學號：B07902143　系級： 資工三　姓名：陳正康

請實做以下兩種不同 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。

e. 根據助教時間的公式表示，(1) 代表 p = 9x18+1 而(2) 代表 p = 9*1+1

以下以 private+public 表示 kaggle 分數

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

(2) 的結果其實不用多做什麼就相當的準確(5.39818+5.11703)，可能再加個 2 次 3 次項就能取得滿好的分數(5.01158+4.83891)，而(1)則容易有滿大的 loss

原因推測是單一 feature 本身就有很大的相關性，以及 pm2.5 是跟時間有關，包括季節、那陣子的天氣、大氣成分等，這些真正會影響 pm2.5 的資訊本身就隱涵在前 9 天的 pm2.5 數值中，因此不用多做 feature engineering 就能做為很清楚的指標

而使用(1)則因為有很多無用的資訊在其中，需要一一去挑出來去掉，若沒有好好做，其中的雜訊會嚴重影響模型的正確性。完全沒有做 feature selection 的(1)分數為 (6.34185+6.20666)

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

2-i.

首先看訓練資料的標準差，前 2 欄分別是 train_data_0, train_data_1 的標準差

SO2	1.44	19.61	1.15
NO	10.38	16.98	8.92
NOx	16.36	45.84	15.36
NO2	9.07	16.09	8.83
CO	0.35	19.53	0.32
O3	19.00	26.14	18.30
THC	0.32	43.65	0.33

CH4	0.15	13.80	0.17
NMHC	0.25	13.94	0.23
PM10	22.52	27.37	20.81
PM2.5	17.04	23.00	14.90
WS	0.39	30.85	0.39
WD	108.60	107.96	108.00
AT	5.43	30.60	5.09
RH	12.98	18.29	13.15

可以發現 train_data_1 的離散程度明顯巨大，分析後發現裡面有離群的資料

SO2	NO	NOx	NO2	CO	O3	THC	CH4	NMHC	PM10	PM2.5	WS	WD	AT	RH
744		2232		744	744	2232			744	744	1488		1488	
0		0		0	0	0			0	0	0		0	
0		0		0	0	0			0	0	0		0	
734		2205		732	733	2190			737	734	1482		1482	
744		2232		744	744	2232			744	744	1488		1488	
0.986559		0.987903		0.983871	0.985215	0.981183			0.990591	0.986559	0.995968		0.995968	
0.7	0.9	13.2	12.3	0.41	35.1	2.2	2.1	0.1	27	14	0.8	41	19.6	68.6
1	1.4	13.8	12.4	0.39	30.2	2.3	2.2	0.2	37	21	0.4	216	19.8	72.5
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.5	1.5	12.9	11.4	0.24	31.1	2.3	2.2	0.1	32	17	0.6	262	19.1	77
0.5	1.6	9.9	8.3	0.16	34.6	2.2	2.2	-	32	15	0.5	7	19.3	71.9
0.6	1.7	9.7	8	0.42	35.4	2.3	2.2	0.1	38	14	0.9	338	19.1	71.3
0.4	1.8	11.7	9.8	0.42	37.3	2.2	2.2	-	34	14	0.9	23	18.4	76
0.2	3.7	19.8	16.1	0.79	25.3	2.2	2.1	0.1	31	12	0.6	337	18.6	80
0.2	7.9	32.4	24.5	1.09	17.3	2.3	2.1	0.2	21	11	0.6	12	18.2	85.5
0.3	19.3	51.3	32	1.08	10	2.6	2.1	0.4	24	17	0.5	300	18	90
0.3	15.1	46.2	31.1	0.71	11.9	3.4	2.2	1.2	32	18	0.7	308	17.8	91
0.3	16.9	45.3	28.4	0.44	15.6	2.5	2.1	0.4	27	17	0.6	315	18	90
0.3	10.1	37	26.8	0.35	17.5	2.5	2.1	0.4	30	16	0.7	19	18.5	85.9
0.4	15.5	45.5	29.9	0.66	14.3	2.9	2.2	0.7	31	25	0.7	335	19	83.5
0.4	9.4	34.5	25.1	0.42	18.6	2.5	2.2	0.3	29	15	0.7	18	19.3	81.2
0.4	7	29	21.9	0.36	22.7	2.6	2.1	0.4	34	18	0.7	13	19.7	76.8
0.4	10.3	37.6	27.3	0.39	21.7	2.5	2.1	0.4	40	14	0.6	17	19.8	74.2

