學號：B05602052 系級： 電機四 姓名：舒泓諭

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

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

	9x18+1	9x1+1
public	10.46342	5.96943
private	5.95145	5.89724

不論在 public score 還是 private score 中，9x18+1 的模型都來得比 9x1+1，要來的更高。從此可以發現 9x18+1 的模型，因為考慮了太複雜的 model 所以有 overfitting 的情形。

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。

我認為可以藉由挑掉極端值，以降低 RMSE(根據 kaggle public+private 分數)。以下都是使用 p = 9x1+1 的 model。

	沒有挑掉極端值的	有挑掉極端值的
public	7.26793	5.96943
private	6.83433	5.89724

不論在 public score 還是 private score 中，如果有挑掉極端值的模型都會比沒有挑掉極端值的 RMSE 來的更低。

此外，我也有對如何**挑掉極端值的範圍**進行實驗。在統計學中的定義，極端值上界為 $(Q3 + 1.5 \times IQR, \infty)$ ，極端值下界為 $(-\infty, Q1 - 1.5 \times IQR)$ 。在本次的 data 中，極端值上界為 $(72.6, \infty)$，極端值下界為 $(-\infty, -40)$ 。然而因為這次 PM2.5 正常數值最小為 0，所以下界定為 $(-\infty, -0)$。

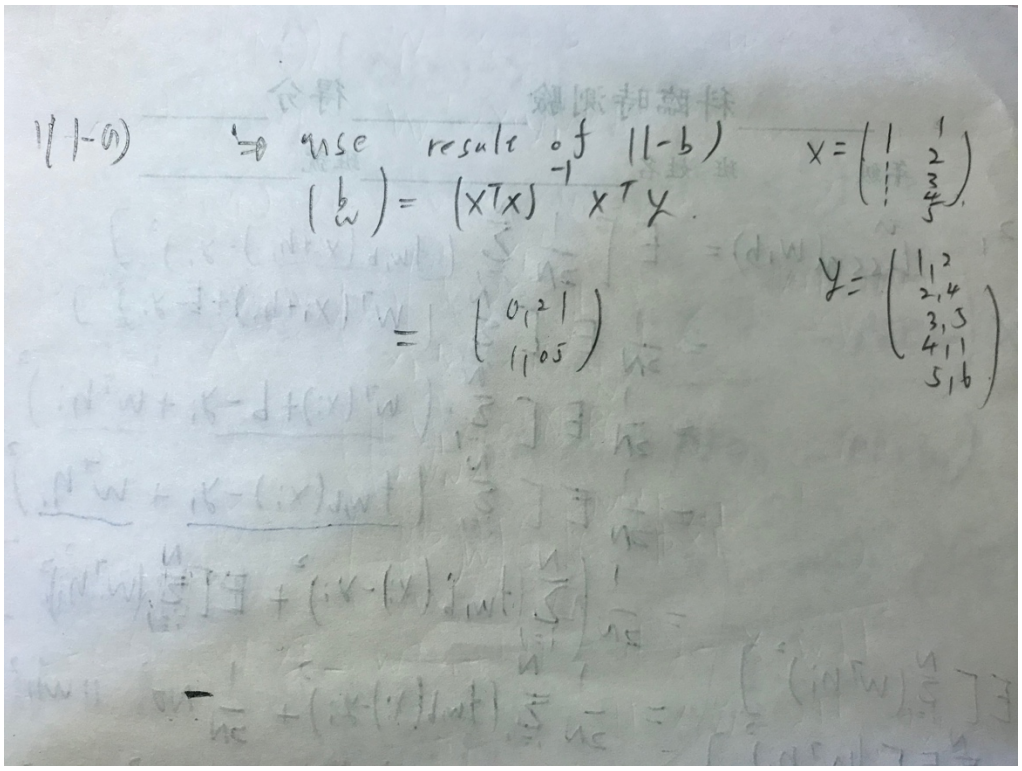
	range(2,100)以外挑掉	range(0,72.6)以外挑掉
public	5.96943	5.96926
private	5.89724	5.85901

從上可以發現雖然在 public score 這個根據統計學定義的極端值沒有什麼大幅的改變，但在 private score 卻上升了 0.04，可見藉由統計定義選出來的極端值，當所測資料量增加時效果才回比較明顯。

3.(3%) Refer to math problem

https://hackmd.io/RFiu1FsYR5uQTrrpdxUvlw?view

1.


