Homework 1 Report - PM2.5 Prediction

學號：b05611038　系級：生機二　姓名：張育堂

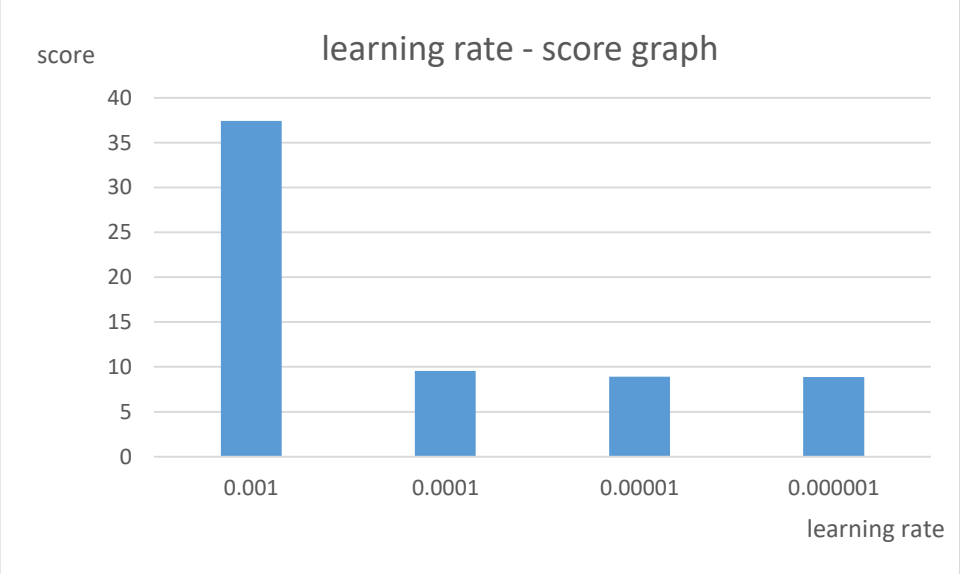
1. (1%) 請分別使用每筆 data9 小時內所有 feature 的一次項（含 bias 項）以及每筆 data9 小時內 PM2.5 的一次項（含 bias 項）進行 training，比較並討論這兩種模型的 root mean-square error（根據 kaggle 上的 public/private score）。

	public	private	Final score
All feature	8.86436	8.87303	8.868695
PM2.5	9.54105	9.67794	9.609495

當所有 feature 都下去 train 的時候，因為所有的參數的權重都會被計算，而當資料有越多的 feature，對未來的預測就更準，而 kaggle 的分數也是合理的體現這個結果，但是當 training 的時候，所花的時間就天差地別，所以同樣演算法的預測模型的精進也意味著計算時間需要長更多。

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。

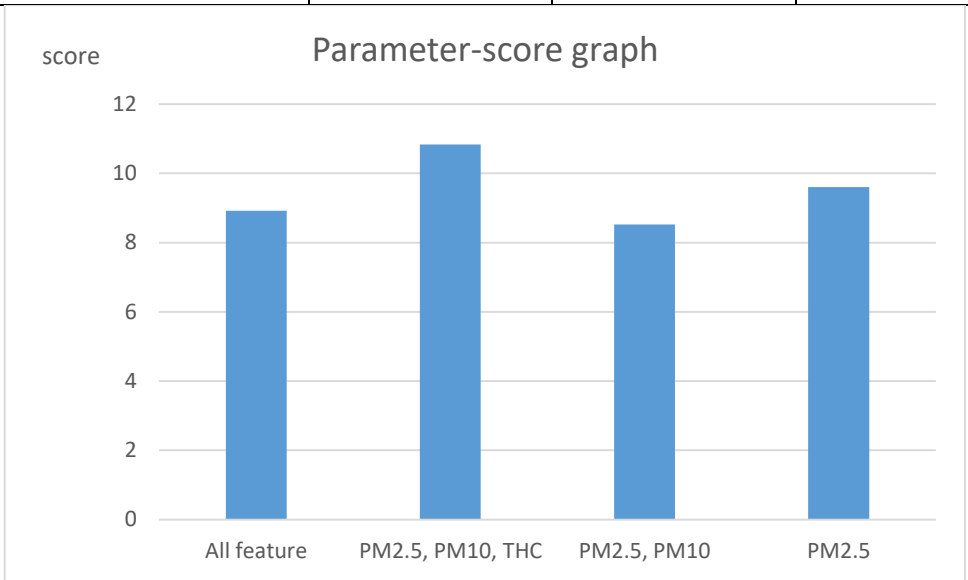
Learning rate	public	private	Final score
0.001	37.88886	36.95167	37.42027
0.0001	9.91317	9.19421	9.55369
0.00001	9.03822	8.80787	8.923045
0.000001	8.86436	8.87303	8.868695



在 train model 的時候，選擇配合自己 model 的 learning rate 是很重要的，而要是 learning rate 過大，常常就會有無法收斂到 local minimum 的狀況，因此在理想狀況，選擇較小的 learning rate 是比較好的選擇。

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）。

parameter	public	private	Final score
All feature,9hr	9.03822	8.80787	8.923045
PM2.5, PM10, THC,9hr	11.71047	9.96643	10.83845
PM2.5, PM10,9hr	8.94169	8.09777	8.51973
PM2.5, 9hr	9.67794	9.54105	9.609495



其實在 training 的時候，因為都是使用 linear regression 的方式，所以參數選擇對於 model 來說非常重要，而今天用 All feature 分數好是正常現象，因為所有的變數都可以被計算進模型裡，而其他的 feature 在挑選的時候，我有盡量挑跟 PM2.5 趨勢相近的，但是三個 feature 選用就是一個標準的錯誤 module，因為 THC 跟 PM2.5 的關聯性過小，在 train 的時候又诶成功把參數降到幾乎不會影響預測的狀態，所以不一定選用越多 feature，module 就會有比較好的結果。

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？）

我的 best 直接使用套件做，用 scikit learn 的 neural network 的 regression，用套件有直接得出 8.011355 的遠超自己寫的 gradient descent 的分數，但實際上我有和同系修過同一堂課的人詢問，其實 data 的預處理也是很重要，無奈的是其實我自己在寫 module 的時候已經沒有多少時間了，而且已經過 strong baseline，所以就沒做這方面的嘗試。