

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

	Public	Private	Sum
(1):all features	7.85423	5.63865	6.83678
(2):only PM2.5	7.26669	5.62639	6.49850

(2)的誤差比(1)還要小，feature(2)為 PM2.5，與預測目標相同，可推論 feature(1)中有和 PM2.5 相關度低的參數，加入後會使得 train 出來的 model 預測效果變差，進而讓誤差變大。

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

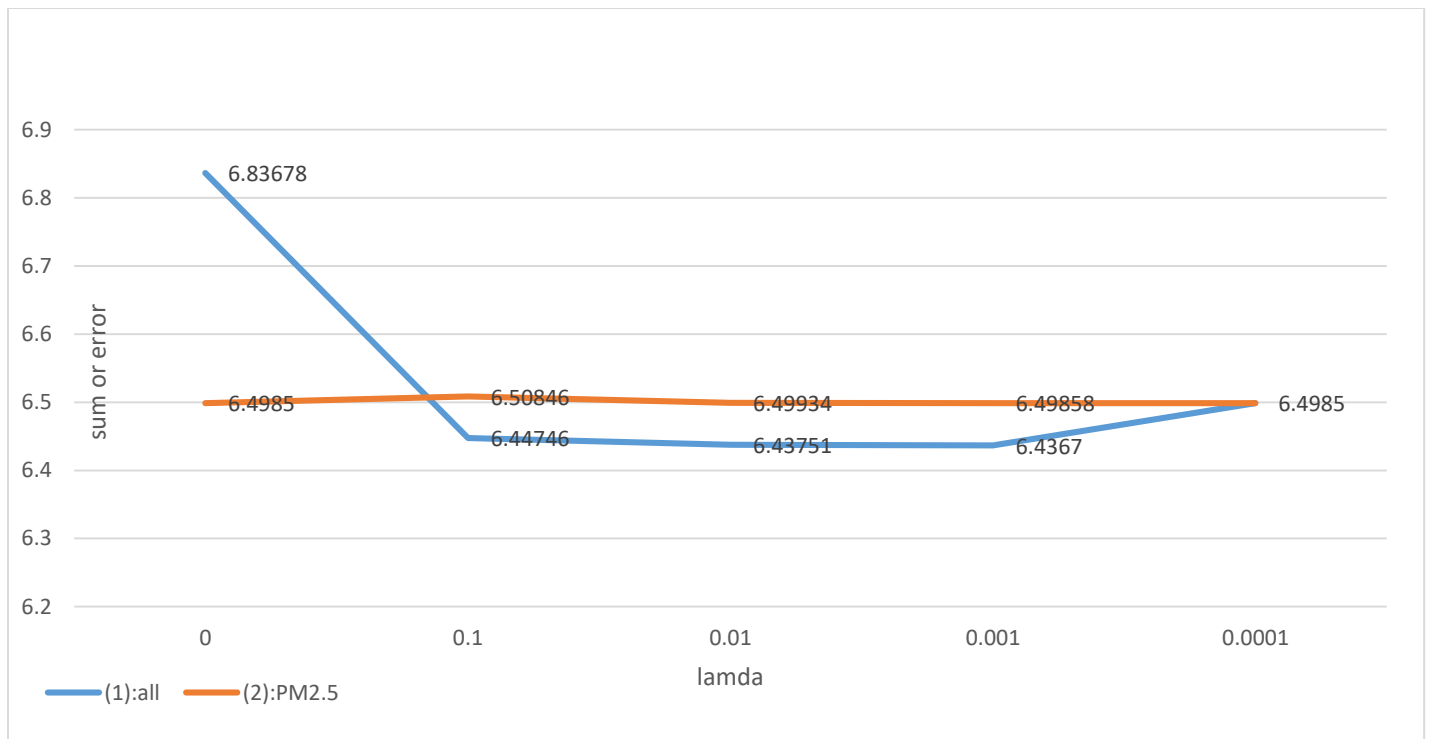
(1):all features	Public	Private	Sum
5 hours	7.64274	5.36036	6.60094
9 hours	7.85423	5.63865	6.83678

對 feature(1)來說，取前五小時的誤差比取前九小時的誤差小，可推論在 feature(1)中，前九到前六個小時的參數和預測目標的相關度較低，故去掉之後使得 train 出來的 model 預測效果變好，讓誤差變小。

(2):only PM2.5	Public	Private	Sum
5 hours	7.41581	5.79723	6.65590
9 hours	7.26669	5.62639	6.49850

對 feature(2)來說，取前五小時的誤差比取前九小時的誤差大，可推論在 feature(2)中，前九個小時的資料和預測目標的相關度高，故去掉之後使得 train 出來的 model 預測效果變差，讓誤差變大。

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



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 $L = \sum_{n=1}^N (y^n - x^n \cdot w)^2$ 。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) $(X^T X)^{-0} X^T y$
- (c) $(X^T X)^{-1} X^T y$
- (d) $(X^T X)^{-2} X^T y$

$$L(w) = (y - X \cdot w)^T \cdot (y - X \cdot w) = y^T y - 2(Xw)^T y + (Xw)^T (Xw)$$

$$\frac{dL(w)}{dw} = -2X^T y + 2X^T X w = 0 \Rightarrow y^T X = X^T X w \Rightarrow w = (X^T X)^{-1} X^T y \Rightarrow \text{選 c}。$$