Homework 1 Report - PM2.5 Prediction

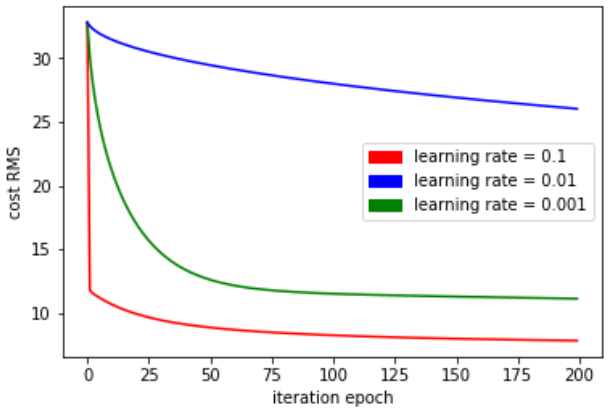
學號：r05543054 系級：應力二 姓名：劉禮榮

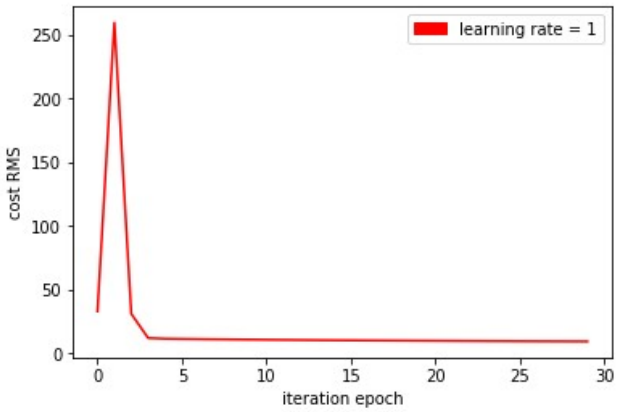
1. (1%) 請分別使用每筆 data9 小時內所有 feature 的一次項（含 bias 項）以及每筆 data9 小時內 PM2.5 的一次項（含 bias 項）進行 training，比較並討論這兩種模型的 root mean-square error（根據 kaggle 上的 public/private score）。

	All features	Only PM2.5
Public score	9.26921	7.51032
Private score	8.95216	8.38806

不論是 public 或是 private 都是”All features”的結果遜於 “Only PM2.5”，在常試過刪除各種其他污染源的後得到的結果，也不會比“Only PM2.5”的分數來的好。

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。





上圖列出 4 個數量級的 learning rate，橫軸及縱軸分別是跌代的次數級 loss function 的值，再大量條參的過程中（其他參數的圖位列於上），大多會在跌代到 200 次以內就趨於穩定收斂，只有在這段區間內 loss function 有顯著變化。

雖然如此，上列的 4 個 case 得到的 model 預測出的 PM2.5 結果再 kaggle 上的到的分數是差不多的（差距約 0.01）。

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	private
0.001	8.47620	8.41146
0.1	8.47620	8.41147
1	8.47626	8.41139
100	8.47167	8.40416

regularization 的效果非常不明顯。
上列列出 4 個數量級的 lambda，看不到任何明顯的改善（和 lambda = 0 相比）

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？）

1. 主要實作 gradient decent 來作 regression，額外預期可改善結果的實作包含 normalization, regularization
2. 對 data 的處理：將不合理的 data 修正掉，例如 PM2.5 的觀測值不應該是負的，另外還有 outlier 也應該做修正。修正的辦法是用簡單的內差法，觀察該項 feature 前後的資料做平均再賦值。異常的資料比例在 1%以下，但是修正後，在 kaggle 上的分數大幅改善。再現實中如果異常的資料比例到的 1%，或許內插法會不再適合，卡爾漫慮波器是可以嘗試的方法。