1/(1-0)
⇒ use result of (1-b)
$\begin{pmatrix} b \\ w \end{pmatrix} = (X^T X)^{-1} X^T Y$
$= \begin{pmatrix} 0.2 \\ 1.05 \end{pmatrix}$
$X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{pmatrix}$
$Y = \begin{pmatrix} 1.2 \\ 2.4 \\ 3.3 \\ 4.1 \\ 5.6 \end{pmatrix}$

(11-b) $f(x_i) = w^T x_i + b \quad w \in R^1 \times R^k$

$f(x_1) - y_1 = w^T x_1 - y_1 + b$

$\Rightarrow f(x_n) - y_n = w^T x_n - y_n + b$

Denote $X = \begin{pmatrix} 1 & x_1^T \\ 1 & x_2^T \\ 1 & x_3^T \\ 1 & x_N^T \end{pmatrix} \in R^{N_1} \times R^{k+1}$

$y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \in R^N$

$\hat{w} = \begin{pmatrix} b \\ w \end{pmatrix} \in R^{k+1}$

$2NL_{SS}(\hat{w}) = L(\hat{w}) = \| y - w^T x_i - b \|^2 = \| y - X \cdot \hat{w} \|^2$

$= \| X \cdot \hat{w} - y \|^2$

$\hat{w}^T X^T y = (y^T X \hat{w})^T$ → both scalar.

$L(\hat{w}) = (X\hat{w} - y)^T (X\hat{w} - y)$

$= \hat{w}^T X^T X \hat{w} - \hat{w}^T X^T y - y^T X \hat{w} + y^T y$

$= \hat{w}^T X^T X \hat{w} - 2 \hat{w}^T X^T y + y^T y$

$\frac{\partial L(\hat{w})}{\partial \hat{w}} = 2 X^T X \hat{w} - 2 X^T y + 0$

minimize $\frac{\partial L(\hat{w})}{\partial \hat{w}} = 0 = 2 X^T X \hat{w} - 2 X^T y$

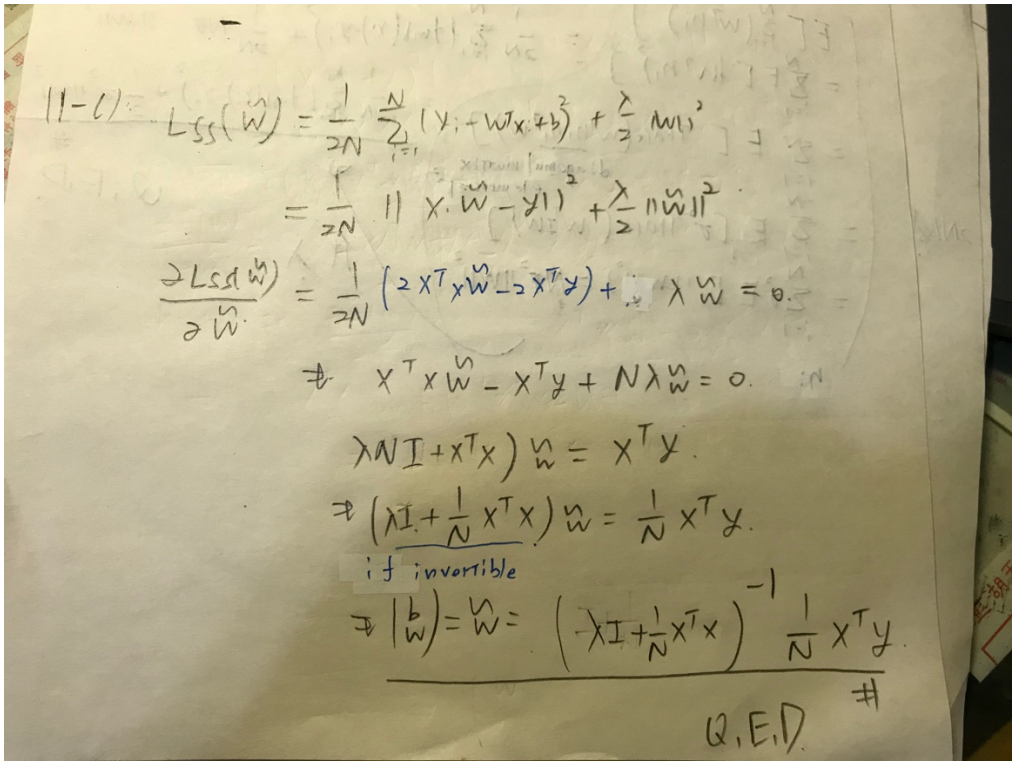
$\Rightarrow X^T X \hat{w} = X^T y \Rightarrow \hat{w} = (X^T X)^{-1} X^T y$

