Homework 1 Report - PM2.5 Prediction

學號：r05922080　系級：資工碩二　姓名：王鵬傑

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）。

Learning Rate = 1

Epoch = 100000

Feature	Public	Private
PM2.5	8.44949	8.38399
ALL	7.42729	7.46793

Public / Private

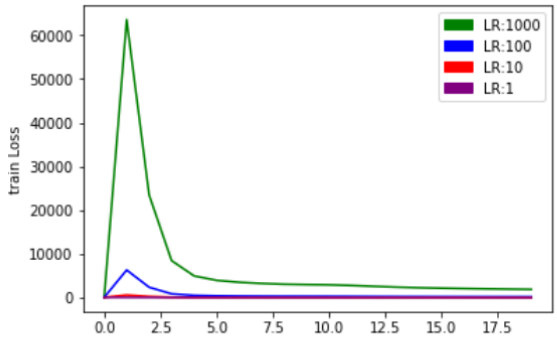
1. 當只選擇 PM2.5 參數時，在第 5000 個 epoch 時就已經收斂，因為使用的 Feature 較少，因此收斂速度很快。
2. 但當選擇所有 Features 時，在第 400000 個 epoch 後才開始慢慢收斂，且 training 時間較長。

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。

Epoch 皆為 10000

LR	1000	100	10	1
ALL Feature	7.48 / 7.73	7.46 / 7.50	7.46 / 7.48	7.45 / 7.48

Public / Private



在第一個 Step 時，隨 Learning rate 調得越大的時候，反而收斂速度會更慢，經過 10000 個 epochs 後，public 分數皆差不多，但 Private 有明顯的差距，因此對於 Private 來說，較小的 Learning Rate 較佳

3. (1%) 請分別使用至少四種不同數值的 regularization 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）。

Epoch: 10000

λ	0.0	0.1	0.01	0.001	0.0001
ALL Feature	7.45 / 7.48	7.46 / 7.48	7.45 / 7.48	7.45 / 7.48	7.45 / 7.48

Public / Private

Regularization parameter 的加入並沒有很大的差別，因為可能還沒有 Overfitting 的問題。

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？）

Feature 的選擇：

1. 除了所有 18 項參數以外，先將所有資料與 PM2.5 做 Correlation，其中挑出較相關的幾項，因此額外 Concatenate CO**2、PM10**2、PM2.5**2、WD_Direct**2，共 198 個 Features。

Preprocessing：

1. PM2.5 的值將出現-1 的值都改成 0，因為出現負數並不合理。
2. 有幾筆資料超過兩三百的也不合理，因此若九天當中出現該筆資料，或者第十天出現不合理值的資料，皆捨棄不使用。
3. 將資料中出現 NR 的資料改為 0。
4. AMB_TEMP 的資料中，有幾筆資料為負數或者零，我們將過大的資料設為該筆資料前後 10 天的平均值。
5. CH4 的資料中，有資料過小，也將其值設為該筆資料前後 30 天的平均值。
6. 將所有參數做 Normalization。
7. 將資料做 random 的排序，Random Seed 為 10。

訓練時：

1. EPOCH：100000
2. 使用 Adam 來優化
3. Regulization parameter λ：0.0