Homework1 Report – PM2.5 Predction

學號:r06921095 系級:電機碩一 姓名:吳睿哲 kaggle:r06921095_chris

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）。

Ans1:

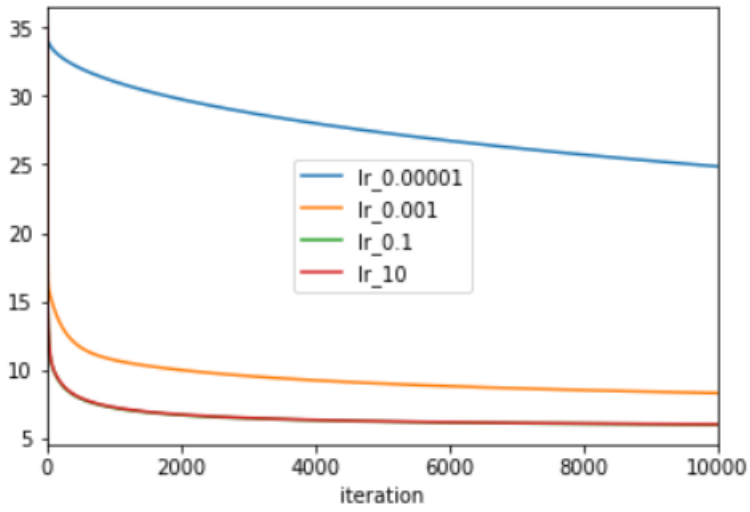
以下紀錄 testing public,private 的誤差,分成 9 小時內所有汙染源 9*18 個 features 及 pm2.5 一個 feature 的一次項,共兩組。

	public	Private
9*18+1(全部特徵)	6.62095	6.21026
9+1(僅 pm2.5)	6.74013	6.69617

取 9*18+1 個 features 特徵空間為度比較大,所以預測出來的 model 會比 9*1+1 個 feature 更接近真實的情況,且有許多 features 跟 pm2.5 的 correlation coefficient 絕對值是很高的,代表這些 features 跟 pm2.5 的高低是有關連性的,所以有用的 features 的 w 就會比較大,沒用的 features 的 w 就會比較小,所以運用所有的 features 預測出來的 model 應該是會比較準的。

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。

Ans2:

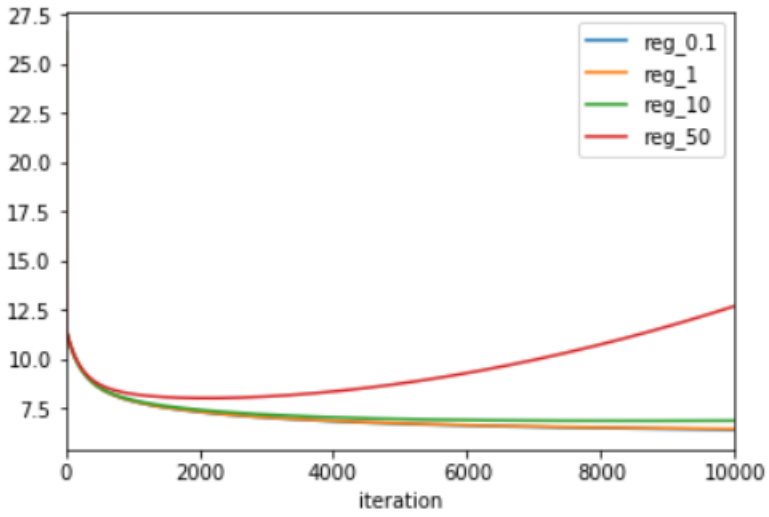


在其他條件都相同的情況下,learning rate 的大小會決定他最後收斂的值,以 0.00001 及 0.1 為例,剛開始 root mean square 的下降速度就差很多了,同樣都 iterate 一萬次的結果也差很多,由此可得知,learning 的大小會決定他一開始收斂的速度以及他最後的收斂結果。

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）。

Ans3:

	Reg=50	Reg=10	Reg=1	Reg=0.1
Public error	8.19620	7.18368	7.27964	7.29122
Private error	7.76669	6.53928	6.59561	6.60364



regularization 的大小主要會影響到最後收斂的結果,在 training data 上 ,如果 regularization 調太大的話,最後的 mean square error 會比較大,甚至發散,在 testing data 上,也會是一樣的結果,但是如果小到一定的值以下,結果就不會差很多了。

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？）

Ans4:

這次的作業中,我覺得影響結果最大的應該是 preprocessing ,mean square error 可以差到 5 甚至以上,我對 pm2.5 及 pm10 preprocessing 的方法是一樣的,如果值小於 0 或是大於 300 的話,我就讓它等於前一天的值,這樣可以確保所有的值都會落在 1~300 之間,對於其他 features 的處理方式則是如果小於 0 的話就讓它等於前一天的值。

features 的選擇方式則是比較每個 features 之間的相關係數,如果大於 0.5 或者小於-0.5 的話,我就用它當 features,在許多的 try and error 之後,發現 train 出來的 model 中最好的結果是選擇 pm2.5 跟 pm10 這 2 個 features 的時候,所以我就拿這兩個 features 的一次像當我的 best.sh。

至於參數的選用我也是使用 try and error 的方式去做嘗試,以 learning rate 為例,我會把每 iterate 一次的 root mean square error 都 print 出來,如果爆掉的話,我就會

把 learning rate 調小一點,如果更新太慢的話,我就會把 learning rate 調大一點,最後找到一個最適合的參數。