學號：R06922097　系級： 資工碩一　姓名：鄭雅文

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

model 1

用 gradient descent + adagrad 重複 50000 次

private: 5.28834　　public: 7.50106　　average: 6.3947

model 2

用 gradient descent + adagrad 重複 4978 次(再重複下去 RMSE 改變不大)

private: 5.64664　　public: 7.47158　　average: 6.55911

由兩種 model 可得知 model 作為預測 kaggle testing set 整體平均較差，可能是因為 feature 較少所以預測較不精準。

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

model 1：用 gradient descent + adagrad 重複 50000 次

private: 7.07302　　public: 7.55768　　average: 7.31535

model 2：用 gradient descent + adagrad 重複 10000 次(因 RMSE 已將近 0.00)

private: 7.07342　　public: 7.55755　　average: 7.315485

可能因為資料量太少，所以上傳到 kaggle 時，表現不如取 9 個小時的 feature 的 model。

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

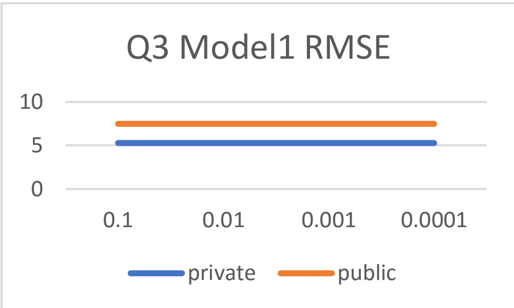
Model 1:

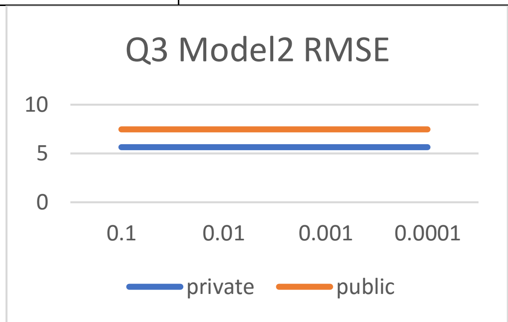
0.1	5.28912	7.49501
0.01	5.28912	7.49501

0.001	5.28912	7.49501
0.0001	5.28912	7.49501

Model 2:

0.1	5.64622	7.47100
0.01	5.64620	7.47101
0.001	5.64622	7.47107
0.0001	5.64622	7.47107





因為 w 數值本身就很小，且 lambda 又讓 w 的影響更小，所以做 gradient descent 幾乎沒有什麼改變。

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。

a. $(X^TX)X^Ty$

b. $(X^TX)^{-0}X^Ty$

c. $(X^TX)^{-1}X^Ty$

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

Answer: c.

最小化 $\|y - X \cdot w\|^2$

$= (y - X \cdot w)^T(y - X \cdot w)$

$= y^Ty - w^TX^Ty - y^TXw + w^TX^TXw$

因內積為純量，所以$w^TX^Ty = y^TXw$

$= y^Ty - 2w^TX^Ty + w^TX^TXw$

對w微分並另微分後的方程式為零以符合 First-order condition

$-2X^Ty + 2(X^TX)w = 0 \quad \therefore w = (X^TX)^{-1}X^Ty$