學號：R06944094 系級： 網媒碩一 姓名：黃敬庭

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

抽全部污染源一次做，利用 adam iterate 100000 次之後

result : 6.56272

只抽 pm2.5，利用 adam iterate 100000 次之後

result : 6.59624

只利用 pm2.5 與利用全部的 feature 所做出來的結果看起來差不多，可見其 feature 對 pm2.5 的預測能力的影響其實遠不如 pm2.5 本身高

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

抽全部污染源一次前五小時做，利用 adam iterate 100000 次之後

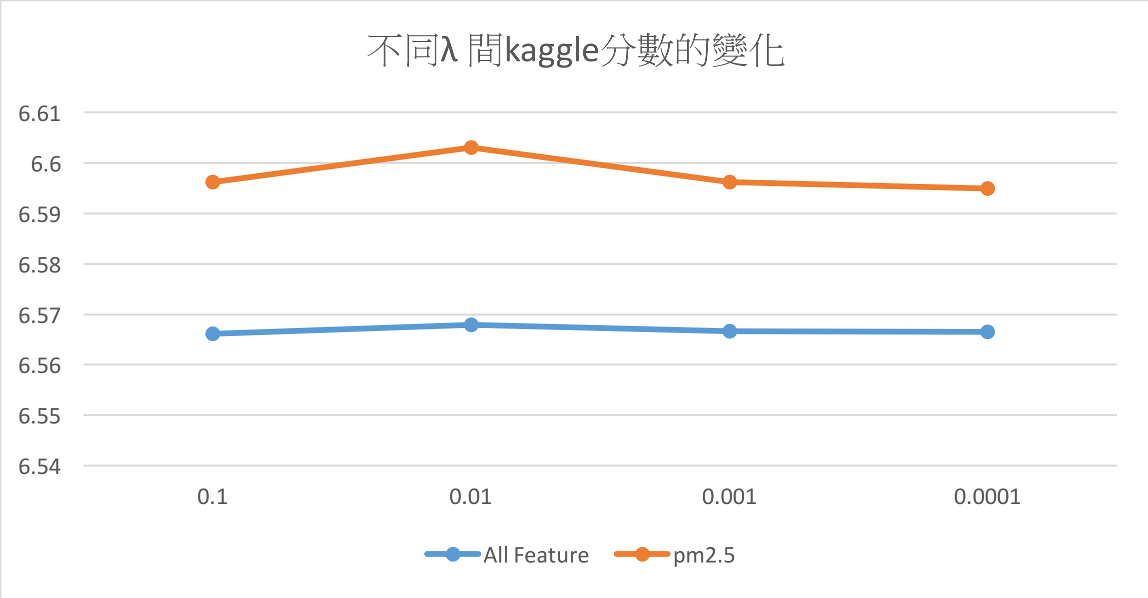
result : 6.64094

只抽 pm2.5 前五小時，利用 adam iterate 100000 次之後

result : 6.66092

雖然整體來講並沒有抽 9 小去的數據表現的這麼好，不過差距也不大，所以我猜想前面 4 小時的數據的影響力可能也並不高

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$\sum_{□=1}^{□}(□^{□} - □^{□}\cdot □)^2$ 。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)

因為 y = XW 不存在解,所以我們要的是一個 quadratic minimization ,他的 obj function 是我們的 loss function,把它寫成 :

$$(y-XW)^T(y-XW)$$

將它對W做微分找為 0 的時候可以使 loss func 最小 :

$$-X^Ty + (X^TX)W = 0$$

運算一下之後得到 :

$$(X^TX)W = X^Ty$$

兩邊同乘$(X^TX)^{-1}$:

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