

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)

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

我取用learning_rate = 10，迭代次數 = 10000 搭配 adagrad 算法的model

當取用全部污染物的feature => RMSE = 6.76048

當取用pm2.5作為feature => RMSE = 6.58502

我們可以發現只取用PM2.5作為feature似乎比取用所有污染物作為feature的誤差值還要好一些，原因可能是18種污染物裡面絕大部分可能都跟PM2.5的預測是沒有太大關係的數據，這些數據的干擾會造成預測上的誤差產生，而PM2.5本身當然會是跟預測PM2.5有高度關係的數據，因此只取它本身作為feature造成的誤差也比較小。

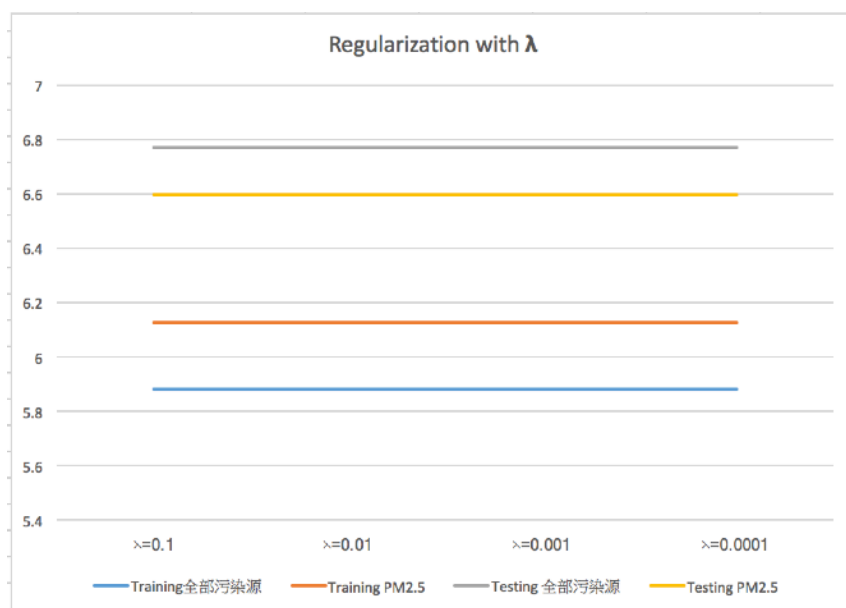
2. (1%)將feature從抽前9小時改成抽前5小時，討論其變化

把全部污染物的feature只取前5小時 => RMSE = 6.67114

只取用PM2.5作為feature抽取前5小時 => RMSE = 6.74073

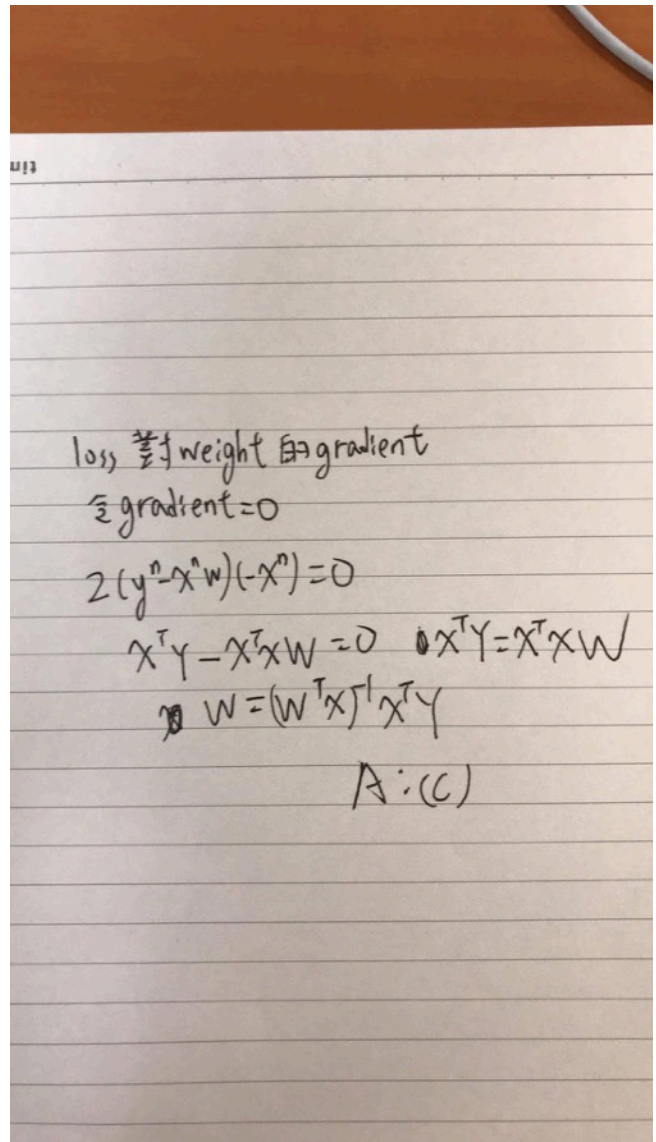
由結果我們可以發現，全部污染物取前五小時作為feature的model，由於減少了將近一半的不太相關的污染物的數據，減少了干擾之後加上參數量也減少，不需要那麼多的迭代次數，導致他的誤差值有所下降。另一方面，只用PM2.5作為feature的model 在改取用前五小時之後，卻因為高度正相關的feature數目減少而導致了誤差值的升高。

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 $\sum_{n=1}^N (y^n - x^n w)^2$ 。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)

1. $(X^T X) X^T y$
2. $(X^T X)^{-1} X^T y$
3. $(X^T X)^{-1} X^T y$
4. $(X^T X)^{-2} X^T y$



Handwritten derivation on lined paper:

$$\begin{aligned} &\text{loss 對 weight 的 gradient} \\ &\hat{=} \text{gradient} = 0 \\ &2(y^n - x^n w)(-x^n) = 0 \\ &x^T y - x^T x w = 0 \quad \Rightarrow x^T y = x^T x w \\ &w = (x^T x)^{-1} x^T y \\ &A: (C) \end{aligned}$$