Machine Learning HW1

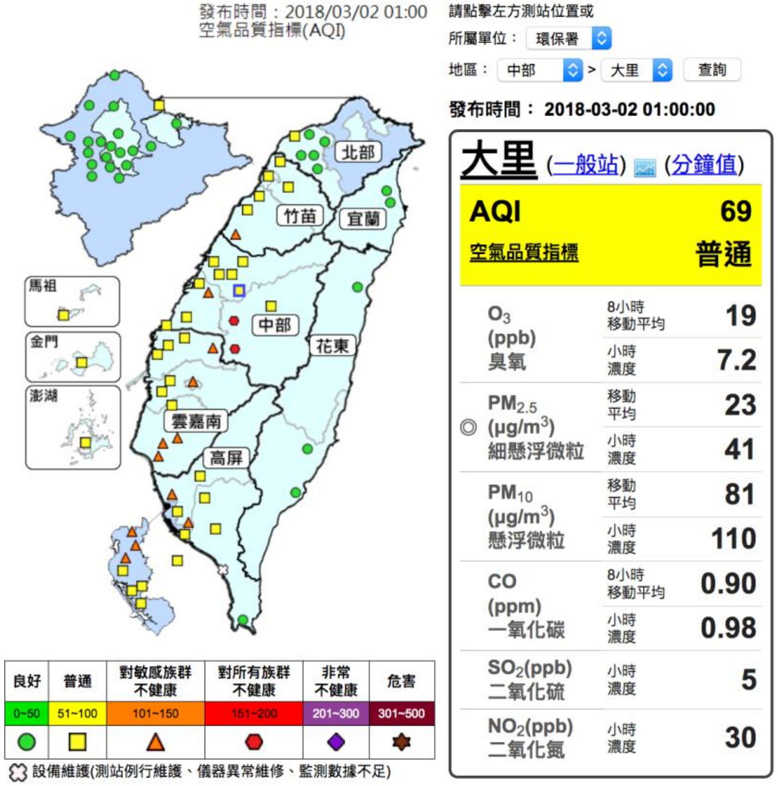
MLTAs
mlta2020fall@gmail.com

Outline

- HW1 Intro - PM2.5 Prediction
    - Tasks Description
    - Training/Testing Data
    - Sample Submission
- Kaggle
- Grading / Assignment Regulation

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。



發布時間：2018/03/02 01:00
空氣品質指標(AQI)

請點擊左方測站位置或
所屬單位：環保署
地區：中部 > 大里 查詢

發布時間： 2018-03-02 01:00:00

大里 (一般站) (分鐘值)

AQI **69**
空氣品質指標 **普通**

O_3 (ppb) 臭氧	8小時 移動平均	19
	小時 濃度	7.2
◎ $PM_{2.5}$ (μg/m^3) 細懸浮微粒	移動 平均	23
	小時 濃度	41
PM_{10} (μg/m^3) 懸浮微粒	移動 平均	81
	小時 濃度	110
CO (ppm) 一氧化碳	8小時 移動平均	0.90
	小時 濃度	0.98
SO_2(ppb) 二氧化硫	小時 濃度	5
NO_2(ppb) 二氧化氮	小時 濃度	30

良好	普通	對敏感族群 不健康	對所有族群 不健康	非常 不健康	危害
0~50	51~100	101~150	151~200	201~300	301~500
●	■	▲	⬣	◆	✶

✣ 設備維護(測站例行維護、儀器異常維修、監測數據不足)

