學號：B03902096 系級： 資工四 姓名：陳柏屹

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

以下數據皆是以初始參數固定下經過 3 次重跑，並使用 SGD 下訓練出的新模型所得到的平均分數，而 iteration 次數為 2,000。

(1) 全部 feature 一次項： 5.81918 (private) + 7.67668(public) = 13.49585

(2) 只取 pm 2.5 一次項： 5.63538 (private) + 7.53421(public) = 13.16959

可以看到在抽取 9 小時的前提下，只取 pm2.5 一次項的成績會些微地比抽取全部 feature 的成績來的要好。我想這是因為全部的 feature 之中存在著許多跟預測 pm2.5 較無關緊要的項，因此將這些項目加入到模型的 training 中反而導致過多額外干擾而造成更多誤差。但其實兩者分數僅有些微差距，很有可能只是因為 testing set 只有 120 筆所造成的偏差，並不一定能夠表示哪個模型一定較適合預測 PM2.5。

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

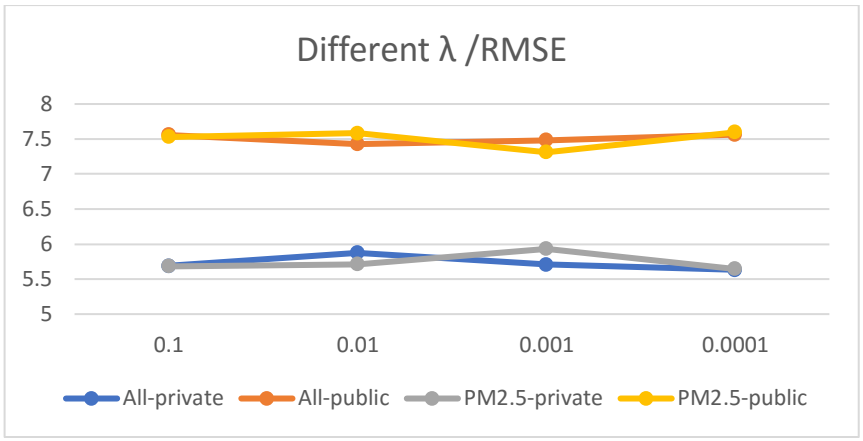
以下數據皆是以初始參數固定下經過 3 次重跑，並使用 SGD 下訓練出的新模型所得到的平均分數，而 iteration 次數為 2,000。

(1) 全部 feature 一次項： 5.60888 (private) + 7.59261(public) = 13.20148

(2) 只取 pm 2.5 一次項： 5.80159 (private) + 7.64643(public) = 13.44802

可以看到在抽取 5 小時的前提下，抽取全部 feature 的成績會些微地比只取 pm2.5 一次項的成績來的要好。我想這是因為單單只有前 5 個小時的 pm2.5 的資訊量太少了，因此這時把全部 feature 都參考進去後，讓其他有用的 feature 諸如 pm10, NO2 等等都得以納入計算。但同樣地和第一題一樣，兩者分數僅有些微差距，並不一定能夠表示哪個模型一定較適合預測 PM2.5。

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



	0.1	0.01	0.001	0.0001
All-private	5.68775	5.87429	5.70559	5.63344
All-public	7.55832	7.42741	7.48199	7.56263
All-Sum	13.24607	13.3017	13.18758	13.19607
PM2.5-private	5.68552	5.71298	5.93108	5.64455
PM2.5-public	7.53259	7.58125	7.31063	7.5952
PM2.5-Sum	13.21811	13.29423	13.24171	13.23975

從圖表中線條們緊密交錯的狀況可以推測 λ 的大小對於這次預測 PM2.5 的模型影響並不大，我想是因為如此簡單的模型比較沒有 Over Fitting 的問題，所以 Regularization 自然也不會有所幫助。

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)為一存量 $\mathbf{y}^n$，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 $\mathbf{X} = [\mathbf{x}^1\ x^2 \ldots x^N]^T$ 表示，所有訓練資料的標註以向量 $\mathbf{y} = [\mathbf{y}^1\ y^2 \ldots y^N]^T$ 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 $\mathbf{X}^T\mathbf{X}$ 為 invertible)

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

$$
\begin{aligned}
Let\ y &= wX \\
w &= X^{-1}y \\
&= X^{-1}((X^T)^{-1}X^T)y \\
&= X^{-1}(X^T)^{-1}X^Ty \\
&= (X^TX)^{-1}X^Ty
\end{aligned}
$$