Machine Learning HW1

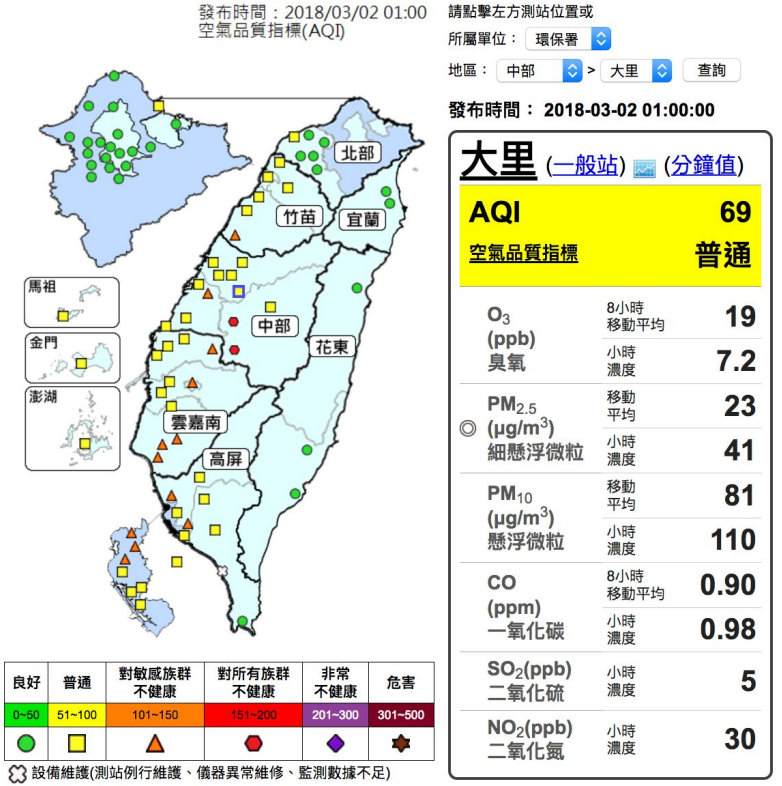
MLTAs
ntumlta2018@gmail.com

Outline

- HW1 Intro - PM2.5 Prediction
  - Tasks Description
  - Training/Testing Data
  - Sample Submission
- Kaggle
- Assignment
- Grading Policy
  - Github
  - Report
  - Others

Task Description

- 本次作業的資料是從行政院環境環保署空氣品質監測網所下載的觀測資料。
- 希望大家能在本作業實作linear regression以及其他方法預測出PM2.5的數值。

