學號：R05943138 系級： 電子所碩二 姓名：賴又誠

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

feature	RMSE(training)	RMSE(public)	RMSE(private)
(1) 9 小時	5.6795	7.46275	5.53423
(2) 9 小時	6.1230	7.44013	5.62719

使用 model(1)在 training 時 RMSE 只有 5.6795，但在 model(2)卻在 training 時 RMSE 有 6.1230，而在 RMSE(public)時 model(2)卻又較好，可能是 model(1)產生 overfitting 造成的結果。

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

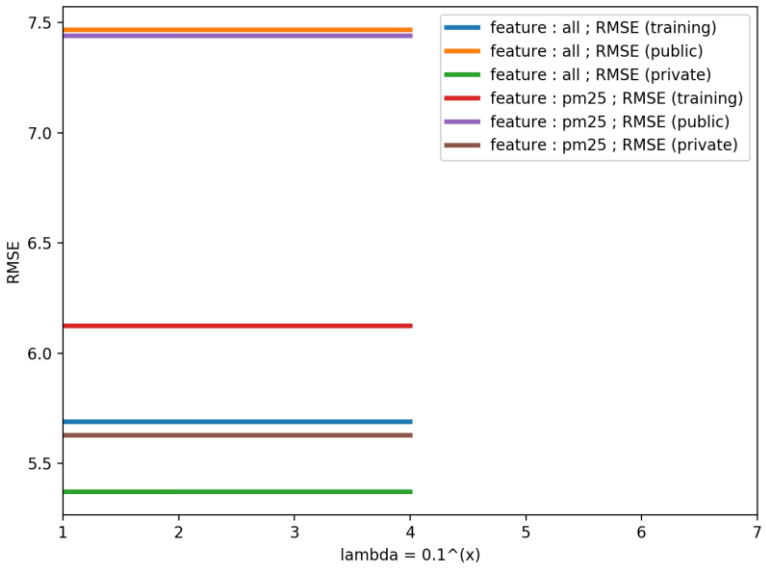
feature	RMSE(training)	RMSE(public)	RMSE(private)
(1) 9 小時	5.6795	7.46275	5.53423
(1) 5 小時	5.8051	7.65098	5.44101
(2) 9 小時	6.1230	7.44013	5.62719
(2) 5 小時	6.2070	7.57904	5.79187

使用 5 小時的 feature 時，不管在(1) or (2)都造成 RMSE(training)、RMSE(public)比使用 9 小時的差，雖然 training data 會變多，可是 feature 數量下降，然而影響較大的 feature 可能在那 4 小時的時間內，所以 train 9 個小時得到的結果不管在哪個 model，都會是比較好的。

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

feature	RMSE(training)	RMSE(public)	RMSE(private)
(1) λ=0.1	5.68865536	7.46711	5.37077
(1) λ=0.01	5.68864692	7.46710	5.37083
(1) λ=0.001	5.68864607	7.46710	5.37084
(1) λ=0.0001	5.68864599	7.46710	5.37084
(2) λ=0.1	6.1230215223407134	7.44012	5.62720
(2) λ=0.01	6.1230215220907871	7.44013	5.62719
(2) λ=0.001	6.1230215220882878	7.44013	5.62719
(2) λ=0.0001	6.1230215220882620	7.44013	5.62719

因為使用了 adagrad 在更新 weight 上較有效率，而且 learning rate 會隨著離目標越近而越慢，而第一份作業是一個 convex 的情況，故在於使用不同的 lambda 下造成的影響不大，或者是因為都使用一次項，還未發生 overfitting 的情況，使得 regularization 不管在 model(1) or (2)對於 Kaggle 上的 public 和 private 分數都沒有太大的影響，也有可能是發生 overfitting 但是 lambda 給不夠大，造成 regularization 效果不好。



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) $(X^TX)^{-2}X^Ty$

Ans : Loss functuin = $(y - Xw)^T(y - Xw)$

找 w 最小值，對 loss function 微分得 :

$$0 = \frac{\partial}{\partial w}(y - Xw)^T(y - Xw) \quad (1)$$

$$0 = \frac{\partial}{\partial w}(y^Ty - w^TX^Ty - y^TXw + w^TX^TXW) \quad (2)$$

$$0 = -X^Ty - X^Ty + 2WX^TX \quad (3)$$

$$X^Ty = WX^TX \quad (4)$$

$$X^T \cdot y = (X^T \cdot X) \cdot w \quad (5)$$

$$(X^TX)^{-1}X^Ty = w \quad (6)$$

故選(c)