Homework 1 Report - PM2.5 Prediction

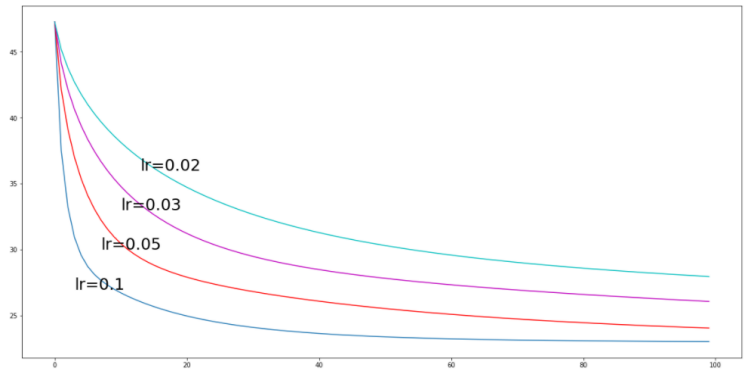
學號：b03902130 系級：資工四 姓名：楊書文

1. (1%) 請分別使用每筆 data 9 小時內所有 feature 的一次項（含 bias 項）以及每筆 data 9 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）。

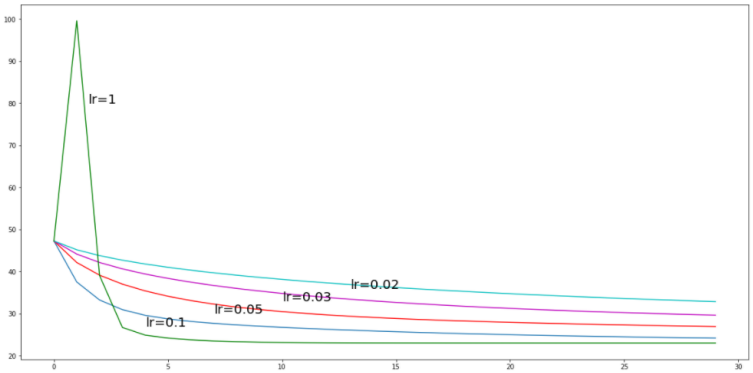
Model	Public	Private
全部 feature 九小時	9.18992	9.37728
僅 PM2.5 九小時	8.10318	7.82952

會有這樣的差距我覺得是因為 linear model 畢竟是相對簡單的 model，如果參數太多而參數間的關係又沒有很明確時，就容易 overfitting。不如取少量但都和 PM2.5 相關性高的 feature，反而會有更好的結果。我自己在選 feature 時也有碰到類似的情形，當我選同樣 feature，但取的小時數越少時，效果通常會越好。

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。



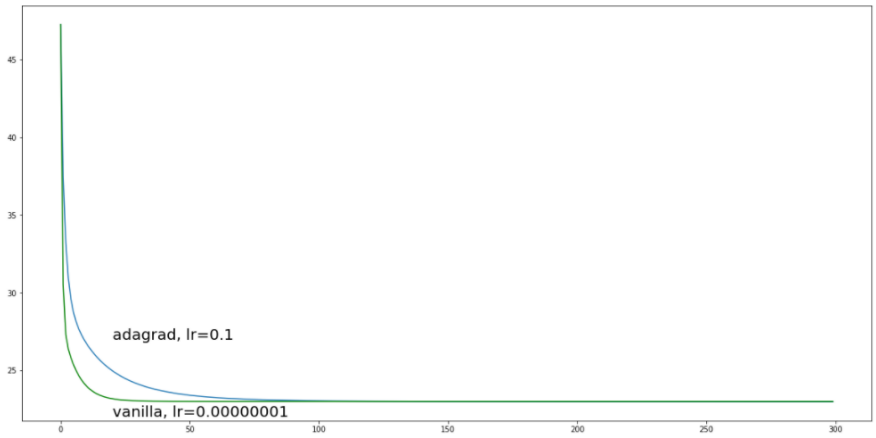
上圖都是使用 adagrad 的前 30 epoch。較大的初始學習率能走比較快。



