學號：B06902052 系級： 資工二 姓名：張集貴

以下將抽全部的 feature（加 bias ）的稱為模型 (1)，抽 pm2.5 的一次項當作 feature（加 bias ）的稱為模型 (2)。因為 kaggle public 跟 private 各用了 120 筆資料，兩者的分數合併應 $\sqrt{\frac{120\ public^{\ 2}+120\ private^{\ 2}}{240}} = \sqrt{\frac{public^{\ 2}+\ private^{\ 2}}{2}}$ 。故以下 kaggle 分數均用合併後的分數。

1. (2%)記錄誤差值 (RMSE)(根據kaggle public+private分數)，討論兩種feature的影響

模型 (1)	6.56009
模型 (2)	6.59620

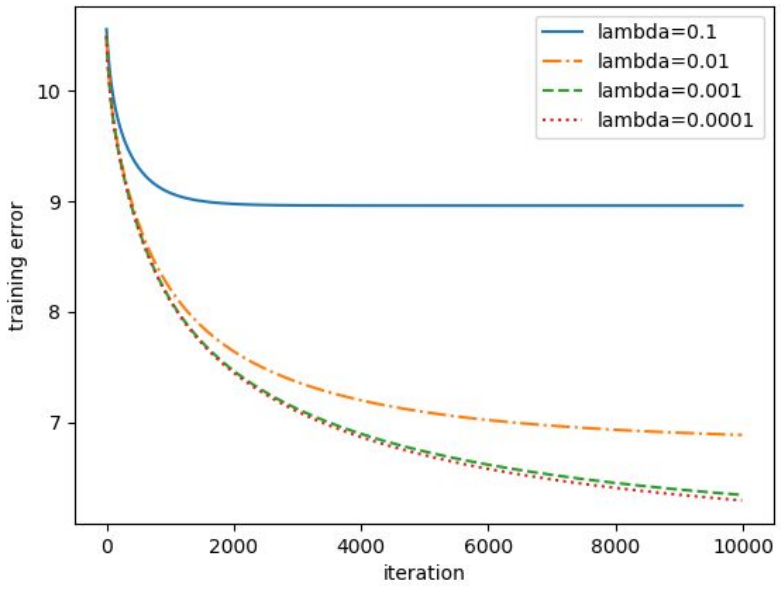
可以發現模型 (2) 的 error 較模型 (1) 高，因為模型 (2) 其實是模型 (1) 的子集，故在 training data 上的 error 不會小於模型 (1)，由於 testing data 的分佈跟 training data 近似，所以理論上模型 (2) 的 kaggle 分數不會小於模型 (1)。
模型 (1) 的訓練結果沒有 overfitting 也是一個原因。

2. (1%)將feature從抽前9小時改成抽前5小時，討論其變化

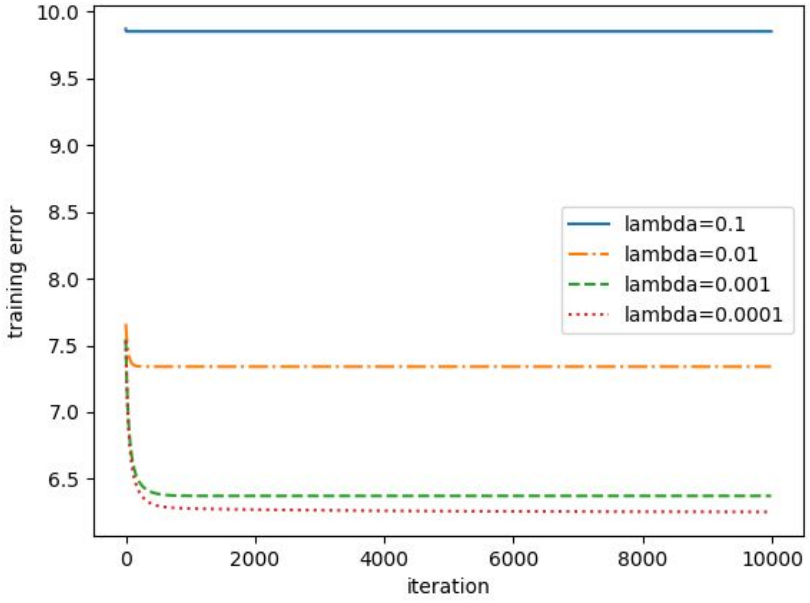
模型 (1)	6.60907
模型 (2)	6.74451

可以發現模型 (2) 的 error 還是較模型 (1) 高，且模型 (1) 跟模型 (2) 的 kaggle 分數都較上一題高。由於這一題的模型都是上一題對應模型的子集， 故在 training data 上的 error 不會小於上一題。由於 testing data 的分佈跟 training data 近似，所以理論上這一題的模型的分數都不會小於上一題對應模型的分數。

3. (1%)Regularization on all the weight with λ=0.1、0.01、0.001、0.0001，並作圖
因為用原始公式作圖看不出不同 λ 的差異，這邊做 gradient descent 用的 error function 是 $Error(w) = \frac{1}{N}\sum_{i=1}^{N}(y_i - w\cdot x_i)^2 + 10000\lambda w\cdot w$ 。作圖時的 training error 用的公式是 $Error(w) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - w\cdot x_i)^2}$ 。



產生模型 (1)



模型 (2)

可以看出 λ 越大 error 越大，這是因為 regularization 會讓演算法避免挑模型裡的參數絕對值較大的那些 function，所以 λ 越大會產生類似讓模型集合變小的效果。

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

Ans: (c)