} = LHS.$$

Q.E.D.