

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

	全部污染源 feature 的一次項	全部 PM2.5 的一次項
RMSE	6.51025	6.60643

根據上表的結果看來，採用「全部污染源 feature 的一次項」這個模型比起採用「全部 PM2.5 的一次項」的模型有更低的 RMSE，也就是有更好的預測表現，其原因可能為僅採取 PM2.5 作為 feature 較不能更全面的了解環境因子對未來 PM2.5 走勢的影響，故採取全部污染源作為 feature 在此情形下對於預測來說才有更佳表現。

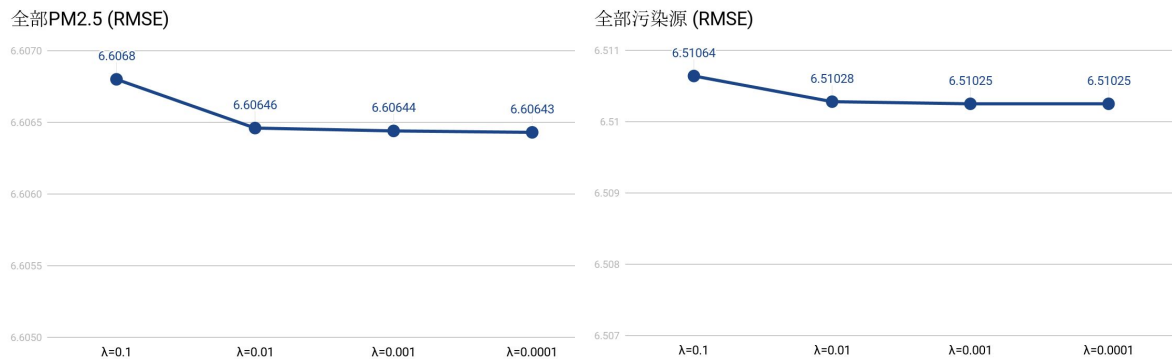
2. (1%)將feature從抽前9小時改成抽前5小時，討論其變化

RMSE	全部污染源 feature 的一次項	全部 PM2.5 的一次項
抽前 9 小時	6.51025	6.60643
抽前 5 小時	6.61879	6.75858

由上表的結果來看，不論是以「全部污染源」或「僅 PM2.5」作為 feature，抽前 5 小時都較抽前 9 小時有稍差的表現，其原因可能為僅抽取前 5 小時的 feature 比起抽取前 9 小時較不能獲得更完整的連續走勢，進而影響預測的準確度。

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

RMSE	全部污染源 feature 的一次項	全部 PM2.5 的一次項
$\lambda=0.1$	6.51064	6.60680
$\lambda=0.01$	6.51028	6.60646
$\lambda=0.001$	6.51025	6.60644
$\lambda=0.0001$	6.51025	6.60643



\* 註：第 1, 2, 3 小題所用來測試的 code 皆有使用 adagrad，且資料有進行過標準化。

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)  $(X^T X)^{-1} X^T y \leftarrow \text{answer}$
- (d)  $(X^T X)^{-2} X^T y$

算式：  $w = [w^1 \ w^2 \ \dots \ w^M]^T$

$$\begin{aligned}
 \text{Loss} &= (y - X \cdot w)^T (y - X \cdot w) \\
 &= (y^T - w^T X^T)(y - X \cdot w) \\
 &= y^T y - y^T X w - w^T X^T y + w^T X^T X w \\
 &= y^T y - 2w^T X^T y + w^T X^T X w
 \end{aligned}$$

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

$$\Rightarrow X^T X w = X^T y$$

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