Machine Learning HW1

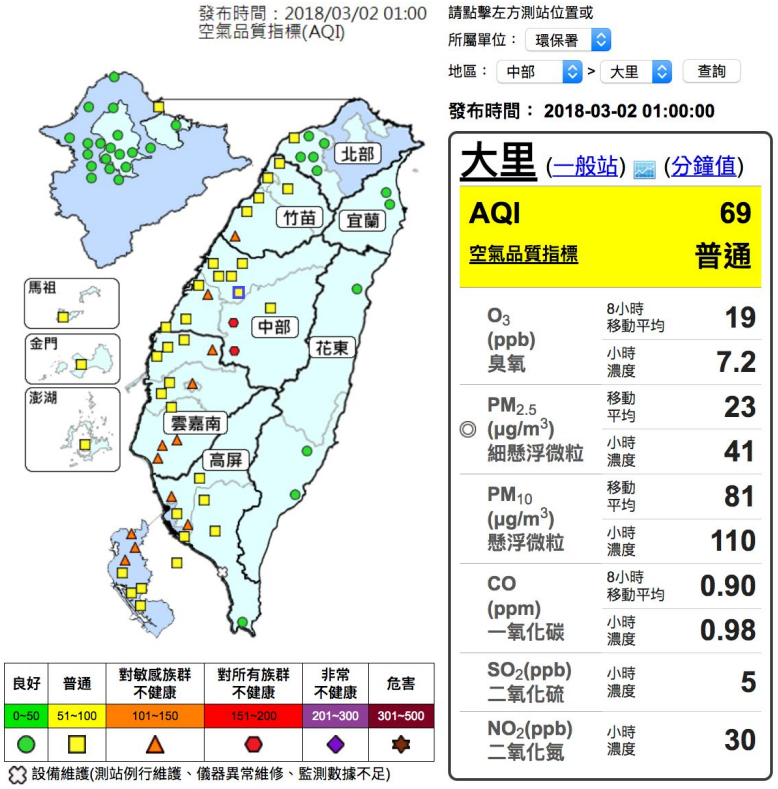
MLTAs
ntumlta2019@gmail.com

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feauure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

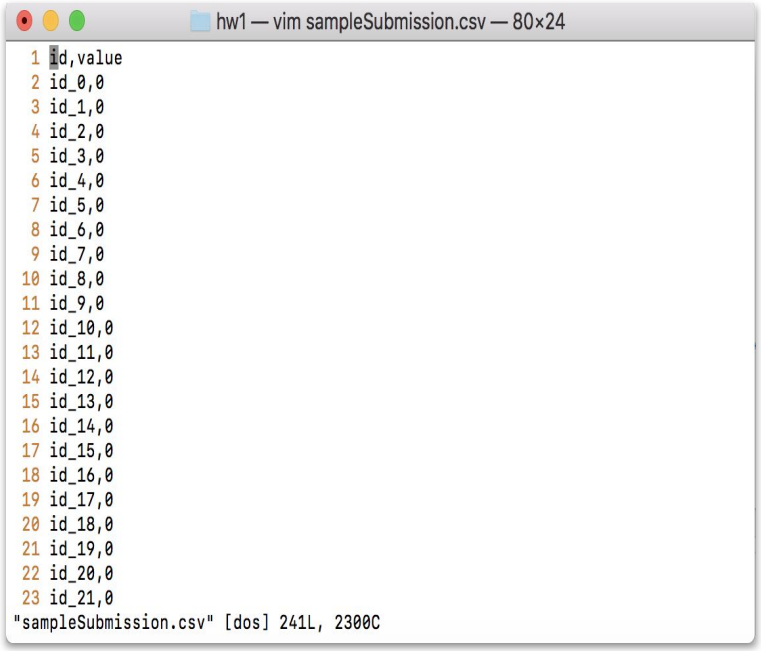
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	日期	測站	測項	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2	2014/1/1	豐原	AMB_TEM	14	14	14	13	12	12	12	12	15	17	20	22	22	22	22
3	2014/1/1	豐原	CH4	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8
4	2014/1/1	豐原	CO	0.51	0.41	0.39	0.37	0.35	0.3	0.37	0.47	0.78	0.74	0.59	0.52	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		

