Homework 1 Report - PM2.5 Prediction

學號：b04505021 系級：工海三 姓名：黃廉弼

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）。

在固定 learning rate 為 0.5
Iteration number 為 100000 的情況下

用全部 feature:

output.csv 4 days ago by zander363 do some data selection	7.31451	7.61433

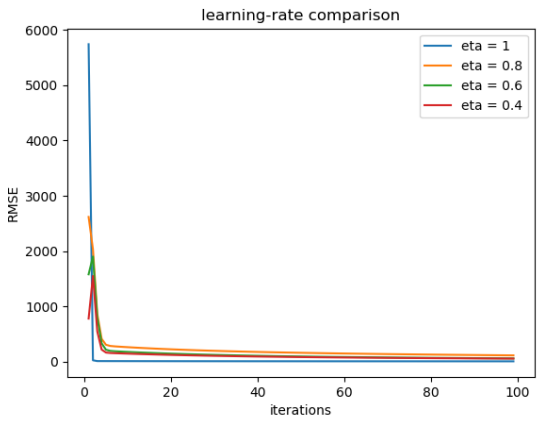
僅用 PM2.5:

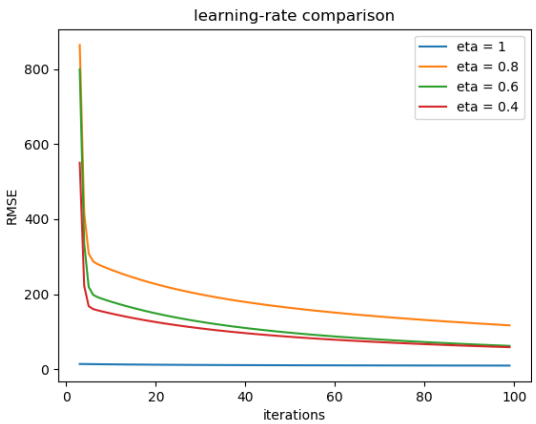
output_comp1.csv a day ago by zander363 To compare with 18 features This one only have PM2.5 this feature	8.47726	8.62201

很明顯可以看出 僅用 PM2.5 下去 train 出來的結果 不論是 public score 還是 private score 都硬生生比用所有 feature 的表現差了一截。
這樣的結果 可說是合情合理，畢竟 PM2.5 的數值可能受很多不同 feature 影響，因此直接排除其他 feature 勢必會降低預測的準確性。
但話又說回來，在捨棄了 17 項的 feature 後，在 RMSE 上卻只掉 1 左右，其實也算是表現非常不俗了。可見 PM2.5 本身對後續 PM2.5 值的影響是十分可觀的。

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。

執行結果如下圖，討論請見下頁：





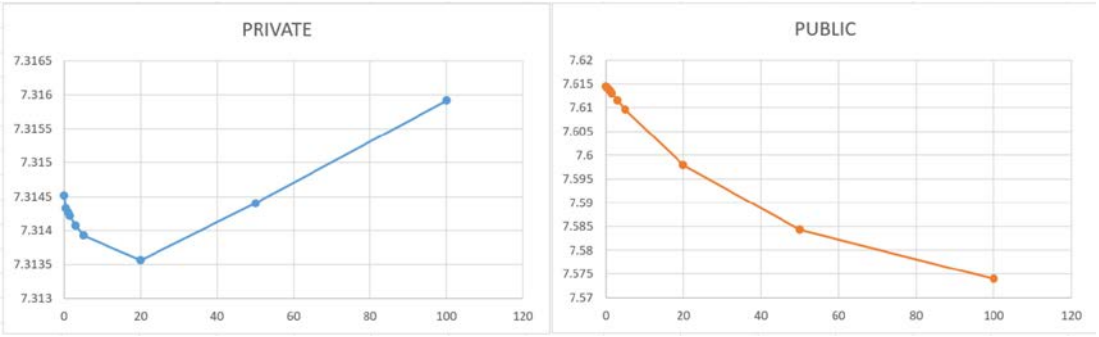
左圖為取 learning rate 在 1, 0.8 ,0.6 ,0.4 等四種不同 learning rate 在 100 個 iteration 中所得到的 iteration-RMSE 折線圖。
從圖中可以看出，最一開始 4 個不同線各自散佈在不同的點，接下來在 5 個 iteration 內 eta=1 的線便掉到最低，看似趨於收斂的位置。
而為了方便觀察，我從第 3 個 iteration 開始作圖得到右圖，可看出在經過一翻抖動後 eta=0.4 的紅線的 error 竟小於 eta=0.8 與 eta=0.6，但好景不常，然而在差不多經過 80 次 iteration 時，出現了 0.4 與 0.6 的黃金交叉，這點就展現出了 learning 較大的優勢，但 0.8 那條可能是一開始噴得太可怕了，因此還要好一段時間才能超英趕美。
因此可以得出兩個結論:

- Eta 較小 一開始暴增的範圍也比較小
- Eta 較大 Loss decay 的速度較快

3. (1%) 請分別使用至少四種不同數值的 regulization parameter λ 進行 training（其他參數需一致），討論其 root mean-square error（根據 kaggle 上的 public/private score）。

由於取點較多，故不逐一附圖，僅表列結果並無條件捨去至小數點後四位

lambda	0	0.5	1	1.5	3	5	20	50	100
Private	7.3145	7.3143	7.3142	7.3142	7.3140	7.3139	7.3135	7.3144	7.3159
Public	7.6143	7.6140	7.6135	7.6130	7.6114	7.6095	7.5978	7.5842	7.5739



由圖可觀察出隨 lambda 增加，RMSE 會逐漸下降，但下降幅度也漸緩。
而 private score 在 lambda 大於 20 後會暴增，而 public socre 卻不會。看起來是 public 的最低點還要再往後才會碰到。
此外，雖說 regulization 會幫助減少 Error，但很明顯其減少的幅度相當有限，因此此方法僅能當作優化的一部份，最其根本還是要設定一個有效的 model 才能使 Error 有效降低。

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？）

這次實作主要是用課堂上教到的 linear regression
此外，原本為了判斷各 feature 的 order，在圖像化顯示出各項 feature 與 PM2.5 關係時，意外發現似乎有一些點是不太正常的。因此借了老師上課提到的 3-folder cross 的觀念，來去除包含不正常數據的資料，雖然會因此損失掉能 training 的 data 量，但在用 training data 測試 以及 kaggle 評估結果都有亮眼的表現，大大的降低了 error。
此外，原本從散布圖看來 O3 的分布集中在兩條線上，原本判斷該用二次項，但實際 run 起來似乎沒有什麼幫助，因此猜想或許是 O3 實際上是更高次方項。