Homework 1 Report - PM2.5 Prediction

學號：b03901015 系級：電機四 姓名：梅希聖

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）。

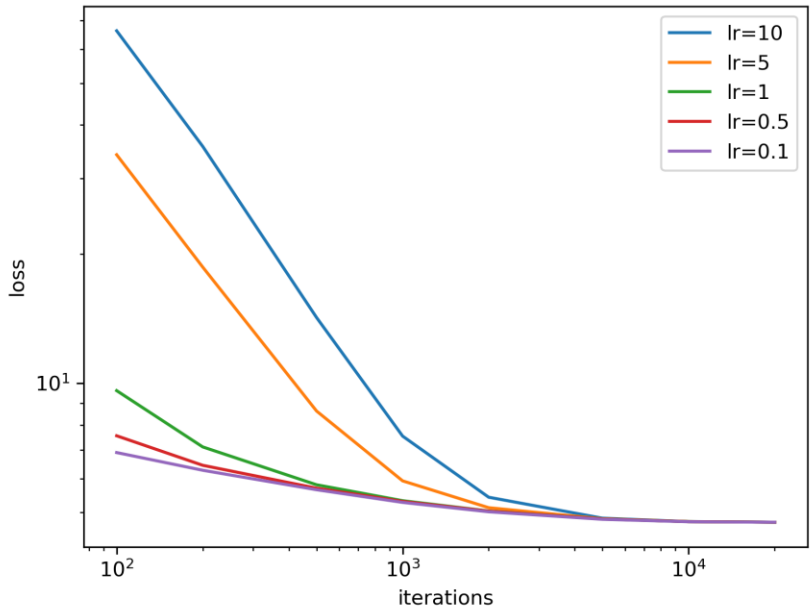
common parameters: learning rate = 1, iteration = 10000

	Private score	Public score	RMSE
All features	8.70431	8.38633	8.54680
Only PM2.5	8.75301	9.04381	8.90188

結果顯示，使用所有 feature 的 RMSE 會比只使用 PM2.5 的小一些，或許可以代表其他汙染源對於 PM2.5 之預測確實有影響。

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。

使用所有 feature 進行 training，max iteration = 20000



結果顯示，learning rate 愈大，一開始之 loss 愈大，收斂速度也愈快。
到了最後，loss 均收斂至相同的值，代表 5 個 model 都很大可能 fit 在同樣的 local optimum 的地方。

3. (1%) 請分別使用至少四種不同數值的 regulization parameter λ 進行 training（其他參數需一致），討論其 root mean-square error（根據 kaggle 上的 public/private score）。

common parameters: learning rate = 1, iteration = 20000

λ	Private score	Public score	RMSE
10	8.78763	8.62407	8.70623
1	8.81660	8.57625	8.69726
0.1	8.82773	8.58006	8.70478
0.01	8.82892	8.58052	8.70561
0	8.82905	8.58058	8.70570

結果顯示，本次作業中加入不同大小的 λ 進行 training 並未對 RMSE 產生太大影響。

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？）

與 hw1.sh 相同，使用 linear regression 與 adagrad，feature 使用 9 天內 18 種汙染源加上 bias 項，共 163 個 feature 進行 training，data 均有經過 normalize 處理。原始 data 中，有些 PM2.5 的值疑似有誤，有些超過正常值許多，有些則是負數。排除此種 data 後開始 training。

training 分成 5 次進行，每次隨機選擇 4/5 的 data 後 train 出該次的 model 結果，最後將 5 次 model 取平均做為最後的 model。

在 model.npy 中使用 learning rate = 10，iterations = 20000。然而，在 deadline 後一天多次上傳後才發現，iteration 在 5000 ~ 10000 間的 model，在 kaggle 上的分數比 20000 的還要高，或許代表 20000 次 iteration 已經讓 model overfit 了。