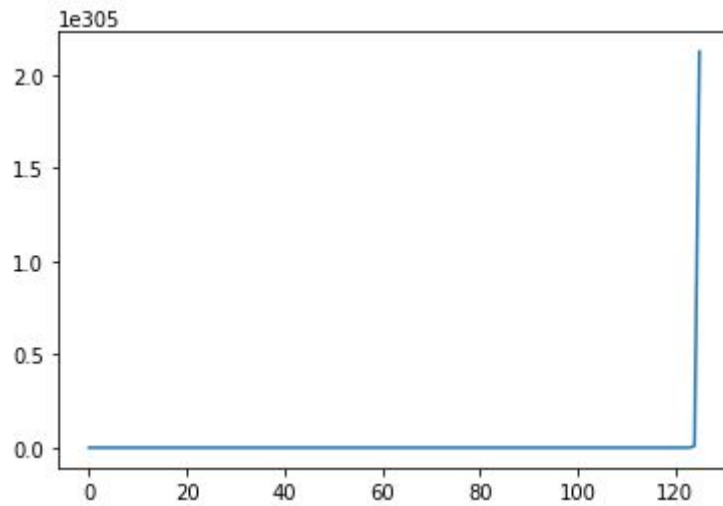


Homework 1 Report - PM2.5 Prediction

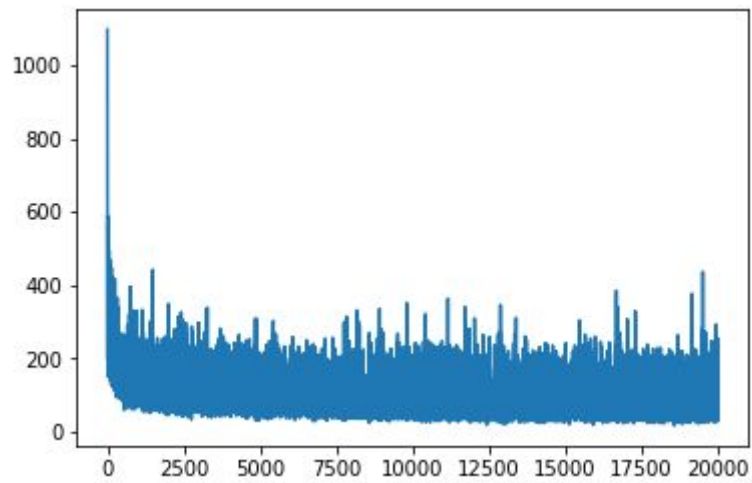
學號：r07942086 系級：電信碩二 姓名：顏宏宇

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。

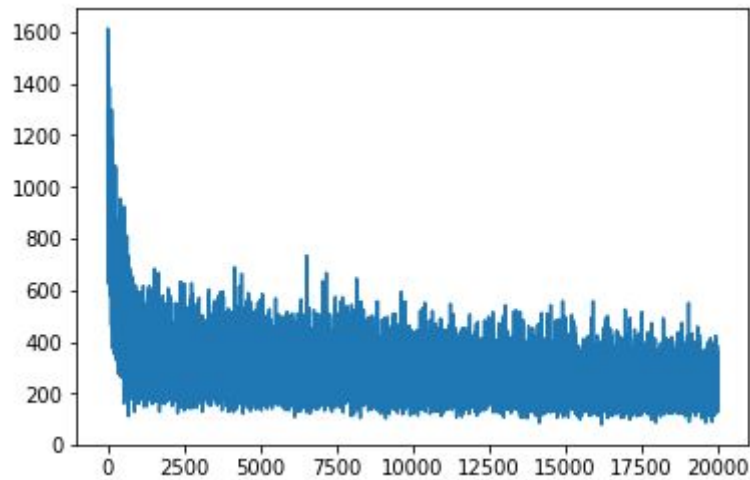
learning rate 10^{-5}



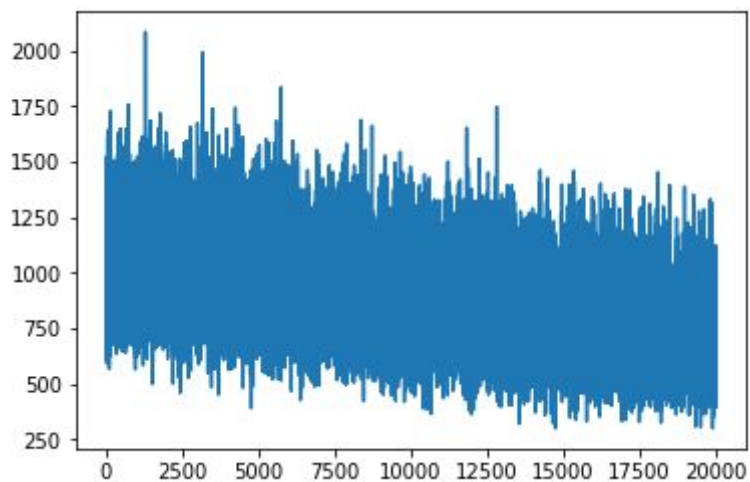
learning rate 10^{-7}



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_hyeyn

only 9 pm2.5 model

9.86511

9.68834

[ans_\[77.035192\].csv](#)

5 days ago by r07942086_hyeyn

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

以我的模型來說，沒有regularization($\lambda=0$)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regularization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. (a)

$$\begin{aligned}
 4. (a) \quad W^* &= \arg \min \sum_{n=1}^N r_n(t_n - W^T x_n)^2 \\
 SSE &= \frac{1}{2} (W^T Y R X^T W - 2 W^T X R Y^T + Y R Y^T) \\
 &= \frac{1}{2} [W^T X - Y] R [W^T X - Y]^T \\
 &= \frac{1}{2} [W^T X R - Y R] [W^T X - Y]^T \\
 &= \frac{1}{2} [W^T X R W^T X - Y R X^T W - W^T X R Y^T + Y R Y^T] \\
 &= \frac{1}{2} [W^T X R W^T X - 2 W^T X R Y^T + Y R Y^T] \\
 \frac{d}{dW} \cdot \frac{1}{2} [W^T X R W^T X - 2 W^T X R Y^T + Y R Y^T] \\
 &= (X R X^T) W - X R Y^T \\
 \Rightarrow X R X^T W - X R Y^T &= 0 \quad X R X^T W = X R Y^T \\
 W &= (X R X^T)^{-1} X R Y^T \quad \#
 \end{aligned}$$

4. (b)

$$\begin{aligned}
 (b) \quad & \left(\begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 5 & 1 \\ 5 & 6 \end{bmatrix} \right)^{-1} \begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 10 \\ 5 \end{bmatrix} \\