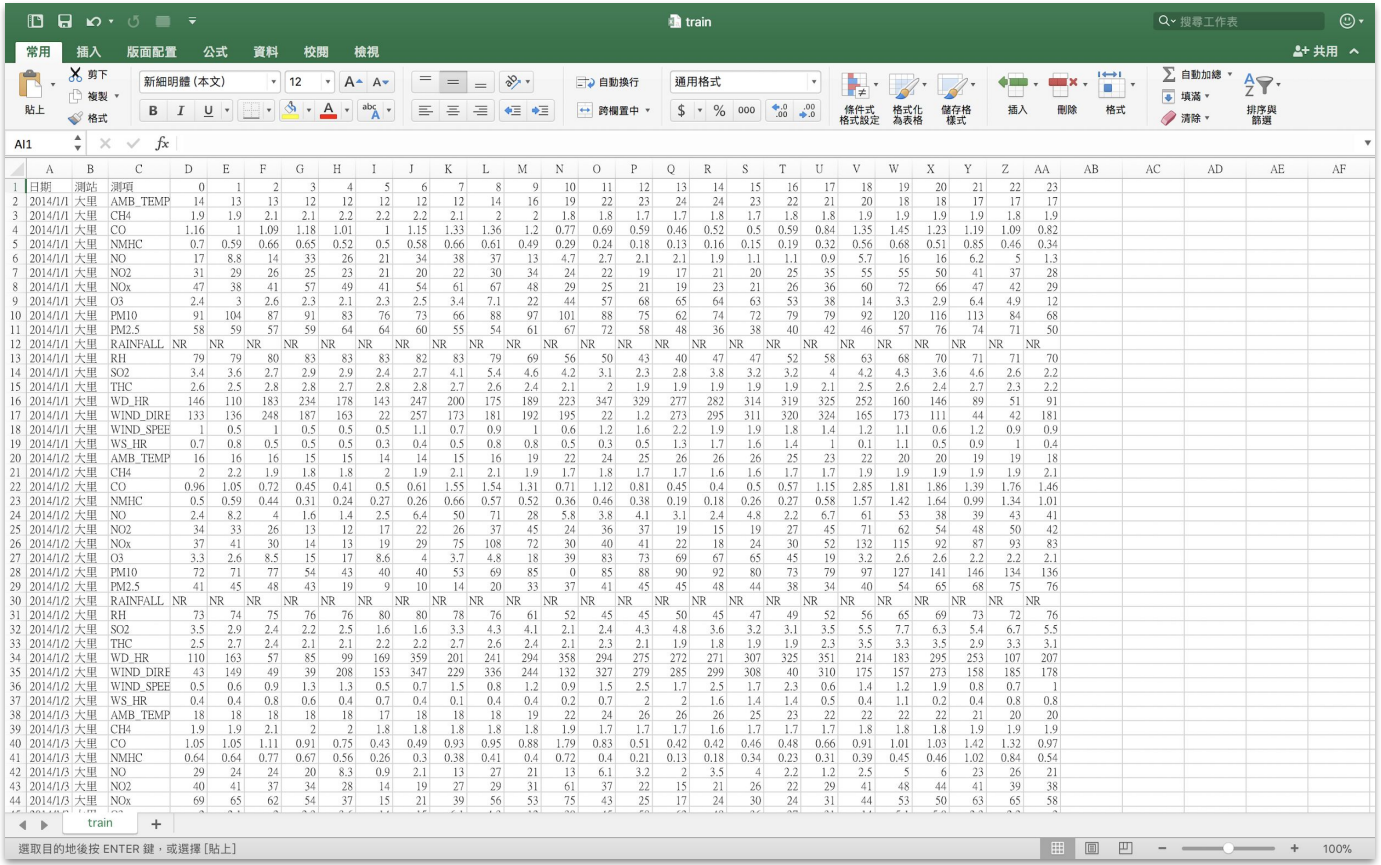


Data Description

- 本次作業使用大里站整年的觀測記錄，分成train.csv跟test.csv。
  - train.csv：使用data中每個月前二十天作為training data。
  - test.csv ：從剩餘的data中抽出260筆以連續10小時為單位且不重複的資料，前9小時所有觀測資料作為feature，第10小時的PM2.5值為answer。
- Data含有18項觀測數據 AMB_TEMP, CH4, CO, NHMC, NO, NO2, NOx, O3, PM10, PM2.5, RAINFALL, RH, SO2, THC, WD_HR, WIND_DIREC, WIND_SPEED, WS_HR。

