Machine Learning - 2021Fall - HW1

學號：R10921098 系級：電機所碩一 姓名：楊仁傑

請實作以下兩種不同feature的模型，回答第 1 ～ 2 題：

(1) 抽全部9小時內的污染源feature當作一次項(加bias)

(2) 抽全部9小時內pm2.5的一次項當作feature(加bias)

備註：

a. NR請皆設為0，其他的非數值(特殊字元)可以自己判斷

b. 所有 advanced 的 gradient descent 技術(如: adam, adagrad 等) 都是可以用的

c. 第1～2題請都以題目給訂的兩種model來回答

d. 同學可以先把model訓練好，kaggle死線之後便可以無限上傳。

1.(1%)記錄誤差值 (RMSE)(根據kaggle public+private分數)，討論兩種feature的影響

ANS:

features	Private Score	Public Score
only PM2.5 (epochs=500)	13.890	13.102
All features (epochs=500)	10.372	10.818
only PM2.5 (epochs=1000)	8.1400	6.9559
All features (epochs=1000)	9.2539	9.8904

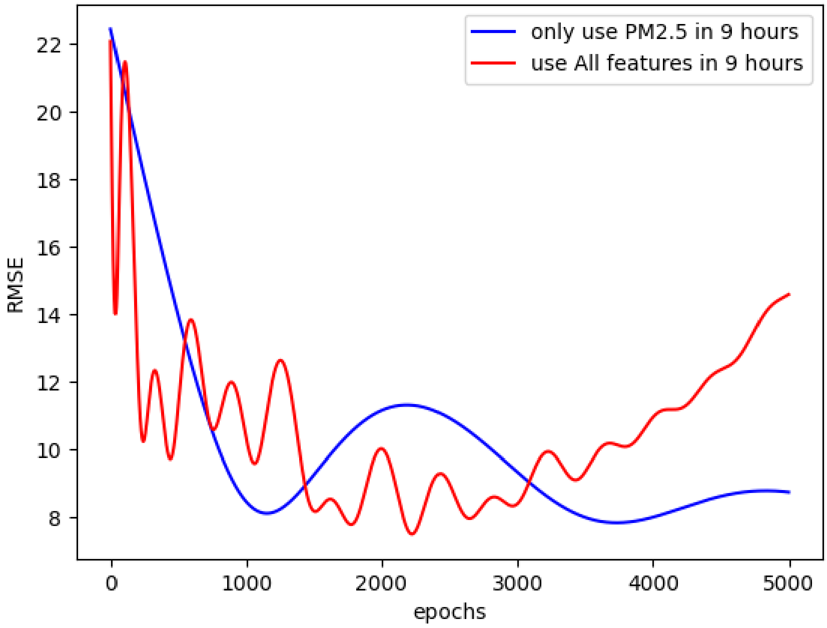

Figure1

[Discussion]

It is difficult to assess which features perform better when the data is presented in a table alone. So, I plotted the RMSE against the test data for different training rounds.

As can be seen from Figure 1, the data composition of only PM2.5 is used as a feature is simple, so the curve of the RMSE is quite smooth, but the data composition of the all features is relatively complex, so the curve of the RMSE is very oscillating compared to the only PM2.5 is used as a feature.

In addition, it can be seen from the graph that when only PM2.5 is used as a feature, the best solution can be converged by the 1000th round; however, when all features are used, the best solution can only be converged after 2,000 rounds for the same reason as above.

2.(1%)解釋什麼樣的data preprocessing可以improve你的training/testing accuracy，

e.g., 你怎麼挑掉你覺得不適合的data points。請提供數據(RMSE)以佐證你的想法。

ANS:

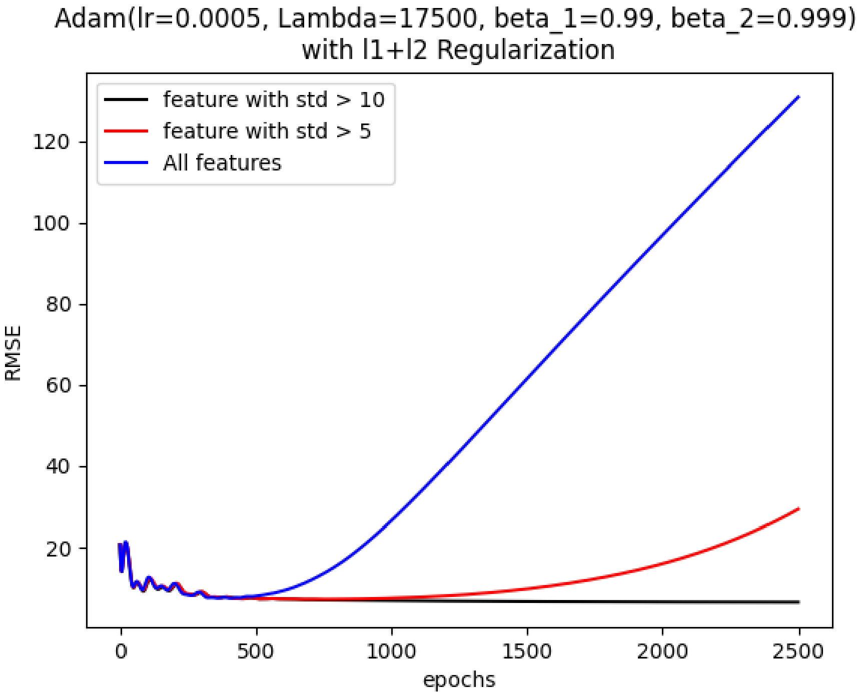

Figure2

[Discussion]

As the standard deviation represents the dispersion of the data, the more disperse the data are, it is better for the model to distinguish the differences in the data.

And the data are divided into three groups, using all features, using features with a std > 5, and using features with a std > 10.

As can be seen from Figure 2, the use of features with std > 10 allows for better model convergence than the other two models.