if $X^T X$ is invertible

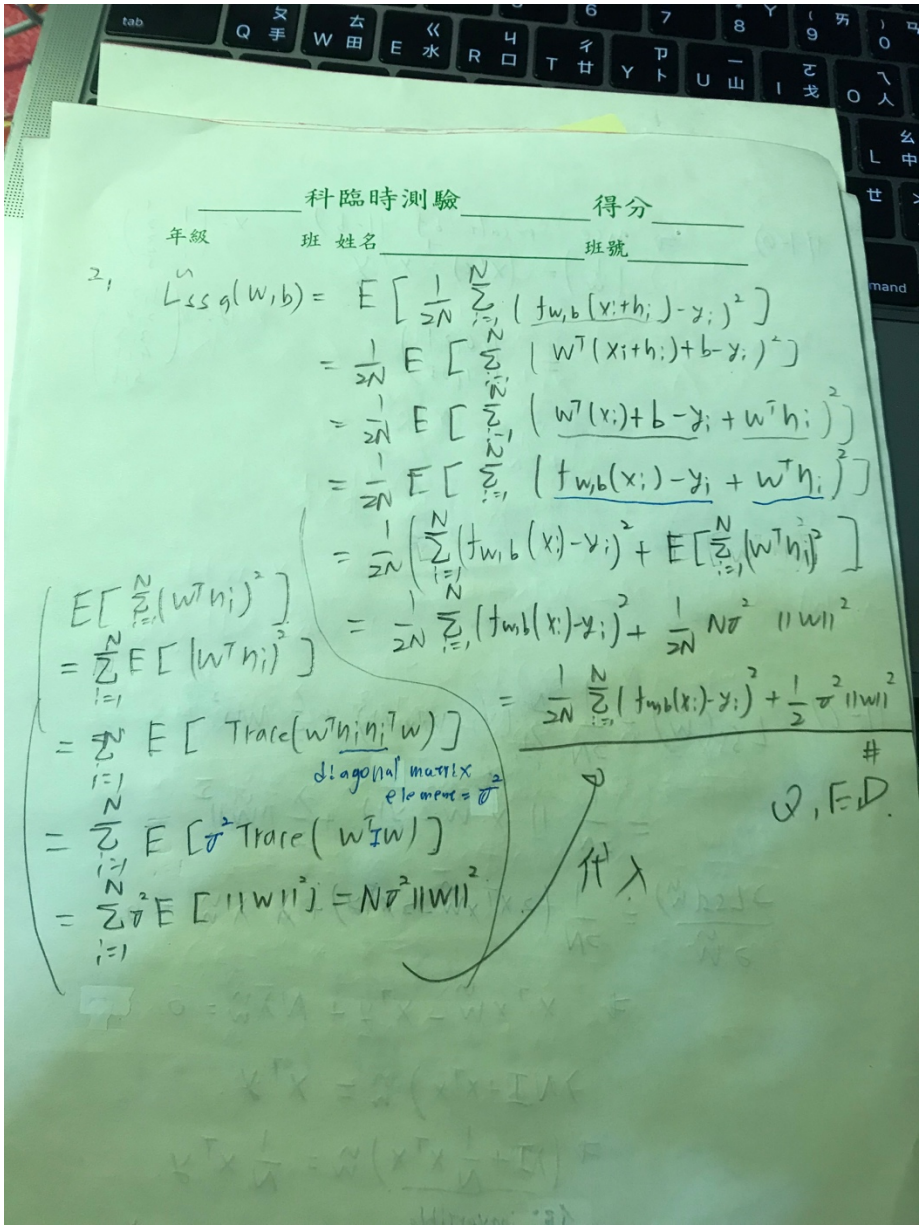
#

$\Rightarrow \begin{pmatrix} b \\ w \end{pmatrix} = (X^T X)^{-1} X^T y$

#


11-(1)
$L_{SS}(\vec{w}) = \frac{1}{2N}\sum_{i=1}^{N}(y_i + w^T x_i + b)^2 + \frac{\lambda}{2}\|w\|^2$
$= \frac{1}{2N}\|X\vec{w} - y\|^2 + \frac{\lambda}{2}\|\vec{w}\|^2$
$\frac{\partial L_{SS}(\vec{w})}{\partial \vec{w}} = \frac{1}{2N}(2X^TX\vec{w} - 2X^Ty) + \lambda\vec{w} = 0$
$\Rightarrow X^TX\vec{w} - X^Ty + N\lambda\vec{w} = 0$
$(N\lambda I + X^TX)\vec{w} = X^Ty$
$\Rightarrow (\lambda I + \frac{1}{N}X^TX)\vec{w} = \frac{1}{N}X^Ty$
if invertible
$\Rightarrow \begin{pmatrix} b \\ \vec{w} \end{pmatrix} = \vec{w} = (\lambda I + \frac{1}{N}X^TX)^{-1}\frac{1}{N}X^Ty$
#
Q.E.D.

2.


科臨時測驗 得分
年級 班 姓名 班號
2,
$L_{ridge}(w,b) = E\left[\frac{1}{2N}\sum_{i=1}^{N}(f_{w,b}(x_i+h_i)-y_i)^2\right]$
$= \frac{1}{2N}E\left[\sum_{i=1}^{N}(w^T(x_i+h_i)+b-y_i)^2\right]$
$= \frac{1}{2N}E\left[\sum_{i=1}^{N}(w^T(x_i)+b-y_i+w^Th_i)^2\right]$
$= \frac{1}{2N}E\left[\sum_{i=1}^{N}(f_{w,b}(x_i)-y_i+w^Th_i)^2\right]$
$= \frac{1}{2N}\left(\sum_{i=1}^{N}(f_{w,b}(x_i)-y_i)^2 + E\left[\sum_{i=1}^{N}(w^Th_i)^2\right]\right)$
