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）。

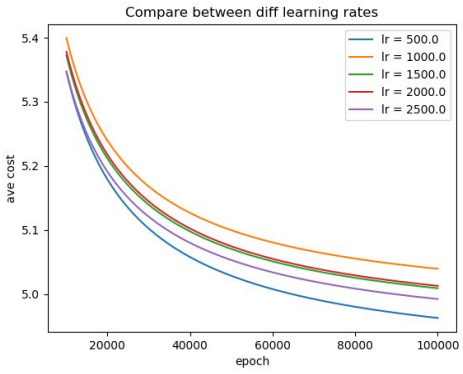
	public score	private score
所有feature一次項	6.08400	6.51652
只用PM2.5一次項	6.80953	7.15000

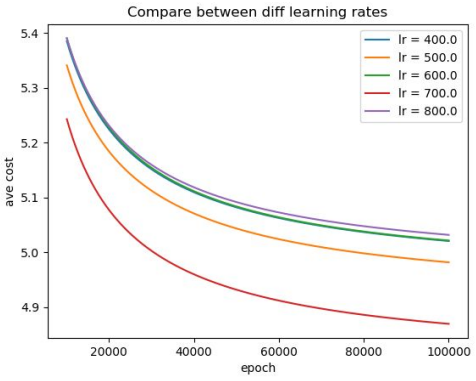
結論：用其他項data果然還是有幫助的，不能只用PM2.5阿!

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))`

測500~2500(間隔500)以及400~800(間隔100)實驗結果在w_lr=700時收斂最快

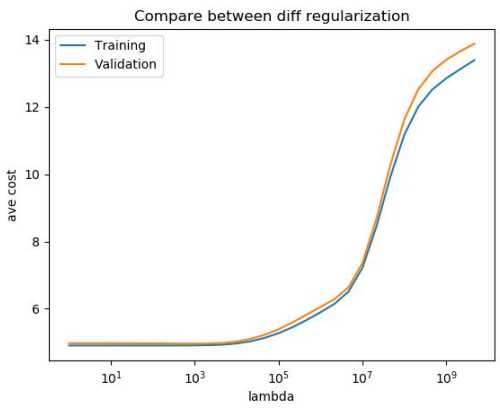


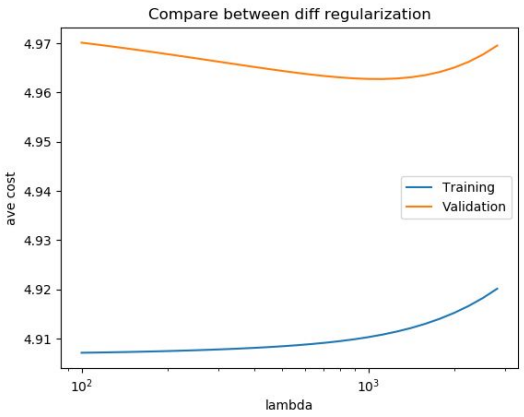


3. (1%) 請分別使用至少四種不同數值的regulization parameter λ進行training（其他參數需一致），討論其root mean-square error（根據kaggle上的public/private score）。

測了10^5附近等間距30個點以及10^2~10^3附近等間距30個點。

實驗後發現regulization效果不顯著，只有在λ=1000附近略為有效果(測得λ=1078效果最好)，用λ=1078重train之後在kaggle上分數為public: 5.98585 / private: 6.28816，原本同參數沒regulization的得分是public: 5.96379 / private: 6.25364，沒比較好。





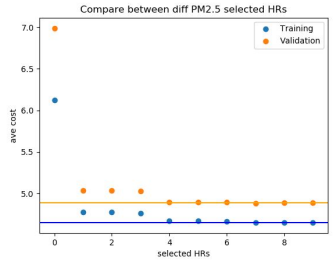
4. (1%) 請這次作業你的best_hw1.sh是如何實作的？（e.g. 有無對Data做任何Preprocessing？Features的選用有無任何考量？訓練相關參數的選用有無任何依據？）

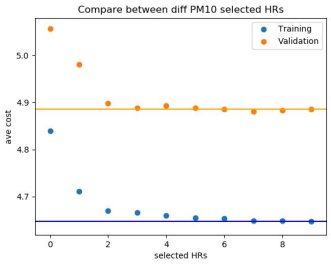
best_hw1.sh 是用以前大一線代教的最小平方和矩陣公式：

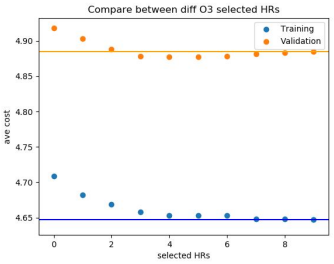
```
w = np.matmul(np.matmul(inv(np.matmul(x.T, x)), x.T), y)
```

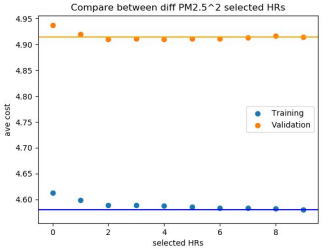
Data Preprocessing: 原本train的結果score一直在9附近，後來發現train.csv裡面有很多地方缺資料補0(感覺是儀器問題)，所以就手動把這些地方全都抓出來去掉，再用剩下的好資料重新train，score就變成7附近了。
後來發現test.csv裡面也有一些資料是壞掉的(突然變0或是負數)，原本是用人工肉眼檢查+腦補data解決，但這樣助教測試code的時候我就無法preprocessing，會GG。因此後來改用pandas.Series的內插法函數pd.interpolation()來做那些爛資料的補齊，另外我也手寫了一些rule針對test data中的一些極端狀況分開處理。在test data都有經過preprocessing之後kaggle成績就降到6附近了。

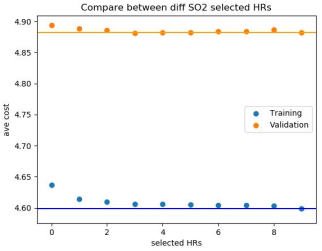
Features: 取feature方面我做了一個實驗，是把所有feature的一次項先都放進來，用迴圈對其中一個feature改取不同小時(距離現在最近的0~9hr)，再去做五份validation再總平均的結果，圖中的藍色橘色兩條線是沒有忽略任何feature的對照值(就是取了9hr)。
結果：PM2.5、PM10、O3、PM2.5^2、SO2影響最大(取的小時數少會導致cost變大)。











實驗後又經過一些trial and error，發現Rainfall及NO2雖然繪圖結果不明顯，但加入後仍有幫助，決定保留*NO2:* 2hr *O3:* 2hr *PM10:* 9hr *PM2.5:* 9hr *PM2.5^2:* 9hr *Rain:* 2hr *SO2:* 2hr
最後kaggle成績：public: 5.96379 private: 6.25364

BUT...在private score公布之後... **才發現，之前實驗時有一筆是完全照我作圖結果來取feature的，是看validation圖形最低點出現在哪來決定取幾小時(只取*O3:* 5hr *PM10:* 7hr *PM2.5:* 9hr *PM2.5^2:* 4hr *SO2:* 3hr，Rainfall及NO2都不取)，原本在public score表現不太好，是 6.28659 ，所以就被我忽視了，但private score竟然只有 6.13604 ，是裡面表現最好的，如果當初有選他做最後評分就好了...，結論：不要不相信實驗結果阿QAQ**