去除這些列

再把 NaN 都用平均值取代(test data 也會出現 0，就把它取代成 train data 中對應欄的平均)

處理完後 train_data_0 和 train_data_1 聯集的標準差如上表的第 3 欄

同時我們可以觀察到標準差最小的幾項依序是：CH4, NMHC, CO, THC

這些 feature 因為 variance 不大，雜訊的影響就高，因此考慮去除

2-ii.

把 feature normalize 後加上 Lasso Regularization 做 regression，把 weight 整理成如下 9*15 的矩陣，代表前 9 天的 15 個 feature：

```
   SO2,  NO,   NOx,  NO2,  CO,   O3,   THC,  CH4,  NMHC, PM10, PM2.5,WS,   WD,  AT,   RH
[[ 1.37  0.15  0.01  0.07  0.    0.01 -1.91 -0.   -0.    0.11  0.17  0.   -0.01 0.24  0.06]
 [-0.37  0.04  0.   -0.05  0.    0.03  0.    0.    0.    0.05  0.12  1.44 -0.01 0.11  0.02]
 [-0.4  -0.01 -0.01 -0.07  0.    0.04  0.    0.    0.    0.07  0.16  2.35 0.01  0.06  0.05]
 [ 0.    0.09 -0.04 -0.15  0.    0.   -0.   -0.   -0.   -0.03  0.12  1.22 0.    0.01 -0.05]
 [-0.03  0.   -0.04 -0.06  0.   -0.01 -0.   -0.   -0.    0.03  0.15  0.   0.    0.16 -0.01]
 [-0.3  -0.03 -0.05 -0.13  0.   -0.02  0.   -0.    0.   -0.02  0.03  0.02 0.02 -0.06 -0.07]
 [-0.    0.13  0.03 -0.   -0.    0.02 -0.   -0.   -0.   -0.05  0.07  0.   -0.01 0.11 -0.03]
 [ 0.7  -0.09  0.02 -0.04 -0.    0.1   0.   -0.    0.    0.03  0.14 -0.   0.   -0.48 -0.03]
 [-0.02  0.19  0.    0.11 -0.   -0.11  0.   -0.    0.    0.05  0.07 -0.   0.02 -0.13  0.2 ]]
```

把顯然接近 0 的欄位去除，也就是 CO, CH4, NMHC, THC

恰好也是 2-i.看到的那幾個 feature
但去除這些項後，分數沒有太大的 improve: 5.40587+5.00254-> 5.39818+4.95364

2-iii.
去除 label (PM2.5)是 NaN 的資料
再去除 feature 中 NaN 數量超過 50%的資料(以取 11 個 feature 為例，11*9*50% = 4.9，那超過 5 個 NAN 的要去除)
然後把 NaN 取代成該 feature 在 training data 中的平均值
分數有些許 improve: 從 5.39818+4.95364 變成 5.03497+4.85440

2-iv. 將 feature 標準化
這在使用 gradient descent 時會有明顯的影響，使得在有限的 iteration 下更容易找到極值
但因為 regression 的 loss function 是 convex，而且在使用 closed-from 解的情況下，分數反而變差，因此最後未採用

2-v. 高次項
本次在 public 和 private 都最高的分數，是加到 3 次項的模型
其中 2 次項採用 cross product，也就是 feature 間的兩兩乘積
這種模型可以 fit 得更好，但同時出現了一點 overfitting 的問題
加入 L2 Regularization 就能解決
分數: 4.99990+4.82539