$= \frac{1}{2N}\sum_{i=1}^{N}(f_{w,b}(x_i)-y_i)^2 + \frac{1}{2N}N\sigma^2\|w\|^2$
$= \frac{1}{2N}\sum_{i=1}^{N}(f_{w,b}(x_i)-y_i)^2 + \frac{1}{2}\sigma^2\|w\|^2$
#
Q.E.D.
$E\left[\sum_{i=1}^{N}(w^Th_i)^2\right]$
$= \sum_{i=1}^{N}E\left[(w^Th_i)^2\right]$
$= \sum_{i=1}^{N}E\left[\mathrm{Trace}(w^Th_ih_i^Tw)\right]$
diagonal matrix element = σ^2
$= \sum_{i=1}^{N}E\left[\sigma^2\mathrm{Trace}(w^Tw)\right]$
$= \sum_{i=1}^{N}\sigma^2E\left[\|w\|^2\right] = N\sigma^2\|w\|^2$
代入

3.

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

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

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

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

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

b) $L = \min \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{K} a_k g_k(x_i) - y_i\right)^2$

$L = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{K} a_k g_k(x_i) - y_i\right)^2$ ⇒ 目標對 L_{ridge} 作偏微分

找 $\frac{\partial L}{\partial a_k} = 0$ (⇒ minimize Loss

$= \frac{1}{N}\sum_{i=1}^{N}\left(\left(\sum_{k=1}^{K} a_k g_k(x_i)\right)^2 - 2\sum_{k=1}^{K} a_k g_k(x_i) y_i + y_i^2\right)$