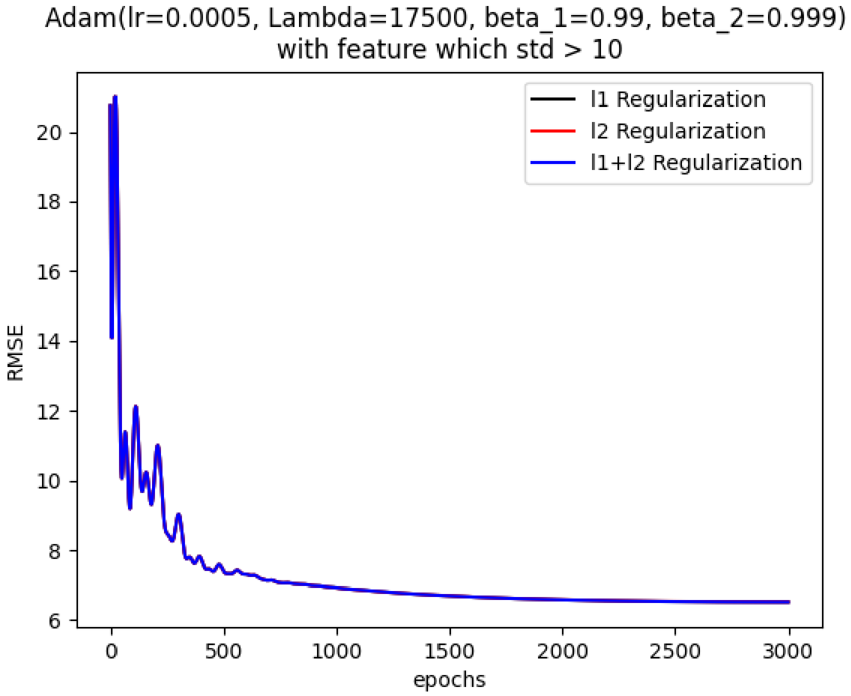

Figure3

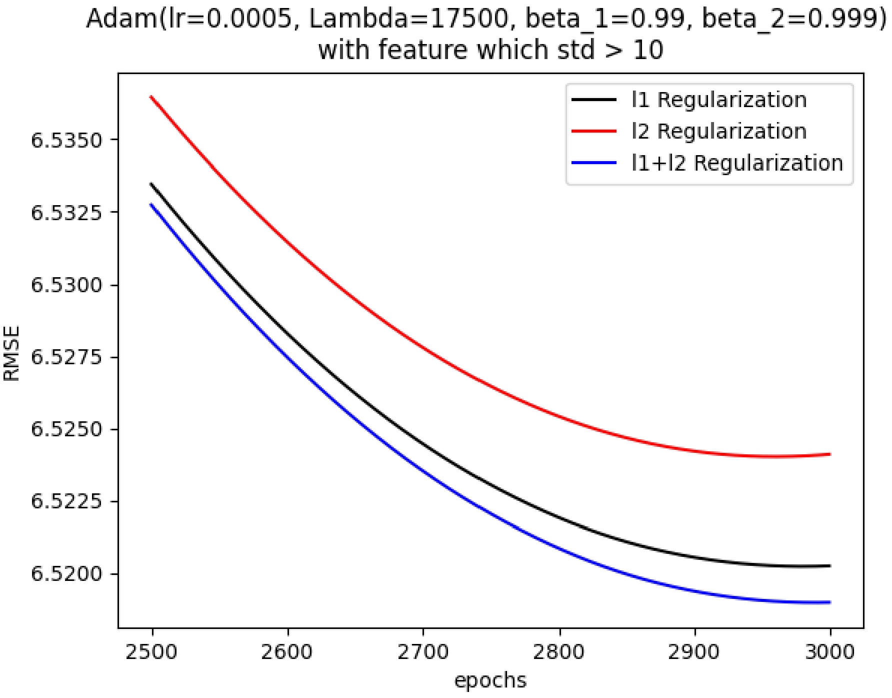

Figure4

[Discussion]

As can be seen from Figure 3 and Figure 4, although using l1+l2 Regularization has a very similar curve in RMSE to using l1 Regularization or l2 Regularization, if we zoom in on the epochs, we can see that l1+l2 Regularization achieves better results.

