Homework 1 Report - PM2.5 Prediction

學號：r06725035 系級：資管所一 姓名：陳廷易

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）。

其他固定參數說明：(採用撰寫於 train.py 當中的 code)

- 在每跑完 1000 個 iterations 為一輪，會將 loss 過大的作為 outlier 剔除。每 1000 個 iteration 做為一輪剔除 outlier 後再跑一輪
- Initial bias = 1，initial weight = -0.01，learning rate =1，lr decay rate = 0.995，l2 regularization lambda =0.05，採用 adagrad
- 將資料中小於 0 的數值剔除

Feature 種類	所有 feature	僅 PM2.5 feature
Training Cost/ Valid Cost	5.25/8.73	4.37/ 7.58
Private Score/ Public Score	8.87/9.09	8.59/ 8.33

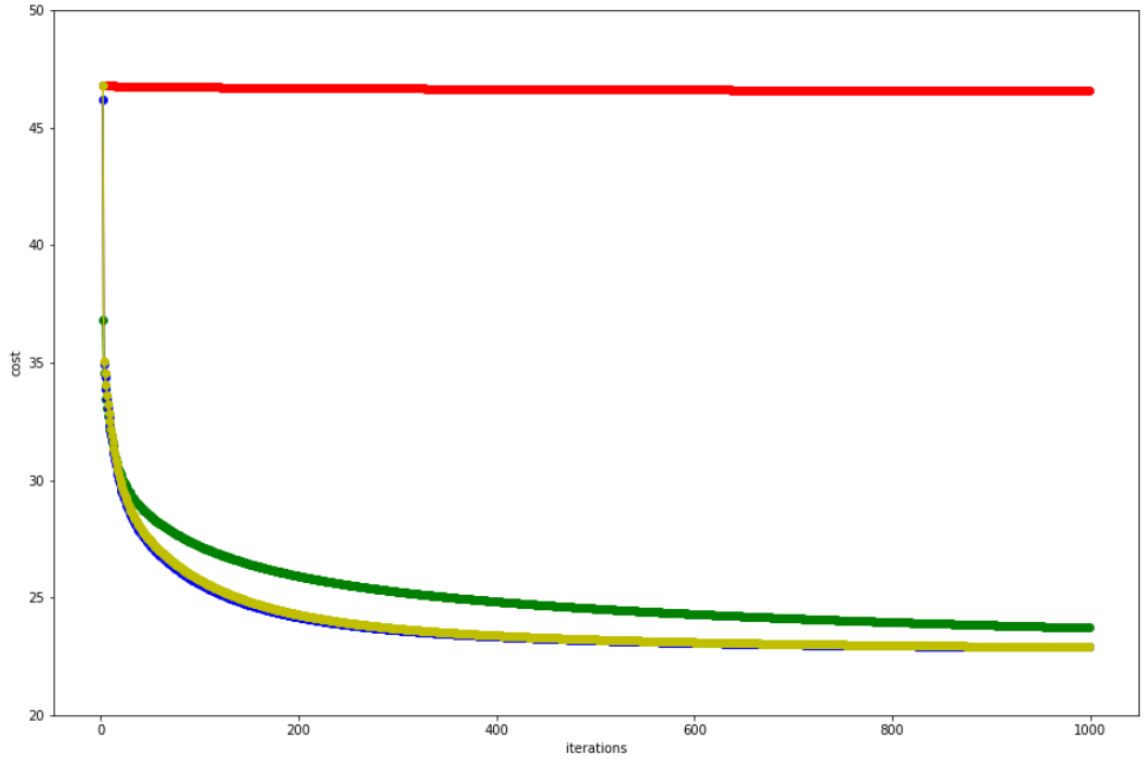
現象觀察與心得討論：

- 在進行所有 feature 訓練時，收斂較慢，每個 iteration 訓練時間較長，且效果較不佳。反而若僅取 PM2.5 feature 效果較好。
- 或許是因為 PM2.5 feature 在預測 PM2.5 時所具有的相關性較大，若訓練所有 feature 反而會使 model 難以找到較好的 function。

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。

其他固定參數說明：(採用撰寫於 report.py 的 code)

- 使用全部的 feature 並採用九小時資料來預測第十小時 PM2.5 值，無其他資料前處理步驟(僅將“NR”值替換為 0)
- 具 bias 但無加入平方項，採用 adagrad，並未使用 regularization
- 四種 learning rate 皆跑 1000 個 iteration
- 繪圖 Y 軸(cost)範圍設置於 20~50 區間中以方便觀察



▲最上方紅線 lr=1e-6；下方綠線 lr=0.01；下方黃線 lr=0.1；最下方藍線 lr=10

現象觀察與心得討論：

- Learning rate 在設置較小時，收斂速度較慢，甚至幾乎很難進步。
- Learning rate 設置越大，在初始的收斂速度就會越快。
- 在此 task 當中，或許是因為 Adagrad 特性的緣故，能自適應對不同參數會有不同的更新量，因此得以使所有參數的收斂速度盡可能一致，所以縱使 learning rate 較大也不太會有收斂到壞掉的情況。

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）。

其他固定參數說明：(採用撰寫於 train.py 當中的 code)

- 在每跑完 1000 個 iterations 為一輪，會將 loss 過大的作為 outlier 剔除
- 每 1000 個 iteration 做為一輪剔除 outlier 後再跑一輪
- initial bias = 1，initial weight = -0.01，learning rate =1，lr decay rate = 0.995
- 僅採用 PM2.5 的 feature 訓練，將資料中小於 0 的數值剔除
- 採用 adagrad

regularization parameter λ	1e-6	1e-3	0.1	0.99

Training Cost/ Valid Cost	4.31/7.59	4.29/7.59	4.28/7.58	4.38/7.64
Private Score/ Public Score	8.61/8.31	8.61/8.31	8.62/8.3	8.77/8.32

現象觀察與心得討論：

- 雖在此 task 中尚不明顯，但可發現 λ 值大雖可是 model 較為 smooth 約束力較長，但也會使模型變得較為難訓練，loss 也較難下降，甚至可能會有 under fitting 的情況發生，使表現較差。
- 經多次嘗試在僅用 PM2.5 的情況下，λ 設在 1e-3 到 0.05 之間會取得比較好的效果

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？）

在 best model 中，參考一些測量、預測 PM2.5 值的論文將較為無關的 feature 屏除(rainfall、temp、RH、CH4、ws_hr、wd_hr)，並對資料進行 scaling、normalize，同時把各項的離群值以前後小時的值進行內插法遞補，而後將此清洗過的資料丟入 fully-connected network 進行訓練。另一方面，將所有 feature 但僅去掉<0 的資料丟進去 Adaboost(DecisionTreeRegressor,n_estimator=200)、RandomForest(min_samples_leaf=3,min_samples_split=13,n_estimators=250)、XGBoost(booster='gbtree',learning_rate=0.1,min_child_weight=2,max_depth=2)進行訓練。最後選擇兩個 model validation 成績較好的(adaboost、RandomForest)，將 testing 出來的值進行 ensemble。