Homework 1 Report - PM2.5 Prediction

學號: R06942018, 姓名: 何適楷, 系級: 電信碩一

1 Problem 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	private
PM2.5	8.39306	8.38727
ALL	7.36575	7.40026

使用所有 Feature 所 train 出來的 RMSE 都比只有用 PM2.5 所得到的低，很直覺的原因就是除了 PM2.5 本身，其他物質也含有 PM2.5 的相關因素，所以所有 feature 下去 train 的資訊量多，自然得到的 RMSE 就會少。從物理的角度去看，排放 PM2.5 的來源不可能只排放純 PM2.5，一定伴隨其他物質，所以其他物質跟 PM2.5 一起擴散，自然會挾帶 PM2.5 的資訊了。

2 Problem 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。

我使用了六種不同的 Learning Rate 來檢視不同情況的收斂過程，並且都疊代 3000 次，下圖是疊代次數與 Loss 的作圖。使用方法是最原始的 gradient descent。

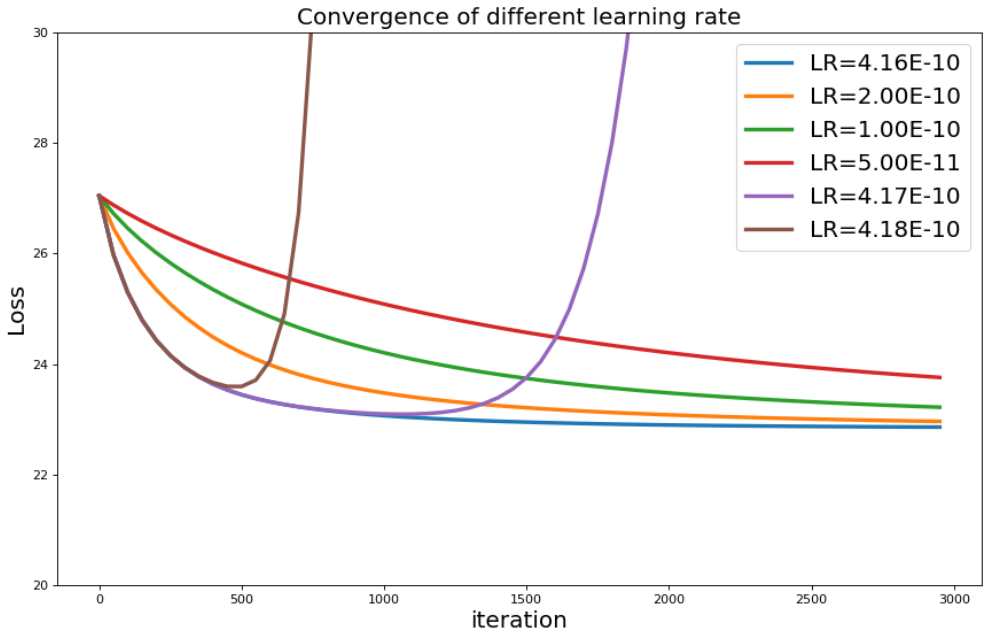

Figure 1: 疊代次數對 Loss 的變化

可以發現 $LearningRate = 4.16 \times 10^{-10}$ 趨近於臨界值，超過這個數值之後 ($LearningRate = 4.17 \times 10^{-10}$) 時，Loss 降到一定程度之後就會發散，另外，$LearningRate < 4.16 \times 10^{-10}$，都會收斂，但是收斂速度都比臨界值 $LearningRate = 4.16 \times 10^{-10}$ 慢。

3 Problem 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)。

linear model	public	private
$\lambda = 0$	7.39410	7.36525
$\lambda = 10$	7.39799	7.36239
$\lambda = 100$	7.61795	7.33514
$\lambda = 1000$	8.08020	7.80975
$\lambda = 10000$	10.40934	10.58393
$\lambda = 10000$	14.69773	14.36112

我使用的是 linear model，可以看出加了 λ 之後 private 和 public 的 RMSE 都是上升趨勢，代表原本的 model 的'accuracy' 已經非常，增加 λ 只是徒增'bias' 而已。

second order model	public	private
$\lambda = 0$	13.09529	9.52875
$\lambda = 10$	12.78238	7.36465
$\lambda = 100$	12.08264	7.62782
$\lambda = 1000$	9.75414	7.97926
$\lambda = 10000$	7.94495	8.17507
$\lambda = 10000$	9.27351	8.31054

這個 model 加了二次項，明顯的，一開始隨著 λ 的增加，RMSE 開始變小，代表一開始有 overfitting 的情形，但是加到一個程度之後，RMSE 又開始增加了，代表這個 model 的 bias 太大，已無法精確描述這些 data 了。

4 (collaborator: 陳致維, b04901165)Problem 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，我先使用所有原始的資料去做，得到

$$RMSE = 22.69823 \tag{1}$$
$$iteraion = 268000 \quad times \tag{2}$$
$$\tag{3}$$

進一步，我先做 normalzie，得到

$$RMSE = 22.65505 \tag{4}$$
$$iteraion = 136000 \quad times \tag{5}$$
$$public = 9.07673 \tag{6}$$

RMSE 幾乎都沒有動，但是疊代次數如同上課所說，normalize 之後 Loss function 比較會接近高維的球面，比起橢圓面更快能達到收斂。這時候 public score 得到 9.07 的分數，與自己的 train data 有明顯的差異，經過觀察資料，發現有些資料點明顯示測站量測錯誤，所以過濾以下資料：

