學號：B03902084 系級： 資工四 姓名：王藝霖

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

參數設定(1): fixed learning rate = 0.000001, iteration = 10000, feature 沒有scale

(2): fixed learning rate = 0.00001, iteration = 10000

模型(1)RMSE: 7.199387104695093

模型(2)RMSE: 6.658049915196641

模型(1)使用 learning rate 較小的原因是因為在模型(1)中有些 feature 的值比較大，同樣的 learning rate 之下相較於(2)會走比較大步（更新值調整太多），因此把 learning rate 調小一點比較容易用 gradient decent 的方式找到最佳解。

模型(1)雖然有較模型(2)多的資訊，但因為每個資訊的空間不太一樣，例如 pm2.5 介於 -1 和 109 之間，NMHC 則介於 0 和 1.3 之間。因為在 gradient decent 調整 weight 的時候都是用同一個 learning rate，因此很容易受值最大的項影響，而值最大的項可能不是最好的能夠預測的 feature，可能會導致結果整個偏掉。因此反而模型(2)的結果比較好。

2. (1%)將feature從抽前9小時改成抽前5小時，討論其變化

參數設定和上一題皆相同

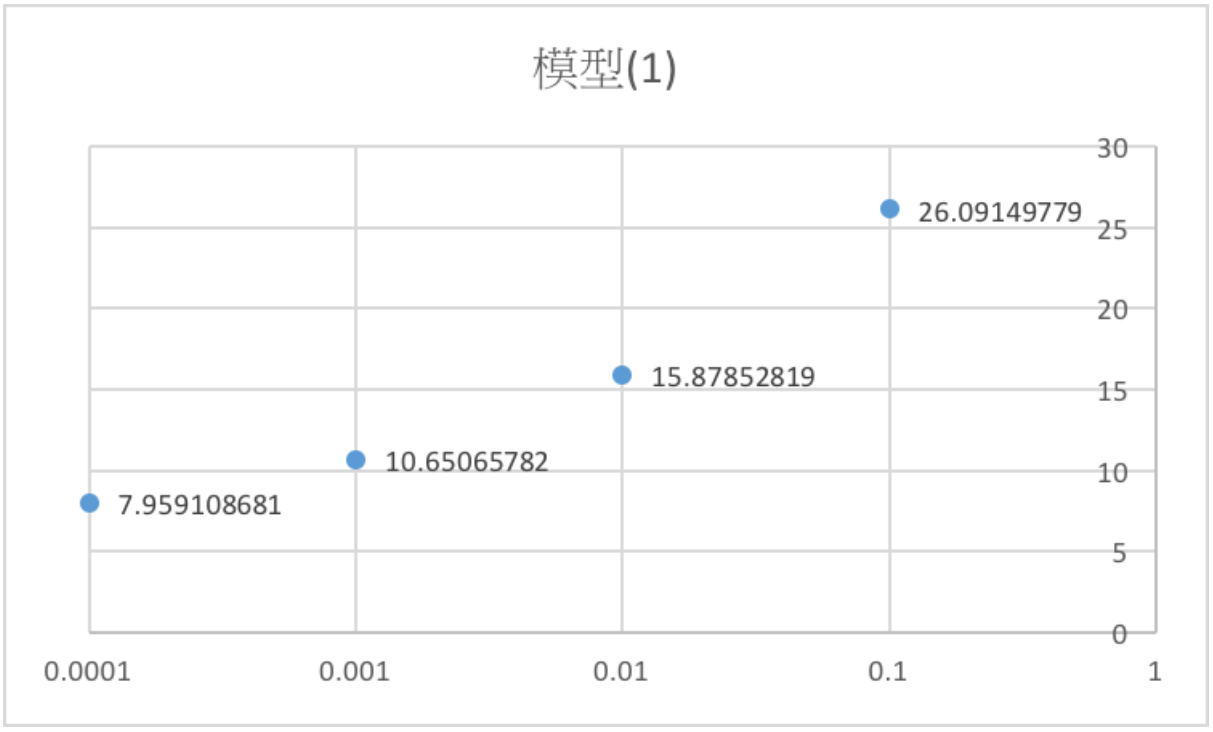
模型(1)RMSE:7.144184737382566

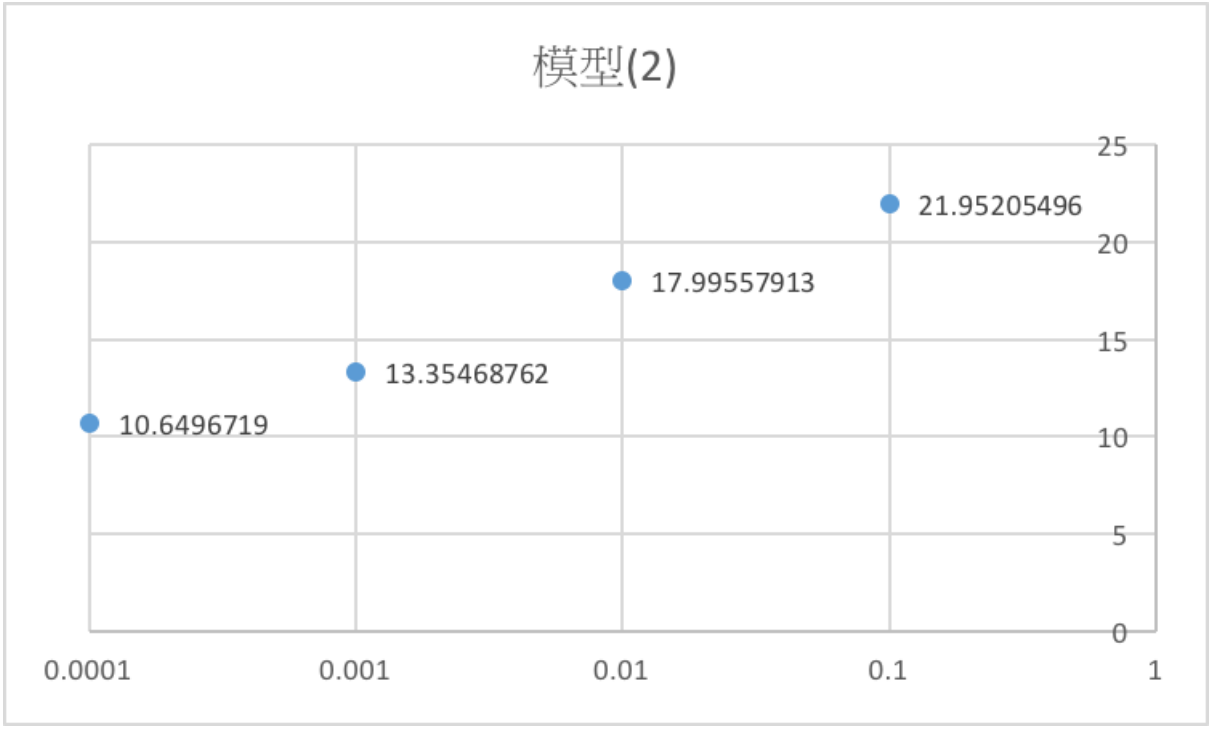
