

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

9小時內的污染源feature, $RMSE=5.46691+7.45974=12.92665$

9小時內的pm2.5feature, $RMSE=5.62719+7.44013=13.06732$

從kaggle的分數來看，差距不大(抽全部的feature稍微好一點點)，可能的原因是只靠一次項不夠，需要更複雜的model。

2. (1%)將feature從抽前9小時改成抽前5小時，討論其變化

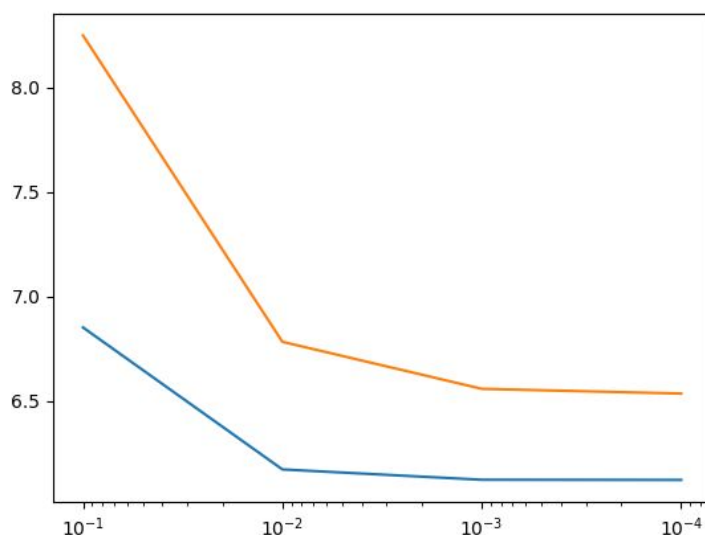
5小時內的污染源feature, $RMSE=5.43346+7.65771=13.09117$

5小時內的pm2.5feature, $RMSE=5.79427+7.57651=13.37078$

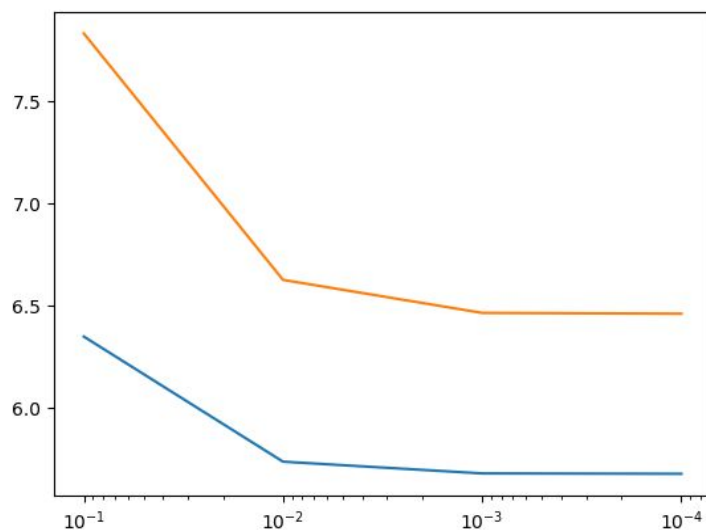
跟9小時相較，結果較差，但差距仍然不大。

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

9小時內的污染源feature + Regularization (藍線:training 橘線:testing)



9小時內的pm2.5feature + Regularization (藍線:training 橘線:testing)



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

loss function 可以改寫成matrix的形式：

$$|Xw - y|^2 = w^T X^T X w - 2w^T X^T y + y^T y$$

取微分 = 0 則

$$\nabla |Xw - y|^2 = X^T X w - X^T y = 0$$

$$X^T X w = X^T y$$

因 $X^T X$ 可逆 故

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

答案為 (c)