Homework 1 Report - PM2.5 Prediction

學號：b03901098 系級：電機四 姓名：王建翔

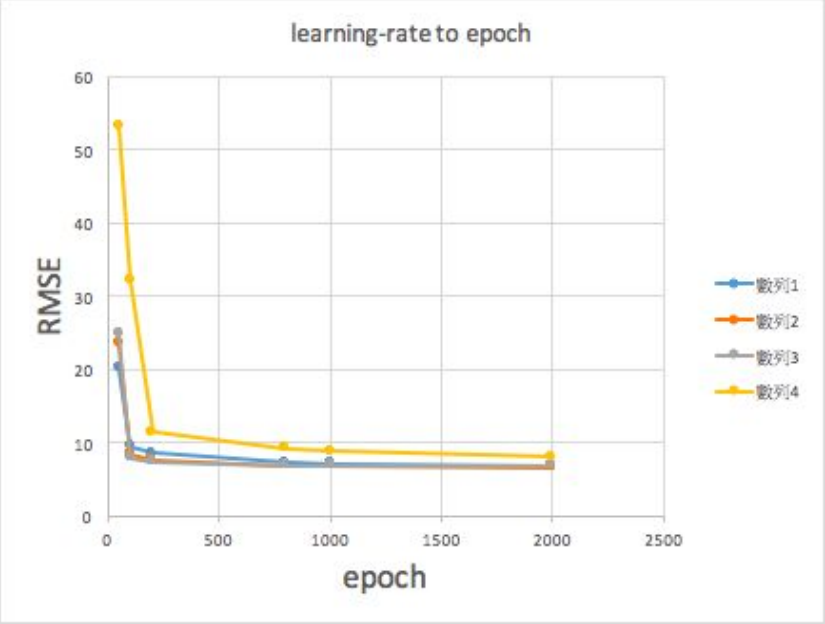
1. (1%) 請分別使用每筆data9小時內所有feature的一次項（含bias項）以及每筆data9小時內PM2.5的一次項（含bias項）進行training，比較並討論這兩種模型的root mean-square error（根據kaggle上的public/private score）。

資料的pre-process為將NR設為0，而兩方PM2.5為<=0者，以前後兩者取平均代替，若為邊界ex:為第九小時，以前一小時代替，epoch為1000，learning rate為0.3，並以十二月的資料為valid data，其他月份皆保留

	Average RMSE
Only PM2.5 features used	8.40693
All features used	9.07241

由此可知，在資料所提供的18種feature和PM2.5的相關性並不大，而單純只使用連續九小時的PM2.5作為feature即已經有大部分需要的資訊

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。
feature取法同於問題(4)，數列1為learning rate:0.3，數列2:0.5，數列3:0.7，數列4:0.1



由圖可知，learning rate較大時，能夠較快降低RMSE。而於同一learning rate達到最低RMSE時，隨著epoch的上升亦會收斂於相近的值，甚至再更高的epoch會造成overfitting，而使error增大。

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）。
epoch為1000，feature取法為使用PM2.5的連續九小時

lambda	Public	Private
100	8.40892	8.35953
10	8.40713	8.35820
1	8.40695	8.35807
0.1	8.40693	8.35805

做了regulization之後，可以避免overfitting的狀況，但是於取PM2.5為feature的情況中，並沒有使score有大的差異，我認為是該feature以epoch 1000來train並沒有太嚴重的overfitting狀況，所以做正規化之後才沒有太大影響。

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？）

這次的best主要是透過對Data的pre-process做的，先將data裡PM2.5,PM10,O3,NO,NO2,NOx取出，並且以每筆連續9小時取，其中，將波動起伏較大的8月以及12月做移除（有很多0），再對每個連續的九小時個別對PM2.5和PM10取倒數五個小時（影響較大），加上bias後的feature去做linear regresssion，weight的update則是透過adagrad的方式。