

學號：R06921066 系級：電機碩一 姓名：劉宇閔

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	Public + private
全考慮	7.59076	5.52561	6.63897
只有 PM2.5	7.84189	6.29361	7.11001

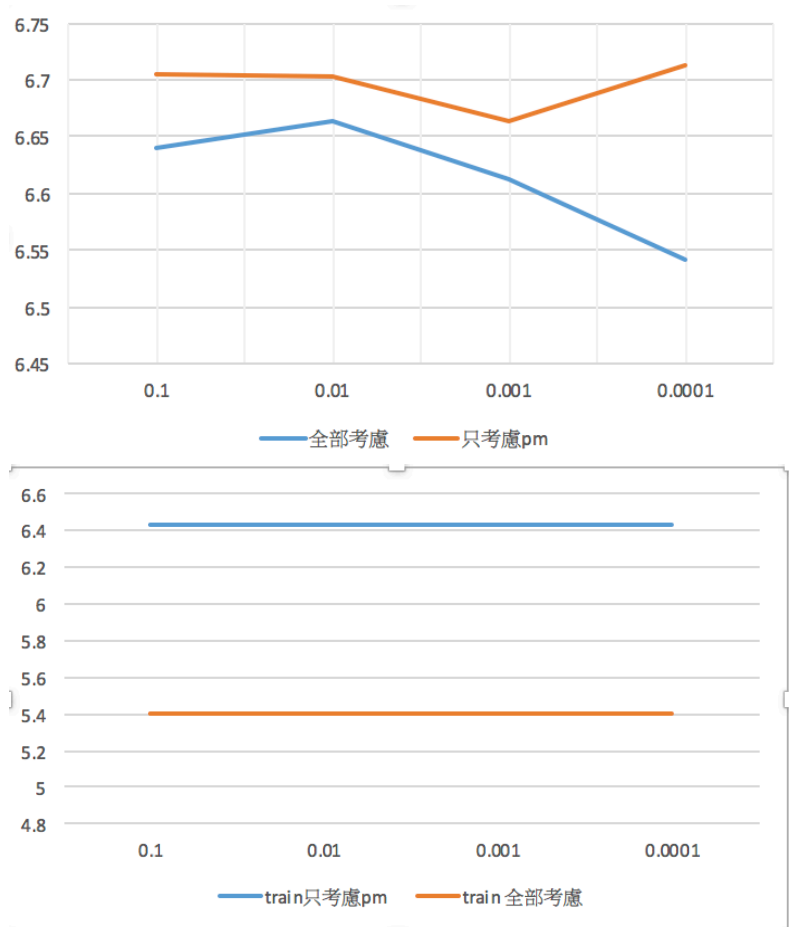
若只單考慮一種污染物的影響在整體結果來說都是較差的，有可能是因為忽略了某些預測時重要的參數

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

	public	private	Public + private
全考慮	9.24716	6.73486	8.0891
只有 PM2.5	9.73083	6.48558	8.2689

當考慮的 feature 數較少時，此時產生誤差較大的原因可能為 P M2.5 不單單只受到這 5 個小時的影響，所以當忽略了其他四小時資訊時產生了相當大的影響

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



第一張圖為 kaggle 上的直計算出的 RMSE 而第二張則是 train 的時候 validation 的誤差

6.426798	6.426351	6.42631	6.42631
5.403014	5.402591	5.402551	5.402541

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)  $(X^T X)^{-1} X^T y$
- (c)  $(X^T X)^{-1} X^T y$
- (d)  $(X^T X)^{-2} X^T y$

$y = Xw$  為無解,其中  $w'$  為最佳近似解

故  $y - Xw'$  屬於  $N(X^T)$ , 可得  $X^T (y - Xw') = 0$

亦等於  $X^T y = X^T X w'$

可知  $w' = (X^T X)^{-1} X^T y$  得證，故答案為 c