學號：B04611041 系級： 工海三 姓名：簡暐晉

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

Loss	全部 feature9 小時	Pm2.5 9 小時
Public	7.48225	7.44013
private	5.2898	5.62719
Average $\sqrt{(\text{public}^2 + \text{private}^2)/2}$	6.4794	6.59624

抽了全部的 feature 在 public 的成績較差，但在 private 和總平均誤差都比較好，有可能是其他 feature 也含有其他會影響 pm2.5 的因子，因此做出來的結果較好。

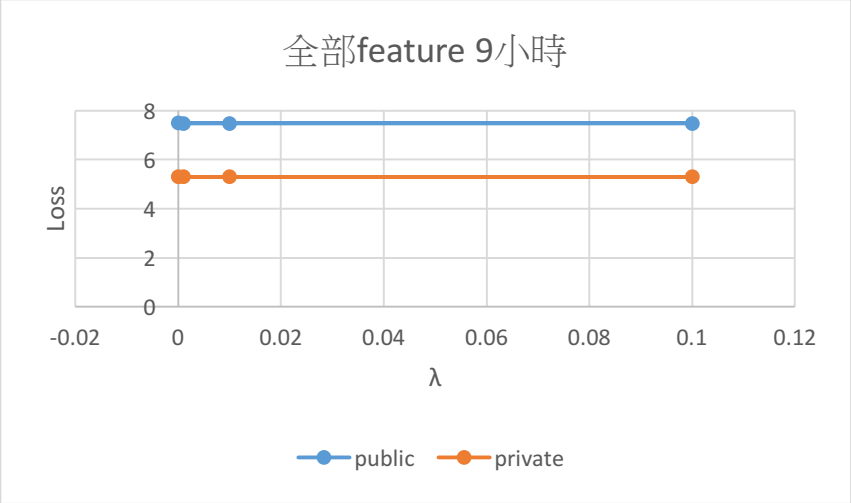
2. (1%)將 feature 從抽前 9 小時改成抽前 5 小時，討論其變化

Loss	全部 feature 5 小時	Pm2.5 5 小時
public	7.66487	7.57904
private	5.32828	5.79187
Average	6.60078	6.74490

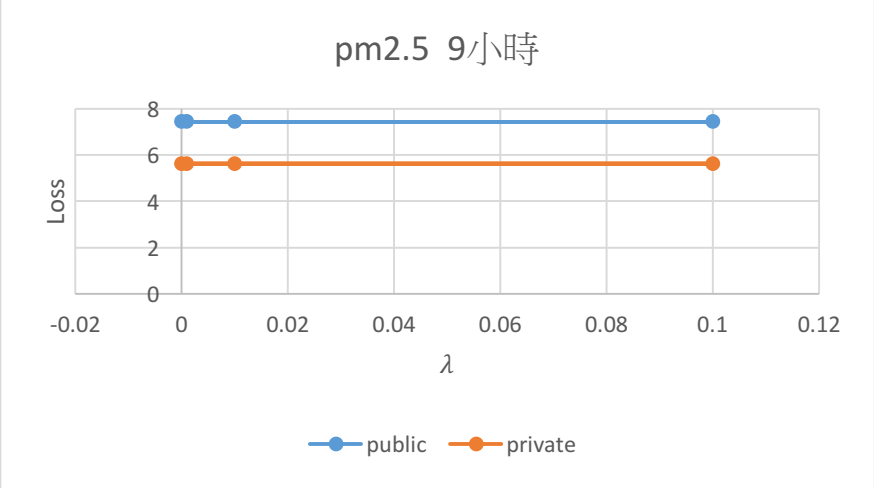
如果只抽前 5 小時的 feature 兩種結果都變差，表示 6~9 小時前的數據對 pm2.5 還是有一定程度的影響，因此取較少的 feature 結果會比較差，並且取全部的 feature 相較於只取 pm2.5 還是會得到比較好的結果。

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

全部 feature 9 小時				
λ	0	0.1	0.01	0.001
public	7.47125	7.47125	7.47125	7.47125
private	5.29283	5.29283	5.29283	5.29283



Pm2.5				
λ	0	0.1	0.01	0.001
public	7.44013	7.44013	7.44013	7.44013
private	5.62719	5.62719	5.62719	5.62719



由於 regularization 是在原來的 loss function 再加上$\lambda\sum(w_i)^2$，但因為我訓練出來的 w 都非常小(0.01 以下)，再平方相加乘上λ就變的非常小，因此對結果幾乎沒有影響。

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 **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 $\mathbf{X} = [\mathbf{x}^1\ x^2 \dots x^N]^T$ 表示，所有訓練資料的標註以向量 $\mathbf{y} = [\mathbf{y}^1\ y^2 \dots y^N]^T$ 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 $\mathbf{X}^T\mathbf{X}$ 為 **invertible**)

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

因為預測的結果是$X \cdot w$，要找的是w使loss function $L = (Y - X \cdot w)^T(Y - Xw)$為最小，而 L 對 w 做微分的值為 $2 \cdot (y - Xw) \cdot (-X)$，令他為零再移項得到答案為 (c) $w=(X^TX)^{-1}X^Ty$