Machine Learning HW1

TAs ntu.mlta@gmail.com

Outline

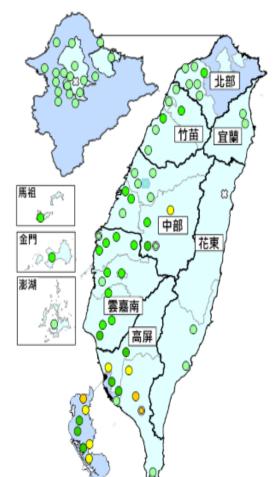
- ♦ hw1介紹
- train/test data
- Kaggle
- ◆ 作業規定、繳交格式、批改方式
- ❖ 配分
- FAQ
- ❖ Github (請看hw0影片)

Task - Predict PM2.5

本次作業的資料是從中央氣象局網站下載的 真實觀測資料,希望大家利用linear regression或其他方法預測PM2.5的數值。

健康影響與活動建議

發布時間: 2017/02/25 23:00 即時細懸浮微約指標



請點擊左方測站位置或

所屬單位: 環保署▼

地區: 中部 ▼ > 忠明▼ 査詢

發布時間: 2017-02-25 23:00:00



單位 : μg/m³,微克/立方公尺

ND : 未檢出(表示數據低於偵測極限2 微

克/立方公尺)

PM_{2.5}移動平均值計算方式: 0.5 × 前12小 時平均 + 0.5 × 前4小時平均(前4小時3筆

有效,前12小時8筆有效)

低	低	低	ф	ф	中	驷	迴	硘	非常高
1	2	3	4	5	6	7	8	9	10

- ◎ 監測車
- 設備維護(測站例行維護、儀器異常維修、監測數據不足)

Data 簡介

- ◆ 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_X.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。

Training Data

	Α	В	С	D	Е	F	G	Н	1	J	K	L	M	N	0	Р	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	03	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	RAINFALI	JR	NR	NR 1	VR.	NR	NR	NR 1	NR.	NR	NR	NR	NR	NR	NR 1	NR N
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_DII	35	79	2.4	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
-04	004440	Ižiži I 	arri	- 40		1.0	1.0	- 10		1.0			1.0		- 10	1.0		1.0

Testing Data

	Α	В	С	D	E	F	G	Н	1	J	K	L	М
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	03	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		<u> </u>
11	id_0	RAINFALI	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_DII	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		

Submission format

- ◆ 預測test set中的240筆PM2.5,上傳至Kaggle。
 - ➤ 上傳格式為csv
 - ➤ 第一行必須是 id, value
 - ➤ 第二行開始,每行分別為id及預測數值,以逗點分開
- ❖ 範例格式:

```
sampleSubmission.csv ×

id, value

id_0,0

id_1,0

id_2,0

id_3,0

id_4,0

id_5,0

id_5,0

id_6,0

id_7,0
```

Kaggle

- ◆ 網址: https://www.kaggle.com/c/ml-2017fall-hw1
- ❖ 請至kaggle創帳號登入 (務必使用@ntu.edu.tw信箱)
- ❖ 個人進行、不須組隊
- ❖ 隊名:學號_任意名稱 (ex.b02901000_gai),旁聽同學則避免使用此命名原則
- ❖ 每日上傳上限5次
- ❖ test.csv的240筆資料分為:120筆public、120筆private
- ❖ Leaderboard上顯示的是public的分數,在死線前可選擇兩份答案作為private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 deadline : 10/12/2017 11:59:00 PM (GMT+8)
- github code & report deadline : 2017/10/13 11:59:00 (GMT+8)

請填寫github url表單:<u>https://goo.gl/forms/LOX5W6ByDPSNWA6N2</u>

github url回應:https://ppt.cc/fTfDNx

#請每位修課同學務必填寫,hw0填過的也再填一次

作業規定

- Only Python 3.5+ available
- ❖ 請實作linear regression,方法限定使用Gradient Descent。
- ❖ 若想嘗試其他方法也可以,但是仍然需實作linear regression。
- ❖ 不能使用現成套件,只能使用numpy、scipy以及pandas。 (Standard library可以) (numpy.linalg.lstsq是不可以用的!!!) 若需要使用其他套件,請在Deadline前寄信至助教信箱詢問,並請簡述原因。

❖ 版本要求:

Python 3.5+ numpy 1.13

繳交格式

1. Kaggle deadline:10/12/2017 11:59:00 PM (GMT+8)
Github code & report deadline:2017/10/13 11:59:00 (GMT+8)
Github commit為local端時間,請注意你電腦時間,並且上Github確認助教會在Deadline一到就clone所有程式,並且不再重新clone任何檔案