3.(3%) Refer to math problem: https://hackmd.io/RFiu1FsYR5uQTrrpdxUvlw?view

1-(a)

Let $X = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}$

$\hat{y} = \begin{pmatrix} 1.2 \\ 2.4 \\ 3.5 \\ 4.1 \\ 5.6 \end{pmatrix}$

$w' = \begin{pmatrix} w \\ b \end{pmatrix}$

$\Rightarrow L_{ssq}(w, b) = L_{ssq}(w') = \frac{1}{10} \| \hat{y} - X^T w' \|^2$

$\Rightarrow$ the optimal w'^* occurs when $X^T w'$ is the projection of $\hat{y}$ on $Col(X^T)$, by linear alg.

$\Rightarrow w'^* = (XX^T)^{-1} X\hat{y} = \begin{pmatrix} 1.05 \\ 0.21 \end{pmatrix}$

$\Rightarrow \begin{cases} w = 1.05 \\ b = 0.21 \end{cases}$

1-(b)

$L_{seg}(w,b) = \frac{1}{2N}\sum_{i=1}^{N}\left(y_i - (w^T x_i + b)\right)^2$, where $y_i, b \in R$; $w, x_i \in R^k$

let: $w' = \begin{pmatrix} w \\ b \end{pmatrix} = \begin{pmatrix} w_1 \\ \vdots \\ w_k \\ b \end{pmatrix}$, $\hat{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$, $X = \begin{pmatrix} x_1 & \cdots & x_N \\ 1 & \cdots & 1 \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{N1} \\ \vdots & \ddots & \vdots \\ x_{1k} & \cdots & x_{Nk} \\ 1 & \cdots & 1 \end{pmatrix}$

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

$= \frac{1}{2N}(\hat{y} - X^T w')^T(\hat{y} - X^T w')$

$= \frac{1}{2N}(\hat{y}^T - w'^T X)(\hat{y} - X^T w')$

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

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

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

To minimize $L_{seg}(w')$, $\nabla L_{seg}(w') = 0$

$\Rightarrow XX^T w' = X\hat{y} \Rightarrow \boxed{w' = (XX^T)^{-1}X\hat{y} = \begin{pmatrix} w \\ b \end{pmatrix}}$ (by lin.alg., XX^T is symmetric and invertible)

1-(c)

$$L_{seq}(w,b) = \frac{1}{2N}\sum_{i=1}^{N}\left(y_i - (w^T x_i + b)\right)^2 + \frac{\lambda}{2}\|w\|^2$$

let w', $\hat{y}$, X defined as in 1-(b)

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

$$\Rightarrow \nabla L_{seq}(w') = \frac{1}{N}(XX^T w' - X\hat{y}) + \lambda w$$

To minimize L_{seq}, $\nabla L_{seq} = 0$

$$\Rightarrow \frac{1}{N}XX^T w' + \lambda w' = \frac{1}{N}X\hat{y}$$

$$\Rightarrow (XX^T + N\lambda I)w' = X\hat{y}$$

$$\Rightarrow \begin{pmatrix} w \\ b \end{pmatrix} = w' = (XX^T + N\lambda I)^{-1}X\hat{y}$$, assume $(XX^T + N\lambda I)$ is invertible

2.

(a) Show: $E[f_{w,b}(x_i+\eta_i)] = f_{w,b}(x_i)$

$\Rightarrow LHS = E[w^T(x_i+\eta_i)+b]$

$= E[(w^Tx_i+b)+w^T\eta_i]$

$= E[f(x_i) + w_1\eta_{i1} + \cdots + w_k\eta_{ik}]$

$= f(x_i) + w_1E[\eta_{i1}] + \cdots + w_kE[\eta_{ik}]$ ($E[\eta_{ij}]=0$)