到網站上爬出正確資料拿來做參考也將視為作弊，請務必注意！！！

Training Data



train

常用 插入 版面配置 公式 資料 校閱 檢視

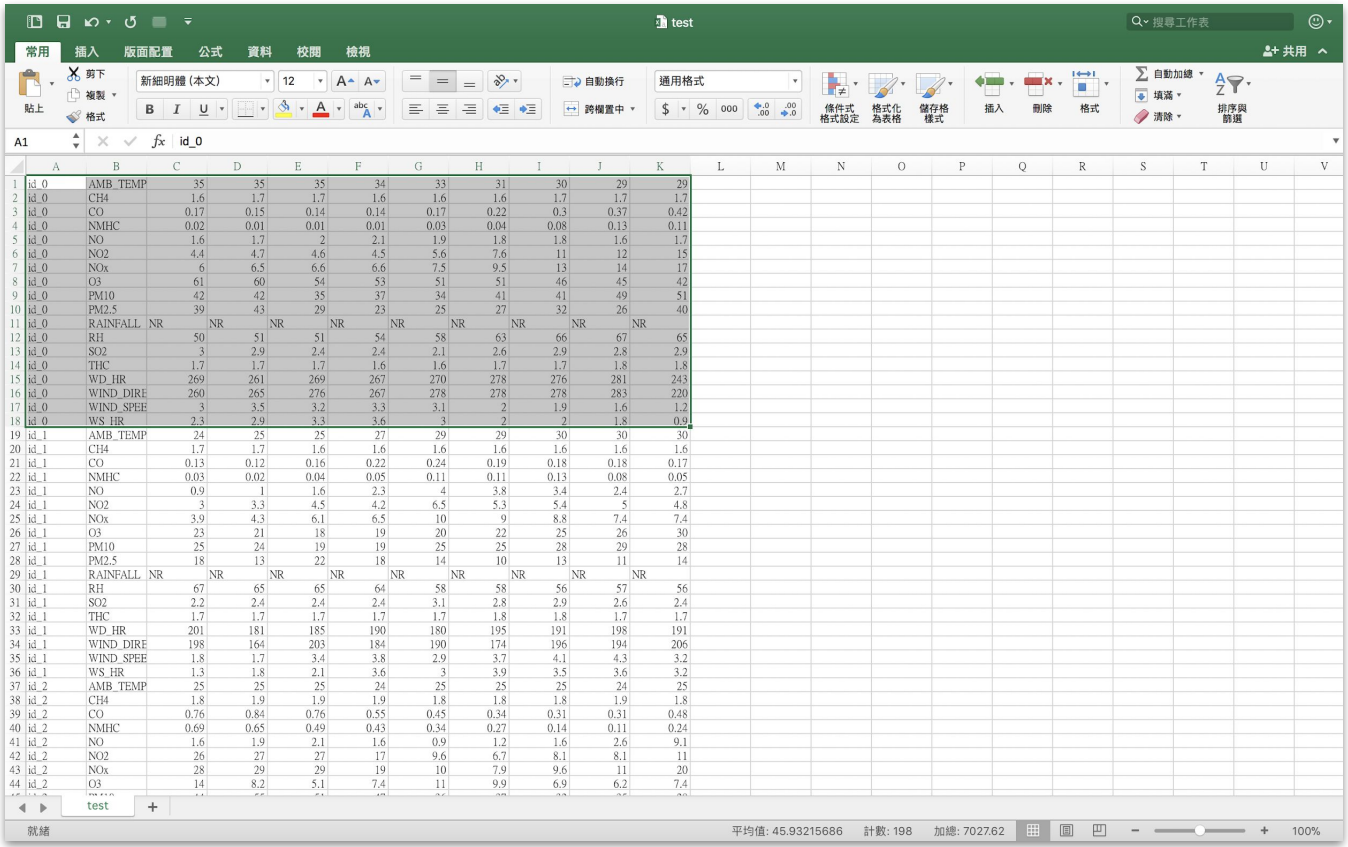
日期	測站	測項	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
2014/1/1	大里	AMB_TEMP	14	13	13	12	12	12	12	12	14	16	19	22	23	24	24	23	22	21	20	18	18	17	17	17
2014/1/1	大里	CH4	1.9	1.9	2.1	2.1	2.2	2.2	2.2	2.1	2	2	1.8	1.8	1.7	1.7	1.8	1.7	1.8	1.8	1.9	1.9	1.9	1.9	1.8	1.9
2014/1/1	大里	CO	1.16	1	1.09	1.18	1.01	1	1.15	1.33	1.36	1.2	0.77	0.69	0.59	0.46	0.52	0.5	0.59	0.84	1.35	1.45	1.23	1.19	1.09	0.82
2014/1/1	大里	NMHC	0.7	0.59	0.66	0.65	0.52	0.5	0.58	0.66	0.61	0.49	0.29	0.24	0.18	0.13	0.16	0.15	0.19	0.32	0.56	0.68	0.51	0.85	0.46	0.34
2014/1/1	大里	NO	17	8.8	14	33	26	21	34	38	37	13	4.7	2.7	2.1	2.1	1.9	1.1	1.1	0.9	5.7	16	16	6.2	5	1.3
2014/1/1	大里	NO2	31	29	26	25	23	21	20	22	30	34	24	22	19	17	21	20	25	35	55	55	50	41	37	28
2014/1/1	大里	NOx	47	38	41	57	49	41	54	61	67	48	29	25	21	19	23	21	26	36	60	72	66	47	42	29
2014/1/1	大里	O3	2.4	3	2.6	2.3	2.1	2.3	2.5	3.4	7.1	22	44	57	68	65	64	63	53	38	14	3.3	2.9	6.4	4.9	12
2014/1/1	大里	PM10	91	104	87	91	83	76	73	66	88	97	101	88	75	62	74	72	79	79	92	120	116	113	84	68
2014/1/1	大里	PM2.5	58	59	57	59	64	64	60	55	54	61	67	72	58	48	36	38	40	42	46	57	76	74	71	50
2014/1/1	大里	RAINFALL	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR
2014/1/1	大里	RH	79	79	80	83	83	83	82	83	79	69	56	50	43	40	47	47	52	58	63	68	70	71	71	70
2014/1/1	大里	SO2	3.4	3.6	2.7	2.9	2.9	2.4	2.7	4.1	5.4	4.6	4.2	3.1	2.3	2.8	3.8	3.2	3.2	4	4.2	4.3	3.6	4.6	2.6	2.2
2014/1/1	大里	THC	2.6	2.5	2.8	2.8	2.7	2.8	2.8	2.7	2.6	2.4	2.1	2	1.9	1.9	1.9	1.9	1.9	2.1	2.5	2.6	2.4	2.7	2.3	2.2
