Homework 1 Report - PM2.5 Prediction

學號：b04901067 系級：電機三 姓名：陳博彥

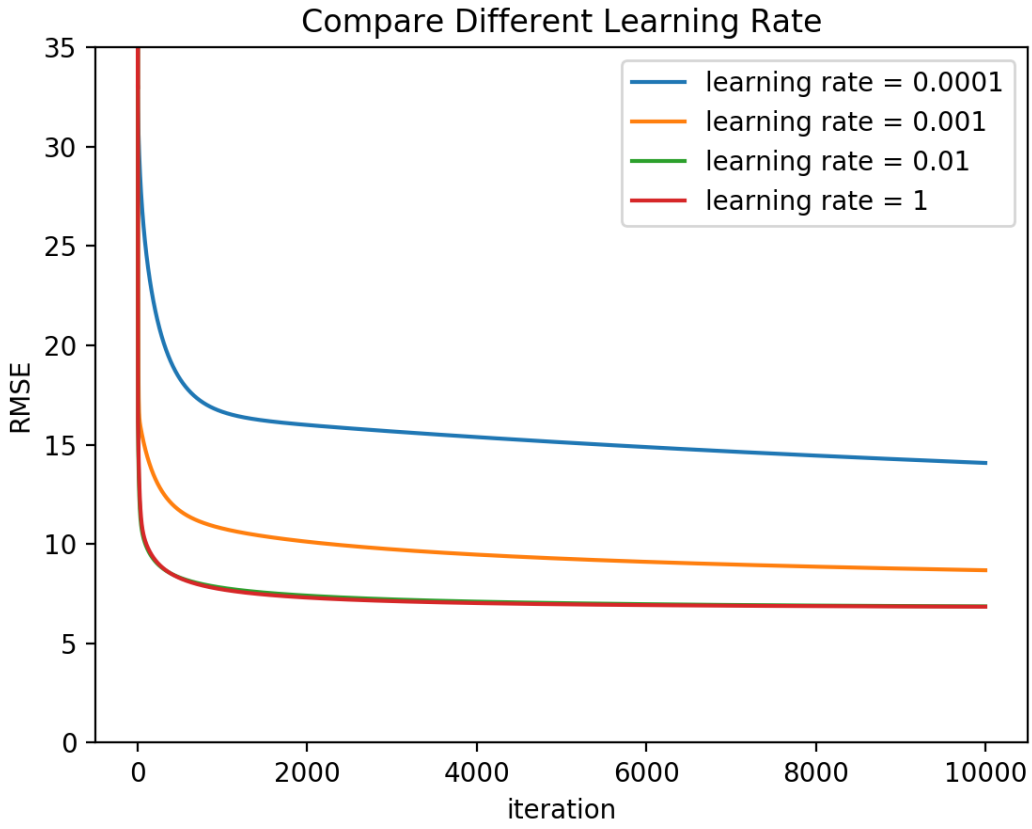
1. (1%) 請分別使用每筆data9小時內所有feature的一次項（含bias項）以及每筆data9小時內PM2.5的一次項（含bias項）進行training，比較並討論這兩種模型的root mean-square error（根據kaggle上的public/private score）。

注：我使用pseudo inverse的做法做這題，且有刪去training data中的極端值，詳情可以看第四題

	public score	private score
所有feature	6.20251	6.45000
只用PM2.5	6.83272	7.22518

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。

注：我有用adagrad



如圖，我採用初始learning rate分別為0.0001、0.001、0.01、1來做linear regression。
四者在第一個iteration中的training cost大約都為32左右，但是第二個iteration都會爆炸到一個非常大的數字（和learning rate呈正相關），之後又急速下降，最終應會降到六點多。
由圖可見，learning rate越大收斂越快，learning rate設為0.0001和0.01，training cost都來不及在一萬次iteration中收斂，而learning rate設為0.01和1效果幾乎一模一樣，在圖表中幾乎完全重疊（綠色的被紅色蓋掉了）。若將learning rate再調大（10、100、1000），其圖形都與0.01和1的重疊。

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）。

lambda	public score	private score
10000	7.89845	7.63715
1000	7.87273	7.60911
100	7.87211	7.60341
10	7.87228	7.60276
0	7.87230	7.60269

由於我是用linear model，lambda值對於分數的影響不大，除非加到一萬以上的數量級才會對分數有比較明顯的影響，且是負面影響。實驗得知，lambda設在0~100之間都可以。
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？）
我的best_hw1.sh主要做了以下幾件事：
(1)刪去極端值：
將PM2.5數據中，超過200的數據點刪除，改成離它最近且數值小於200的數據點。因為經過觀察得知，這些數據是自然界不可能發生的（可能是儀器有問題，或者助教亂加的？），且偏離正常數據太多，會大幅影響training的準確度。在testing時也會對testing data做相同處理。
(2)刪去小於零的PM2.5值：
理論上PM2.5含量不應小於零，所以我把小於零的數據，改成離它最近且大於等於零的數據。在testing時也會對testing data做相同處理。
(3)只採用與PM2.5相關係數大於0.25的數據：
我先用pandas算出各個數據之間的相關係數，發現只有CO、NO2、PM10、PM2.5四者與PM2.5的相關係數大於0.25。雖然，把18種數據都拿來用，training cost會最小，但是只採用以上四種數據，testing cost最小，最後也採用此作法。
(4)用pseudo inverse直接求出weight和bias：

其實所有training的過程，只要用pseudo inverse就可以一行code搞定了

```
w = np.dot(np.linalg.pinv(x),y)
```

根據線性代數中pseudo inverse的原理，這樣可以產生w，使得loss的絕對值最小，這個w就是一般training過程中最終將收斂到的那個值，但是這樣算保證可以達到最低點，且計算時間不到半秒。