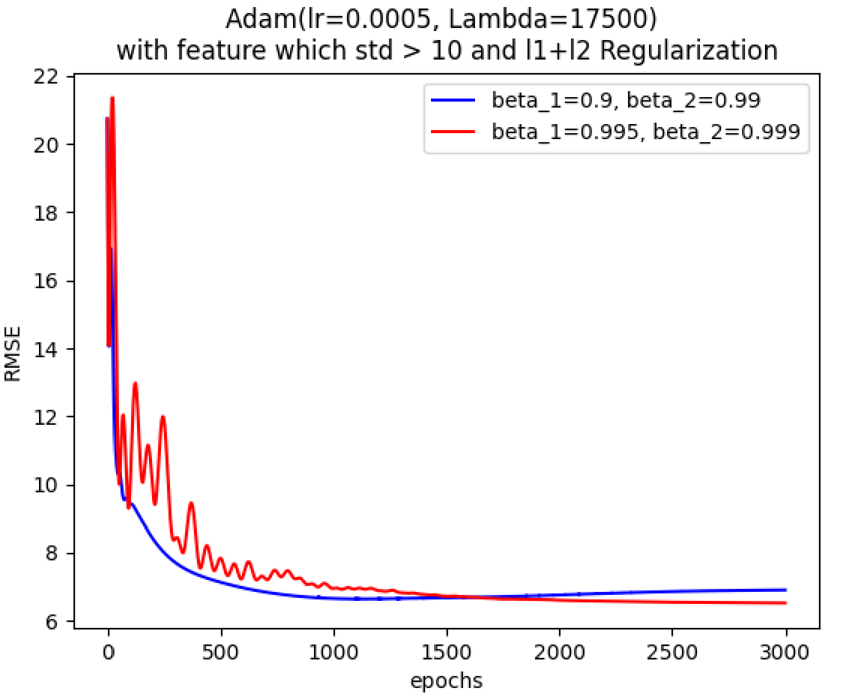

Figure5

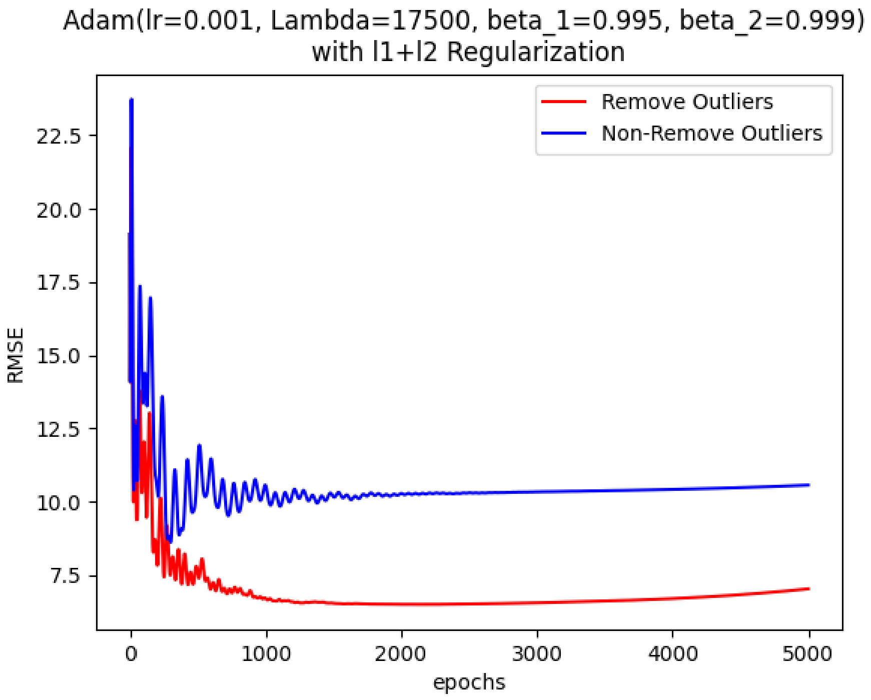
Adam(lr=0.001, Lambda=17500, beta_1=0.995, beta_2=0.999) with l1+l2 Regularization

Figure6

[Discussion]

As can be seen from Figure 5, a proper adjustment of beta_1 and beta_2 will result in better convergence. Also, as can be seen in Figure 6, if the outliers are not removed, RMSE will be poorer.

[Conclusion]

After the above discussion, I using the following as my model：

1. Using features with a std > 10

1. Remove Outliers

3. l1+l2 Regularization

4. beta_1=0.995, beta_2=0.999

3.(4%) Refer to math problem

[(https://hackmd.io/@GfOkB4kgS66YhhM7j6TJew/BykqpjhEK)](https://hackmd.io/@GfOkB4kgS66YhhM7j6TJew/BykqpjhEK)

ANS:

1.(a)

$b = 1,\ w_1 = -1,\ w_2 = 2,\ w_3 = -1,\ w_4 = 5$

$x_1 = 7,\ x_2 = 0,\ x_3 = 0,\ x_4 = 10$

$$\begin{aligned} f_{w,b}(x) = P_{w,b}(C_1|x) &= \sigma(\sum_i w_i x_i + b) \\ &= \sigma(\sum_{i=1}^{4} w_i x_i + b) \\ &= \sigma((-1)*7 + 2*0 + (-1)*3 + 5*10) + 1) \\ &= \sigma(-41) \\ &= \frac{1}{1+e^{-41}} \\ &\approx 1 \end{aligned}$$

1.(b)

$L(w, b) = f_{w,b}(x^1) f_{w,b}(x^2)(1 - f_{w,b}(x^3)) \cdots f_{w,b}(x^N)$

Take the $-ln$ of $L(w, b)$:

$$\begin{aligned} -lnL(w,b) &= -ln(f_{w,b}(x^1)) - ln(f_{w,b}(x^2)) - ln(1 - f_{w,b}(x^3)) \cdots f_{w,b}(x^N) \\ &= \sum_{i=1}^{N} -[\hat{y}^i ln(f_{w,b}(x^i)) + (1-\hat{y}^i) ln(f_{w,b}(x^i))] \end{aligned}$$

$$\hat{y}^n = \begin{cases} 1, & \text{for Class 1} \\ 0, & \text{for Class 2} \end{cases}$$