 &= \left(\begin{bmatrix} 4 & 5 & 15 \\ 6 & 1 & 18 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 5 & 1 \\ 5 & 6 \end{bmatrix} \right)^{-1} \begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 10 \\ 5 \end{bmatrix} \\
 &= \begin{bmatrix} 108 & 107 \\ 107 & 127 \end{bmatrix}^{-1} \begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 10 \\ 5 \end{bmatrix} \\
 &= \frac{1}{2267} \begin{bmatrix} 127 & -107 \\ -107 & 108 \end{bmatrix} \begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 10 \\ 5 \end{bmatrix} \\
 &= \frac{1}{2267} \begin{bmatrix} 127 & -107 \\ -107 & 108 \end{bmatrix} \begin{bmatrix} 125 \\ 100 \end{bmatrix} = \frac{1}{2267} \begin{bmatrix} 5175 \\ -2575 \end{bmatrix}
 \end{aligned}$$

6.

6. $A \in \mathbb{R}^{n \times n}$ & λ one of the elements of A

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

LHS:

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

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

RHS

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

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

$$\Rightarrow \textcircled{1} \text{Tr}(Q^{-1}D^{-1}Q \frac{dQ}{d\lambda} DQ^{-1}) = \text{Tr}(DQ^{-1}QD^{-1}Q^{-1} \frac{dQ}{d\lambda}) = \text{Tr}(Q^{-1} \frac{dQ}{d\lambda})$$

$$\textcircled{2} \text{Tr}(QD^{-1}Q^{-1} Q \frac{dD}{d\lambda} Q^{-1}) = \text{Tr}(Q^{-1}QD^{-1}Q^{-1} Q \frac{dD}{d\lambda}) = \text{Tr}(D^{-1} \frac{dD}{d\lambda})$$

$$\textcircled{3} \text{Tr}(QD^{-1}Q^{-1} QD \frac{dQ^{-1}}{d\lambda}) = \text{Tr}(Q \frac{dQ^{-1}}{d\lambda})$$

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

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

Q.E.D.