

# Homework 1 Report - PM2.5 Prediction

學號：r06521601 系級：土木碩一 姓名：黃伯凱

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）。

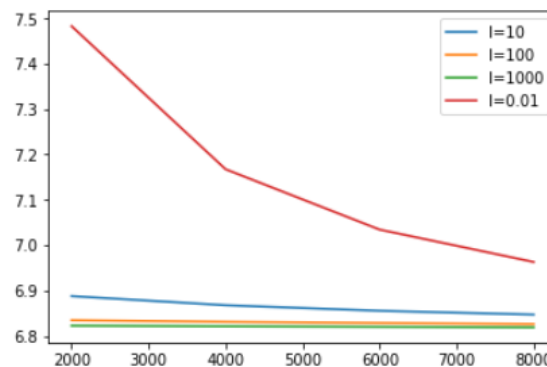
測試結果如下表所示：

	public	private
18 污染物	7.88325	7.58716
PM2.5	9.56307	9.69286

在不變動任何訓練參數的情況下，比較起來考慮 18 種污染物的表現較好，但也不是考慮越多種污染物越好，其實有些項目之間根本毫無關係，只是我未必知道，救我聽說，超強神人算命師群他們並非全部污染物皆考慮，應該有配對訓練出較好的組合。

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。

作圖如下：



我測試了 0.01、10、100 以及 1000 四種 learning rate，並每兩千擷取一點，從圖中可見，learning rate 越小即步伐越緩慢，需要花較多的時間收斂，而 learning rate 越大收斂速度也快，但相對也有可能遇到因步伐太大一直在谷間震盪，無法進到谷底的現象。

3. (1%) 請分別使用至少四種不同數值的 regularization parameter  $\lambda$  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）。

測試結果如下表所示：

lambda	public	private
0	7.88325	7.58716
10	7.88318	7.58722
100	7.88259	7.58778
1000	7.88006	7.59286

一樣在不改變任何訓練參數僅變動 lambda 參數的情況下，整體呈現 lambda 越大 public 表現越好，但 private 卻表現越差，regularization 因為讓曲線變得更加平滑，也能降低 overfitting 的現象，但我的 lambda 參數其實是按照林軒田老師推導的公式下去計算，並且盡可能讓  $(1 - \text{learning} * \text{lambda} / \text{adagrad})$  趨近於 1。

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？）

一開始單純使用 gradient descent 時，準確度並不是相當高，加上 adagrad 之後就過了 simple baseline，漸漸地開始嘗試把老師投影片中說到的技巧加到 code 中，最後最佳的成果是我將 train data 中有許多離譜的值剔除(包括爆表的 pm2.5 以及一大堆連續好幾小時 18 個污染物數值都是 0)，最後再對讀進來的 train data 及 test data 作 feature scaling(我使用 standard deviation normalization, 並且以 train data 為標準)，最終成果有突破 strong baseline，差點沒哭出來。

這次作業總結發現，train data 才是真正的訓練關鍵，品質的好壞會影響訓練成果太多，這也是在 fb 社團有人提出這個疑問時，助教回覆他說，這就是實務上會碰到的狀況。我看到真的嚇歪，老師跟助教真的很用心在這門課上，不單教我們原理，還悄悄的把實務會碰到的狀況塞進作業中，真的很感謝啊!以後到了業界一定會注意這個問題。