

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

train 300000 次/ learning rate = 10 (有使用 **adagrad**)

- All feature 9 hours => Public: 8.42 / Private: 5.71 / RMSE: 7.19
- Only PM2.5 9 hours => Public: 7.53 / Private: 5.49 / RMSE: 6.58

發現使用較少 **feature** 會得到較為準確的預測，應該是因為取全部 **feature** 時會取到太多不相干的 **feature**，進而影響了預測結果。

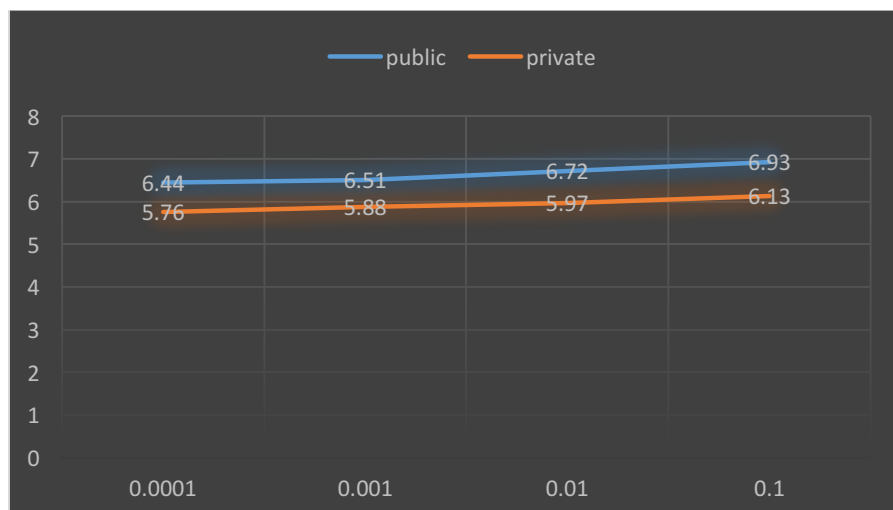
2. (1%)將 **feature** 從抽前 9 小時改成抽前 5 小時，討論其變化

train 300000 次/ learning rate = 10 (有使用 **adagrad**)

- All feature 5 hours => Public: 8.35 / Private: 5.66 / RMSE: 7.13
- Only PM2.5 5 hours => Public: 7.64 / Private: 5.61 / RMSE: 6.70

可以發現 All feature 預測更準了，可能是因為如此一來不相干的 **feature** 影響力變小(因維度變小)，而相反的，只用 PM2.5 的預測卻變差，應該是因為 PM2.5 是很關鍵的一個 **data**，所以歷史以來的數據很重要。

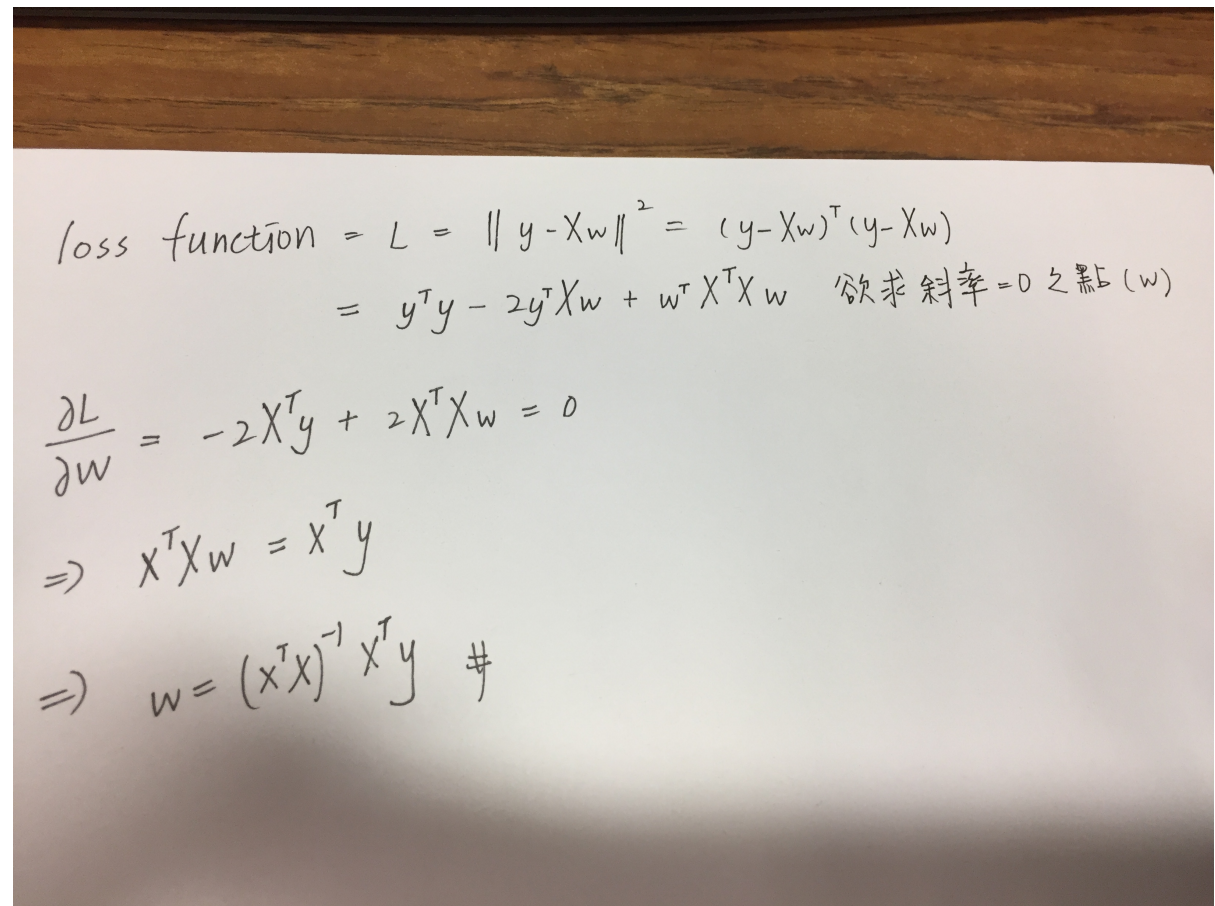
3. (1%)Regularization on all the weight with $\lambda=0.1$ 、0.01、0.001、0.0001，並作圖



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 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 $\sum_{n=1}^N (\hat{y}^n - \mathbf{x}^n \cdot \mathbf{w})^2$ 。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 $\mathbf{y} = [y^1 y^2 \dots y^N]^T$ 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)

- (a) $(\mathbf{X}^T \mathbf{X}) \mathbf{X}^T \mathbf{y}$
- (b) $(\mathbf{X}^T \mathbf{X})^{-0} \mathbf{X}^T \mathbf{y}$
- (c) $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$
- (d) $(\mathbf{X}^T \mathbf{X})^{-2} \mathbf{X}^T \mathbf{y}$

答案：C



Handwritten derivation of the linear regression solution:

$$\begin{aligned} \text{loss function} = L &= \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 = (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) \\ &= \mathbf{y}^T \mathbf{y} - 2\mathbf{y}^T \mathbf{X}\mathbf{w} + \mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} \quad \text{欲求斜率} = 0 \text{ 之點 } (\mathbf{w}) \end{aligned}$$

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

$$\Rightarrow \mathbf{X}^T \mathbf{X} \mathbf{w} = \mathbf{X}^T \mathbf{y}$$

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