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	Training	Valid	Public Test	Private Test
M1 (163維)	5.62916	6.17541	7.52838	5.33626
M2 (10維)	6.12302	6.5835	7.35774	5.71235

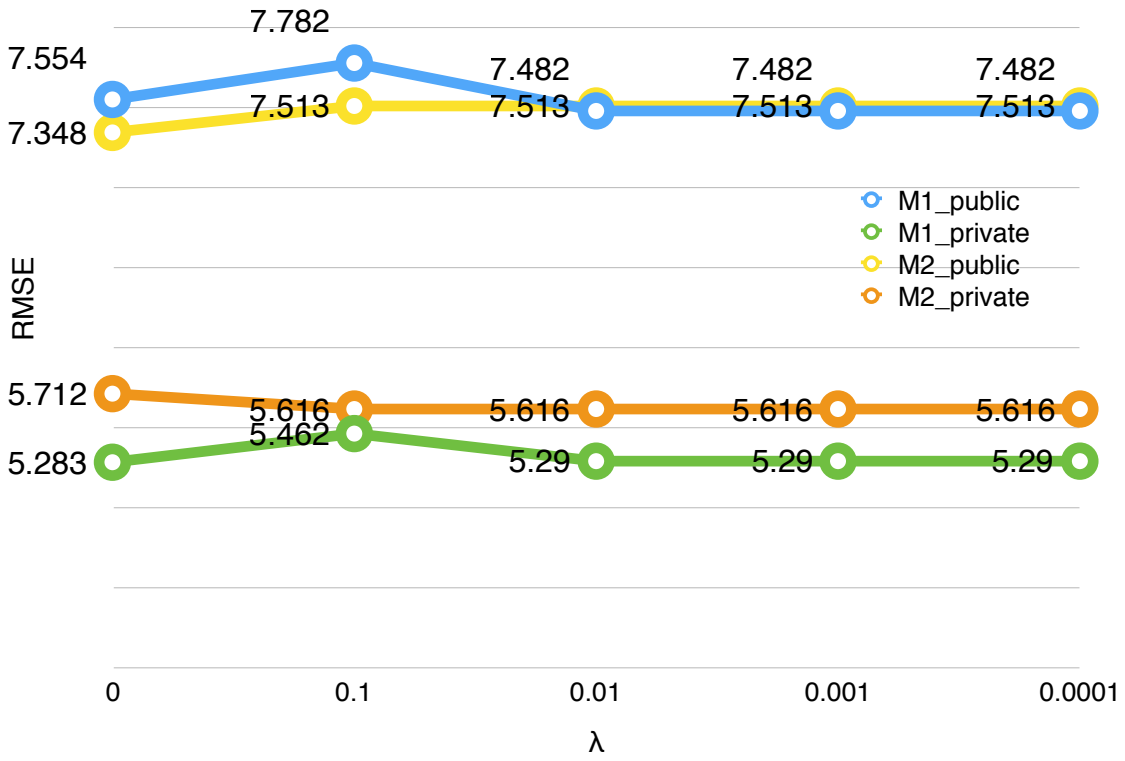
以上皆有使用 adagrad，Iterations 數皆為 5 萬。
Training 跟 Valid Data 分別有 4692 跟 948 筆。
可以發現當 feature 較多時（M1），Training 的 Loss 會較低，因為 M2 其實算是 M1 的子集，較複雜的 model 能比較 fit Training 算是很合理。但在 Public Test 上，M1 反而表現得較差，直覺上會猜想是不是 M1 太過複雜導致了 overfitting？但從 Valid 來看又似乎不是如此，畢竟 M2 在 Valid 也表現較差，甚至從 Private Test 也可以看見 M2 比較差的結果。
我的推測是，兩種 Test 都只有各 120 筆測資太少了，很有可能 Test Data 其實是很 bias 的，這樣未必能從 Test Data 的成績好壞判斷 Model 的優劣了。（事實上，在我自己的嘗試中，如果只切 120 筆當 Valid，切的資料不同能讓 RMSE 從 5 浮動到 8，可見其影響之劇烈）。

2. (1%)將feature從抽前9小時改成抽前5小時，討論其變化。

Model	Training	Valid	Public Test	Private Test
M1 (91維)	5.77981	6.17824	7.49682	5.28841
M2 (6維)	6.1902	6.59914	7.28201	5.8186

改為抽5小時後，最直觀的影響就是Training Data變多了（變為4740筆），但 Training 跟 Valid 都變差了一點，有可能是因為 feature 維度下降，變相降低了 model的 複雜度。
至於 Testing 的成績改變有高有低，推測跟第 1 題狀況類似，即 Test 太少不容易判斷 Model 改變對其造成的影響。

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



可以發現 regularization 對於這次的 work 沒有什麼幫助，甚至可能造成傷害。推測 regularization 的效果有被 adagrad 抵銷、因此影響不大，至於為什麼反而可能變差，除了推測 model 並沒有 overfit 外，也可能跟 Test Data 的性質有關。

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)為一純量 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)

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

Let $y - X \cdot w = 0$,
Then $w = X^{-1}y = X^{-1}((X^T)^{-1}X^T)y = X^{-1}(X^T)^{-1}X^Ty = (X^TX)^{-1}X^Ty$