

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
(考慮 training 次數為 15000 次，且 learning_rate 為 0.1)

All feature → Public : 8.44 / Private : 5.64 / RMSE : 7.18

PM2.5 → Public : 7.44 / Private : 5.63 / RMSE : 6.6

可以發現到如果只使用 PM2.5 的話，會得到較正確的預測值。可能是因為所有 feature 之中，有些比較不相關的 feature 會影響到預測的準確性。

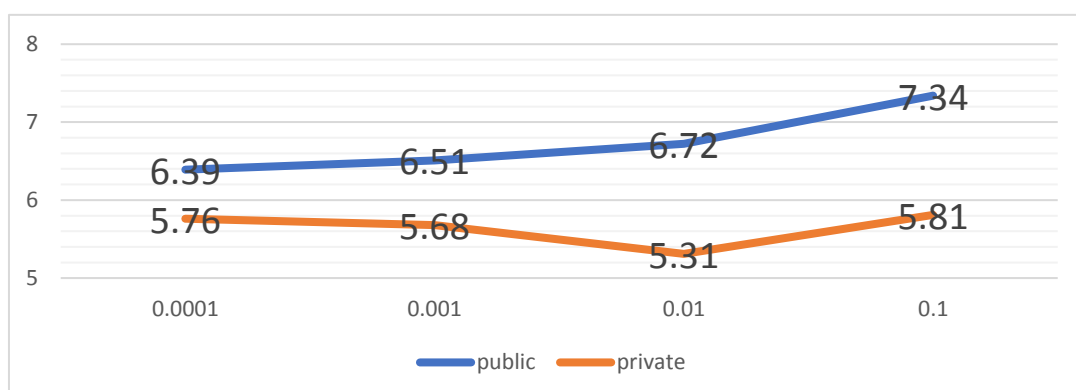
2. (1%)將 feature 從抽前 9 小時改成抽前 5 小時，討論其變化
(考慮 training 次數為 15000 次，且 learning_rate 為 0.1)

All feature → Public : 8.41 / Private : 5.56 / RMSE : 7.13

PM2.5 → Public : 7.58 / Private : 5.8 / RMSE : 6.75

發現到 All feature 的準確率提升了，有可能是因為維度減少，把較不相關的資料也一併除去了，然而，對於 PM2.5 來說，準確率卻下降了，代表著 PM2.5 可能是一項非常相關而且關鍵的資料。

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

Proof:

$$\text{Loss function} = L = \|y - Xw\|^2 = (y - Xw)^T (y - Xw) = y^T y - 2y^T Xw + w^T X^T Xw$$

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

$$X^T Xw = X^T y$$

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