2. 你的Github上至少需要有下列三個檔案:

ML2017FALL/hw1/report.pdf (請按照page. 13提供之templete撰寫)

ML2017FALL/hw1/hw1.sh

ML2017FALL/hw1/hw1 best.sh

如果你有其他程式檔案,請一併上傳,e.g. xxx.py等等。另外,請不要上傳與作業無關檔案以及data

3. 請各位自行跑training部分,儲存訓練完的模型參數,一併上傳至github,hw1.sh & hw1_best.sh僅執行testing部分。

批改方式

- 1. test data會shuffle過,請勿直接輸出事先存取的答案
- 2. 助教批改改程式時,會用下列的方法執行:

bash hw1.sh [input file] [output file] bash hw1_best.sh [input file] [output file]

input file : 助教提供test.csv的路徑

output file : 助教提供output的路徑

e.g. 如果為 bash hw1.sh ./data/test.csv ./result/res.csv

則hw1.sh 最後需要產生一個 res.csv的檔案在result資料夾中

- 3. hw1.sh與hw1_best.sh皆需要在3分鐘內跑出結果,否則不會拿到分數
- 4. 請勿在程式內寫死路徑、並且只能output在助教給定的路徑,不然會有permission denied的問題!

配分 (5%) - kaggle

❖ Kaggle Rank :

- ➤ (0.8%) 超過public leaderboard的simple baseline分數
- ➤ (0.8%) 超過public leaderboard的strong baseline分數
- ➤ (0.8%) 超過private leaderboard的simple baseline分數
- ➤ (0.8%) 超過private leaderboard的strong baseline分數
- ➤ (0.8%) 10/5 23:59 前超過public simple baseline
- ➤ (1%) kaggle排名前五名(且願意上台跟大家分享的同學)
- ♦ hw1.sh的結果必須超過public simple baseline,且hw1_best.sh需要重現在 kaggle上面的成績,否則程式部分0分,report部分也會是部分給分

配分 (5%) - report

❖ report.pdf: PDF檔!!!!!!!

(限制:中文(非中文母語者用英文)&不能超過2頁、請使用template作答)

- ➤ 請實做以下兩種不同feature的模型,回答第(1)~(3)題:
 - 抽全部9小時內的污染源feature的一次項(加bias)
 - 抽全部9小時內pm2.5的一次項當作feature(加bias)
- ➤ (1) 記錄誤差值 (RMSE)(根據kaggle public+private分數), 討論兩種feature的影響
- ➤ (2) 將feature從抽前9小時改成抽前5小時,討論其變化
- ➤ (3) Regularization on all the weight with λ=0.1、0.01、0.001、0.0001, 並作圖顯示訓練後的差別(with RMSE)
- ➤ (4) 證明linear regression的close form
- Report template:

https://ppt.cc/fbdOZx

Other policy

- ❖ script 錯誤,kaggle分數直接0分。若是格式錯誤,請在公告時間內找助教修好,修完kaggle分數*0.7。
- ❖ Kaggle超過deadline可以繼續上傳但不計入成績。
- ❖ Github遲交一天(*0.7),不足一天以一天計算,不得遲交超過兩天,有特殊原因 請找助教。
- ❖ Github遲交表單: https://goo.gl/forms/mq0F6u82AKiw4tt33 (遲交才必需填寫) 遲交請「先上傳程式」至Github再填表單,助教會根據表單填寫時間當作繳交時間。

FAQ

1. 如果只有做一個方法是否需要繳交兩份script?

Ans.

是的。如果只有做linear regression,kaggle上的分數也是linear regression 的話, 也麻煩交兩份script。

2. 表單填錯怎麼辦?

Ans.

請直接重新填即可,會以最新的表單為準。

提醒,表單只是蒐集各位github repo url,不需每次git push都填一次表單。

再次提醒,修課同學一定都要填寫!

小老師制度 (手把手教學)

- ◆ 在10/4以前超過simple baseline並願意在10/6在上課時間教導同學撰寫作業一程式,請填寫一下表單:https://goo.gl/forms/epctvoGTRxeDsFIH2
- ◆ 10/5將公布小老師名單在作業網頁,人數太多將以符合以下標準的同學為主:
 - 1. 沒有當過小老師
 - 2. kaggle成績排名較高
- ❖ 小老師當次成績 +1%