2014/1/1	大里	WD_HR	146	110	183	234	178	143	247	200	175	189	223	347	329	277	282	314	319	325	252	160	146	89	51	91
2014/1/1	大里	WIND_DIRE	133	136	248	187	163	22	257	173	181	192	195	22	1.2	273	295	311	320	324	165	173	111	44	42	181
2014/1/1	大里	WIND_SPEE	1	0.5	1	0.5	0.5	0.5	1.1	0.7	0.9	1	0.6	1.2	1.6	2.2	1.9	1.9	1.8	1.4	1.2	1.1	0.6	1.2	0.9	0.9
2014/1/1	大里	WS_HR	0.7	0.8	0.5	0.5	0.5	0.3	0.4	0.5	0.8	0.8	0.5	0.3	0.5	1.3	1.7	1.6	1.4	1	0.1	1.1	0.5	0.9	1	0.4
2014/1/2	大里	AMB_TEMP	16	16	16	15	15	14	14	15	16	19	22	24	25	26	26	26	25	23	22	20	20	19	19	18
2014/1/2	大里	CH4	2	2.2	1.9	1.8	1.8	2	1.9	2.1	2.1	1.9	1.7	1.8	1.7	1.7	1.6	1.6	1.7	1.7	1.9	1.9	1.9	1.9	1.9	2.1
2014/1/2	大里	CO	0.96	1.05	0.72	0.45	0.41	0.5	0.61	1.55	1.54	1.31	0.71	1.12	0.81	0.45	0.4	0.5	0.57	1.15	2.85	1.81	1.86	1.39	1.76	1.46
2014/1/2	大里	NMHC	0.5	0.59	0.44	0.31	0.24	0.27	0.26	0.66	0.57	0.52	0.36	0.46	0.38	0.19	0.18	0.26	0.27	0.58	1.57	1.42	1.64	0.99	1.34	1.01
2014/1/2	大里	NO	2.4	8.2	4	1.6	1.4	2.5	6.4	50	71	28	5.8	3.8	4.1	3.1	2.4	4.8	2.2	6.7	61	53	38	39	43	41
2014/1/2	大里	NO2	34	33	26	13	12	17	22	26	37	45	24	36	37	19	15	19	27	45	71	62	54	48	50	42
2014/1/2	大里	NOx	37	41	30	14	13	19	29	75	108	72	30	40	41	22	18	24	30	52	132	115	92	87	93	83
2014/1/2	大里	O3	3.3	2.6	8.5	15	17	8.6	4	3.7	4.8	18	39	83	73	69	67	65	45	19	3.2	2.6	2.6	2.2	2.2	2.1
2014/1/2	大里	PM10	72	71	77	54	43	40	40	53	69	85	0	85	88	90	92	80	73	79	97	127	141	146	134	136
2014/1/2	大里	PM2.5	41	45	48	43	19	9	10	14	20	33	37	41	45	45	48	44	38	34	40	54	65	68	75	76
