Homework 1 Report - PM2.5 Prediction

學號：r06521705  系級：土木系營管組碩一 姓名：陳思愷

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/privat score）。

|  |  |  |
| --- | --- | --- |
| RSME | ALL-FEATURE | ONLY-PM2.5 |
| PUBLIC | 8.90088 | 9.55819 |
| PRIVATE | 8.68778 | 9.69185 |

使用所有feature的模型預測精準度上不管是public或private都較為精準，

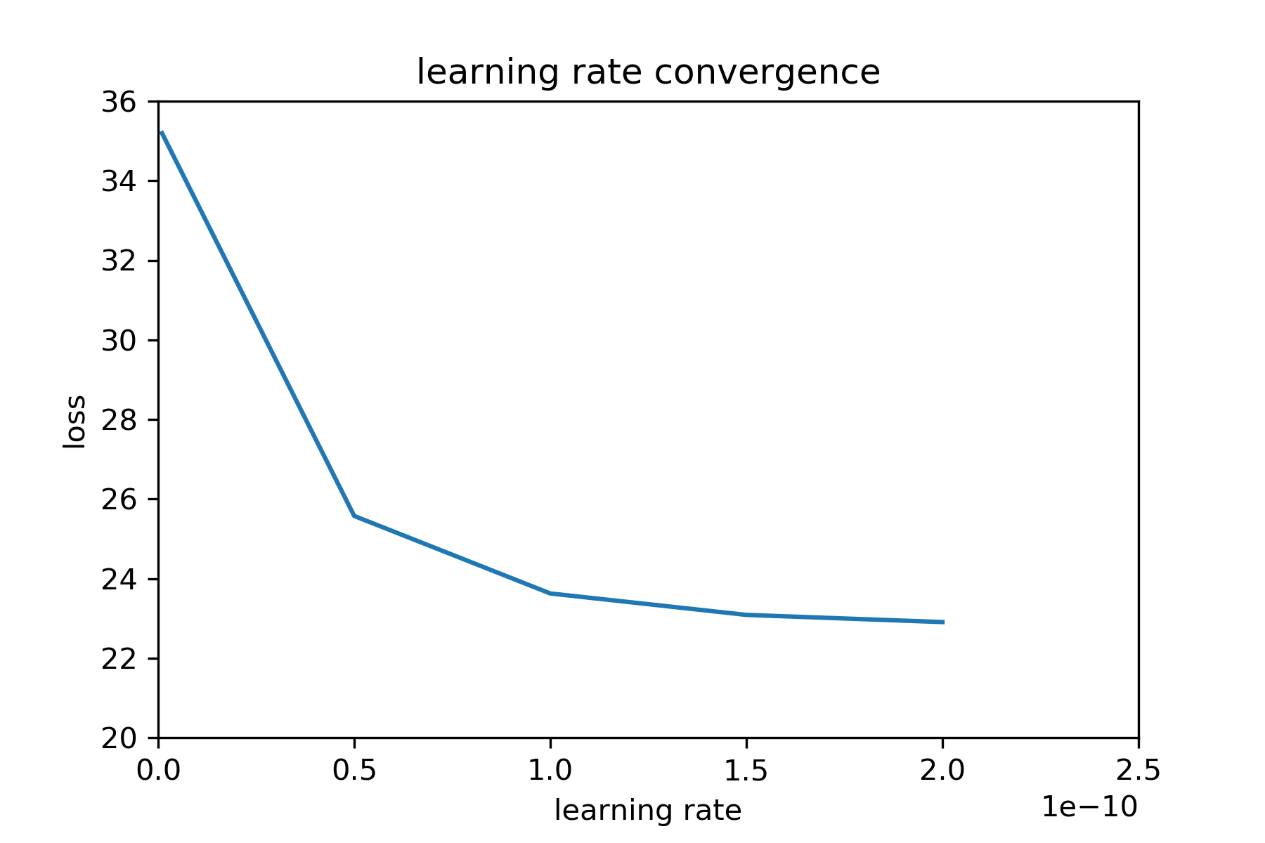
推測是18項feature當中，有實際對於pm2.5影響程度相當重大的因子，從一些

Pm2.5的相關文獻回顧當中可以發現，不管是NOx或SOx等等化合物，以及溫度、濕度、風速等等都會對pm2.5的濃度有一定程度上的影響，所以對於預測精準度來說，這些feature是需要考慮進去的。

1. (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 | 1e-12 | 5e-11 | 1e-10 | 1.5e-10 | 2e-10 |
| loss | 35.195 | 25.573 | 23.624 | 23.085 | 22.905 |

當繼續調整learning rate 以期達到更小的loss又不會使loss急速變大，最後會發現loss逐漸收斂到22.7左右



1. (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）。

(依照hw1.py去時做加入regulization parameter λ進行training)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | λ = 0.1 | λ=0.01 | λ=0.001 | λ=0.0001 |
| public | 8.83322 | 8.83322 | 8.83322 | 8.83322 |
| private | 8.68740 | 8.68740 | 8.68740 | 8.68740 |

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？）

有關於這次hw1\_best，有鑑於我唯一學過方式就是Gradient Descent，所以training方法來說並沒有和hw1不同的地方。

所以我便從feature的數量開始著手去試，feature越多，所涉及的維度越大，空間將變得更加寬廣，當資料量不是那麼足夠的時候，提升維度可能會造成資料的稀疏化，使其train出來的模型沒有辦法相當精準，所以我從取前9小時的資料改成取前2小時的資料來預測，精準度有所提升。