Homework 1 Report - PM2.5 Prediction

學號：r06522828　系級：機械碩一　姓名：王榆昇

1. (1%) 請分別使用每筆 data9 小時內所有 feature 的一次項（含 bias 項）以及每筆 data9 小時內 PM2.5 的一次項（含 bias 項）進行 training，比較並討論這兩種模型的 root mean-square error（根據 kaggle 上的 public/private score）。

* 以下結果都對資料做過前處理再進行訓練，詳細會在第四題說明。

Features	Training Error (RMSE)	Public Score	Private Score
PM2.5	5.96531	6.74597	7.23362
All	5.96531	6.03145	6.27773

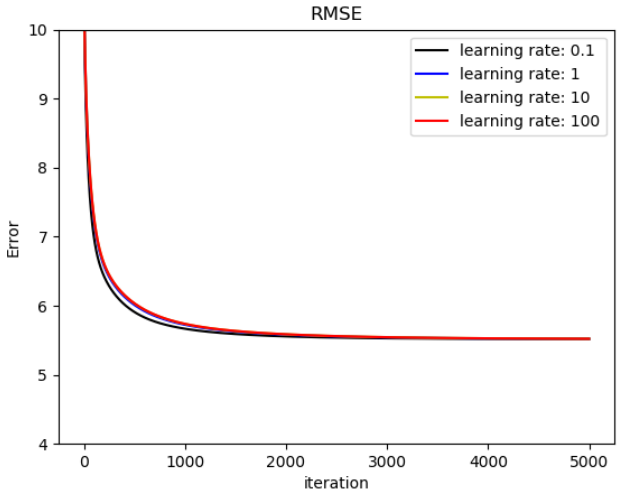
learning rate = 10；iteration times = 100000

取全部汙染物作為特徵在 public 和 private 都得到較低的 RMSE(相比於只取 pm2.5 做特徵)，推測是其他汙染物與 pm2.5 的變化其實有高度相關，在預測時能提供更多資訊，使 model 更精確。

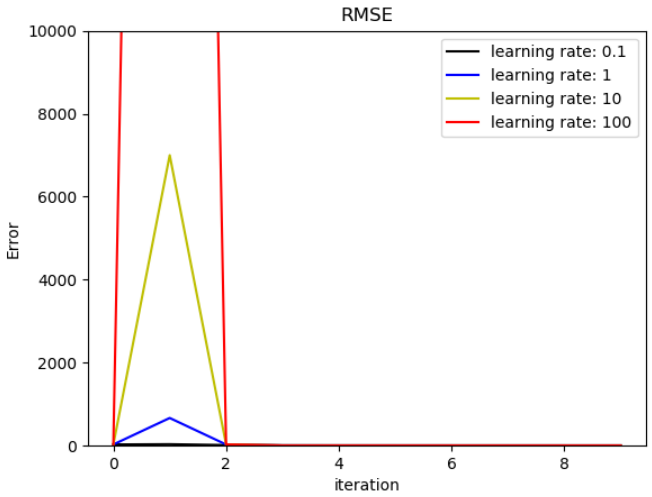
2. (2%) 請分別使用至少四種不同數值的 learning rate 進行 training（其他參數需一致），作圖並且討論其收斂過程。

Features: NO2、NOx、PM10、PM2.5、SO2、WS_HR (9hr)

分別採用 learning rate = 0.1、1、10、100，擷取不同的 iteration times 作圖。



由上圖可見，不論 learning rate 為何，都在 iteration times = 500 次時大約收斂至 5.7，但 learning rate = 0.1 時收斂較快，將前 10 次的訓練過程放大來看:



由圖就可清楚看見，當 learning rate 偏小時(0.1)，weight 不會一次改變太多，error 的變化幅度也不大，但 learning rate 越大時，weight 第一步就改變越大，導致 error 也會急遽改變。

3. (1%) 請分別使用至少四種不同數值的 regulization parameter λ 進行 training（其他參數需一至），討論其 root mean-square error（根據 kaggle 上的 public/private score）。

Features: NO2、NOx、PM10、PM2.5、SO2、WS_HR (9hr)

Learning rate: 10；iteration times: 10000

Lambda	Training Error	Public Score	Private Score
10	5.61601	6.12400	6.35140
1	5.60421	6.12415	6.35146
0.1	5.49292	6.12147	6.35147
0.01	5.48166	6.12417	6.35147
0	5.48041	6.12147	6.35147

加入 regulization parameter 後 test error 的 RMSE 只微微下降 0.001 以下，說明在這個 linear model 中 regulization term 影響非常小，把 weight matrix 拿出來看，可以發現 weight 的值大多在 0.1 以下，最大值是 0.8，本身數值就很小，因此 regulization term 沒有甚麼影響。

4. (1%) 請這次作業你的 best_hw1.sh 是如何實作的？（e.g. 有無對 Data 做任何 Preprocessing？Features 的選用有無任何考量？訓練相關參數的選用有無任何依據？）

Data Preprocessing:
在 train.csv 中，存在許多為 0 的資料點，甚至有連續好幾小時所有測值為 0，但除了 RF 外我認為不該會有此現象，判斷是機器失靈的因素，因此在遇到資料點為 0 時會把前兩小時的資料作平均取代。

$$\text{if x[i]} == 0: x[i] = \frac{(x[i-2] + x[i-1])}{2}$$

PM2.5 中有數個點測值到達 900 多，遠高於平常 1~100 的正常值，判斷同是機器失靈，做上式同樣之處理。

在 test.csv 中同樣也存在這個問題，當不正常 0 值在第一小時，會平均此筆資料(9 小時)取代；當在第二小時，直接取第一小時的值做取代；在 3 到 9 小時會做與 train.csv 一樣的處理。

Features:
從網路資料中，我發現 pm2.5 和一些特定汙染粒子有正向關係，其中 NO2、NOx、SO2、PM10 影響特別大，除此之外 PM2.5 也會與風向、風速高度相關。但風向是分類問題，資料以 0 ~ 360 度做紀錄，在 linear regression 裡很難用上(舉例來說，1 度和 359 度都是北風，但放進 linear regression model 裡卻是兩個極端值)，有機會誤導模型因此捨棄風向特徵。

我在 best 模型裡只選用了前 5 小時的資料，我認為 9 小時會包含太多無用的資料，導致模型被誤導。