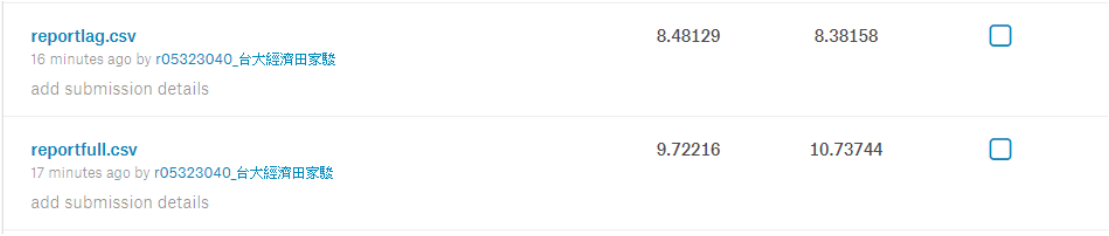
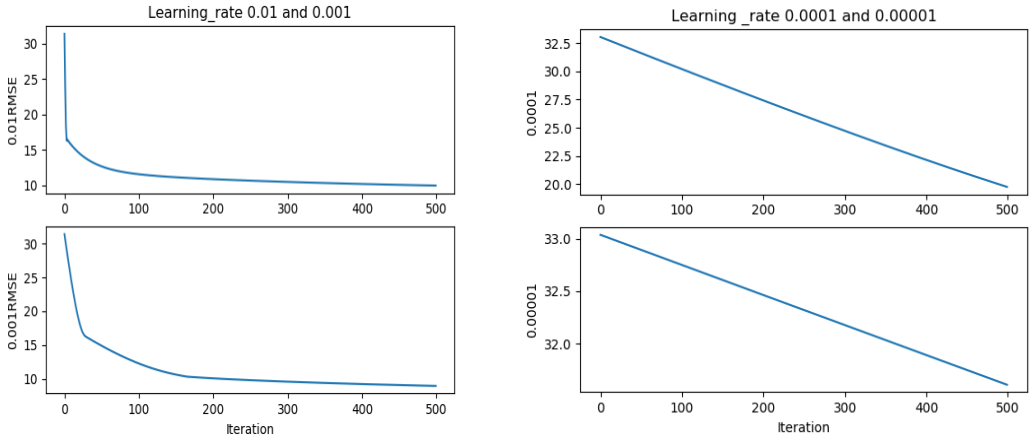
Homework 1 Report - PM2.5 Prediction

學號：r05323040　系級：經濟碩二　姓名：田家駿

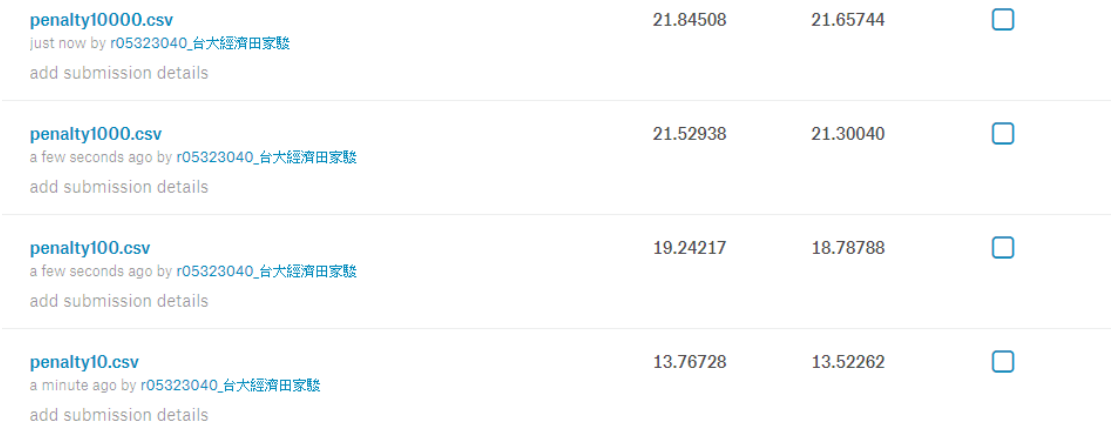
1. (1%) 請分別使用每筆 data9 小時內所有 feature 的一次項（含 bias 項）以及每筆 data9 小時內 PM2.5 的一次項（含 bias 項）進行 training，比較並討論這兩種模型的 root mean-square error（根據 kaggle 上的 public/private score）
我們可以看到只用前九期 PM2.5 作為變數做預測表現比前九期所有 Feature 還要好，我們可以看到當 feature 太多時，可能會有 overfitting 的問題。



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。
可以看到左邊的 learning rate 收斂速度明顯大於右邊的，因此 learning rate 太小會影響收斂速度，太大則有可能會爆掉。



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）。
越大的 regulization parameter 雖然讓參數越平滑，但預測表現也可能越差，因此我們在做 regulization parameter 時，需要用 Cross validation 去 tune 參數，如果是在 L1-norm 情況下除了可以用 CV，也有一些統計系的學者導出 Lasso regulization parameter 的理論值。



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？）

這次的 best model 主要是最後兩天發現資料有 missing value 的問題，像是溫度是零度，PM2.5 為負值，觀察 PM2.5 負值的資料，我們可以看到前後期可能都七八十，不太可能為負值，於是我把一些負值有問題的資料用前後期代掉，再跑迴歸讓我的 Public score 降到 6.08，在 best model 中，因為 PM2.5 呈現雙峰分配，想試著去分大於 60 以及小於 60 的族群，去造一個 Dummy variable，讓不同族群有不同截距向，在 Public board 上表現似乎很好，然而似乎出現 overfitting 的問題，Private board 掉到 6.48，所以 hw1 的表現才會比 hw1_best 好。