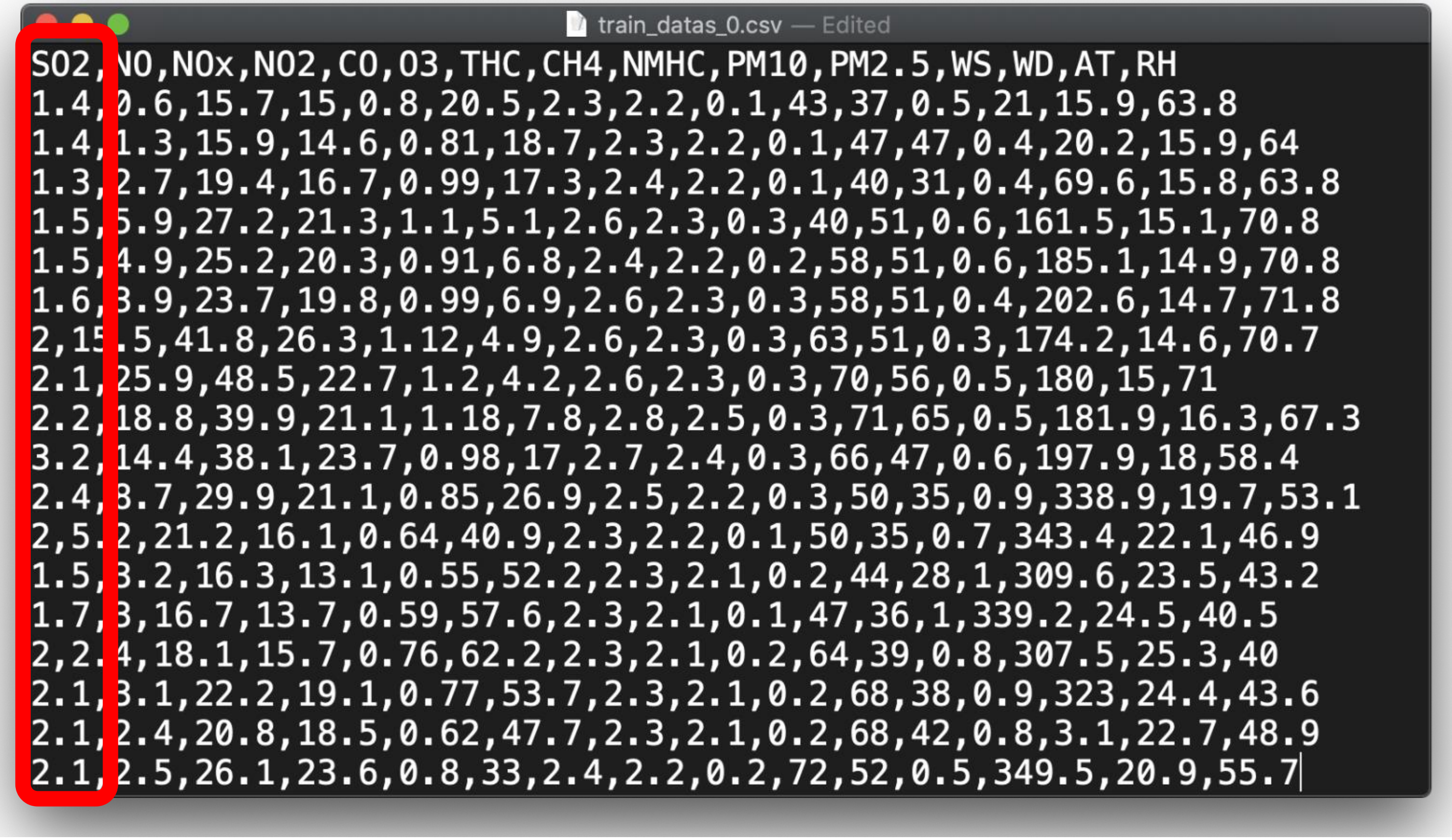
Data Description

- 本次作業使用的觀測記錄，分成train set跟test set，train set是兩年份的所有資料。test set則是一年份中的資料中取樣出來。
    - training data: 某連續兩年的觀測資料。
    - testing data : 第三年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500筆不重複的test data，請根據feature預測這500筆的PM2.5。
- Data含有15項觀測數據 SO2, NO, NOx, NO2, CO, O3, THC, CH4, NMHC, PM10, PM2.5, WS, WD, AT, RH。
- Data 中數字後面若有特殊符號，如0.3#, 0.3x, 0.3*請parse成0.3(只有數字)

