Homework 1 Report - PM2.5 Prediction

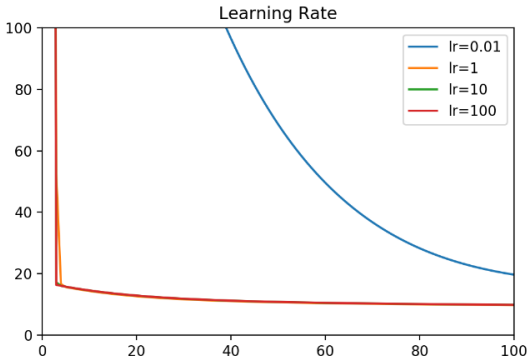
學號：b04505026 系級：工海三 姓名：蔡仲閔

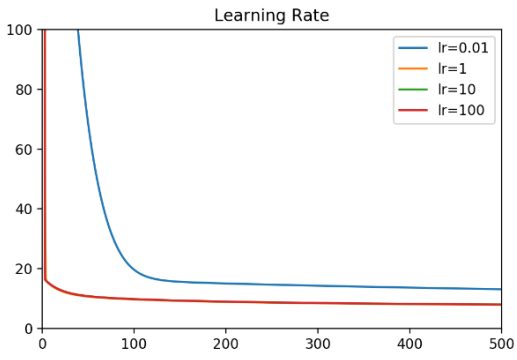
1. (1%) 請分別使用每筆 data9 小時內所有 feature 的一次項（含 bias 項）以及每筆 data9 小時內 PM2.5 的一次項（含 bias 項）進行 training，比較並討論這兩種模型的 root mean-square error（根據 kaggle 上的 public/private score）。

Submission and Description	Private Score	Public Score
output_1.csv a few seconds ago by b04505026_zmtsai add submission details	8.37096	8.39316
output_18.csv 7 minutes ago by b04505026_zmtsai add submission details	7.43071	7.44861

從上圖結果可以看到，使用所有 feature 的 model 得到較佳的結果，而僅使用 PM2.5 的 model 則有不小的落差。和老師上課提到的一樣(寶可夢分類)，這與我們預期的結果相同，若使用較多的 feature，在沒有 noise 的前提下，應該可以 train 出較佳的結果。

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。





上面兩個分別用了 learning rate=0.01,1,10,100(以下稱 lr)進行作圖，並取不同的 iteration 做觀察，可以看到 lr=0.01 時明顯收斂的比較慢，甚至在 iteration 到 500 次時仍然沒辦法到達其他組的水準，而其他三組的行為則相當接近，即便放到 100 次內也沒有看出太明顯的區別。。

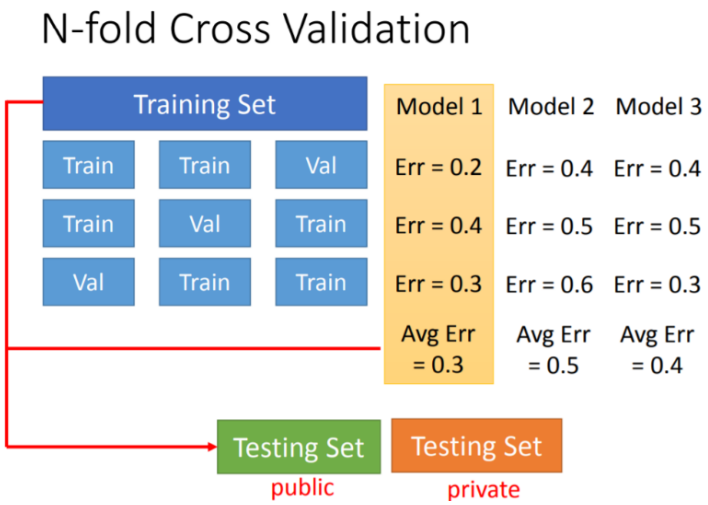
3. (1%) 請分別使用至少四種不同數值的 regulization parameter λ 進行 training（其他參數需一至），討論其 root mean-square error（根據 kaggle 上的 public/private score）。

Submission and Description	Private Score	Public Score
output_ita_2.csv 3 hours ago by b04505026_zmtsai add submission details	12.26638	12.20687
output_ita_1.csv 3 hours ago by b04505026_zmtsai add submission details	11.09248	11.20569
output_ita_05.csv 3 hours ago by b04505026_zmtsai add submission details	10.06456	10.27533
output_ita_0.csv 3 hours ago by b04505026_zmtsai add submission details	7.63631	7.79871

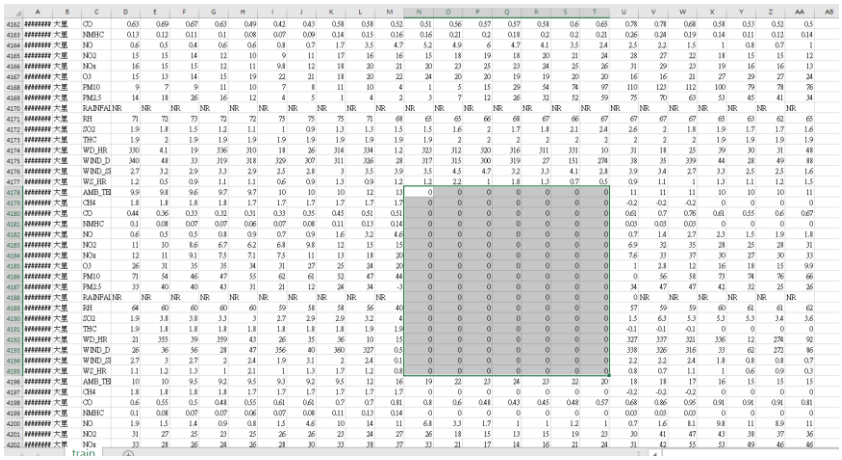
上圖從下到上 lamda 分別從 0, 0.5, 1 到 2，我們可以觀察隨著 lamda 增加在相同的 iteration 中收斂的速度是越來越慢，這也印證 lamda 雖會讓我們得到一個 smooth 的線，但若是 lamda 過大則會造成 under fitting 的結果。

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？）

在這次作業讓我對 machine learning 有了第一步的認識，從剛開始完全無從著手，到後來看著 Sample code 慢慢推敲並改善，其中讓我順利通過 strong baseline 的方法是資料的挑選，首先我將資料分組，進行 training 並找出較佳的 model。



在過程中發現某些月份出現異常的極端值，上網查完資料後，確定那是不應出現的數值，便將資料從 data set 中去除，便順利通過 strong baseline。之後更發現某些月份的資料出現異常的連續的 0，



此時再將這些資料去除後，便得到較佳的結果。這次的經驗也讓我明白資料整理及挑選的重要性，而對於該領域的 domain knowledge 也是相當重要，若是拿到 data 便開始 train 往往會得到不好的結果。