Machine Learning HW1

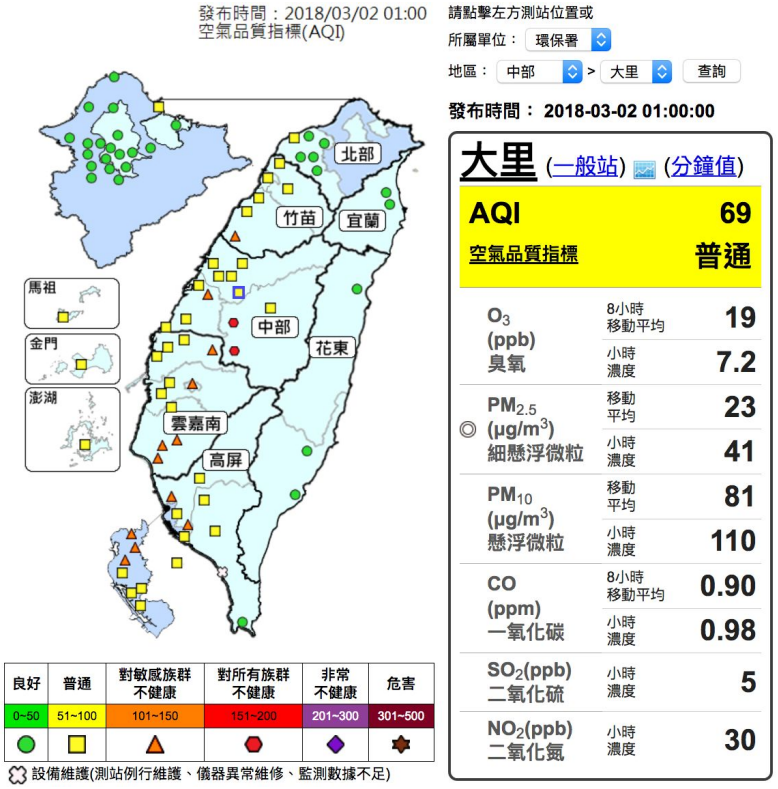
ML TAs
ntu-ml-2020spring-ta@googlegroups.com

Outline

- HW1 Intro - PM2.5 Prediction
  - Tasks Description
  - Training/Testing Data
  - Sample Submission
- Kaggle
- Assignment Regulation
- 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 預測出 PM2.5 的數值。



Data Description

- 本次作業使用豐原站的觀測記錄，分成 train set 跟 test set，train set 是豐原站每個月的前 20 天所有資料。test set 則是從豐原站剩下的資料中取樣出來。
  - train.csv: 每個月前 20 天的完整資料。
  - test.csv : 從剩下的資料當中取樣出連續的10 小時為一筆，前九小時的所有觀測數據當作 feature，第十小時的 PM2.5 當作 answer。一共取出 240 筆不重複的 test data，請根據 feature 預測這 240 筆的 PM2.5。
- 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

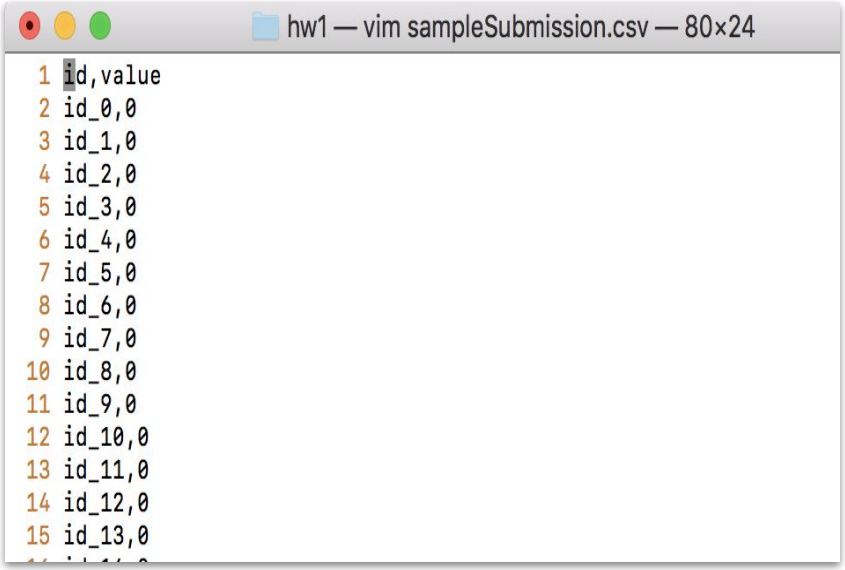
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	日期	測站	測項	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2	2014/1/1	豐原	AMB_TEM	14	14	14	13	12	12	12	12	15	17	20	22	22	22	22
3	2014/1/1	豐原	CH4	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8
4	2014/1/1	豐原	CO	0.51	0.41	0.39	0.37	0.35	0.3	0.37	0.47	0.78	0.74	0.59	0.52	0.41	0.4	0.37
5	2014/1/1	豐原	NMHC	0.2	0.15	0.13	0.12	0.11	0.06	0.1	0.13	0.26	0.23	0.2	0.18	0.12	0.11	0.1
6	2014/1/1	豐原	NO	0.9	0.6	0.5	1.7	1.8	1.5	1.9	2.2	6.6	7.9	4.2	2.9	3.4	3	2.5
7	2014/1/1	豐原	NO2	16	9.2	8.2	6.9	6.8	3.8	6.9	7.8	15	21	14	11	14	12	11
8	2014/1/1	豐原	NOx	17	9.8	8.7	8.6	8.5	5.3	8.8	9.9	22	29	18	14	17	15	14
9	2014/1/1	豐原	O3	16	30	27	23	24	28	24	22	21	29	44	58	50	57	65
10	2014/1/1	豐原	PM10	56	50	48	35	25	12	4	2	11	38	56	64	56	57	52
11	2014/1/1	豐原	PM2.5	26	39	36	35	31	28	25	20	19	30	41	44	33	37	36
12	2014/1/1	豐原	RAINFALL	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR
13	2014/1/1	豐原	RH	77	68	67	74	72	73	74	73	66	56	45	37	40	42	47
14	2014/1/1	豐原	SO2	1.8	2	1.7	1.6	1.9	1.4	1.5	1.6	5.1	15	4.5	2.7	3.5	3.6	3.9
15	2014/1/1	豐原	THC	2	2	2	1.9	1.9	1.8	1.9	1.9	2.1	2	2	2	1.9	1.9	1.9
16	2014/1/1	豐原	WD_HR	37	80	57	76	110	106	101	104	124	46	241	280	297	305	307
17	2014/1/1	豐原	WIND_DIR	35	79	24	55	94	116	106	94	232	153	283	269	290	316	313
18	2014/1/1	豐原	WIND_SPI	1.4	1.8	1	0.6	1.7	2.5	2.5	2	0.6	0.8	1.6	1.9	2.1	3.3	2.5
19	2014/1/1	豐原	WS_HR	0.5	0.9	0.6	0.3	0.6	1.9	2	2	0.5	0.3	0.8	1.2	2	2.6	2.1
20	2014/1/2	豐原	AMB_TEM	16	15	15	14	14	15	16	16	17	20	22	23	24	24	24

