學號：R07922163 系級： 資工所二 姓名：徐浩翔

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)

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

	Private Score	Public Score
只用 PM2.5	5.98013	6.06713
用所有 feature	5.91050	5.76324

單單 PM2.5 這個 feature RMSE 就可以壓在 6 左右表示 PM2.5 這個 feature 是很重要的。
但保留所有 feature 的模型整體而言 RMSE 會比只考慮 PM2.5 的模型還要小，表示除了 PM2.5 這個 feature 還有其他因素也需要考慮。

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

我去查了和 pm2.5 相關的空污指標，得 CO、NO、NO2、O3、PM10、PM2.5、SO2 與其正相關，我使用七項 feature 做 Gradient descent，效果比放所有的 18 項 feature 來得好
選這七種 feature 都可以過 public 和 private 的 strong base line，另外有一些極端值像是超過 100 的我會把他們設成 80 小於零的把他設為 10 因為我算出了全年 pm2.5 平均大概是 16 所以就設了一個比較有可能的數字，這樣可以讓 traing data 變得比較多如果直接把他去掉會少好幾千筆 data

項目	Private Score	Public Score
取 7 個 feature	5.56758	5.72966
只用 PM2.5	5.98013	6.06713
用所有 feature	5.91050	5.76324

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

(1)-(a)

$\min\|Aw-y\|^2 \Rightarrow w=(A^TA)^{-1}A^Ty$ 其中 A為 $\begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_i & 1 \end{bmatrix}$ w為 $\begin{bmatrix} w \\ b \end{bmatrix}$ y為 $\begin{bmatrix} y_1 \\ \vdots \\ y_i \end{bmatrix}$

$A^TA = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \\ 4 & 1 \\ 5 & 1 \end{bmatrix} = \begin{bmatrix} 55 & 15 \\ 15 & 5 \end{bmatrix}$

$A^TA^{-1} = \frac{1}{50}\begin{bmatrix} 5 & -15 \\ -15 & 55 \end{bmatrix}$

$W = \frac{1}{50}\begin{bmatrix} 5 & -15 \\ -15 & 55 \end{bmatrix}\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}\begin{bmatrix} 1.2 \\ 2.4 \\ 3.5 \\ 4.1 \\ 5.6 \end{bmatrix}$

$= \frac{1}{50}\begin{bmatrix} 5 & 15 \\ -15 & 55 \end{bmatrix}\begin{bmatrix} 60.9 \\ 16.8 \end{bmatrix}$

$= \frac{1}{50}\begin{bmatrix} 52.5 \\ 10.5 \end{bmatrix} = \begin{bmatrix} 1.05 \\ 0.21 \end{bmatrix}$ #

(1)-(b)

Let $W' = \begin{bmatrix} w \\ b \end{bmatrix}$ $x'_i = \begin{bmatrix} x_i \\ 1 \end{bmatrix}$

$A = \begin{bmatrix} x'_1 & x'_2 & \cdots & x'_n \end{bmatrix}$ $Y = [y_1 \cdots y_n]$

$\min Lssq(w, b) = \min\|Y - W^TA\|^2$

$\Rightarrow \frac{(Y-W^TA)^T(Y-W^TA)}{\partial W} = 0$

$\frac{Y^TY - Y^TW^TA - A^TWY + A^TWW^TA}{\partial W} = 0$

$\Rightarrow -Y^TA - A^TY + 2A^TAW = 0$

$\Rightarrow 2A^TAW = 2A^TY$ ($\because A^TY = Y^TA$)

$\Rightarrow W = (A^TA)^{-1}A^TY$ #

$\Rightarrow$

(1)-(c)

Let $W' = \begin{bmatrix} w \\ b \end{bmatrix}$ $A = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \\ 1 & 1 & \cdots & 1 \end{bmatrix}$ $Y = [y_1 \cdots y_n]$

$\min L_{reg}(w,b) = \min \| Y - W'^T A \|^2 + \frac{\lambda}{2}\|W\|^2 - b^2$ $(\because W' = [\begin{smallmatrix} w \\ b \end{smallmatrix}])$

$\Rightarrow \frac{(Y - W'^T A)^T (Y - W'^T A) + \frac{\lambda}{2} W^T W - b^2}{\partial W} = 0$

$\Rightarrow \frac{Y^T Y - Y^T W^T A - A^T W Y + A^T W W^T A + \frac{\lambda}{2} W W^T}{\partial W} = 0$

$\Rightarrow -2Y^T A + 2A^T A W + \lambda W = 0$

$(A^T A + I\lambda)^{-1} W = Y^T A$

$\Rightarrow W = (A^T A + \frac{1}{2}\lambda I)^{-1} A^T Y$

(2)

$f_{w,b}(x) = W^T x + b$

$\eta \sim N(0,1)$

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

$2(W^T\eta_i)(W^T x + b - y) = 0$

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

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

$= \frac{1}{2N}\sum_{i=1}^{N} f(w_b(x_i) - y_i)^2 + \frac{1}{2N} E\left[\sum_{i=1}^{N}(W^T\eta_i)^2\right]$

$(E[\eta_{ij}\eta_{i'j'}] = \sigma^2 \quad i=i', j=j')$

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

3-(a)

$$e_k = \frac{1}{N}\sum\Big(\underbrace{g_k(x_i)^2}_{s_k} + \underbrace{y_i^2}_{e_0} - 2g_k(x_i)y_i\Big)$$

$$\Rightarrow \sum g(x_i)y_i = \frac{N}{2}(s_k - e_0 - e_k) \#$$

3-(b)

$$\min L\Big(\sum_{k=1}^{K}\alpha_k g_k\Big) = \min\Big[\frac{1}{N}\sum_{i=1}^{N}\Big(\sum_{k=1}^{K}\alpha_k g_k(x_i) - y_i\Big)^2\Big]$$

$$\frac{\partial L}{\partial \alpha} = \frac{2}{N}\sum_{i=1}^{N}\Big[\Big(\sum_{k=1}^{K}\alpha_k g_k(x_i) - y_i\Big)\sum_{k=1}^{K} g_k(x_i)\Big] = 0$$

$$\Big(\sum_{k=1}^{K}\alpha_k g_k(x_1) - y_1\Big)\sum_{k=1}^{K} g_k(x_1) + \Big(\sum_{k=1}^{K}\alpha_k g_k(x_2) \cdot g_k(x_2)\Big) \cdots \Big(\sum_{k=1}^{K}\alpha_k g_k(x_N) - y_N\Big)\sum_{k=1}^{K} g_k(x_N) = 0$$

Let $G = \begin{bmatrix} g_1(x_1) & g_2(x_1) & \cdots & g_K(x_1) \\ \vdots & & & \\ g_1(x_N) & \cdots & \cdots & g_K(x_N) \end{bmatrix}$ $\alpha = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_K \end{bmatrix}$ $y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$

$$\Rightarrow G^T G\alpha = G^T Y$$

$$\Rightarrow \alpha = (G^T G)^{-1} G^T Y$$