Homework 1 Report - PM2.5 Prediction

學號：b0491020　系級：電機三　姓名：解正平

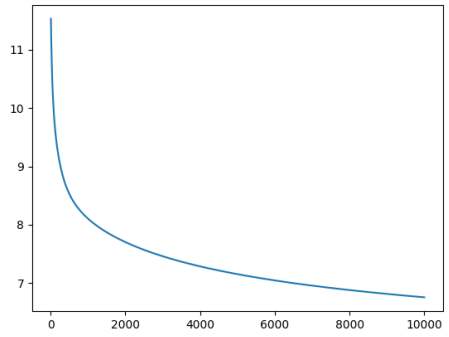
1. (1%) 請分別使用每筆 data9 小時內所有 feature 的一次項（含 bias 項）以及每筆 data9 小時內 PM2.5 的一次項（含 bias 項）進行 training，比較並討論這兩種模型的 root mean-square error（根據 kaggle 上的 public/private score）。

	Public score	Private score
18 項 feature	8.36876	9.07412
1 項 feature	8.62401	8.85752

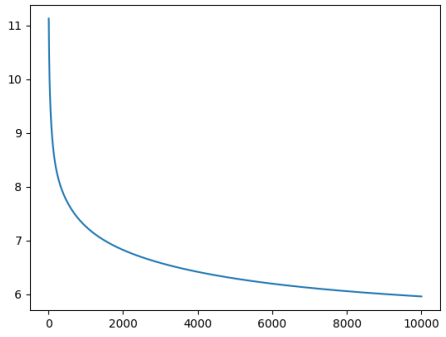
取所有 feature 的 model 可以看出它的 public score 較小，因為考慮比較多參數影響 PM2.5 自然可以較能準確預測，如果只拿前九小時的 PM2.5 feature 來 train 會造成很快就達到 gradient descent 最低點，可是卻很難讓第十小時準確得出相近結果，必須把與 PM2.5 相關的影響因素考慮進去，才可以有好的 model。
可是從 private score 就不一樣了，因為取 18 項 train 太久很容易造成 overfitting，造成 private 成績不盡理想，然而 1 項 score 就低可能是因為可以計算出大概的結果，可是往往跟前後因素無關，只是以前可能 PM2,5 這樣，未來也會這樣，考慮因素太少。

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。

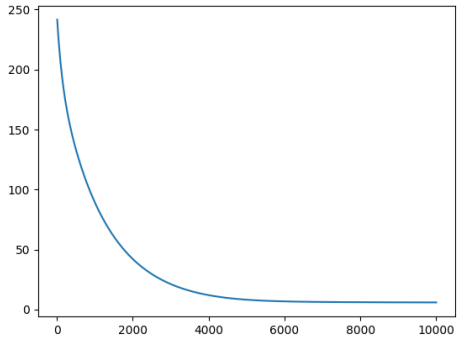
Learning rate = 0.0001



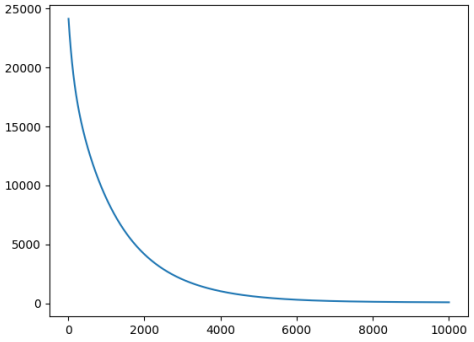
Learning rate = 1



Learning rate = 100



Learning rate = 10000



下兩圖可以明顯看出，當 Learning rate 太高的時候，收斂過程較慢，而且剛開始初始 error 較高，但因腳步也大並不會有很多影響，然而 model 必須需要經過多次 adagram 慢慢修正，足夠的 iterate 次數才能有較低的 training error。
上兩圖是 Learning rate 較小的數據，收斂速度明顯比 learning rate 高還要快，斜率較大，相同 iterate 次數下可以得到比較低的 learning rate，可是也並不是 rate 越低越好，看圖可以發現 rate = 1 的 error 相對小而且收斂速度亦較 rate = 0.0001 的快。
因此得出結論，learning rate 最佳要經過實驗得知，tradeoff 不一定說越大或越小越好。

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 score	Private score
λ = 0	7.39507	8.97930
λ = 1	7.39509	8.97928
λ = 10	7.39519	8.97915
λ = 100	7.39625	8.97791
λ = 1000	7.40421	8.97045
λ = 10000	7.42224	8.97009

從表中可以發現有 λ 的 public score 較沒有 regulization 的還要大，雖然說可以使圖形較為 smooth，使得 model 不太會因為 x 的值而有很大變化，減少一點 overfitting 的狀況，但是本次作業的影響不太明顯，可能是因為我本次 model 的參數並沒有很多，從 18 個參數只選 6 個，因此 λ 沒有很大影響力，不太改變 w 值。
然而 private score 有一點些微下降，可能因為 overfitting 沒那麼嚴重，還是可有效預測。
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？）
首先會先分析去掉不好的 data，比如說 PM2.5 的值為 0 或是超過範圍最大值 500 都會優先去掉，但除了 PM2.5 還有很多數據為 0 的狀況，這些數據我選擇捨棄以免造成很大的 error，因此利用前 9 小時的 feature 只要有出現 0，我就部會放進 training data 當中。接下來我又對風向的數據取 sin 值，比較可以符合相關的角度變化，這樣便可以有效減少 training error，然而還是無法使得 testing error 下降，因此我上網爬了一些 paper 了解 PM2.5 的影響因素，我選用 NO、NO2、NOX、PM10、PM2.5、WD_HR、WIND_DIREC 一共七筆數據來進行訓練，畢竟考慮過多參數很容易造成 overfitting。
在選用訓練參數的時候我做了一些實驗，首先 learning rate 不能太大也不能太小，所以我將 rate 設為 1，另外 iterate 的次數也很重要，太多次容易造成 overfitting，經過實驗我選擇次數為 50000，因為可以使 testing error 較低。