Testing Data

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	id_0	AMB_TEM	15	14	14	13	13	13	13	13	12		
2	id_0	CH4	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8		
3	id_0	CO	0.36	0.35	0.34	0.33	0.33	0.34	0.34	0.37	0.42		
4	id_0	NMHC	0.11	0.09	0.09	0.1	0.1	0.1	0.1	0.11	0.12		
5	id_0	NO	0.6	0.4	0.3	0.3	0.3	0.7	0.8	0.8	0.9		
6	id_0	NO2	9.3	7.1	6.1	5.7	5.5	5.3	5.5	7.1	7.5		
7	id_0	NOx	9.9	7.5	6.4	5.9	5.8	6	6.2	7.8	8.4		
8	id_0	O3	36	44	45	44	44	44	43	40	38		
9	id_0	PM10	51	51	31	40	34	51	42	36	30		
10	id_0	PM2.5	27	13	24	29	41	30	29	27	28		
11	id_0	RAINFALI	NR	NR	NR	NR	NR	NR	NR	NR	NR		
12	id_0	RH	75	71	71	73	74	74	74	74	74		
13	id_0	SO2	1.2	1.2	1.2	1.6	1.5	1.5	1.5	1.6	1.6		
14	id_0	THC	1.9	1.8	1.8	1.9	1.9	1.9	1.9	1.9	1.9		
15	id_0	WD_HR	116	114	112	109	111	104	107	108	104		
16	id_0	WIND_DI	115	113	105	102	106	106	112	113	106		
17	id_0	WIND_SPI	2.6	2.2	2	1.9	2.4	2.4	2.5	2.8	2		
18	id_0	WS_HR	2.1	2.4	2.2	1.9	2.3	2.3	2.5	2.5	2.3		
19	id_1	AMB_TEM	12	12	12	13	14	15	14	14	13		
20	id_1	CH4	1.8	1.8	1.9	1.9	1.8	1.8	1.8	1.8	1.8		

Kaggle & Submission Format

- Link: https://www.kaggle.com/c/ml2020spring-hw1
- 預測 240 筆 testing data 中的 PM2.5 值，並將預測結果上傳至 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。



作業規定 Assignment Regulation

- hw1.sh
  - 請**手刻**實作 linear regression，方法限使用 gradient descent。
  - 禁止使用 numpy.linalg.lstsq
- hw1_best.sh
  - 不限定作法，但套件規定仍必須遵照期初公告。

繳交格式 Submission Format

- GitHub 上的 hw1-<account> 裡至少要有下列 3 類檔案：
  - report.pdf
  - hw1.sh及training、testing相關程式碼
  - hw1_best.sh及training、testing相關程式碼
  - 請勿上傳 train.csv, test.csv！
- 你的 repo 裡可以還有其他檔案：
  - e.g., model.npy
- hw1.sh 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。

批改規則及 Script 格式

- test data 會 shuffle 過，請勿直接輸出事先存取的答案
- 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 及 hw1_best.sh 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。

Reproducing Result

- hw1.sh在reproduce時只需要超過simple baseline即可。
- hw1_best.sh必需要能reproduce出Kaggle上所勾選的成績。
- 請同學確保你上傳的程式所產生的結果，會跟你在 Kaggle 上的結果一致，基本上誤差範圍在 **0.1** 之內都屬於一致，若超過範圍，Kaggle 的部份將不予計分。

Report

- 限制
  - 檔名必須為 report.pdf！（**R**eport.pdf 是不正確的）
  - 撰寫 report 時可使用中文或英文，但助教強烈建議使用中文。
  - 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 template 連結