學號：B04501073 系級： 土木四 姓名：李利元

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

使用 adam 並採全部 9 小時的汙染源作為 feature，得到結果為：7.27922(private)+ 5.77036(public) = 13.04952

只有使用 pm2.5 之變化作為 feature，得到結果為：7.22356(private)+ 5.90263(public) = 13.12619

以結果看來，增加 feature 數量的確可以減少誤差的產生，但兩者的差異並不太明顯，或許也代表 PM2.5 這個 feature 本身對於 y 的真值之間的相關係數非常大，當我進一步求得每一個小時對於 Y 值的相關係數也發現事實確實如此：

小時	相關係數
1	0.434
2	0.469
3	0.506
4	0.548
5	0.595
6	0.65
7	0.704
8	0.817
9	0.914

次高的為 PM10 的相關係數：

小時	相關係數
1	0.396
2	0.433
3	0.47
4	0.509
5	0.552
6	0.604
7	0.659
8	0.714
9	0.757

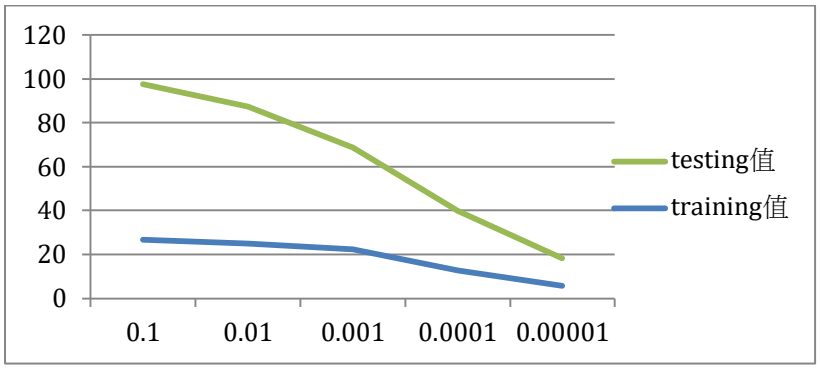
其餘 feature 的相關係數皆介於 0.01 到 0.4 之間，從相關係數的角度而言，的確可以透過 PM2.5 這個 feature 對於 Y 值做到一定的預測。

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

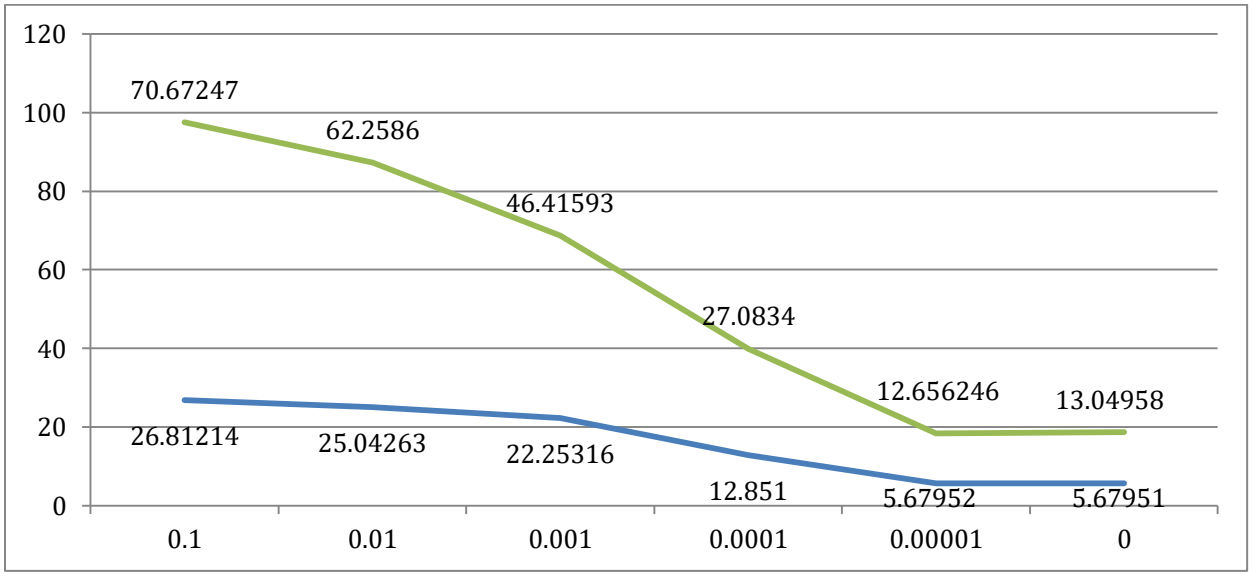
從上一題之圖表不難發現，雖然可以說 PM2.5 具有很大的影響性，但是越後面的值影響才越大，若只取前 5 個小時來做 training 效果會大大降低，且再加上本身的 feature 又少，結果可想而知會不太優秀。
最終結果為：19.91515 (private)+ 17.83639(public) = 37.75154

反之，如果只取後 5 個小時的值，結果為：7.70821 (private)+ 6.90892(public) = 14.61713，的確如圖表所顯示，越後面的值影響越大

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



(縱軸為 RMSE)
以上圖來看，雖然這四個值都明顯 underfitting，但可以看到 λ 越小 RMSE 也越小， 若加入 λ=0. 00001 以及 0 可以觀察到在 0.00001 到 0 之間可能產生了 overfitting，如下圖：



而以作業來講，我使用的 λ 為 0.00004 左右，我認為差不多可以達到平衡。

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) $(X^TX)^{-1}yX^T$

A: c