Homework 1 Report - PM2.5 Prediction

學號：b03505052　系級：工海四　姓名：李吉昌

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	Private
所有 feature	8.34404	9.16944
只有 PM2.5	8.68333	9.49128

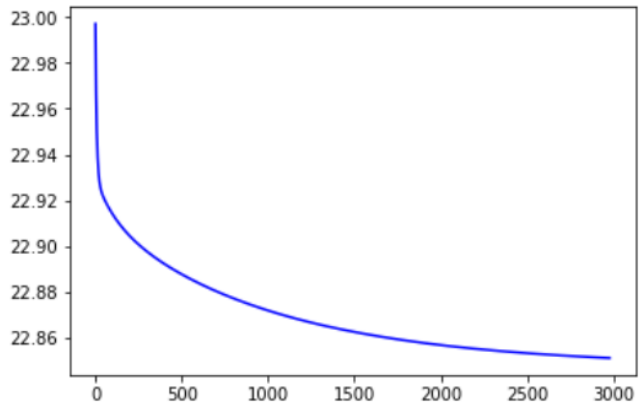
Score，也就是 RMSE，為使用所有 feature 的結果較好，其原因個人推測為如果只有 PM2.5 作為 feature 的 model 過為簡單，而反映實際 PM2.5 的所需要考慮的參數不只是只有 PM2.5，相較於只有 PM2.5 的 model，考慮全部 feature 比較 fit 真正的結果，但以自己最佳的分數來比較的話，稍有 overfitting 的情況

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。

Iterator 的次數設定在 3000 次，從第 20 圈觀察其收斂過程並記錄最小值

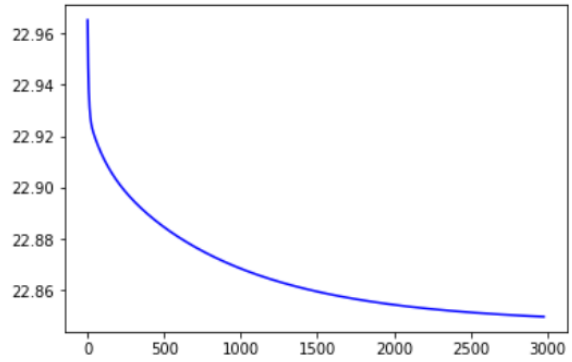
Learning rate = 10：

```
22.85104260436914
```



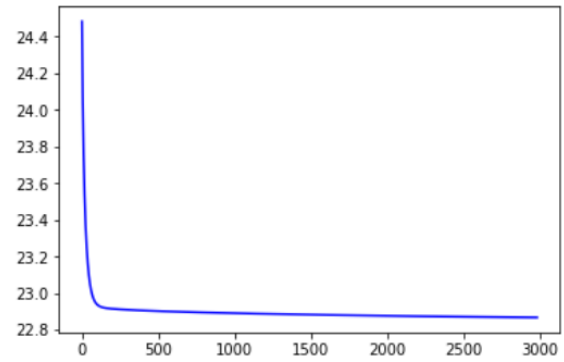
Learning rate = 1：

22.849786966274397



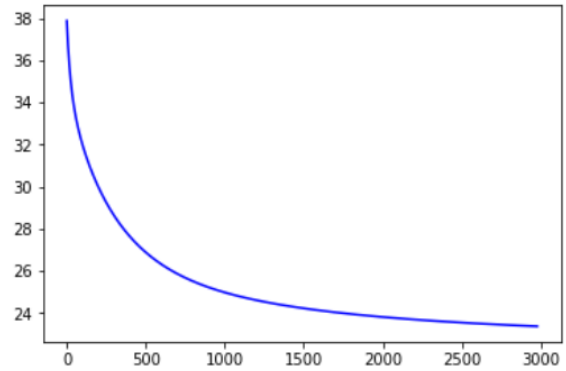
Learning rate = 0.1：

22.86716998103664



Learning rate = 0.01：

23.35888216038887



從 learning rate 為 0.1 開始的時候，第二十圈的 RMSE 很明顯較大，下降的幅度顯著偏低，

另外，若 learning rate 約略在 5~10 時，有可能到達的 valid loss 最低點會比其他量級還低(例如 0.1 或是 100)，個人推測其原因乃實際在做 gradient descent 的時候前進方向為 train data 等高線的垂直方向，但實際 test data 的方向可能稍微偏差，而在適度前進的速度則有可能剛好先走到 test data 的最低點。

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）。

	Public	Private
10	7.86644	7.97288
1	7.86764	7.9734
0.1	7.86782	7.97347
0.01	7.86784	7.97348

Parameter 取 10，1，0.1，0.01 四種參數量級，其 RMSE 在參數值越來越大時有越好的 performance，這表示其 data 在考慮雜訊的程度越高時表現越好。

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 版本的實作與原本相同，對 data 做資料的處理有兩種版本，觀察其 data 後發現 pm2.5 以及 NO 系列等元素有負值，在打電話詢問觀測站人員之後發現其數值為負值時有些為記錄觀測失敗的 label，有些是沒有記錄但是指針偏差產生的誤差，總之，負值是一個失真的值，不得作為判斷實際 PM2.5 的依據，因為在資料處理的時候第一種版本是將本來負值的元素變號，第二種是將其資料放為前一小時的值(其用意在於較接近原本可能觀測到的數值)。

在 feature 選用上，選用 NO 以外且相關係數較高的元素，不選用 NO 的原因在於其元素數值變化比較不規律。

另外，其選用相關參數的依據除了觀察相關係數外亦有採用觀測站人員提供可能影響較大的元素選項，如下雨狀況，並結合自己多方嘗試比對 valid loss 的結果。