Sample Submission

- 預測240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: Machine Learning (2019, SPRING) HW1 - PM2.5 Prediction
- 個人進行、不須組隊
- Team Name:
  - 修課學生：**學號_任意名稱（ex: b08901666_台大谷翔平）**
  - 旁聽：旁聽__任意名稱
- Maximum Daily Submission: 5 times
- Simple Bonus Deadline: 02/28/2019 23:59:59 (GMT+8)
- Kaggle Deadline: 03/07/2019 11:59:59 (GMT+8)
- Github Deadline: 03/08/2019 23:59:59 (GMT+8)
- test.csv的240筆資料分為：120筆public、120筆private
- 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。
- 最後計分排名將將會考慮到public以及private的成績

Kaggle Baselines

- Public Leaderboard
  - 120 out of 240 from the testing dataset
  - Participants receive instant feedback about their performance.
  - Be sure not to overfit on the public leaderboard.
- Private Leaderboard
  - 120 out of 240 from the testing dataset
  - Remain unknown until the end of the competition.

作業規定 Assignment Regulation

- Only Python 3.6 available !!!!
- 開放使用套件
  - numpy >=1.14
  - scipy == 1.2.1
  - pandas >= 0.24.1
  - python standard library
  - numpy.linalg.lstsq是不可以用的!!!
- 請實作linear regression，方法限定使用Gradient Descent。
- hw1_best.sh不限做法，開放以下套件（但有版本限制請注意）
  - pytorch == 1.0.1
  - tensorflow == 1.12.0
  - keras == 2.2.4
  - scikit-learn == 0.20.0
- 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。

作業規定 Assignment Regulation

- hw1.sh
  - Please handcraft "linear regression" using Gradient Descent
  - beat public simple baseline
- hw1_best.sh
  - meet the higher score you choose in kaggle
  - You can use any

繳交格式 Handin Format

- Kaggle deadline：03/07/2019 11:59:59 (GMT+8)
  Github code & report deadline：03/08/2019 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3個檔案（格式必須完全一樣）：
  - ML2019SPRING/hw1/report.pdf
  - ML2019SPRING/hw1/hw1.sh
  - ML2019SPRING/hw1/hw1_best.sh
  - 請勿上傳 train.csv, test.csv等等dataset!!!
- 你的github上可能還有其他檔案：
  - e.g. ML2019SPRING/hw1/model.npy
- 注意!!!hw1.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

- 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。

配分 Grading Criteria - kaggle (5% + Bonus 1%)

- Kaggle Deadline : 03/07/2019 11:59:59 (GMT+8)
- Early Baseline Point - 1%
  - 在 02/28/2019 23:59:59 (GMT+8) 前於 **public scoreboard** 通過 **early baseline** : **1%**
- Private Score Point - 4%
  - 以 03/07/2019 11:59:59 於 **public/private scoreboard** 之分數為準：
    - 超過public leaderboard的simple baseline分數 : **1%**
    - 超過public leaderboard的strong baseline分數： **1%**
    - 超過private leaderboard的simple baseline分數 : **1%**
    - 超過private leaderboard的strong baseline分數 : **1%**
- Bonus - 1%
  - (1.0%) private leaderboard 排名前五名且於助教時間上台分享的同學

Reproduce

- 請務必在訓練過程中，隨時存取參數。
- 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，基本上誤差在**±0.5**之間都屬於一致，若超過以上範圍，kaggle將不予計分。

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：Link
- 截止日期同 Github Deadline: **03/08/2019 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
  - 不得遲交超過一天，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。
  - 表單連結:Link (遲交才必需填寫)
- 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。
  - 可以更改的部分僅限syntax及io的部分，不得改程式邏輯或是演算法，至於其他部分由助教認定為主。
  - 只能在助教面前更改你的script。

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
  - 不得上傳之前的kaggle競賽
  - 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

2/21手把手教學

連結：Link