Homework 1 Report - PM2.5 Prediction

學號B05902042 系級：資工二 姓名：林瑋毅

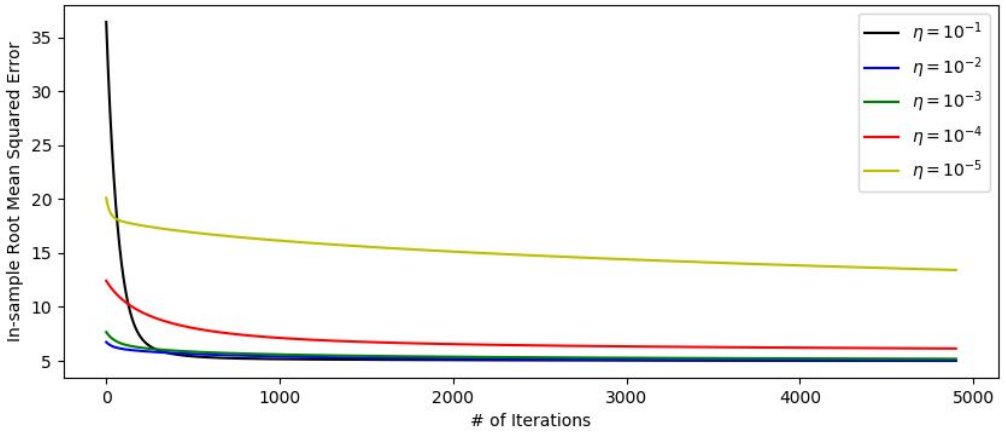
1. (1%) 請分別使用每筆data9小時內所有feature的一次項（含bias項）以及每筆data9小時內PM2.5的一次項（含bias項）進行training，比較並討論這兩種模型的root mean-square error（根據kaggle上的public/private score）。

[所有 feature] public/private = 5.83168 / 6.30522
[PM2.5] public/private = 6.57293 / 7.04001

不論是 public 還是 private ，用所有featur的一次項進行訓練所得的模型誤差較小，原因是其他非PM2.5能提供PM2.5之外的資訊量，因此能得到較好的結果。

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。

此題使用AdaGrad，feature取前9小時的PM2.5。下圖為迭代數量與Ein的關係圖（從迭代數大於100開始），如圖所示，當 learning rate 越高，Ein 收斂的越快，即能用越短的時間逼近最佳解，但須注意若 learning rate 太高反而會造成初期不穩定的狀況（如下圖黑線）。



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）。

此題 feature 取前 9 小時的 PM2.5 及 PM10，且因為加入 regulization term，各 feature有先經過 normalization。

[regulization term = 1] public/private = 10.24932 / 9.98597
[regulization term = 0.1] public/private = 7.62223 / 7.14388

[regulization term = 0.01] public/private = 6.48920 / 6.36257
[regulization term = 0.001] public/private = 6.28848 / 6.28650
[regulization term = 0.0001] public/private = 6.27479 / 6.28499

增加 regularization 的強度反而變差，可知原先的模型並沒有 overfit 。

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？）

1. Preprocessing 過程請參閱附註。
2. 考量各feature對PM2.5的相關係數後，選用 PM2.5 、 PM10 、 CO、NOx、SO2 做為 feature，其中 NOx 及 SO2 的資料不易清理，故取出小於等於零的值之後，直接 quantilization。
3. 用 5-fold Cross Validation 選擇最佳參數為 PM2.5 最近 8 小時、PM10最近 8 小時、CO 最近 4 小時、NOx 最近 7 小時、SO2 最近 6 小時，並以此進行訓練。

附註：以上題目皆在以下的實驗條件進行：

- feature 做過以下的前處理：
  - PM2.5 只取 (0,200) 為有效數值
  - AMB_TEMP 只取 (0, np.inf) 為有效數值
  - CH4 只取 (0, np.inf) 為有效數值
  - CO 只取 (0, np.inf) 為有效數值
  - NMHC 只取 (0, np.inf) 為有效數值
  - NO 只取 (0, np.inf) 為有效數值
  - NO2 只取 (0, np.inf) 為有效數值
  - O3 只取 (0, np.inf) 為有效數值
  - PM10 只取 (0, np.inf) 為有效數值
  - RAINFALL 只取 (-0.1, np.inf) 為有效數值
  - RH 只取 (0, 100) 為有效數值
  - SO2 只取 (0, np.inf) 為有效數值
  - THC 只取 (0, np.inf) 為有效數值
  - WD_HR 只取 (-0.1, 361) 為有效數值
  - WIND_DIREC 只取 (-0.1, 361) 為有效數值
  - WIND_SPEED 只取 (-0.1, np.inf) 為有效數值
  - WS_HR 只取 (-0.1, np.inf) 為有效數值
- 若feature出現無效數值，則不拿來train。
- test.csv 中若出現無效數值，使用線性內插替代，若無法內插，則取最臨近值替代，若無最臨近值，則取0。