1.(c)

$$g(w,b) = \frac{-1}{N} \sum_{i=1}^{N} [\hat{y}^i ln(f_{w,b}(x^i)) + (1-\hat{y}^i) ln(f_{w,b}(x^i))]$$

$$\frac{\partial}{\partial(w,b)} g(w,b) = \frac{1}{N} \sum_{i=1}^{N} [f_{w,b}(x^i) - \hat{y}^i](x_j)^i$$

$$\begin{aligned} &(w,b)_j \leftarrow (w,b)_{j-1} - \eta \frac{\partial}{\partial(w,b)} g(w,b) \\ &\Rightarrow (w,b)_j \leftarrow (w,b)_{j-1} - \eta \frac{1}{N} \sum_{i=1}^{N} [f_{w,b}(x^i) - \hat{y}^i](x_j)^i \end{aligned}$$

2.(a)

Method 1:

$$\begin{aligned} L_{ssq}(w,b) &= \frac{1}{2N} \sum_{i=1}^{N} (y_i - (w^T x_i + b))^2 \\ &= \frac{1}{2*5} \sum_{i=1}^{5} (y_i - (w x_i + b))^2 \end{aligned}$$

To minizie $L_{ssq}(w,b)$

$$\begin{aligned} &\because \frac{\partial L}{\partial w} = \frac{1}{10} \sum_{i=1}^{5} (y_i - (w x_i + b))(x_i) = 0 \\ &\therefore 59.7 - 55w - 15b = 0 - (1) \end{aligned}$$

$$\begin{aligned} &\because \frac{\partial L}{\partial b} = \frac{1}{10} \sum_{i=1}^{5} (y_i - (w x_i + b))(1) = 0 \\ &\therefore 16.8 - 15w - 5b = 0 - (2) \end{aligned}$$

$$\begin{cases} 59.7 - 55w - 15b = 0 & \text{-(1)} \\ 16.8 - 15w - 5b = 0 & \text{-(2)} \end{cases}$$
$$\Rightarrow (w,b) = (0.93, 0.57)$$

Method 2:

Using the proof results of 2.(b), we can know that

$$X = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}, \hat{y} = \begin{bmatrix} 1.5 \\ 2.4 \\ 3.5 \\ 4.1 \\ 5.3 \end{bmatrix}, w' = \begin{bmatrix} w \\ b \end{bmatrix}$$

$$\begin{aligned} &w' = (XX^T)^{-1} X \hat{y} = \begin{bmatrix} 0.93 \\ 0.57 \end{bmatrix} \\ &\Rightarrow (w,b) = (0.93, 0.57) \end{aligned}$$

2.(b)

Let $X = \begin{bmatrix} x_1 \cdots x_n \\ 1 \cdots 1 \end{bmatrix}, \hat{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, w' = \begin{bmatrix} w \\ b \end{bmatrix}$

so, $L_{ssq}(w,b) = \frac{1}{2N} \sum_{i=1}^{N} (y_i - (w^T x_i + b))^2$ can be write as follows:

$$
\begin{aligned}
L_{ssq}(w') &= \frac{1}{2N}||\hat{y} - X^T w'||^2 \\
&= \frac{1}{2N}(\hat{y} - X^T w')^T(\hat{y} - X^T w') \\
&= \frac{1}{2N}(\hat{y}^T - w'^T X)(\hat{y} - X^T w') \\
&= \frac{1}{2N}(\hat{y}^T y - 2(X\hat{y})^T w' + w' X X^T w')
\end{aligned}
$$

To minizie $L_{ssq}(w'), \nabla L_{ssq}(w') = 0$

$$\because \nabla L_{ssq}(w') = \frac{1}{2N}(-2X\hat{y} + 2XX^T w') = \frac{1}{N}(XX^T w' - X\hat{y}) = 0$$

$$
\begin{aligned}
&\therefore XX^T w' - X\hat{y} = 0 \\
&\Rightarrow w' = (XX^T)^{-1} X\hat{y}, \text{ assume } (XX^T) \text{ is invertible.}
\end{aligned}
$$

2.(c)