模型(2)RMSE:11.75649524600125

模型(1)和抽9小時的結果相差不大，但模型(2)卻差很多，推測可能是因為資訊量太少了（只有前五個小時總共5個feature），才會導致這種結果。而(1)因為還有其他很多資訊（例如pm10等等）所以還是可以得到相差不大的結果，甚至因為有可能濾掉了些noise，因此結果甚至稍好一點。

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

參數設定和之前相同，但改成取第1000 iteration 的結果





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)為一存量 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 = $[x^1\ x^2 \ldots x^N]^T$ 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 ？請寫下算式並選出正確答案。(其中X^TX為invertible)

(a) $(X^TX)X^Ty$
(b) $(X^TX)^{-0}X^Ty$
(c) $(X^TX)^{-1}X^Ty$
(d) $(X^TX)^{-2}X^Ty$

(c) 是正確答案
proof:

$$RMSE = E(w) = \sqrt{(Xw - Y)^2/N} = (1/N)\sqrt{w^TX^TXw - 2w^TX^Ty + y^Ty}$$

to minimize RMSE, find w such that $gradient\ E(w) = (2/N)(X^TXw - X^Ty) = 0$

$$\Rightarrow X^TXw = X^Ty \Rightarrow w = (X^TX)^{-1}X^Ty$$