學號：B04901138 系級： 電機三 姓名：張景程

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

9hr	public	private	sum
全部 feature	7.46239	5.53391	12.99630
Only pm2.5	7.44013	5.62719	13.06732

以全部 18 種 feature 來做 training，預測出來的結果較為準確，但可能是因為全部的 feature 中有參雜了部分與 pm2.5 較無關的資訊，所以兩者的相差其實並不大。此外，在 train.csv 中，pm2.5 含有一些 invalid 的值(-1)，若將這些資訊排除或做適當的處理後，預測結果應會更準確。

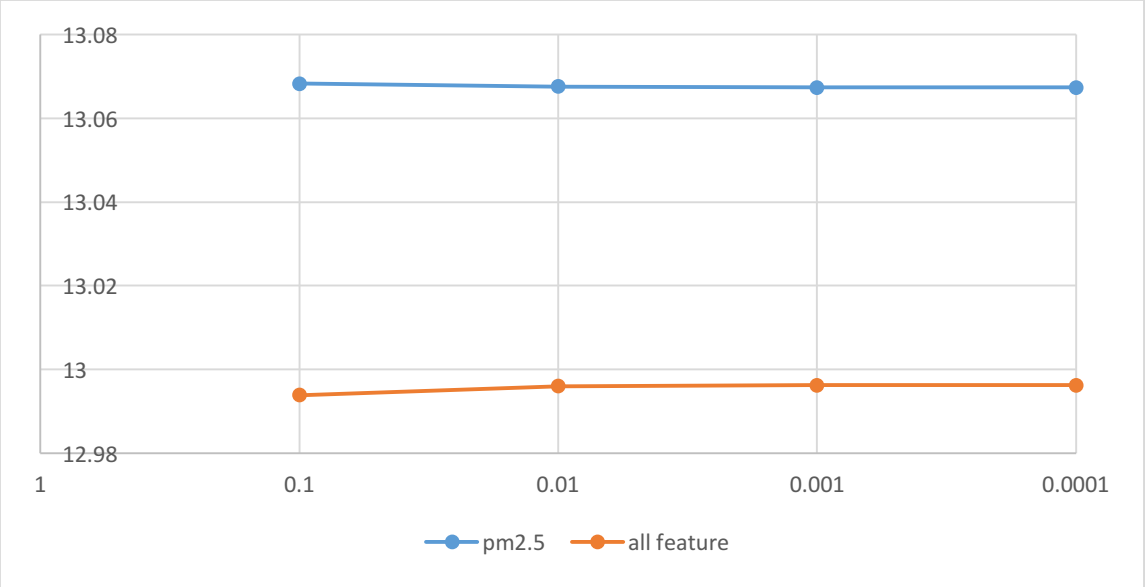
2. (1%)將 feature 從抽前 9 小時改成抽前 5 小時，討論其變化

5hr	public	private	sum
全部 feature	7.65925	5.44091	13.10016
Only pm2.5	7.57904	5.79187	13.37091

整體而言抽前 9 小時比只抽前 5 小時的預測結果要來得好，以直覺來說這樣的結果也是蠻合理的，因為 9 小時用的資料更多，特別是在只用 pm2.5 的地方，從 9 小時改成 5 小時，維度只剩下 5，所以準確度下降得更為明顯。

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

	pm2.5			all feature		
lambda	public	private	sum	public	private	sum
0.1	7.44070	5.6275	13.0682	7.46284	5.53109	12.99393
0.01	7.44018	5.62733	13.06751	7.46243	5.53362	12.99605
0.001	7.44013	5.62719	13.06732	7.46239	5.53388	12.99627
0.0001	7.44013	5.62719	13.06732	7.46239	5.53390	12.99629



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

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

答案：(C)

Let $\varepsilon = y - w \cdot X$
Loss function = $\varepsilon^2 = (y - w \cdot X)^2 = (y - w \cdot X)^T(y - w \cdot X)$
為了得到 Loss function 最小值，所以我們找微分=0 的地方，
$\frac{\partial}{\partial w}(y - w \cdot X)^T(y - w \cdot X) = 0$
$\Longrightarrow -2X^T(y - w \cdot X) = 0$
$\Longrightarrow X^Ty = (X^TX)w$
$\Longrightarrow w = (X^TX)^{-1}X^Ty$ (if (X^TX) is invertible)