Homework 1 Report - PM2.5 Prediction

學號：r06944051 系級：網媒碩一 姓名：郭柏辰

以下題目均先對Data做Preprocessing，將值為NR的資料改為0，PM2.5項中所有小於0的資料全改為0。每筆連續9小時Data中，若含有PM2.5大於200，則刪除該筆Data。故最後有5624筆連續9小時的Data。最後再對5624筆Data做Normalize。

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）。

Ans:

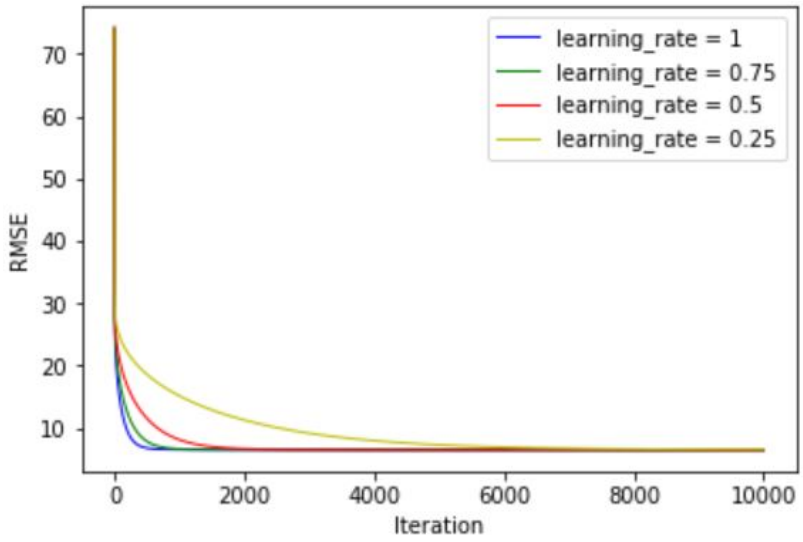
將5624筆Data隨機取4687筆當作training set，剩下937筆當作validation set，將Learning rate設為1 ，執行200000次(Iteration = 200000)。

	Public score	Private score
18 features + bias	7.36475	7.30679
PM2.5 + bias	8.34087	8.35942

由於PM2.5+bias的function包含於18 features + bias的function，故可以預期使用18個features比只使用PM2.5的分數還來得低，預測的較準確。

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。

Ans:

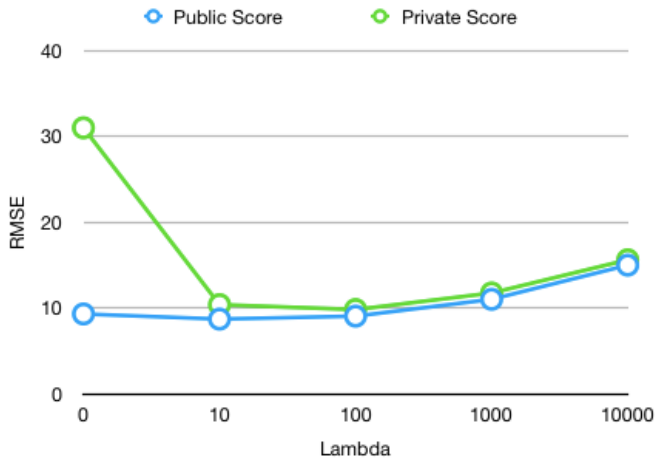


Learning rate越大，weight及bias更新的幅度越大，故收斂較快。Learning rate越小，則weight及bias更新的幅度越小，故收斂較慢。

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）。

Ans:

Lambda	Public score	Private score
0	9.30961	31.00235
10	8.71013	10.39593
100	9.05625	9.82198
1000	11.0159	11.75291
10000	14.9956	15.6217



此題使用每筆data9小時內PM2.5的一次項到五次項(含bias項)，並只用500筆Data當作training set進行training，當 λ = 0時，可以發現在private score分數高達31，可能是因為overfitting所造成的，故使用regulization。可以發現當 λ = 100 時，分數最低，當 λ 值大於100時，function則太過平滑，所以隨著 λ 越大分數也逐漸變高。

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？）

Ans:

本次作業我只實作gradient descent，故best_hw1.sh與hw1.sh一樣。先對Data做Preprocessing，將值為NR的資料改為0，PM2.5項中所有小於0的資料全改為0。每筆連續9小時Data中，若含有PM2.5大於200，則刪除該筆Data。故最後有5624筆連續9小時的Data。最後再對5624筆Data做Normalize。Features的選用，除了使用18項features的一次項外，額外新增AMB_TEMP, O3, PM10, PM2.5, WD_HR, WIND_DIREC, WIND_SPEED, WS_HR 二次項，我認為天氣好壞及風向可能會影響PM2.5高低，天氣好時通常氣溫較高，故新增溫度及風向考量。