學號：R05546014　系級：　工工碩二　姓名：洪紹綺

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

參數設定:Learning Rate = 10000;Iteration = 100000

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

Feature	1.　18 項	2.　PM2.5
Public+Private	7.46632+5.30103	7.44013+5.62719
Total	12.76735	13.06732

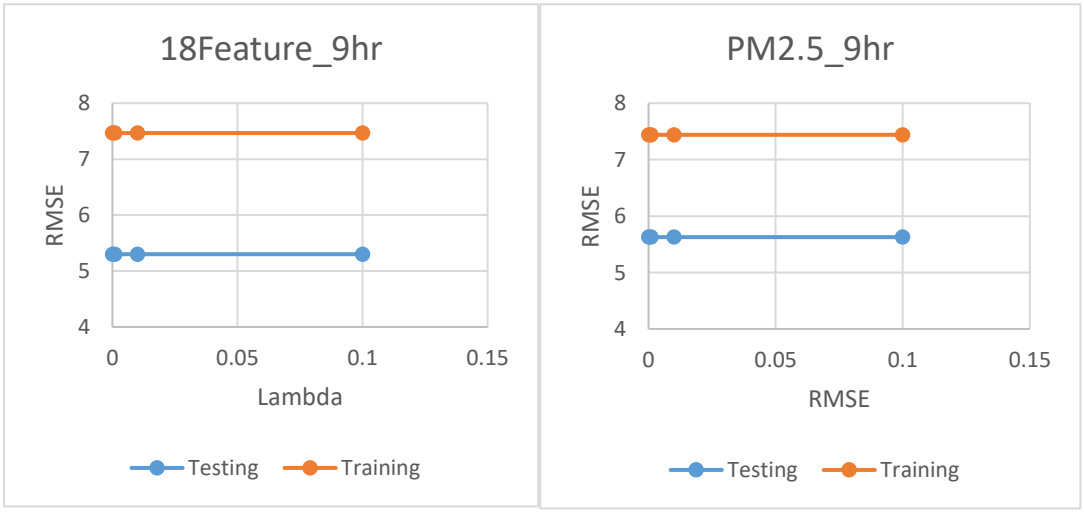
從兩種比較可推測 18 項 Feature 裡也有會影響 PM2.5 的指標，考慮較完整的影響因素，因此誤差較低。

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

Feature	1.　18 項	2.　PM2.5
Public+Private	7.66477+5.32990	7.57904+5.79187
Total	12.99467	13.37091

從此結果可推測，連續時間內檢測到的資料，極有可能與前一分鐘或甚至數個小時前的環境因子有關，因此考慮較多過往資料的模型誤差較低。之後建模可以參考時間序列模型。

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



	ALL of 18 Features		PM2.5	
Lambda	Testing	Training	Testing	Training
0.1	5.30103	7.46632	5.62721	7.44011
0.01	5.30103	7.46632	5.62719	7.44013
0.001	5.30103	7.46632	5.62719	7.44013
0.0001	5.30103	7.46632	5.62719	7.44013

從圖中可以看出誤差降低有限，將 w 的 closed form 打開可以看到 w 數值很小，推測是由於 w 及 Lambda 數值小，因此下降程度有限。

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 n=1Nyn-xnw2 。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^TX 為 invertible)

c. $(X^TX)^{-1}X^Ty$

Proof:

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

$$= y^{n'}y^n - y^{n'}x^nw - w'x^{n'}y^n + w'x^{n'}x^nw$$

$$= y^{n'}y^n - 2w'x^{n'}y^n + w'x^{n'}x^nw \dots (a)$$

(Since $y^{n'}x^nw = w'x^{n'}y^n$)

$$let\ \frac{\partial(\mathrm{a})}{\partial \mathrm{w}} = -2x^{n'}y^n + 2x^{n'}x^nw = 0$$

$$=> \left(x^{n'}x^n\right)\mathrm{w} = x^{n'}y^n$$

$$=> \mathrm{w} = \left(x^{n'}x^n\right)^{-1}x^{n'}y^n$$

As proved#