

學號：R06521608 系級：土木碩一 姓名：陳德元

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

- (1)全部 9 小時內的污染源 **feature** 的一次項(加 **bias**)

Public : 7.46237

Private: 5.53562

相加 = 12.99799

- (2)全部 9 小時內 **pm2.5** 的一次項當作 **feature**(加 **bias**)

Public : 7.44013

Private: 5.62719

相加 = 13.06732

由上面 Public 的結果可看見，並非所有的數據皆倒入去預測會得到較好的結果，而是將真正有相關的當作 **feature** 再去預測效果較好！

但是在 Private 的結果出來後，好像又不是這麼絕對，全部的 **feature** 都加入好像效果還是比只有 **PM2.5** 的好一點點，不過這也代表其中還是有些是無助於預測的 **feature**。

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

- (1)全部 5 小時內的污染源 **feature** 的一次項(加 **bias**)

Public : 7.92778

Private: 6.33048

- (2)全部 5 小時內 **pm2.5** 的一次項當作 **feature**(加 **bias**)

Public : 7.57904

Private: 5.79187

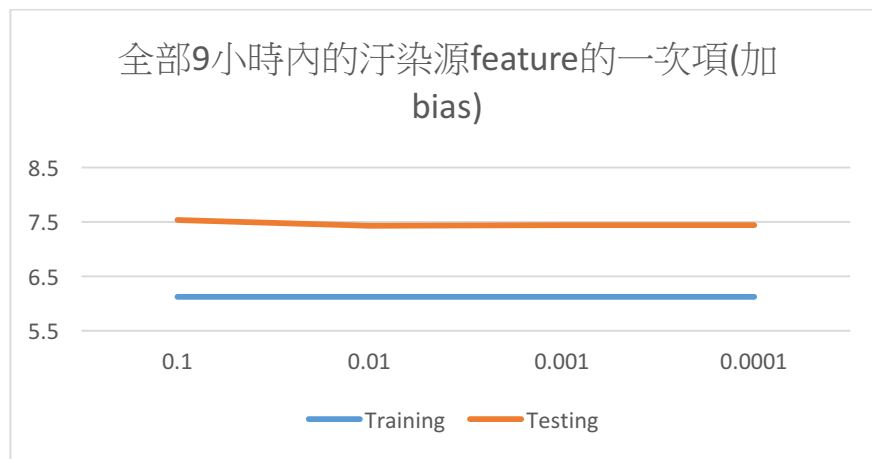
改成抽前 5 個小時後，誤差的值都升高了，這就代表著只有 5 個小時的 **feature** 不夠多，如果想要準確一點，有更多的資料會比較有機會達到。

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

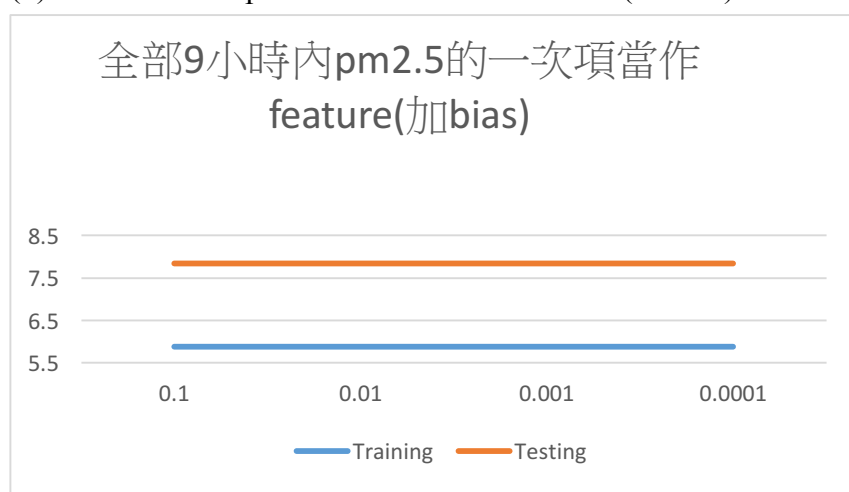
- (1)全部 9 小時內的污染源 **feature** 的一次項 **bias**)

Lambda	Training	Testing
0.1	6.123711	7.53416
0.01	6.123028	7.43181
0.001	6.123022	7.44013
0.0001	6.123022	7.44004

(加



(2)全部 9 小時內 pm2.5 的一次項當作 feature(加 bias)



Lambda	Training	Testing
0.1	5.878284	7.83378
0.01	5.878284	7.83378
0.001	5.878284	7.83378
0.0001	5.878284	7.83378

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 \mathbf{x}^n ，其標註(label)為一存量 \mathbf{y}^n ，模型參數為一向量 \mathbf{w} (此處忽略偏權值 b)，則線性回歸的損失函數(loss function)為 $\sum^N (\mathbf{y}^n - \mathbf{x}^n \cdot \mathbf{w})^2$ 。若將所有訓練資料的特徵值以矩陣 $\mathbf{X} = [\mathbf{x}^1 \mathbf{x}^2 \dots \mathbf{x}^N]^T$ 表示，所有訓練資料的標註以向量 $\mathbf{y} = [\mathbf{y}^1 \mathbf{y}^2 \dots \mathbf{y}^N]^T$ 表示，請問如何以 \mathbf{X} 和 \mathbf{y} 表示可以最小化損失函數的向量 \mathbf{w} ? 請寫下算式並選出正確答案。(其中 $\mathbf{X}^T \mathbf{X}$ 為 invertible)

Ans : (c) $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$

$$L = \sum^N (\mathbf{y}^n - \mathbf{x}^n \cdot \mathbf{w})^2$$

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

$$= \mathbf{y}^T \mathbf{y} - \mathbf{y}^T (\mathbf{X}\mathbf{w}) - (\mathbf{X}\mathbf{w})^T \mathbf{y} + (\mathbf{X}\mathbf{w})^T (\mathbf{X}\mathbf{w})$$

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{y}^T \mathbf{X} - \mathbf{y}^T (\mathbf{X}\mathbf{w}) - (\mathbf{X}\mathbf{w})^T \mathbf{y} + 2\mathbf{X}^T (\mathbf{X}\mathbf{w}) = 0$$

$$2\mathbf{X}^T (\mathbf{X}\mathbf{w}) = 2\mathbf{X}^T \mathbf{y}$$

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