Homework 1 Report - PM2.5 Prediction

學號：r06922086　系級：資工碩 一　姓名：林凡煒

1.

(1.) 每筆 data 9 小時內所有 feature 的一項（含 bias 項）：

validation error (RMSE): 16.90

Kaggle's public/private score: 9.03/9.27

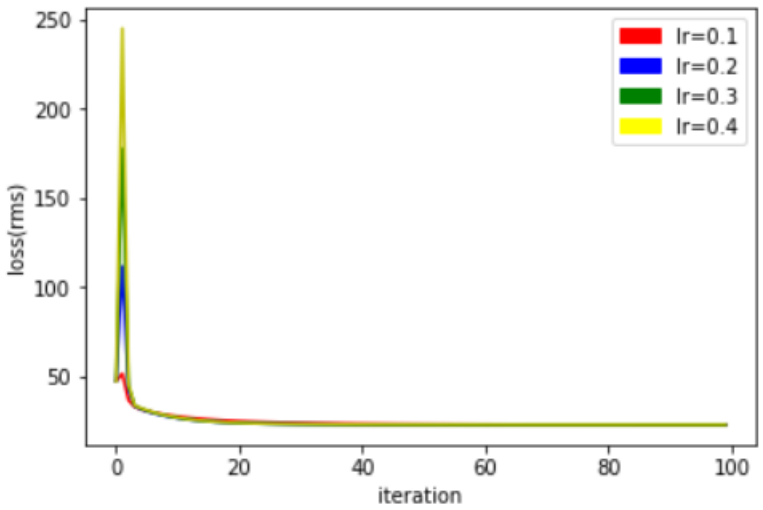
(2.) 每筆 data 9 小時內的 PM2.5 的一次項（含 bias 項）：

validation error (RMSE): 17.09

Kaggle's public/private score: 9.48/9.51

以第一種方法取 features 的 training data 訓練出來的 performance 比較好 。從結果得知，第二種 model 所考慮的因素不夠全面，導致模型不夠複雜，效果比較差。

2.



learning rate 分別取 0.1, 0.2, 0.3, 0.4 並作 training。

可以看出測試的 learning rate 越大時，第一步所產生的誤差會越大，接著快速的收斂一致。猜測是因為採用 adagrad 的關係，learning rate 對 gradient 的因素被 sum of gradient 所取代。

3.

(1) regularization parameter: 0.0

Kaggle’s public/private score: 8.303/7.893

(2) regularization parameter: 1.0

Kaggle’s public/private score: 8.308/7.886

(3) regularization parameter: 10.0

Kaggle’s public/private score: 8.371/7.852

(4) regularization parameter: 100.0

Kaggle’s public/private score: 10.563/9.366

加上一點點的 regularization 可能會有比較好的表現，但因為我這邊所使用的 model 並本身沒有產生 overffiting ，所以 regularization 前後差別不明顯。而如果加的太大，會使得整個 model 產生 underfitting ，導致誤差增加。

4.

Preprocessing: 一個月總共可以產生 20 * 24 - 1 筆 data，並從中去掉 BM25 <= 0 的 data

Features: 第九個小時的資料(除了 RAIN_FALL)加上 bias: 18 features

Method: Linear regression closed form solution