

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 $p = 9 \times 18 + 1$ 而(2) 代表 $p = 9 \times 1 + 1$

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

	$9 \times 18 + 1$	$9 \times 1 + 1$
public	10.46342	5.96943
private	5.95145	5.89724

不論在 public score 還是 private score 中， $9 \times 18 + 1$ 的模型都來得比 $9 \times 1 + 1$ ，要來的更高。從此可以發現 $9 \times 18 + 1$ 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 $p = 9 \times 1 + 1$ 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/RFiu1FsYR5uQTrrpdxUvIw?view>

1.

1/1-0) \Rightarrow use result of 11-b)

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

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

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

$$= \begin{pmatrix} 0.21 \\ 1.05 \end{pmatrix}$$

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

$$f(x_i) - y_i = w^T x_i - y_i + b$$

$$\Rightarrow f(x_m) - y_m = w^T x_m - y_m + b$$

Denote

$$X = \begin{pmatrix} 1 & x_1^T \\ \vdots & \vdots \\ 1 & x_N^T \end{pmatrix} \in \mathbb{R}^{N \times (k+1)}$$

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

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

$$2N L(\tilde{w}) = \|Y - X\tilde{w}\|^2 = \|Y - X\tilde{w}\|^2$$

$$L(\tilde{w}) = (X\tilde{w} - Y)^T (X\tilde{w} - Y) = \|X\tilde{w} - Y\|^2$$

$\tilde{w}^T X^T Y = (Y^T X \tilde{w})^T$
 \Rightarrow both scalar.

$$L(\tilde{w}) = \tilde{w}^T X^T X \tilde{w} - \tilde{w}^T X^T Y - Y^T X \tilde{w} + Y^T Y$$

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

$$\text{minimize } \frac{\partial L(\tilde{w})}{\partial \tilde{w}} = 0 = 2X^T X \tilde{w} - 2X^T Y$$

$$\Rightarrow X^T X \tilde{w} = X^T Y \Rightarrow \tilde{w} = (X^T X)^{-1} X^T Y$$

if $X^T X$ is invertible

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

$$\begin{aligned}
 ||-||: L_{SS}(\tilde{w}) &= \frac{1}{2N} \sum_{i=1}^N (y_i - w^T x_i + b)^2 + \frac{\lambda}{2} ||w||^2 \\
 &= \frac{1}{2N} ||X \tilde{w} - y||^2 + \frac{\lambda}{2} ||\tilde{w}||^2 \\
 \frac{\partial L_{SS}(\tilde{w})}{\partial \tilde{w}} &= \frac{1}{2N} (2X^T X \tilde{w} - 2X^T y) + \lambda \tilde{w} = 0 \\
 \Rightarrow X^T X \tilde{w} - X^T y + N\lambda \tilde{w} &= 0 \\
 \lambda N I + X^T X) \tilde{w} &= X^T y \\
 \Rightarrow \left(\lambda I + \frac{1}{N} X^T X \right) \tilde{w} &= \frac{1}{N} X^T y \\
 \text{if invertible} \\
 \Rightarrow \tilde{w} &= \left(\lambda I + \frac{1}{N} X^T X \right)^{-1} \frac{1}{N} X^T y
 \end{aligned}$$

Q.E.D. #

2.

科臨時測驗

得分

年級

班 姓名

班號

$$\begin{aligned}
 2, \quad L_{SS} g(w, b) &= E \left[\frac{1}{2N} \sum_{i=1}^N (f_{w,b}(x_i + h_i) - y_i)^2 \right] \\
 &= \frac{1}{2N} E \left[\sum_{i=1}^N (w^T(x_i + h_i) + b - y_i)^2 \right] \\
 &= \frac{1}{2N} E \left[\sum_{i=1}^N (w^T(x_i) + b - y_i + w^T h_i)^2 \right] \\
 &= \frac{1}{2N} E \left[\sum_{i=1}^N (f_{w,b}(x_i) - y_i + w^T h_i)^2 \right] \\
 &= \frac{1}{2N} \left(\sum_{i=1}^N (f_{w,b}(x_i) - y_i)^2 + E \left[\sum_{i=1}^N (w^T h_i)^2 \right] \right) \\
 &= \frac{1}{2N} \sum_{i=1}^N (f_{w,b}(x_i) - y_i)^2 + \frac{1}{2N} N \sigma^2 \|w\|^2
 \end{aligned}$$

