Homework 1 Report - PM2.5 Prediction

學號：b05901033 系級：電機二 姓名：莊永松

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：得分=>6.08400

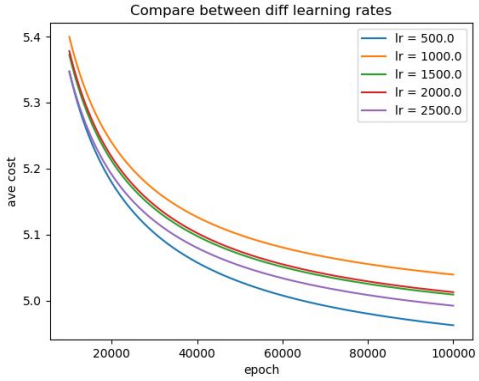
只用PM2.5：得分=>6.80953

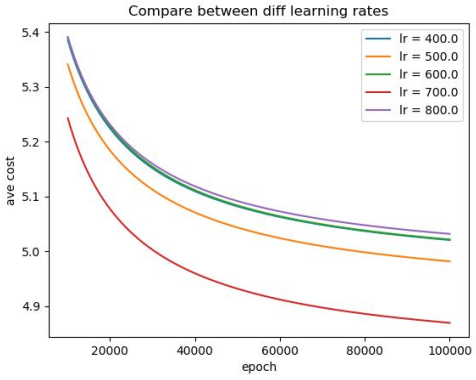
結論：用其他項data果然還是有幫助的，不能只用PM2.5

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。

使用adagrad `w = w - w_lr * (w_grad / np.sqrt(w_grad_sum))`

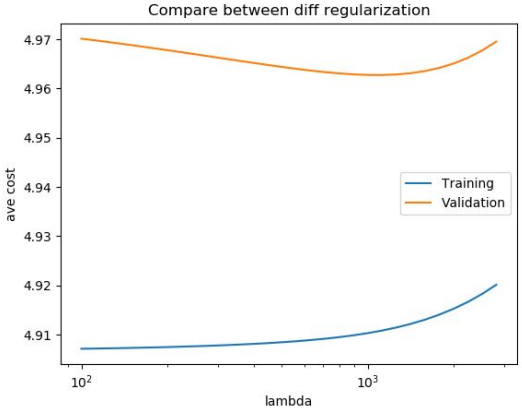
實驗結果，在w_lr=700時收斂最快

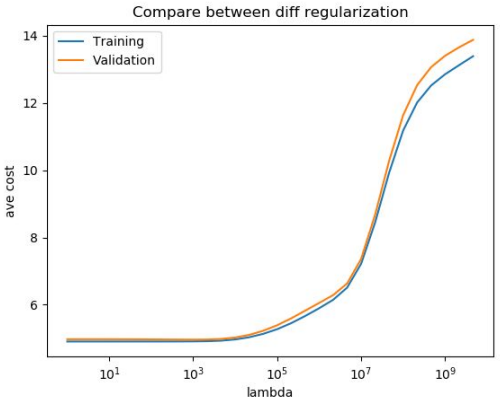




3. (1%) 請分別使用至少四種不同數值的regulization parameter λ進行training（其他參數需一致），討論其root mean-square error（根據kaggle上的public/private score）。

實驗後發現regulization效果不顯著，只有在lambda=1000附近略為有效果，但用lambda重新train之後在kaggle上public的分數並無提升(反而下降一小點)。



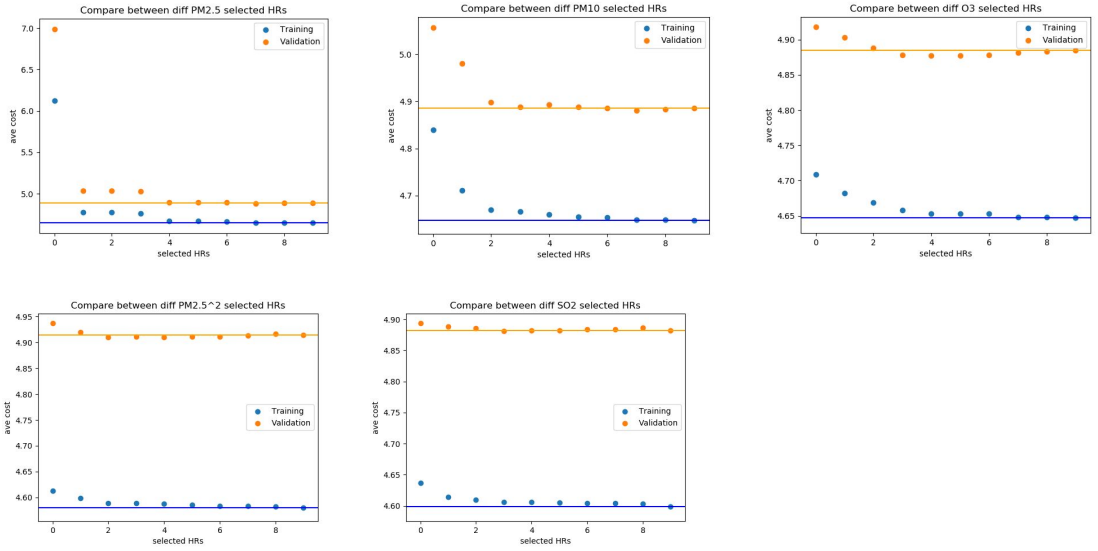


4. (1%) 請這次作業你的best_hw1.sh是如何實作的？（e.g. 有無對Data做任何Preprocessing？Features的選用有無任何考量？訓練相關參數的選用有無任何依據？）

best_hw1.sh 是用以前大一線代教的最小平方和矩陣公式：

```
w = np.matmul(np.matmul(inv(np.matmul(x.T, x)), x.T), y)
```

Features: 取feature方面我做了一個實驗，是把所有feature的一次項先都放進來，用迴圈對其中一個feature改取不同小時(0~9hr)，再去做validation再平均的結果，圖中的藍色橘色兩條線是沒有忽略任何feature的對照值(就是取了9hr)。結果：PM2.5、PM10、O3、PM2.5^2、SO2影響最大(取的小時數少會導致cost變大)。



實驗後又經過一些trial and error，決定保留feature：NO2: 2hr O3: 2hr PM10: 9hr PM2.5: 9hr PM2.5^2: 9hr Rain: 2hr SO2: 2hr

Data Preprocessing: 原本train的結果score一直在9附近，後來發現data裡面有很多地方缺資料補0，所以就手動把這些地方全都抓出來去掉，再用剩下的資料重新train，score就變成7附近了。

後來發現test.csv裡面也有一些資料是壞掉的(突然變0或是負數)，原本是用人工肉眼檢查+腦補data解決，但這樣助教測試code的時候我就無法preprocessing，會GG，因此後來改用pandas.Series的內插法函數pd.interpolation()來做那些爛資料的補齊，令外野手寫了一些rule針對test data中的一些極端狀況分開處理。

經過preprocessing後kaggle成績：5.96379