2014/1/2	大里	RAINFALL	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR
2014/1/2	大里	RH	73	74	75	76	76	80	80	78	76	61	52	45	45	50	45	47	49	52	56	65	69	73	72	76
2014/1/2	大里	SO2	3.5	2.9	2.4	2.2	2.5	1.6	1.6	3.3	4.3	4.1	2.1	2.4	4.3	4.8	3.6	3.2	3.1	3.5	5.5	7.7	6.3	5.4	6.7	5.5
2014/1/2	大里	THC	2.5	2.7	2.4	2.1	2.1	2.2	2.2	2.7	2.6	2.4	2.1	2.3	2.1	1.9	1.8	1.9	1.9	2.3	3.5	3.3	3.5	2.9	3.3	3.1
2014/1/2	大里	WD_HR	110	163	57	85	99	169	359	201	241	294	358	294	275	272	271	307	325	351	214	183	295	253	107	207
2014/1/2	大里	WIND_DIRE	43	149	49	39	208	153	347	229	336	244	132	327	279	285	299	308	40	310	175	157	273	158	185	178
2014/1/2	大里	WIND_SPEE	0.5	0.6	0.9	1.3	1.3	0.5	0.7	1.5	0.8	1.2	0.9	1.5	2.5	1.7	2.5	1.7	2.3	0.6	1.4	1.2	1.9	0.8	0.7	1
2014/1/2	大里	WS_HR	0.4	0.4	0.8	0.6	0.4	0.7	0.4	0.1	0.4	0.4	0.2	0.7	2	2	1.6	1.4	1.4	0.5	0.4	1.1	0.2	0.4	0.8	0.8
2014/1/3	大里	AMB_TEMP	18	18	18	18	18	17	18	18	18	19	22	24	26	26	26	25	23	22	22	22	22	21	20	20
2014/1/3	大里	CH4	1.9	1.9	2.1	2	2	1.8	1.8	1.8	1.8	1.8	1.9	1.7	1.7	1.7	1.6	1.7	1.7	1.7	1.8	1.8	1.8	1.9	1.9	1.9
2014/1/3	大里	CO	1.05	1.05	1.11	0.91	0.75	0.43	0.49	0.93	0.95	0.88	1.79	0.83	0.51	0.42	0.42	0.46	0.48	0.66	0.91	1.01	1.03	1.42	1.32	0.97
2014/1/3	大里	NMHC	0.64	0.64	0.77	0.67	0.56	0.26	0.3	0.38	0.41	0.4	0.72	0.4	0.21	0.13	0.18	0.34	0.23	0.31	0.39	0.45	0.46	1.02	0.84	0.54
2014/1/3	大里	NO	29	24	24	20	8.3	0.9	2.1	13	27	21	13	6.1	3.2	2	3.5	4	2.2	1.2	2.5	5	6	23	26	21
2014/1/3	大里	NO2	40	41	37	34	28	14	19	27	29	31	61	37	22	15	21	26	22	29	41	48	44	41	39	38
2014/1/3	大里	NOx	69	65	62	54	37	15	21	39	56	53	75	43	25	17	24	30	24	31	44	53	50	63	65	58

