學號：R06942082 系級： 電信碩一 姓名：黃釋平

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

Sol:

	public	**private**
取全汙染物	**8.562**	**5.727**
取 **PM2.5**	**7.652**	**5.497**

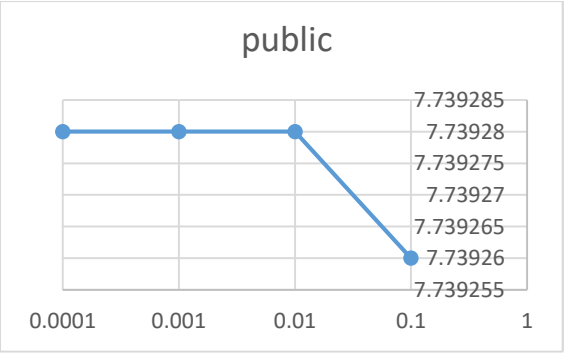
只取 **PM2.5** 的效果明顯比取全部汙染物來的好，可能其中有不少汙染物是與 **PM2.5** 無關，造成 overfitting，而下一小時的 PM2.5 和先前的 PM2.5 值則有很大關係。

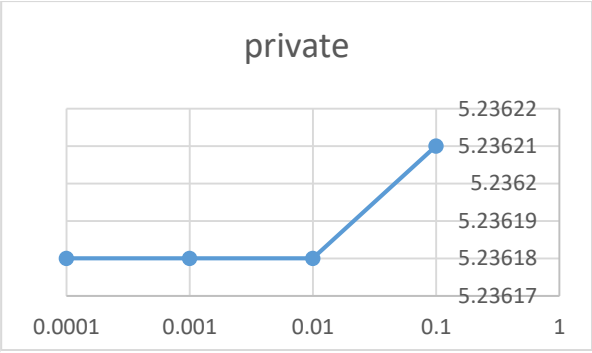
2. (1%)將 feature 從抽前 9 小時改成抽前 5 小時，討論其變化

		public	**private**
取全汙染物	九小時	**8.562**	**5.727**
	五小時	**8.104**	**5.484**
取 **PM2.5**	九小時	**7.652**	**5.497**
	五小時	**7.825**	**5.675**

取全汙染物時，五小時比九小時好，也許因為垃圾資訊較少;而只取 PM2.5 則相反，同上推論，PM2.5 本身資料量越多越好。

3. (1%)Regularization on all the weight with λ=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 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$\sum_{n=1}^{N}(y^n - w \cdot x^n)^2$。若將所有訓練資料的特徵值以矩陣 $X = [x^1\ x^2 \ldots x^N]^T$ 表示，所有訓練資料的標註以向量 $y = [y^1\ y^2 \ldots y^N]^T$ 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) $(X^TX)^{-2}X^Ty$

Sol:

設loss function $E(w) = (y - Xw)^2$

$= (y - Xw)^T(y - Xw)$

$= (y^T - w^TX^T)(y - Xw)$

$= y^Ty - 2y^TXw + w^TX^TXw$

$\min\{E(w)\} \Rightarrow \frac{\partial E(w)}{\partial w} = 0$

$\frac{\partial E(w)}{\partial w} = 2X^TXw - 2X^Ty = 0$

$X^TXw = X^Ty$

因X^TX可逆，兩邊共乘$(X^TX)^{-1}$

得$\mathbf{w} = (\boldsymbol{X^TX})^{-1}\boldsymbol{X^Ty}$