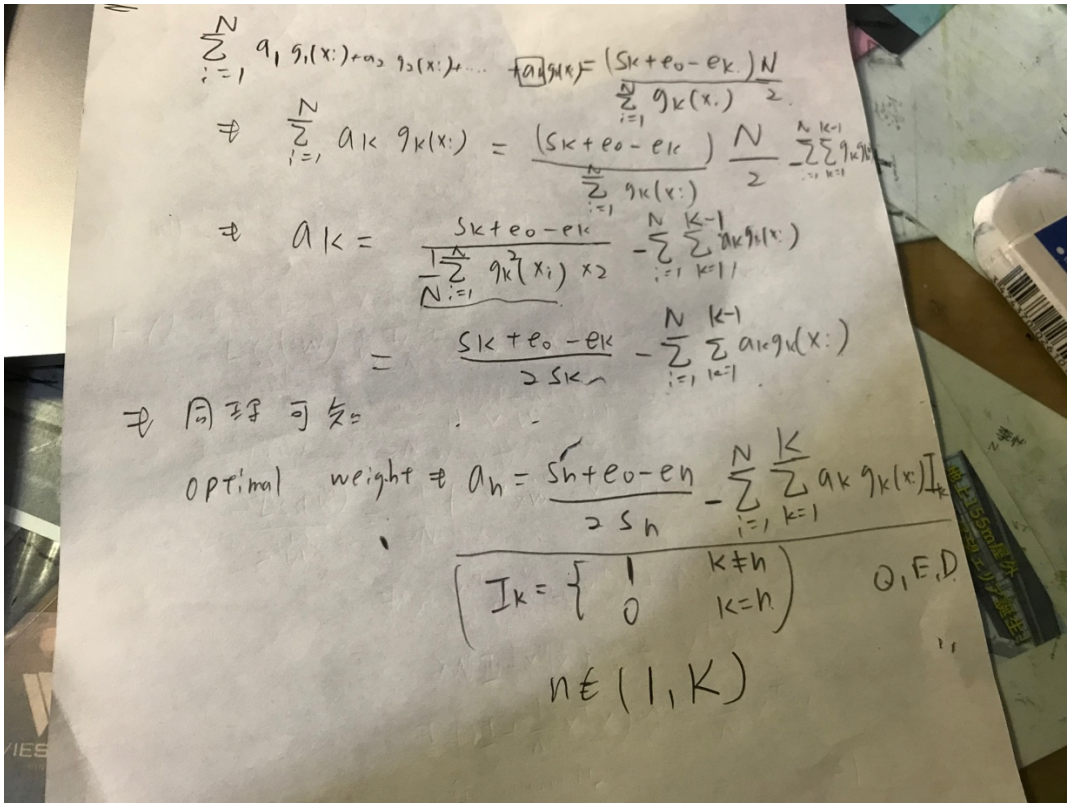
$= \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{K} a_k g_k(x_i)\right)^2 - \frac{2}{N}\sum_{k=1}^{K} a_k \sum_{i=1}^{N} g_k(x_i) y_i + \frac{1}{N}\sum_{i=1}^{N} y_i^2$

$= \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{K} a_k g_k(x_i)\right)^2 - \frac{2}{N}\sum_{k=1}^{K} a_k \frac{N}{2}\left(s_k + e_0 - e_k\right) + e_0$

$= \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{K} a_k g_k(x_i)\right)^2 - \sum_{k=1}^{K} a_k\left(s_k + e_0 - e_k\right) + e_0$

$\frac{\partial L}{\partial a_k} = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial a_k}\left(a_1 g_1(x_i) + a_2 g_2(x_i) + \cdots + a_k g_k(x_i)\right)^2 - \left(s_k + e_0 - e_k\right)$

$= \frac{1}{N}\sum_{i=1}^{N} 2\left(\sum_{k=1}^{K} a_k g_k(x_i)\right) g_k(x_i) - \left(s_k + e_0 - e_k\right) = 0$



$$\sum_{i=1}^{N} a_1 g_1(x_i) + a_2 g_2(x_i) + \cdots + a_K g_K(x_i) = \frac{(s_K + e_0 - e_K)\frac{N}{2}}{\sum_{i=1}^{N} g_K(x_i)}$$

$$\therefore \sum_{i=1}^{N} a_K g_K(x_i) = \frac{(s_K + e_0 - e_K)\frac{N}{2}}{\sum_{i=1}^{N} g_K(x_i)} - \sum_{i=1}^{N}\sum_{k=1}^{K-1} a_k g_k(x_i)$$

$$\therefore a_K = \frac{s_K + e_0 - e_K}{\sum_{i=1}^{N} g_K^2(x_i) \times 2} - \sum_{i=1}^{N}\sum_{k=1}^{K-1} a_k g_k(x_i)$$

$$= \frac{s_K + e_0 - e_K}{2 s_K} - \sum_{i=1}^{N}\sum_{k=1}^{K-1} a_k g_k(x_i)$$

∴ 同理可知

optimal weight ∴ $a_n = \frac{s_n + e_0 - e_n}{2 s_n} - \sum_{i=1}^{N}\sum_{k=1}^{K} a_k g_k(x_i) I_k$

$$\left(I_k = \begin{cases} 1 & k \neq n \\ 0 & k = n \end{cases} \right)$$

Q.E.D

$n \in (1, K)$