

請實做以下兩種不同 feature 的模型，回答第 (1) ~ (3) 題：

- (1) 抽全部 9 小時內的污染源 feature 的一次項(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，其他的數值不要做任何更動
- b. 所有 advanced 的 gradient descent 技術(如: adam, adagrad 等) 都是可以用的

1. (2%)記錄誤差值 (RMSE)(根據 kaggle public+private 分數)，討論兩種 feature 的影響

(1) RMSE：6.57001

(2) RMSE：6.767938

跑所有 feature 對 pm2.5 的預測有太多不相干的因素影響，故只跑 pm2.5 一個 feature 的 model 會比較好。

2. (1%)將 feature 從抽前 9 小時改成抽前 5 小時，討論其變化

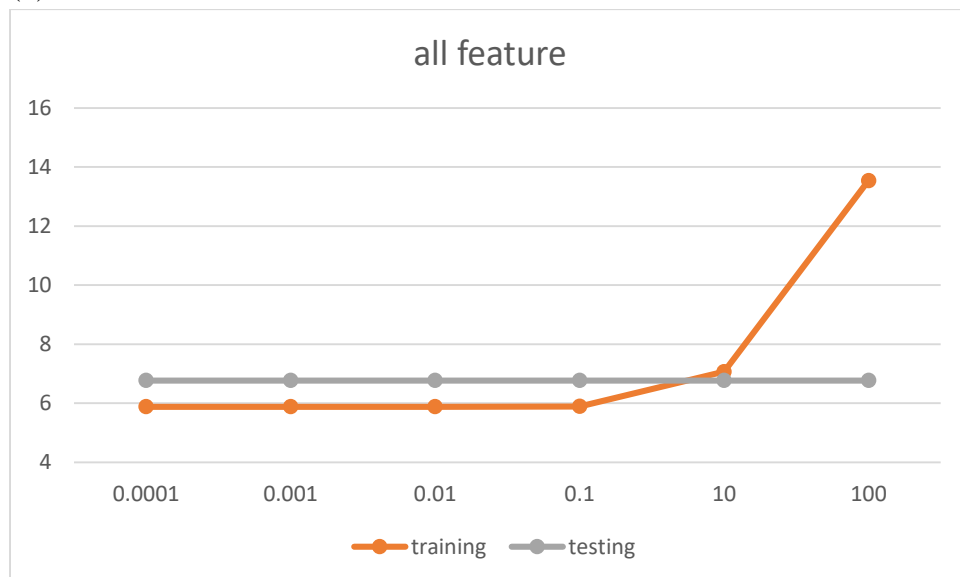
5 小時只傳 PM2.5：5.72630 7.23504=6.524427772

5 小時全傳：5.38403 7.73111=6.661750553

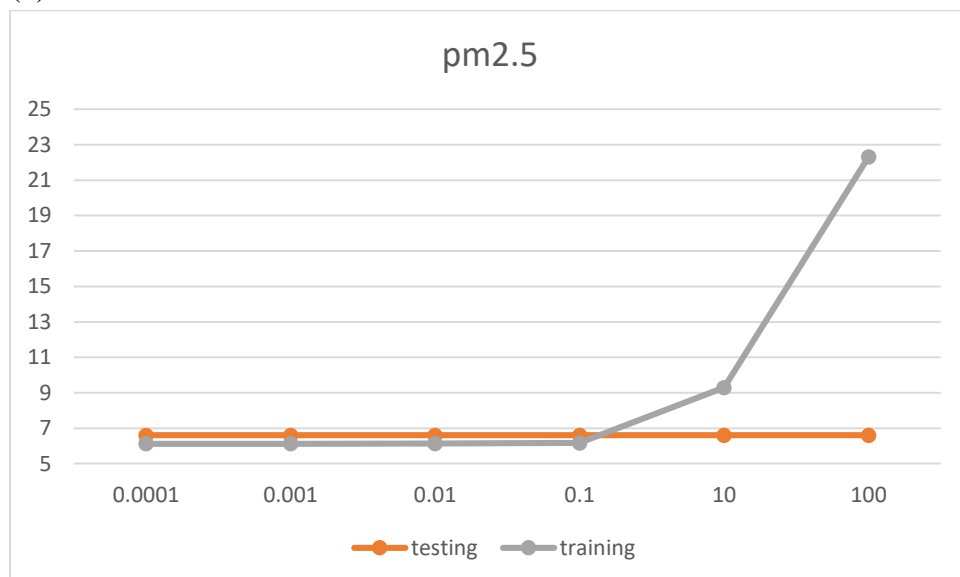
仍然是指傳 pm2.5 一個 feature 的 model 比較好，跟 9 小時的資料相比，5 小時的資料產生的 RMSE 較低。

3. (1%)Regularization on all the weight with  $\lambda=0.1$ 、 $0.01$ 、 $0.001$ 、 $0.0001$ ，並作圖

(1)



(2)



4. (1%)在線性回歸問題中，假設有  $N$  筆訓練資料，每筆訓練資料的特徵 (feature) 為一向量  $x^n$ ，其標註(label)為一存量  $y^n$ ，模型參數為一向量  $w$  (此處忽略偏權值  $b$ )，則線性回歸的損失函數(loss function)為  $\sum_{n=1}^N (y^n - x^n \cdot w)^2$ 。若將所有訓練資料的特徵值以矩陣  $X = [x^1 \ x^2 \ \dots \ x^N]^T$  表示，所有訓練資料的標註以向量  $y = [y^1 \ y^2 \ \dots \ y^N]^T$  表示，請問如何以  $X$  和  $y$  表示可以最小化損失函數的向量  $w$ ？請寫下算式並選出正確答案。(其中  $X^T X$  為 invertible)

(a)  $(X^T X) X^T y$

(b)  $(X^T X)^{-1} X^T y$

(c)  $(X^T X)^{-1} X^T y$

(d)  $(X^T X)^{-2} X^T y$

$$\text{input } X = \begin{bmatrix} x_0^1 & x_1^1 \\ x_0^2 & x_1^2 \\ \vdots & \vdots \\ x_0^m & x_1^m \end{bmatrix} \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \quad w = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$
$$H_w(X) = XW = \begin{bmatrix} x_0^1 & x_1^1 \\ x_0^2 & x_1^2 \\ \vdots & \vdots \\ x_0^m & x_1^m \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \quad H_w(X) - Y = XW - Y = \begin{bmatrix} h_w(x^1) - y^1 \\ h_w(x^2) - y^2 \\ \vdots \\ h_w(x^m) - y^m \end{bmatrix}$$

$$\min_w \sum_{j=1}^m [h_w(x^j) - y^j]^2 \Rightarrow \sum_{j=1}^m [h_w(x^j) - y^j]^2 = (XW - Y)^T (XW - Y)$$

$$J(w) = \frac{1}{2} \sum_{j=1}^m [h_w(x^j) - y^j]^2 = \frac{1}{2} (XW - Y)^T (XW - Y)$$

$$\begin{aligned} \frac{\partial}{\partial w} J(w) &= \frac{1}{2} \frac{\partial}{\partial w} (XW - Y)^T (XW - Y) \\ &= \frac{1}{2} \frac{\partial}{\partial w} (W^T X^T X - W^T X^T Y - Y^T XW + Y^T Y) \\ &= X^T XW - X^T Y = 0 \end{aligned}$$

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