學號：B06502149 系級：資工二 姓名：張琦琛

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 當作一次項(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

c. 第 1-3 題請都以題目給訂的兩種 model 來回答

d. 同學可以先把 model 訓練好，kaggle 死線之後便可以無限上傳。

e. 根據助教時間的公式表示，(1) 代表 p = 9x18+1 而(2) 代表 p = 9*1+1

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 score	Private score
All feature	5.80226	7.36075
PM2.5 Only	5.9362	7.31609

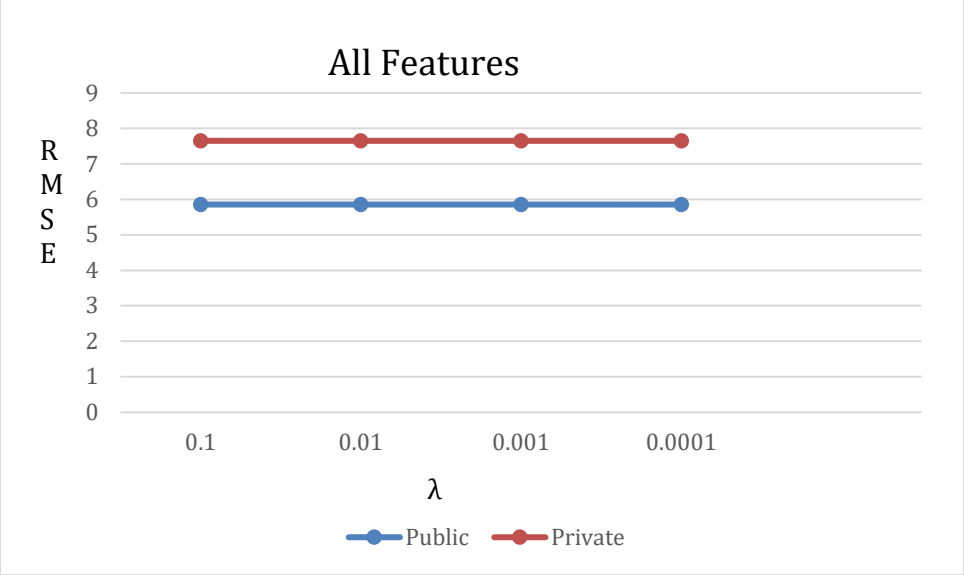
由結果可知在 Public 上抽取所有 feature 的結果較只抽取 PM2.5 較好，可能是所有 feature 的參數較多，考慮較多因素，有更好的預測結果。但是在 Private 上卻比只抽取 PM2.5 差，可能是加入了許多不必要的變數導致 overfitting。

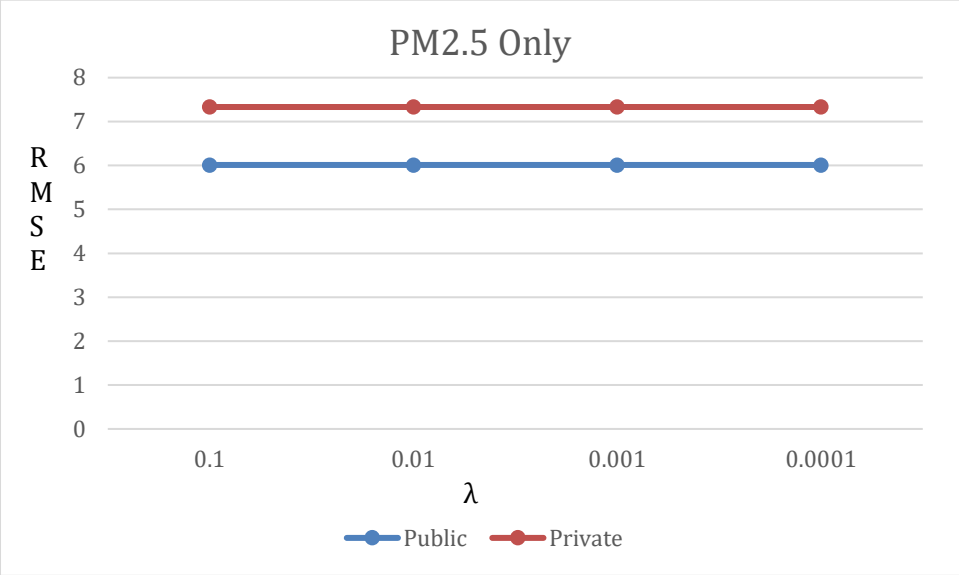
2. (1%)將 feature 從抽前 9 小時改成抽前 5 小時，討論其變化

5 小時	Public score	Private score
All feature	5.77986	7.32266
PM2.5 Only	5.97279	6.42956

若從抽前 9 小時改成抽前 5 小時，PM2.5 的分數變差，可能是 feature 太少導致 underfitting。但是兩者在 Private 上都得到較好的結果，減少參數將 overfitting 的機會下降，較小的維度也更快收斂。

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





若取所有 feature, RMSE 將隨著 lambda 變小而增長，但是沒有顯著變化。若只取 PM2.5，RMSE 則不隨著 lambda 改變而改變，始終為水平線。綜合以上結果，我認為太小的 lambda 值做 regularization，對預測結果並沒有太大幫助。

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 $\mathbf{x}^n$，其標註**(label)**為一純量 $\mathbf{y}^n$，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 $\mathbf{X} = [\mathbf{x}^1\ x^2 \dots x^N]^T$ 表示，所有訓練資料的標註以向量 $\mathbf{y} = [\mathbf{y}^1\ y^2 \dots y^N]^T$ 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** ? 請選出正確答案。(其中 $\mathbf{X}^T\mathbf{X}$ 為 **invertible**)

(a) $(X^TX)X^Ty$
(b) $(X^TX)yX^T$
(c) $(X^TX)^{-1}X^Ty$
(d) $(X^TX)^{-1}yX^T$

ANS: (C)

Loss function = $\sum_{n=1}^{N}(y^n - x^n \cdot w)^2$

$\frac{\partial L}{\partial w} = 2\sum_{n=1}^{N}(y^n - x^n \cdot w)(-x^n) = 0$

$\rightarrow$ $-2X^T(y - Xw) = 0 \rightarrow \ 2X^TXw = 2X^Ty \ \rightarrow w = (X^TX)^{-1}X^Ty$