學號：B04901060 系級： 電機三 姓名：黃文璁

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

9 小時	public	private
全部汙染源	7.46237	5.53562
PM2.5	7.44013	5.62719

9 小時全部汙染源的 public RMSE 比只抽 PM2.5 的 RMSE 來得高（約高 0.02）。

9 小時全部汙染源的 private RMSE 比只抽 PM2.5 的 RMSE 來得低（約低 0.1）。

兩種 feature 就本題來說不容易分出好壞。

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

5 小時	public	private
全部汙染源	7.65925	5.44092
PM2.5	7.57904	5.79187

5 小時全部汙染源的 public RMSE 比只抽 PM2.5 的 RMSE 來得高（約高 0.08）

5 小時全部汙染源的 private RMSE 比只抽 PM2.5 的 RMSE 來得低（約低 0.35）

綜合 public 和 private 分數來看，抽 5 小時全部汙染源的結果較好。

另外值得注意的是，抽 5 小時全部汙染源在 private set 的結果要比抽 9 小時來得好。

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$\sum_{n=1}^{N}(y^n - x^n \cdot w)^2$ 。若將所有訓練資料的特徵值以矩陣 $X = [x^1\ x^2\ \ldots\ x^N]^T$ 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) $(X^TX)^{-2}X^Ty$

(c)

Using matrix calculus:

$L(w, X, y) = \sum_{n=1}^{N}(y^n - x^n \cdot w)^2 = (y - X\mathrm{w})^T(y - X\mathrm{w})$

$\frac{\partial}{\partial w}L(w, x, y) = \frac{\partial}{\partial w}\left[(y - X\mathrm{w})^T(y - X\mathrm{w})\right] = -2(y - X\mathrm{w})^TX = 0$

$(y - X\mathrm{w})^TX = y^TX - (Xw)^TX = y^TX - w^TX^TX = 0$

$w^TX^TX = y^TX$

$(X^TX)^Tw = X^Ty,\ (X^TX)^T = X^TX$

$w = (X^TX)^{-1}X^Ty$