$$\begin{aligned}
 &E \left[\sum_{i=1}^N (w^T h_i)^2 \right] \\
 &= \sum_{i=1}^N E \left[(w^T h_i)^2 \right]
 \end{aligned}$$

$$= \sum_{i=1}^N E \left[\text{Trace}(w^T h_i h_i^T w) \right]$$

$$= \sum_{i=1}^N E \left[\sigma^2 \text{Trace}(w^T w) \right]$$

$$= \sum_{i=1}^N \sigma^2 E \left[\|w\|^2 \right] = N \sigma^2 \|w\|^2$$

diagonal matrix
element = σ^2

$\uparrow \lambda$

\varnothing, E, D

$$\begin{aligned}
 (a) \quad Ne_k &= \frac{1}{N} \sum_{i=1}^N (g_k(x_i) - y_i)^2 \\
 &= \frac{1}{N} \sum_{i=1}^N (g_k(x_i)^2 - 2y_i g_k(x_i) + y_i^2) \\
 \sum_{i=1}^N y_i g_k(x_i) &= \left(\sum_{i=1}^N g_k(x_i)^2 + \sum_{i=1}^N y_i^2 - Ne_k \right) \times \frac{1}{2} \\
 &= \frac{1}{2} (N s_k + N e_0 - Ne_k) \\
 &= \frac{N}{2} (s_k + e_0 - e_k) \quad Q, E, P.
 \end{aligned}$$

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

$$\begin{aligned}
 L &= \frac{1}{N} \sum_{i=1}^N \left(\sum_{k=1}^K a_k g_k(x_i) - y_i \right)^2 \Rightarrow \text{目標對 } L_{\text{loss}} \text{ 求極小值} \\
 &= \frac{1}{N} \sum_{i=1}^N \left(\left(\sum_{k=1}^K a_k g_k(x_i) \right)^2 - 2 \sum_{k=1}^K a_k g_k(x_i) y_i + y_i^2 \right) \Rightarrow \frac{\partial L}{\partial a_k} = 0 \Rightarrow \text{minimize Loss} \\
 &= \frac{1}{N} \sum_{i=1}^N \left(\sum_{k=1}^K a_k g_k(x_i) \right)^2 - 2 \sum_{k=1}^K a_k \frac{1}{N} \sum_{i=1}^N g_k(x_i) y_i + \frac{1}{N} \sum_{i=1}^N y_i^2 \\
 &= \frac{1}{N} \sum_{i=1}^N \left(\sum_{k=1}^K a_k g_k(x_i) \right)^2 - 2 \sum_{k=1}^K a_k \frac{N}{2} (s_k + e_0 - e_k) + e_0 \\
 &= \frac{1}{N} \sum_{i=1}^N \left(\sum_{k=1}^K a_k g_k(x_i) \right)^2 - \sum_{k=1}^K a_k (s_k + e_0 - e_k) + e_0 \\
 \frac{\partial L}{\partial a_k} &= \frac{1}{N} \sum_{i=1}^N \frac{\partial}{\partial a_k} \left(a_1 g_1(x_i) + a_2 g_2(x_i) + \dots + a_k g_k(x_i) \right)^2 - (s_k + e_0 - e_k) \\
 &= \frac{1}{N} \sum_{i=1}^N 2 \left(\sum_{k=1}^K a_k g_k(x_i) \right) g_k(x_i) - (s_k + e_0 - e_k) = 0
 \end{aligned}$$

$$\sum_{i=1}^N g_1(x_i) + a_0 g_2(x_i) + \dots + a_K g_K(x_i) = \frac{(s_K + e_0 - e_K) N}{\sum_{i=1}^N g_K(x_i)} \frac{1}{2}$$

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

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

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

⇒ 同法可证

$$\text{Optimal weight } a_h = \frac{s_h + e_0 - e_h}{2 s_h} - \sum_{i=1}^N \sum_{k=1}^{K-1} \frac{a_k g_k(x_i)}{g_h(x_i)} I_k$$

$$I_k = \begin{cases} 1 & k \neq h \\ 0 & k = h \end{cases} \quad Q.E.D$$

$$n \in (1, K)$$