學號：R06921037 系級： 電機所碩一 姓名：許哲瑋

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**

**Ans:**

| | Public | Private | All |
|---|---|---|---|
| With all features: | 7.46472 | 5.32300 | 6.482923 |
| With only pm2.5: | 7.44013 | 5.62719 | 6.596245 |

兩者都是 training 100000 次，可以發現兩者在 public 上面的表現相近，但是在 private 上面只有 pm2.5 的表現較差，可能是因為資訊量不足的原因，導致在預測結果上面稍微有落差。考慮所有 testing data 算出來的平均，只有考慮一項資訊的預測(單變數)比考慮多項資訊(多變數)的預測稍差，但是考慮太多不必要(不相干)的資訊也可能降低預測準確率。

另外，在 training 只有 pm2.5 的 model 時也比較快收斂。

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

**Ans:**

All feature: 6.689942

Only pm2.5: 6.673713

兩者都是 training 100000 次。

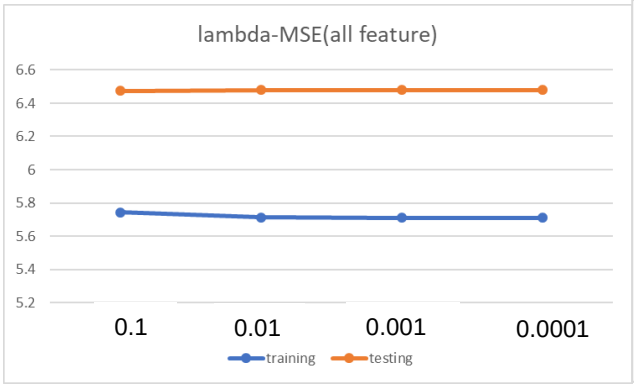
利用全部 240 筆 testing data 測出來，發現 all feature 和只有 pm2.5 的 model 的表現相近。

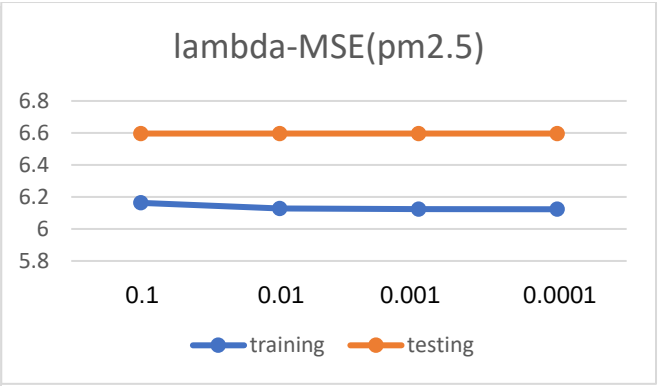
與抽前九小時的 model 互相比較，可以發現預測的誤差稍微變大，可能是因為只有前五小時的資訊，資訊量會稍嫌不足夠。

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

**Ans:**

| | **With all features** | | **Only pm2.5** | |
|---|---|---|---|---|
| | **Training = 50000** | **Testing** | **Training = 50000** | **Testing** |
| **λ=0.1** | **5.742831** | **6.474839** | **6.163088** | **6.596245** |
| **λ=0.01** | **5.712809** | **6.479311** | **6.127040** | **6.596245** |
| **λ=0.001** | **5.709799** | **6.479582** | **6.123423** | **6.596245** |
| **λ=0.0001** | **5.709497** | **6.479582** | **6.123062** | **6.596245** |




**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 $X^TX$ 為 invertible)**

**(a) $(X^TX)X^Ty$**
**(b) $(X^TX)^{-0}X^Ty$**
**(c) $(X^TX)^{-1}X^Ty$**
**(d) $(X^TX)^{-2}X^Ty$**

**Ans: (c)**

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

$$\frac{\partial L}{\partial w_j} = -2 \sum_{n=1}^{N} (y^n - w_0 \cdot x_0^n - w_1 \cdot x_1^n - \dots - w_d \cdot x_d^n) \cdot x_{n,j} = 0 \ = -2X^T (y - X \cdot w)$$

$$X^T y = X^T X w$$

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