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

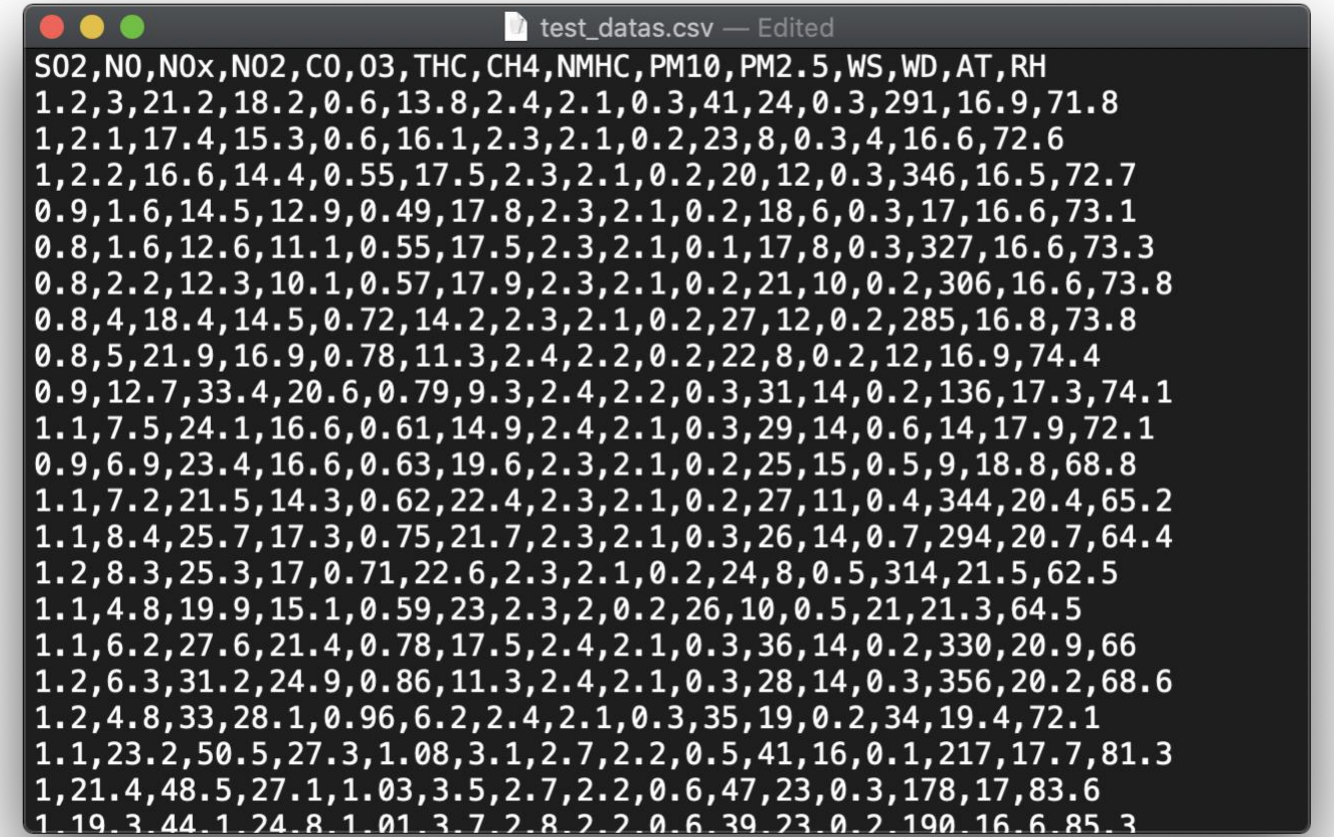
每一筆資料都是相鄰的
以0~8筆去預測第9筆
以1~9筆去預測第10筆…



```
SO2,NO,NOx,NO2,CO,O3,THC,CH4,NMHC,PM10,PM2.5,WS,WD,AT,RH
1.4,0.6,15.7,15,0.8,20.5,2.3,2.2,0.1,43,37,0.5,21,15.9,63.8
1.4,1.3,15.9,14.6,0.81,18.7,2.3,2.2,0.1,47,47,0.4,20.2,15.9,64
1.3,2.7,19.4,16.7,0.99,17.3,2.4,2.2,0.1,40,31,0.4,69.6,15.8,63.8
1.5,5.9,27.2,21.3,1.1,5.1,2.6,2.3,0.3,40,51,0.6,161.5,15.1,70.8
1.5,4.9,25.2,20.3,0.91,6.8,2.4,2.2,0.2,58,51,0.6,185.1,14.9,70.8
1.6,3.9,23.7,19.8,0.99,6.9,2.6,2.3,0.3,58,51,0.4,202.6,14.7,71.8
2,15.5,41.8,26.3,1.12,4.9,2.6,2.3,0.3,63,51,0.3,174.2,14.6,70.7
2.1,25.9,48.5,22.7,1.2,4.2,2.6,2.3,0.3,70,56,0.5,180,15,71
2.2,18.8,39.9,21.1,1.18,7.8,2.8,2.5,0.3,71,65,0.5,181.9,16.3,67.3
3.2,14.4,38.1,23.7,0.98,17,2.7,2.4,0.3,66,47,0.6,197.9,18,58.4
2.4,8.7,29.9,21.1,0.85,26.9,2.5,2.2,0.3,50,35,0.9,338.9,19.7,53.1
2,5.2,21.2,16.1,0.64,40.9,2.3,2.2,0.1,50,35,0.7,343.4,22.1,46.9
1.5,3.2,16.3,13.1,0.55,52.2,2.3,2.1,0.2,44,28,1,309.6,23.5,43.2
1.7,3,16.7,13.7,0.59,57.6,2.3,2.1,0.1,47,36,1,339.2,24.5,40.5
2,2.4,18.1,15.7,0.76,62.2,2.3,2.1,0.2,64,39,0.8,307.5,25.3,40
2.1,3.1,22.2,19.1,0.77,53.7,2.3,2.1,0.2,68,38,0.9,323,24.4,43.6
2.1,2.4,20.8,18.5,0.62,47.7,2.3,2.1,0.2,68,42,0.8,3.1,22.7,48.9
2.1,2.5,26.1,23.6,0.8,33,2.4,2.2,0.2,72,52,0.5,349.5,20.9,55.7
```

