學號：R04921040 系級： 電機碩一 姓名：李中原

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

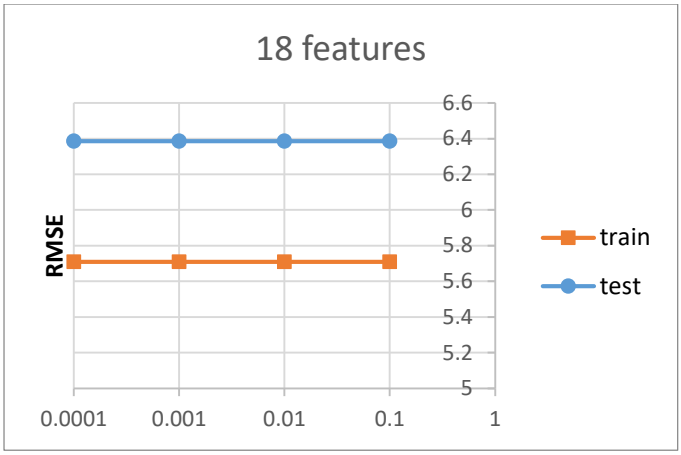
18 項前 9 小時 train 50000 次後，RMSE = (7.48248 + 5.28983)/2 = 6.386155
只取 PM2.5，前 9 小時 train 50000 次後，RMSE = (7.44013 + 5.62719)/2= 6.53366
兩種方法做出來的結果並沒有太大的差距，不過在 train 時收斂的速度就差很多，猜測是因為每個的 PM2.5 本身就已經包含了前面時間其他 feature 所影響的因素在，所以直接透過 PM2.5 做預測的結果跟全部 feature 相比較沒差

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

18 項前 5 小時 train 50000 次後，RMSE = (7.66521 + 5.32875)/2 = 6.49698
只取 PM2.5，前 5 小時 train 50000 次後，RMSE =(7.57904 + 5.79187)/2 =6.685455
前 9 小時與前 5 小時相比，RMSE 算出來結果比較有差，但是參數少了很多所以 train 的時候收斂的速度相比起來取 5 小時的會快許多，至於取 18 項 feature 跟只取 PM2.5 比起來沒有很顯著的差距

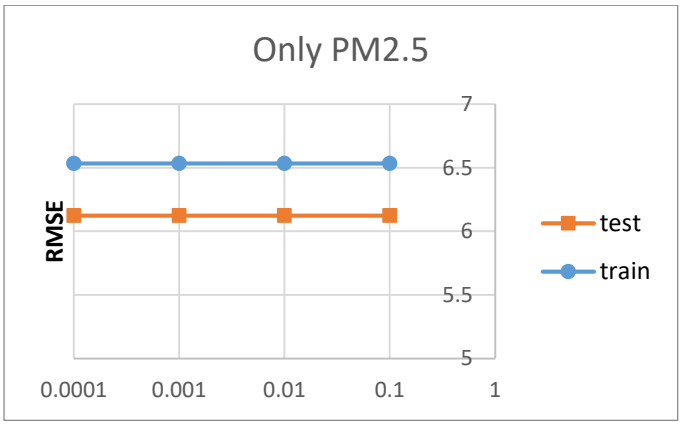
3. (1%)Regularization on all the weight with λ=0.1、0.01、0.001、0.0001，並作圖
18 features

lambda	train	test
0.1	5.70947	6.386155
0.01	5.709464	6.386155
0.001	5.709464	6.386155
0.0001	5.709464	6.386155



Only PM2.5

lambda	train	test
0.1	6.123029	6.53366
0.01	6.123022	6.53366
0.001	6.123022	6.53366
0.0001	6.123022	6.53366



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 $X = [x^1\ x^2 \cdots x^N]^T$ 表示，所有訓練資料的標註以向量 $y = [y^1\ y^2 \cdots y^N]^T$ 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)

(a) $(X^TX)X^Ty$

(b) $(X^TX)^{-0}X^Ty$

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

(d) $(X^TX)^{-2}X^Ty$

$$Loss = (y - Xb)^T(y - Xb)$$

$$0 = \frac{d\ Loss}{dw} = \frac{d}{dw}(y^Ty - w^TX^Ty - y^TXw + w^TX^TXw)|_{w=\hat{w}} = -2X^Ty + 2X^TX\hat{w}$$

$$\hat{w} = (X^TX)^{-1}X^Ty$$