Homework 1 Report - PM2.5 Prediction

學號：b04705043 系級：資管三 姓名：張凱庭

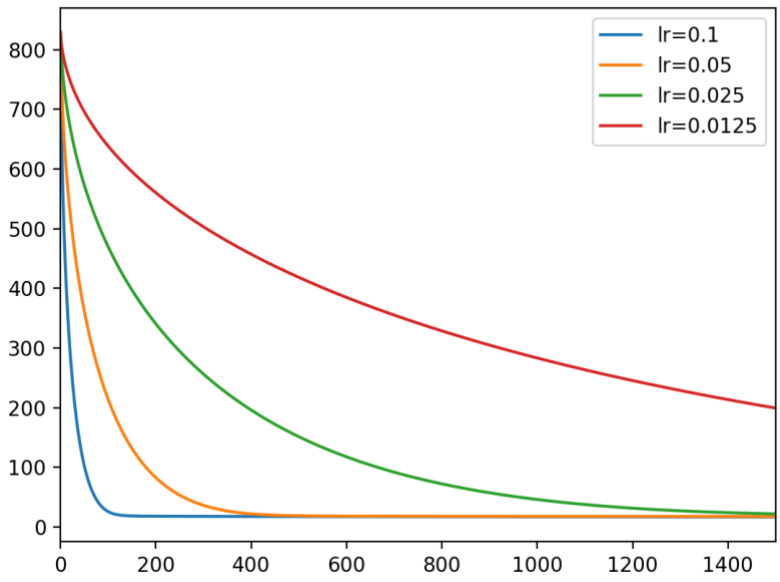
1.(1%) 請分別使用每筆 data9 小時內所有 feature 的一次項（含 bias 項）以及每筆 data9 小時內 PM2.5 的一次項（含 bias 項）進行 training，比較並討論這兩種模型的 root mean-square error（根據 kaggle 上的 public/private score）。

Learning rate:1 ,訓練次數 350000

	Private socre	Public score
9 小時所有	7.52764	7.75042
9 小時 pm2.5	8.51741	8.67107

由分數可以看到矢勇全部的 feature 的 root mean-square error 的 model 明顯比只使用 pm2.5 單一種 feature 的 model 好。原因可能是因為 pm2.5 是會受到其他的因素的影響，只考慮 pm2.5 的情況下，無法反映出其他 feature 帶來的影響，所以預測結果 root mean-square error 較高。而考慮所有 feature 的 model 較為複雜，比較可以反映出其他 feature 的影響，預測結果較準確

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。



由圖可看出，基本上 learning rate 越低，到達收斂所要花的次數越多，learning rate = 0.1 時，不到 100 次就收斂了，learning rate = 0.0125 超過 1500 次都還未收斂

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)** 。

Lambda	Private socre	Public score
100	8.15276	8.44770
10	8.21638	8.41352
1	8.22514	8.41021
0.1	8.22604	8.40988
0.01	8.22614	8.40985

從上表數據可看出使用正規化對於預測 pm2.5 並沒有顯著的影響 λ，主因應該是正規化主要是用來修正 overfit 的問題，但是這個 model 本身並沒有 overfitk 的問題，所以正規化並沒有帶來什麼影響

4.(1%) 請這次作業你的 **best_hwl.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 **?)**

(1)將 data 中的 outlier 的移除對 root mean-square error 有極大的影響，將離平均>13 個標準差的資料（共 791 筆）刪除後分數有顯著的進步

(2)feature 最後只留下'PM10', 'PM2.5', '03', 'CO', 'SO2', 'RAINFALL' 依據 loss 的結果決定保留這些 feature 特別是將與 wd_相關的 feature 移除後 model 就有小幅的進步

(3)最後的 model 決定不採用連續的 9 個小時 ，而改採用連續 7 個小時，預測時也只用連續 7 個小時，也得到相當的進步，直觀上來看，9 個小時中各項數值的變化實在太大，很難說明受到太前面的時段影響，因此縮小為 7 個小時作為預測 model