Homework 1 Report - PM2.5 Prediction

學號：r06725041 系級：資管碩一 姓名：彭証鴻

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）。

由下面的 root mean-square error 可知，若只選 PM2.5 單一個 feature 作預測，有 under-fitting 的現象，這代表 PM2.5 可能會被其他的 hidden factor 所影響，所以就得重新設計 model，將更多的 feature 放進去試試看。

- 所有 feature 一次項：
  - public score: 7.59318
  - private score: 7.40184
- PM2.5 的一次項：
  - public score: 8.45218
  - private score: 8.38641

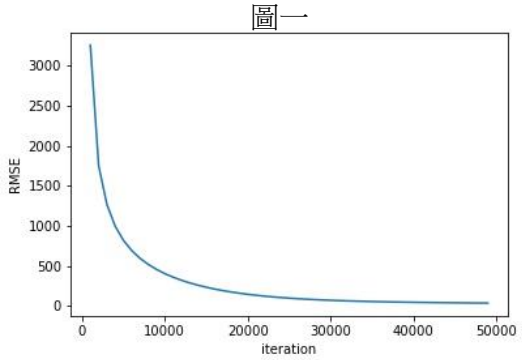
2. (2%) 請分別使用至少四種不同數值的 learning rate 進行 training（其他參數需一致），作圖並且討論其收斂過程。

我們由下面圖表可發現，當我們的 leaning rate 過大時(圖一)，得到的 loss 是相對大的(最後 error 為 30 以上)，learning rate 過小時(圖四)，趨近 local minimum 的速度較圖二、圖三慢很多。

下列四張圖，固定其 lambda 皆為 0.001，且 w, b 的 initial value 皆一樣：

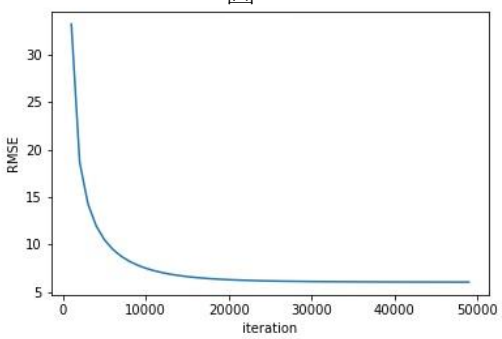
1. learning rate = 1000

圖一



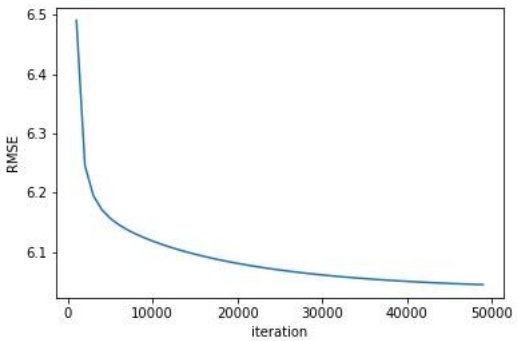
2. learning rate = 10

圖二



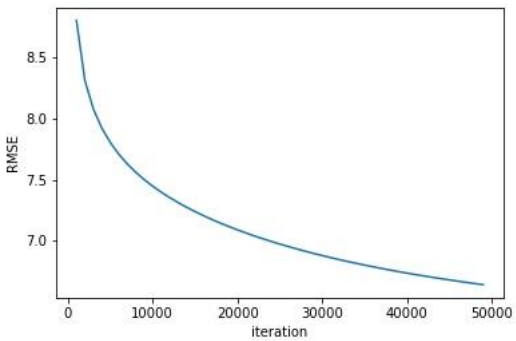
3. learning rate = 0.5

圖三



4. learning rate = 0.001

圖四



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）。

由下面的 root mean-square error 可知，當 λ 值越接近 0 時，我的 private error 就越低，這就代表比較能應付新的 data。

固定 learning rate、w 和 b 的 initial value：

1. λ = 100
   public score: 7.59794
   private score: 7.41322
2. λ = 1
   public score: 7.58183
   private score: 7.39421
3. λ = 0.1
   public score: 7.58130
   private score: 7.39355
4. λ = 0.001
   public score: 7.58124
   private score: 7.39347

4. (1%) 請這次作業你的 best_hwl.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？）

Feature 的選用，通常涉及豐富的 domain knowledge，所以有先上網找相關資料，並參考相關論文的實驗，最後選擇其中 8 個 feature 的一次項，作為我最終 model 的 feature。

此外，有發現我挑的 feature 中，有些 feature 有離譜的 outlier，我就將利用內插法將其值重新調整。