Let $X = \begin{bmatrix} x_1 \cdots x_n \\ 1 \cdots 1 \end{bmatrix}, \hat{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, w' = \begin{bmatrix} w \\ b \end{bmatrix}$

so, $L_{reg}(w, b) = \frac{1}{2N} \sum_{i=1}^{N} (y_i - (w^T x_i + b))^2 + \frac{\lambda}{2}||w||^2$ can be write as follows:

$$L_{reg}(w') = \frac{1}{2N}||\hat{y} - X^T w'||^2 + \frac{\lambda}{2}||w||^2$$

To minizie $L_{reg}(w'), \nabla L_{reg}(w') = 0$

$$\because \nabla L_{reg}(w') = \frac{1}{2N}(-2X\hat{y} + 2XX^T w') + \lambda w' = \frac{1}{N}(XX^T w' - X\hat{y}) + \lambda w' = 0$$

$$
\begin{aligned}
&\therefore \frac{1}{N}(XX^T w' - X\hat{y}) + \lambda w' = 0 \\
&\Rightarrow (XX^T + N\lambda)w' = X\hat{y} \\
&\Rightarrow w' = (XX^T + N\lambda)^{-1} X\hat{y}, \text{ assume } (XX^T + N\lambda) \text{ is invertible.}
\end{aligned}
$$

3.

First, we prove $E[f_{w,b}(x_i + \eta_i)] = f_{w,b}(x_i)$

$$
\begin{aligned}
E[f_{w,b}(x_i + \eta_i)] &= E[w^T(x_i + \eta_i) + b] \\
&= E[(w^T x_i + b) + w^T \eta_i] \\
&= E[f_{w,b}(x_i) + w^T \eta_i] \\
&= f_{w,b}(x_i) + E[w^T \eta_i] \\
&= f_{w,b}(x_i) + \sum_{i=1}^{N} w^T \underbrace{E[\eta_i]}_{E[noise]=0} \\
&= f_{w,b}(x_i)
\end{aligned}
$$

Then, we prove $E[f_{w,b}(x_i + \eta_i)^2] = f_{w,b}^2(x_i) + \sigma^2||w||^2$

$$
\begin{aligned}
E[f_{w,b}(x_i + \eta_i)^2] &= E[(w^T(x_i + \eta_i) + b)^2] \\
&= E[(f_{w,b}(x_i) + w^T \eta_i)^2] \\
&= E[(f_{w,b}^2(x_i)] + 2E[(f_{w,b}(x_i) w^T \eta_i] + E[(w^T \eta_i)^2] \\
&= f_{w,b}^2(x_i) + \sigma^2||w||^2
\end{aligned}
$$

4.(a)

we know that

$$
\begin{aligned}
e_k &= \frac{1}{N}\sum_{i=1}^{N}(g_k(x_i) - y_i)^2 \\
s_k &= \frac{1}{N}\sum_{i=1}^{N} g_k(x_i)^2 \\
e_0 &= \frac{1}{N}\sum_{i=1}^{N} y_i^2
\end{aligned}
$$

and,

$$\because (a-b)^2 = a^2 - 2ab + b^2$$
$$\therefore ab = \frac{-1}{2}[(a-b)^2 - a^2 - b^2]$$

so, $\sum_{i=1}^{N} g_k(x_i)y_i$ will be

$$\begin{aligned}\sum_{i=1}^{N} g_k(x_i)y_i &= \frac{-1}{2}\sum_{i=1}^{N}[g_k(x_i)-y_i)^2 - g_k(x_i)^2 - y_i^2] \\ &= \frac{-1}{2}[Ne_k - Ns_k - Ne_0] \\ &= \frac{N}{2}[s_k - e_0 - e_k]\end{aligned}$$

4.(b)

Let $X = \begin{bmatrix} g_1(x_1) \cdots g_k(x_1) \\ \vdots \ddots \vdots \\ g_1(x_1) \cdots g_k(x_1) \end{bmatrix}, \hat{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \alpha = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_k \end{bmatrix}$

$$L_{test}(\sum_{k=1}^{K}\alpha_k g_k) = \frac{1}{N}\sum_{i=1}^{N}[(\sum_{k=1}^{K}\alpha_k g_k(x_i)) - y_i]^2$$
$$\Rightarrow L_{test}(\alpha) = \frac{1}{N}||X\alpha - \hat{y}||^2$$

To minize $L_{test}, \nabla L_{test} = 0$

$$\nabla L_{test} = \frac{2}{N}X^T(X\alpha - \hat{y}) = 0$$
$$\Rightarrow \alpha = (X^TX)^{-1}x^T\hat{y}$$