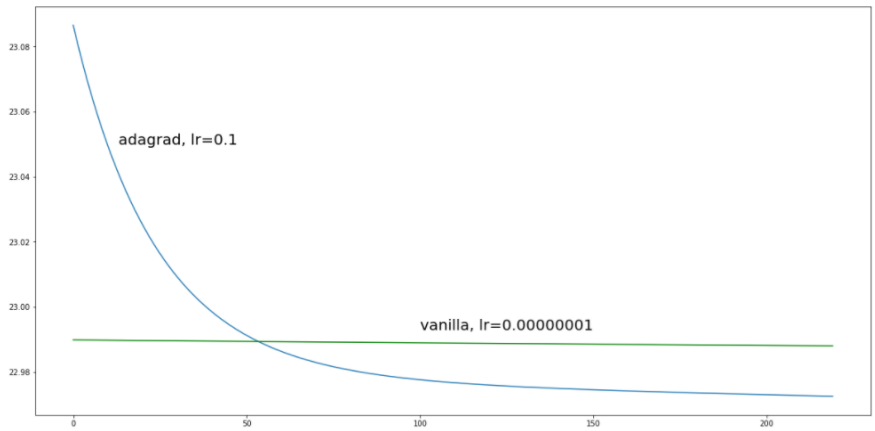
上圖也都是使用 adagrad 的前 30 epoch。當初始學習率太大(lr = 1)時，第一步就爆掉了，但由於 adagrad 能動態調整學習率讓下一 epoch 的步伐立刻變小，所以能穩定的走下去。如果換成傳統 vanilla gradient descent 的 fixed lr = 1 走法，loss 的變化如下：

```
47.265064149
2475329424.47
1.61847923368e+17
inf
inf
nan
```

太大的步伐會爆掉，且 loss 會一路飆升直到 overflow，第六個 epoch 就 nan 了。如果使用合理的 fixed lr 還是能穩定走下去，比方 lr = 0.00000001 如下圖 (1~300 epoch)



可以發現 fixed lr 甚至收斂得比 adagrad 快。但如果放大兩條線的交點來看：(80~300 epoch)



會發現 fixed lr 會因為固定的學習率而在最後階段走不下去(lr 太大)，另一邊 adagrad 則因為動態調整學習率而能在 train 的各個階段都能走得不錯。

3. (1%) 請分別使用至少四種不同數值的 regularization parameter λ 進行 training（其他參數需一致），討論其 root mean-square error（根據 kaggle 上的 public/private score）。

Regularization parameter	Public RMSE	Private RMSE
0	7.84244	7.83096
0.1	7.84258	7.83098
10	7.85680	7.83369
100	7.95370	7.86497
10000	8.37227	8.07771

從表格可以發現，Regularization 似乎完全沒有幫助。從 0.1 到 10000，parameter 越大則結果會穩定的越差。同時不只是 public 變差，private 也穩定的變差，因此這裡 regularization 對於 generalize 並沒有幫助。

4. (1%) 請問這次作業你的 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？）

Data preprocessing 部分我做了兩個處理，一是把 training data 中第十個小時 PM2.5 小於等於 0 的訓練資料砍掉，二是把剩下的負值全部變成正值。前者是因為我同學打去氣象局詢問後發現，負值通常代表儀器有問題而沒有量測到真正的數值，所以該筆訓練資料等於無效，硬要 fit 它可能會 overfitting。而負變正則是實驗的結果。原本試過把負的填成 0，但效果沒比較好。也試過把負值填成該筆訓練資料的前一筆資料中同 feature 的值，因為覺得如果儀器當時出問題沒測到，那前一個小時的值應該可以拿來逼近。但結果是在某些 CV 下變好，某些 CV 下變差，並沒有很穩定的 generalize。

Features 的選用則是依據 cross validation 的結果。為了保險起見我每次選用 feature 時都會用四種 CV 檢驗，在每種 CV 下該 feature 選法都會比較好才相信真的會比較好。而實驗結果和我原本想的滿不一樣，我原本以為把相關係數高的當參數一定會比較好，但這種選法只對一半。當我選了 PM2.5 前第一個小時(相關係數約 0.7)，效果會非常好。但如果又加上 PM2.5 前第二個小時(相關係數約 0.6)，反而沒有比加上 PM10 前第一個小時(相關係數約 0.4)好。另外我原本會直接取某些 feature 的前九個小時或前五個小時，後來卻發現當我小時數越取越少，CV 出來的結果竟然穩定的越來越好。因此後來 feature 的選用我就不太敢依靠感覺了，主要都是用 CV 選。最後是取 PM2.5、PM10、CO、SO2、O3 和 WIND_SPEED，都只有前第一個小時。