Homework 1 Report - PM2.5 Prediction

學號：r06521705　系級：土木系營管組碩一　姓名：陳思愷

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/privat score）。

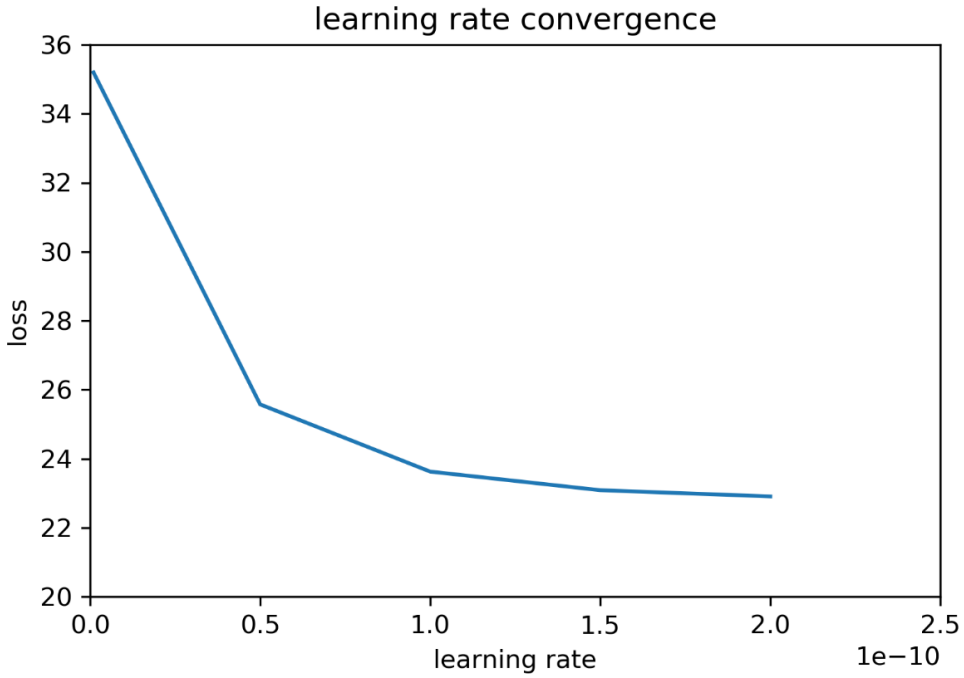
RSME	ALL-FEATURE	ONLY-PM2.5
PUBLIC	8.90088	9.55819
PRIVATE	8.68778	9.69185

使用所有 feature 的模型預測精準度上不管是 public 或 private 都較為精準，推測是 18 項 feature 當中，有實際對於 pm2.5 影響程度相當重大的因子，從一些 Pm2.5 的相關文獻回顧當中可以發現，不管是 NOx 或 SOx 等等化合物，以及溫度、濕度、風速等等都會對 pm2.5 的濃度有一定程度上的影響，所以對於預測精準度來說，這些 feature 是需要考慮進去的。

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。

Learning rate	1e-12	5e-11	1e-10	1.5e-10	2e-10
loss	35.195	25.573	23.624	23.085	22.905

當繼續調整 learning rate 以期達到更小的 loss 又不會使 loss 急速變大，最後會發現 loss 逐漸收斂到 22.7 左右



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）。

(依照 hw1.py 去時做加入 regulization parameter λ 進行 training)

	λ = 0.1	λ=0.01	λ=0.001	λ=0.0001
public	8.83322	8.83322	8.83322	8.83322
private	8.68740	8.68740	8.68740	8.68740

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？）

有關於這次 hw1_best，有鑑於我唯一學過方式就是 Gradient Descent，所以 training 方法來說並沒有和 hw1 不同的地方。
所以我便從 feature 的數量開始著手去試，feature 越多，所涉及的維度越大，空間將變得更加寬廣，當資料量不是那麼足夠的時候，提升維度可能會造成資料的稀疏化，使其 train 出來的模型沒有辦法相當精準，所以我從取前 9 小時的資料改成取前 2 小時的資料來預測，精準度有所提升。