train

選取目的地後按 ENTER 鍵，或選擇 [貼上]

100%

Testing Data



test

常用 插入 版面配置 公式 資料 校閱 檢視 共用

A1 id_0

	A	B	C	D	E	F	G	H	I	J	K
1	id_0	AMB_TEMP	35	35	35	34	33	31	30	29	29
2	id_0	CH4	1.6	1.7	1.7	1.6	1.6	1.6	1.7	1.7	1.7
3	id_0	CO	0.17	0.15	0.14	0.14	0.17	0.22	0.3	0.37	0.42
4	id_0	NMHC	0.02	0.01	0.01	0.01	0.03	0.04	0.08	0.13	0.11
5	id_0	NO	1.6	1.7	2	2.1	1.9	1.8	1.8	1.6	1.7
6	id_0	NO2	4.4	4.7	4.6	4.5	5.6	7.6	11	12	15
7	id_0	NOx	6	6.5	6.6	6.6	7.5	9.5	13	14	17
8	id_0	O3	61	60	54	53	51	51	46	45	42
9	id_0	PM10	42	42	35	37	34	41	41	49	51
10	id_0	PM2.5	39	43	29	23	25	27	32	26	40
11	id_0	RAINFALL	NR	NR	NR	NR	NR	NR	NR	NR	NR
12	id_0	RH	50	51	51	54	58	63	66	67	65
13	id_0	SO2	3	2.9	2.4	2.4	2.1	2.6	2.9	2.8	2.9
14	id_0	THC	1.7	1.7	1.7	1.6	1.6	1.7	1.7	1.8	1.8
15	id_0	WD_HR	269	261	269	267	270	278	276	281	243
16	id_0	WIND_DIRE	260	265	276	267	278	278	278	283	220
17	id_0	WIND_SPEE	3	3.5	3.2	3.3	3.1	2	1.9	1.6	1.2
18	id_0	WS_HR	2.3	2.9	3.3	3.6	3	2	2	1.8	0.9
19	id_1	AMB_TEMP	24	25	25	27	29	29	30	30	30
20	id_1	CH4	1.7	1.7	1.6	1.6	1.6	1.6	1.6	1.6	1.6
21	id_1	CO	0.13	0.12	0.16	0.22	0.24	0.19	0.18	0.18	0.17
22	id_1	NMHC	0.03	0.02	0.04	0.05	0.11	0.11	0.13	0.08	0.05
23	id_1	NO	0.9	1	1.6	2.3	4	3.8	3.4	2.4	2.7
24	id_1	NO2	3	3.3	4.5	4.2	6.5	5.3	5.4	5	4.8
25	id_1	NOx	3.9	4.3	6.1	6.5	10	9	8.8	7.4	7.4
26	id_1	O3	23	21	18	19	20	22	25	26	30
27	id_1	PM10	25	24	19	19	25	25	28	29	28
28	id_1	PM2.5	18	13	22	18	14	10	13	11	14
29	id_1	RAINFALL	NR	NR	NR	NR	NR	NR	NR	NR	NR
30	id_1	RH	67	65	65	64	58	58	56	57	56
31	id_1	SO2	2.2	2.4	2.4	2.4	3.1	2.8	2.9	2.6	2.4
32	id_1	THC	1.7	1.7	1.7	1.7	1.7	1.8	1.8	1.7	1.7
33	id_1	WD_HR	201	181	185	190	180	195	191	198	191
34	id_1	WIND_DIRE	198	164	203	184	190	174	196	194	206
35	id_1	WIND_SPEE	1.8	1.7	3.4	3.8	2.9	3.7	4.1	4.3	3.2
36	id_1	WS_HR	1.3	1.8	2.1	3.6	3	3.9	3.5	3.6	3.2
37	id_2	AMB_TEMP	25	25	25	24	25	25	25	24	25
38	id_2	CH4	1.8	1.9	1.9	1.9	1.8	1.8	1.8	1.9	1.8
39	id_2	CO	0.76	0.84	0.76	0.55	0.45	0.34	0.31	0.31	0.48
40	id_2	NMHC	0.69	0.65	0.49	0.43	0.34	0.27	0.14	0.11	0.24
41	id_2	NO	1.6	1.9	2.1	1.6	0.9	1.2	1.6	2.6	9.1
42	id_2	NO2	26	27	27	17	9.6	6.7	8.1	8.1	11
43	id_2	NOx	28	29	29	19	10	7.9	9.6	11	20
44	id_2	O3	14	8.2	5.1	7.4	11	9.9	6.9	6.2	7.4