1. PM2.5 超過 600
2. PM2.5 小於等於 0
3. Rainfall, NO, NO2, NOx 以外有超過 5 個以上的非正數

這時候我們得到 error 明顯下降

$$RMSE = 4.73346 \tag{7}$$
$$iteraion = 121000 \quad times \tag{8}$$
$$public = 8.51940 \tag{9}$$
$$private = 8.77867 \tag{10}$$

觀察資料發現，有些資料點可能跟 PM2.5 的濃度無關，所以我計算各個 feature 與 PM2.5 預測值的相關係數，把相關係數 <0.22 的資料濾掉，得到

	0	1	2	3	4	5	6	7	8
AMB_TEM	-0.1101	-0.1066	-0.1031	-0.1013	-0.0971	-0.0919	-0.0846	-0.0765	-0.0685
CH4	0.0858	0.0891	0.0975	0.1071	0.1124	0.1187	0.1276	0.1296	0.1257
CO	0.192	0.2008	0.2229	0.2481	0.2781	0.3065	0.322	0.3193	0.3035
NMHC	0.1437	0.1522	0.1698	0.1859	0.2061	0.2272	0.2505	0.2563	0.2437
NO	0.0693	0.0824	0.1048	0.1229	0.1418	0.154	0.1496	0.1332	0.1195
NO2	0.2035	0.2069	0.2171	0.2326	0.2546	0.2784	0.3024	0.3076	0.2912
NOx	0.1757	0.1845	0.2019	0.2213	0.2458	0.2681	0.283	0.2794	0.2615
O3	0.0312	0.0318	0.0277	0.0238	0.0226	0.0255	0.0341	0.0478	0.0648
PM10	0.3551	0.3663	0.3784	0.3918	0.404	0.4192	0.4406	0.4646	0.4768
PM2.5	0.2253	0.3208	0.4167	0.5123	0.5024	0.4034	0.4111	0.5207	0.7366
RAINFALI	-0.0466	-0.044	-0.0434	-0.0425	-0.0441	-0.0469	-0.046	-0.0418	-0.0378
RH	-0.0274	-0.0246	-0.0191	-0.0138	-0.0098	-0.0091	-0.0104	-0.0143	-0.0153
SO2	0.1786	0.1835	0.1892	0.1987	0.2125	0.2285	0.2481	0.2606	0.2588
THC	0.1323	0.1396	0.1558	0.1697	0.1851	0.2019	0.22	0.224	0.2146
WD_HR	0.0715	0.059	0.0698	0.073	0.0714	0.071	0.0657	0.0546	0.0576
WIND_DI	0.059	0.0457	0.0388	0.05	0.0542	0.0636	0.0631	0.0545	0.0556
WIND_SPI	-0.0609	-0.0648	-0.0661	-0.0683	-0.0735	-0.0752	-0.0796	-0.073	-0.065
WS_HR	-0.0542	-0.0549	-0.0545	-0.0571	-0.0604	-0.0654	-0.0657	-0.0593	-0.0549

Figure 2: 相關係數列表

$$RMSE = 4.83 \tag{11}$$
$$iteraion = 10000 \quad times \tag{12}$$
$$public = 8.77385 \tag{13}$$
$$private = 8.96263 \tag{14}$$

很失望地，結果竟然變差了，我的猜測是 linear regression 自然會把某些無相關的 data' 平均' 掉，所以加工的 data 反而會遺失一些資訊，導致 error 變差。而且，我的資料應該是 overfitting 了，所以我進一步進行 regularize，加上 $\lambda = 1000$

$$RMSE = 5.78 \tag{15}$$
$$iteraion = 50000 \quad times \tag{16}$$
$$public = 7.77929 \tag{17}$$
$$private = 7.53077 \tag{18}$$
$$\lambda = 1000 \tag{19}$$

結果變好了。另一方面，我覺得 PM2.5 可能與二次項有關，所以我增加了二次項得到

$$RMSE = 5.00318 \tag{20}$$
$$iteraion = 280000 \quad times \tag{21}$$
$$public = 7.55813 \tag{22}$$
$$private = 7.92696 \tag{23}$$
$$\lambda = 100 \tag{24}$$

因為二次項很容易 overfitting，所以我也實做了一個 cross validation，用來找適當了 λ

```
set 0    train: 6.61261 test 7.51206
set 1    train: 6.65750 test 7.28893
set 2    train: 6.82466 test 6.43016
set 3    train: 6.82599 test 6.57384
set 4    train: 6.80907 test 6.45316
set all  train: 6.56021
```

Figure 3: cross validation

因為事後才得知相關係數會不好，所以將相關係數去掉：

$$RMSE = 5.41866 \tag{25}$$
$$iteraion = 30000 \quad times \tag{26}$$
$$public = 7.54058 \tag{27}$$
$$private = 7.48302 \tag{28}$$
$$\lambda = 400 \tag{29}$$

結果更好了，另外放寬了塞選資料的條件把"Rainfall, NO, NO2, NOx 以外有超過 5 個以上的非正數" 中的 5 改為 10，而且聽說同學用 Linear 就做得比我好了，所以就嘗試看看，結果也變好了!

$$RMSE = 6.25529 \tag{30}$$
$$iteraion = 5000 \quad times \tag{31}$$
$$public = 7.28675 \tag{32}$$
$$private = 7.33939 \tag{33}$$
$$\lambda = 0 \tag{34}$$

只能說用盡心機，繞了一大圈，還是繞回 linear regression，從 public、private 分數來看，linear regression 還是略勝有二次項的 model，不過看到有 RMSE 小於 7 的同學，想必有其他 model 的方式吧，希望助教能在助教課多加指點了。