$= f_{w,b}(x_i) = RHS$

(b) Show: $E[(f_{w,b}(x_i+\eta_i))^2] = f^2_{w,b}(x_i) + \sigma^2\|w\|^2$

$\Rightarrow LHS = E[(w^T(x_i+\eta_i)+b)^2]$

$= E[(f_{w,b}(x_i) + w^T\eta_i)^2]$

$= E[f^2_{w,b}(x_i)] + 2\underbrace{E[f_{w,b}(x_i)w^T\eta_i]}_{\text{as (a)}=0} + E[(w^T\eta_i)^2]$

$= f^2_{w,b}(x_i) + 0 + E[(w_1\eta_{i1}+\cdots+w_k\eta_{ik})^2]$

$= f^2_{w,b}(x_i) + w_1^2\underbrace{E[\eta_{i1}^2]}_{=\sigma^2} + \cdots + w_k^2\underbrace{E[\eta_{ik}^2]}_{=\sigma^2} + 2\sum_{r\neq s} w_rw_sE[\eta_{ir}\eta_{is}]$ ($E[\eta_{ir}\eta_{is}]=0$)

$= f^2_{w,b}(x_i) + \sigma^2\|w\|^2$

(c)

$$L_{sq}(w,b) =$$

$$= E\left[\frac{1}{2N} \sum_{i=1}^{N} \left(f_{w,b}(x_i + \eta_i) - y_i \right)^2 \right]$$

$$= \frac{1}{2N} \sum_{i=1}^{N} \left\{ E\left[f_{w,b}^2(x_i + \eta_i) \right] - 2y_i E\left[f_{w,b}(x_i + \eta_i) \right] + y_i^2 \right\}$$

$$= \frac{1}{2N} \sum_{i=1}^{N} \left\{ \overbrace{f_{w,b}^2(x_i) + \sigma^2 \|w\|^2}^{\text{by (b)}} - 2y_i \overbrace{f_{w,b}(x_i)}^{\text{by (a)}} + y_i^2 \right\}$$

$$= \frac{1}{2N} \sum_{i=1}^{N} \left(f_{w,b}(x_i) - y_i \right)^2 + \frac{\sigma^2}{2} \|w\|^2$$

know:

$S = \{(x_i, y_i)\}_{i=1}^{N}$, $x_i \in R^d$, $y_i \in R$

$g_0, g_1, \cdots g_K$, $g_k : R^d \to R$, $k = 0, \cdots K$

$g_0(x) = 0$

$e_k = \frac{1}{N} \sum_{i=1}^{N} \left(g_k(x_i) - y_i \right)^2$, $k = 0, 1 \cdots K$

$s_k = \frac{1}{N} \sum_{i=1}^{N} g_k^2(x_i)$

3-(a)

$$\sum_{i=1}^{N} g_k(x_i) y_i = \frac{-1}{2} \sum_{i=1}^{N} \left[\left(g_k(x_i) - y_i \right)^2 - g_k^2(x_i) - y_i^2 \right]$$

$$= \frac{-1}{2} \left(N e_k - N s_k - N e_0 \right)$$

$$= \frac{N}{2} \left(s_k + e_0 - e_k \right)$$

3-(b)

Let $X = \begin{pmatrix} g_1(x_1) & \cdots & g_K(x_1) \\ \vdots & \ddots & \vdots \\ g_1(x_n) & \cdots & g_K(x_n) \end{pmatrix}$, $\alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_K \end{pmatrix}$, $\hat{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$

$$L_{test}\left(\sum_{k=1}^{K} \alpha_k g_k \right) = \frac{1}{N} \sum_{i=1}^{N} \left[\left(\sum_{k=1}^{K} \alpha_k g_k(x_i) \right) - y_i \right]^2$$

$$\Rightarrow L_{test}(\alpha) = \frac{1}{N} \| X\alpha - \hat{y} \|^2$$

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

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

$$\Rightarrow \alpha = (X^T X)^{-1} \underbrace{X^T \hat{y}}_{\text{by 3-(a)}}$$

$$= \frac{N}{2} (X^T X)^{-1} \begin{pmatrix} e_0 + s_1 - e_1 \\ e_0 + s_2 - e_2 \\ \vdots \\ e_0 + s_K - e_K \end{pmatrix}$$