test

就緒 平均值: 45.93215686 計數: 198 加總: 7027.62 100%

Testing Data

id_0	AMB_TEMP	35	35	35	34	33	31	30	29	29
id_0	CH4	1.6	1.7	1.7	1.6	1.6	1.6	1.7	1.7	1.7
id_0	CO	0.17	0.15	0.14	0.14	0.17	0.22	0.3	0.37	0.42
id_0	NMHC	0.02	0.01	0.01	0.01	0.03	0.04	0.08	0.13	0.11
id_0	NO	1.6	1.7	2	2.1	1.9	1.8	1.8	1.6	1.7
id_0	NO2	4.4	4.7	4.6	4.5	5.6	7.6	11	12	15
id_0	NOx	6	6.5	6.6	6.6	7.5	9.5	13	14	17
id_0	O3	61	60	54	53	51	51	46	45	42
id_0	PM10	42	42	35	37	34	41	41	49	51
id_0	PM2.5	39	43	29	23	25	27	32	26	40
id_0	RAINFALL	NR	NR	NR	NR	NR	NR	NR	NR	NR
id_0	RH	50	51	51	54	58	63	66	67	65
id_0	SO2	3	2.9	2.4	2.4	2.1	2.6	2.9	2.8	2.9
id_0	THC	1.7	1.7	1.7	1.6	1.6	1.7	1.7	1.8	1.8
id_0	WD_HR	269	261	269	267	270	278	276	281	243
id_0	WIND_DIRE	260	265	276	267	278	278	278	283	220
id_0	WIND_SPEE	3	3.5	3.2	3.3	3.1	2	1.9	1.6	1.2
id_0	WS_HR	2.3	2.9	3.3	3.6	3	2	2	1.8	0.9

?

Sample Submission

- 預測260筆testing data中的PM2.5值，將預測結果上傳至kaggle
  - Upload format : csv file
  - 第一行必須是 id,value
  - 第二行開始，每行分別為id值及預測PM2.5數值，以逗號隔開
- 範例格式：

Kaggle Info

- 請自行到kaggle創建帳號（務必使用ntu信箱）
- Link: https://goo.gl/CJxCLv
- Team Name: 學號_任意名稱（ex: b03901666_台大谷翔平）
- Maximum Daily Submission: 5
- Simple Bonus Deadline: 3/14/2018 23:59:59 (GMT+8)
- Kaggle Deadline: 3/21/2018 23:59:59 (GMT+8)
- Github Deadline: 3/22/2018 23:59:59 (GMT+8)
- Leaderboard上所顯示為public score，在Kaggle Deadline前可以選擇2份submission作為private score的評分依據。

- github url form: https://goo.gl/forms/4TTE24hLviFXaCqz2

Kaggle Baselines

- Public Leaderboard
  - 130 out of 260 from the testing dataset
  - Participants receive instant feedback about their performance.
  - Be sure not to overfit on the public leaderboard.
- Private Leaderboard
  - 130 out of 260 from the testing dataset
  - Remain unknown until the end of the competition.

作業規定 Assignment Regulation

- Only Python 3.5+ available !!!!
- 開放使用套件
  - numpy 1.13
  - scipy 0.19
  - pandas 0.21
- hw1.sh(train.py)必須實作linear regression，方法限定gradient descent。
- hw1_best.sh不限作法、不限套件（但有版本限制請注意）。
  - Tensorflow 1.4.0
  - Pytorch 0.3.0
  - Keras 2.0.8
  - Scikit-learn 0.19.0
- 若需使用其他套件，請儘早寄信至助教信箱詢問，並請闡明原因。

