學號：B05901063 系級： 電機三 姓名：黃世丞

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

RMSE:		private	public
all-feature	:	7.88702	6.46614
pm2.5-only	:	7.24293	5.92501

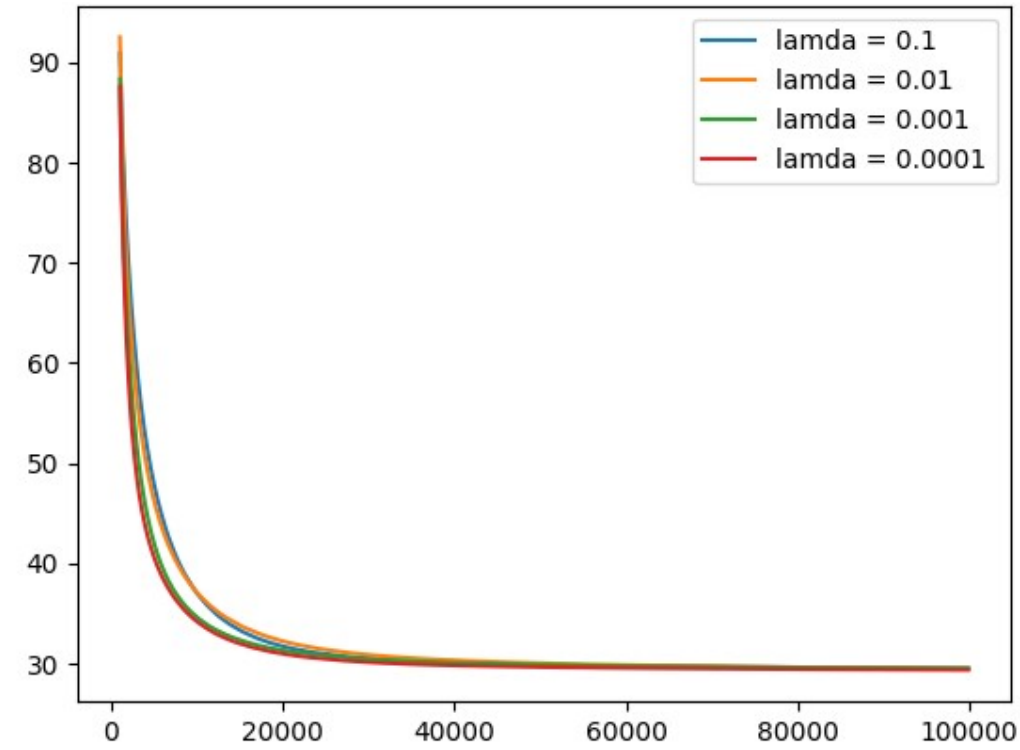
只取一個 feature 的情況下，不管是 public 還是 private 的表現都遠勝取所有的 features。猜測是因為有很多觀測資料其實不乾淨，有許多不合理的負值，影響 model 的正確性，而單靠 pm2.5 就可以達到不錯的預測結果，推測其他 feature 的影響並沒有那麼大，而且拋棄不合理的資料也對預測有幫助。

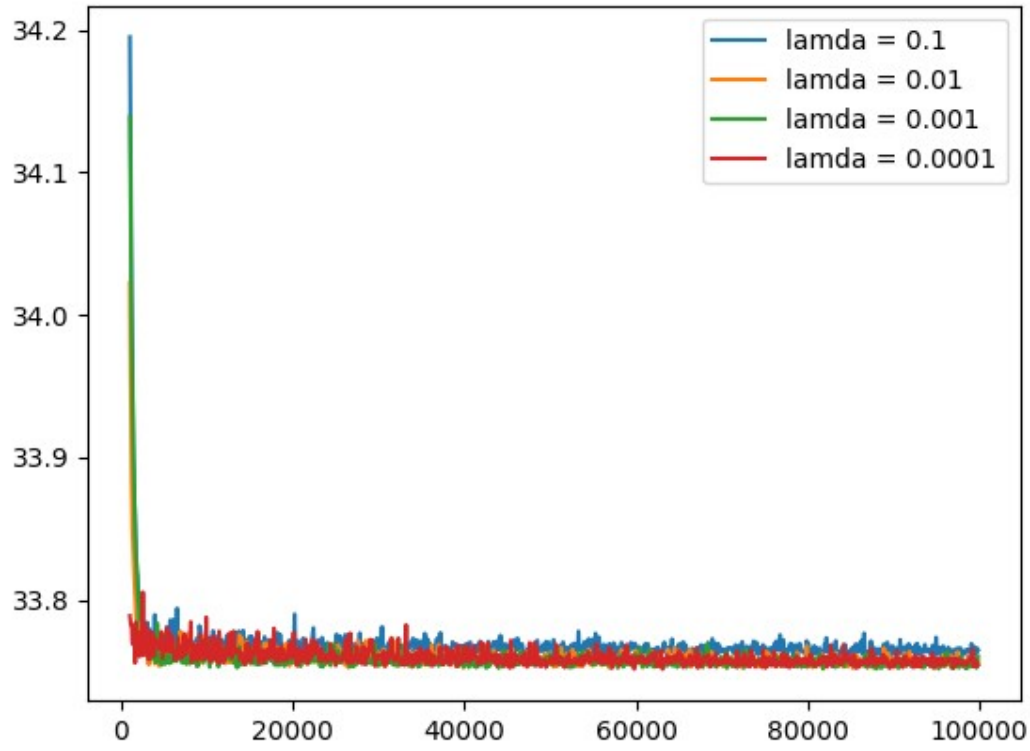
2. (1%)將 feature 從抽前 9 小時改成抽前 5 小時，討論其變化

RMSE(5 hours):		private	public
all-feature	:	7.19548	6.24248
pm2.5-only	:	7.22552	6.22732

取全部 feature 的模型無論是 private 還是 public 都有不小的進步，可能是因為較少 feature 相較之下較容易收斂，可以有效降低不重要的 feature 的權重。
只取 pm2.5 的 private 有些微進步，但 public 卻退步，這個結果就沒什麼特別的趨向，因為這裡 feature 太少，只訓練一下就走到了 local minima，而且 loss 就沒有再改變了。

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


lamda = 0.1
lamda = 0.01
lamda = 0.001
lamda = 0.0001
90
80
70
60
50
40
30
0
20000
40000
60000
80000
100000


lamda = 0.1
lamda = 0.01
lamda = 0.001
lamda = 0.0001
34.2
34.1
34.0
33.9
33.8
0
20000
40000
60000
80000
100000

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$\sum_{n=1}^{N}\left(y^n - x^n \cdot w\right)^2$ 。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？ 請選出正確答案。(其中 X^TX 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)