學號：R05229016　系級： 大氣碩二　姓名：羅章碩

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) Rmse = 6.665 (private = 5.5, public = 7.83)

(2) Rmse = 6.5 (private = 5.6, public = 7.4)

個別來看的話，抽全部的汙染源在 private 的結果比較好一點點，而在 public 的情況就會較差一些，但是如果看兩個的平均的話其實不會差很多，因此我覺得用全部九個 feature 的一次項和只用 pm2.5 的情形差不多。

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

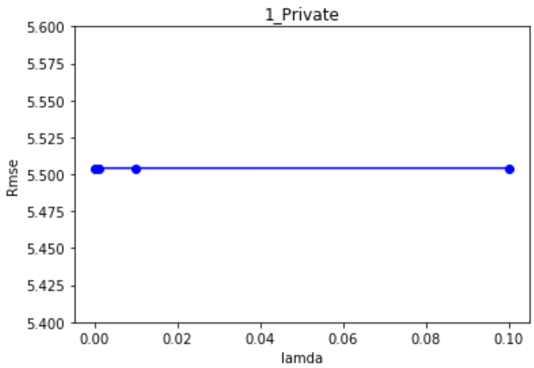
(1) Rmse = 6.56 (private = 5.38, public = 7.74)

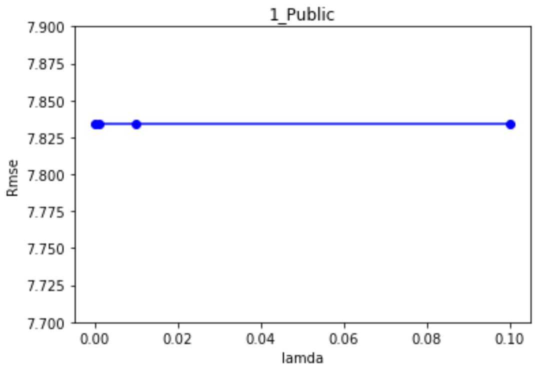
(2) Rmse = 6.685 (private = 5.79, public = 7.58)

從九小時變成五小時可以看到在(1)的結果好像有變好的情形，但是(2)的情形好像就有點變差，所以代表在比較多的 feature 之下，使用較長的時間結果會比較差，因為這段時間內有太多可能的變化，只要有一個變數有比較大的變動就會影響到結果，而如果只有 pm2.5 存在的話，有較長的時間來做預測會更準確些。

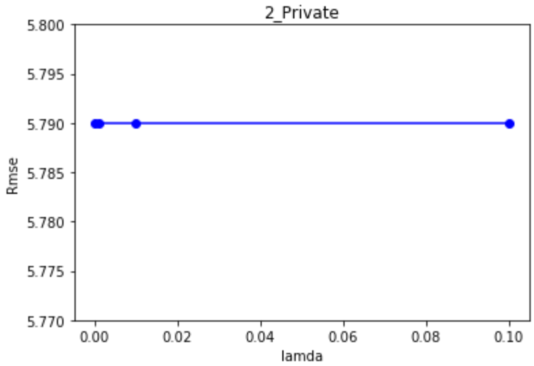
3. (1%)Regularization on all the weight with λ=0.1、0.01、0.001、0.0001，並作圖

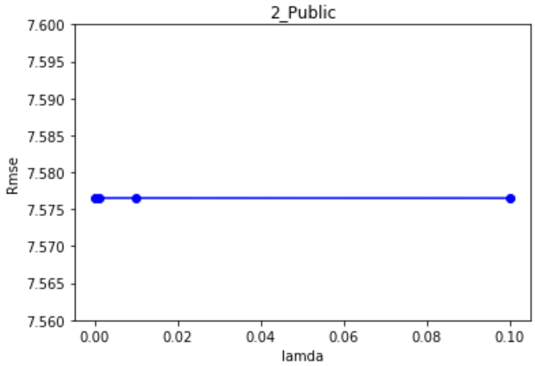
(1) 做了 Regularization 之後發現，用這四個大小的 λ，RMSE 幾乎都沒有差別，代表這個函數可能對於這些資料，不管怎麼樣 train 線條的平滑程度都差不多。





(2)第二組的情況也是一樣的，不管 λ 調多少出來的 Rmse 結果都一樣，跟上述情形一樣。





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)$(X^TX)^{-1}X^Ty$ →正確解答
(d)$(X^TX)^{-2}X^Ty$

假設此方程式為：$h_w(x) = w_0x_0 + w_1x_1 + \cdots + w_nx_n$ ，因為 w 為一向量，所以可以把方程式寫為：$h_w(x) = w^Tx$，而令一大寫 X 為 N 行的矩陣(題目所說)，可將 loss function 改寫成矩陣的型式：$\mathrm{L(w)} = (Xw - y)^T(\mathrm{Xw} - \mathrm{y})$，然後用一些矩陣的運算方法，改寫成 $\mathrm{L(w)} = ((Xw)^T - y^T)(\mathrm{Xw} - \mathrm{y})$，然後乘開得到 $\mathrm{L(w)} = (Xw)^T\mathrm{Xw} - (Xw)^Ty - y^T\mathrm{Xw} + y^Ty$，將第一項展開得：$\mathrm{L(w)} = w^TX^T\mathrm{Xw} - (Xw)^Ty - y^T\mathrm{Xw} + y^Ty$，之後我們知道 Xw 和 y 都是一個向量，在互相乘的時候不管誰乘誰得到的結果都一樣，因此我們知道 L(w)中的第二和第三項的結果會相同，可以將 L(w)寫成：$\mathrm{L(w)} = w^TX^T\mathrm{Xw} - 2(Xw)^Ty + y^Ty$，之後因為要最小化 loss function，所以我們取它的一次微分並等於 0 ($\partial\mathrm{L(w)}/\partial\mathrm{w} = 0$)，會得到：

$$\frac{\partial\mathrm{L(w)}}{\partial\mathrm{w}} = 2X^T\mathrm{Xw} - 2X^Ty = 0$$

然後移項：

$$X^T\mathrm{Xw} = X^Ty$$

最後可得：

$$\mathrm{w} = (X^T\mathrm{X})^{-1}X^Ty$$

(C)選項為正確解答