繳交格式 Handin Format

- Github Deadline: 3/22/2018 23:59:59 (GMT+8)
- 請注意github commit為local端之時間，務必注意本機的電腦時間設定，助教群將在deadline一到就clone所有程式以及報告，並且不再重新clone任何檔案
- 你的github上至少有下列四個檔案（格式必須完全一樣）：
  - ML2018SPRING/hw1/Report.pdf
  - ML2018SPRING/hw1/hw1.sh
  - ML2018SPRING/hw1/hw1_best.sh
  - ML2018SPRING/hw1/train.py
  - 請勿上傳dataset!!!
- 你的github上可能還有其他檔案：
  - e.g. ML2018SPRING/hw1/model.npy
- 注意!!!hw1.sh以及hw1_best.sh將只執行testing，請自行跑完training部分並且儲存相關模型參數並上傳至github

批改方式 Script Policy

- 助教在批改程式部分時，會執行以下指令：
  - bash hw1.sh [input file] [output file]
  - bash hw1_best.sh [input file] [output file]
  - [input file]為助教提供的test.csv路徑
  - [output file]為助教提供的output file路徑
  - E.g. 如果助教執行了bash hw1.sh ./data/test.csv ./result/ans.csv，則應該要在result資料夾中產生一個檔名為ans.csv的檔案
- hw1.sh與hw1_best.sh皆需要在3分鐘內執行完畢，否則該部分將以0分計算。
- 切勿於程式內寫死test.csv或者是output file的路徑，否則該部分將以0分計算。
- Script所使用之模型，如npy檔、pickle檔等，可以於程式內寫死路徑，助教會cd進hw1資料夾執行reproduce程序。
- What your shell script should look like:
  - python3 hw1.py $1 $2

配分 Grading Criteria - kaggle (5%)

- Kaggle Ranking(before 3/21/2018 23:59:59 GMT+8):
  - (0.8%) 超過public simple baseline
  - (0.8%) 超過public strong baseline
  - (0.8%) 超過private simple baseline
  - (0.8%) 超過private strong baseline
  - (0.8%) 3/14/2018 23:59:59 前超過public simple baseline
- Final Ranking:
  - (1.0%) private leaderboard 排名前五名且於助教時間上台分享的同學
- Note:
  - hw1.sh必須能夠reproduce出超過public simple baseline的結果
  - hw1_best.sh必須能夠reproduce出最後作為submission兩份檔案中private分數較高者的結果（+/-0.1內都算成功reproduce）
  - 以上兩者必須都達到否則kaggle部分成績將以0分計算

配分 Grading Criteria - report (5%)

- 限制
  - 檔名必須為 Report.pdf !!!
  - 檔名必須為 Report.pdf !!!
  - 檔名必須為 Report.pdf !!!
  - 請用中文撰寫report（非中文母語者可用英文）
  - 請標明系級、學號、姓名，並按照report模板回答問題，切勿隨意更動題號順序
  - 若有和其他修課同學討論，請務必於題號前標明collaborator（含姓名、學號）
- Report模板連結
  - 連結：https://goo.gl/H1abpq
- 截止日期同 Github Deadline: 3/22/2018 23:59:59 (GMT+8)

其他規定 Other Policy

- Lateness
  - Github每遲交一天(不足一天以一天計算)hw1所得總分將x0.7
  - 不接受程式or報告單獨遲交
  - 不得遲交超過兩天，若有特殊原因請儘速聯絡助教
  - Github遲交表單：遲交請先上傳遲交檔案至自己的github後再填寫遲交表單，助教群會以表單填寫時間作為繳交時間手動clone檔案
- Script Error
  - 當script格式錯誤，造成助教無法順利執行，請在公告時間內寄信向助教說明，修好之後重新執行所得kaggle部分分數將x0.7。
  - 不接受任何py檔的coding錯誤更改。

其他規定 Other Policy

- Cheating
  - 抄code、抄report （含之前修課同學）
  - 開設kaggle多重分身帳號註冊competition
  - 於訓練過程以任何不限定形式接觸到testing data的正確答案
  - 填寫前人的github repo url
  - 教授與助教群保留請同學到辦公室解釋coding作業的權利，請同學務必自愛
- 小老師制度
  - 於hw1 hands-on tutorial到場協助助教回答同學問題
  - 報名方法：通過simple baseline並填寫小老師志願表單
  - 篩選標準：未當過小老師且kaggle排名較前面
  - Bonus：該次作業成績+1%