Testing Data

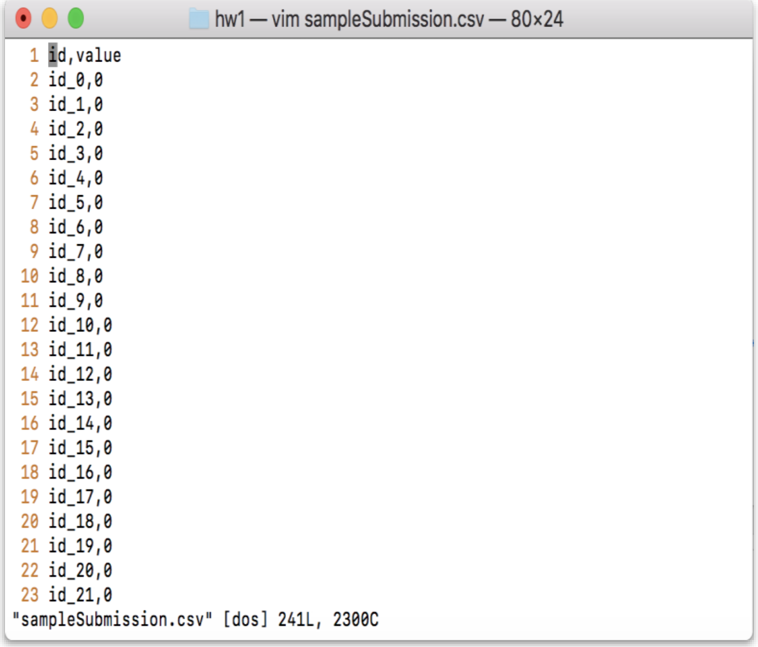
格式和 training data一樣
但請以0~8筆去預測 id_0
以9~17筆去預測 id_1



```
SO2,NO,NOx,NO2,CO,O3,THC,CH4,NMHC,PM10,PM2.5,WS,WD,AT,RH
1.2,3,21.2,18.2,0.6,13.8,2.4,2.1,0.3,41,24,0.3,291,16.9,71.8
1,2.1,17.4,15.3,0.6,16.1,2.3,2.1,0.2,23,8,0.3,4,16.6,72.6
1,2.2,16.6,14.4,0.55,17.5,2.3,2.1,0.2,20,12,0.3,346,16.5,72.7
0.9,1.6,14.5,12.9,0.49,17.8,2.3,2.1,0.2,18,6,0.3,17,16.6,73.1
0.8,1.6,12.6,11.1,0.55,17.5,2.3,2.1,0.1,17,8,0.3,327,16.6,73.3
0.8,2.2,12.3,10.1,0.57,17.9,2.3,2.1,0.2,21,10,0.2,306,16.6,73.8
0.8,4,18.4,14.5,0.72,14.2,2.3,2.1,0.2,27,12,0.2,285,16.8,73.8
0.8,5,21.9,16.9,0.78,11.3,2.4,2.2,0.2,22,8,0.2,12,16.9,74.4
0.9,12.7,33.4,20.6,0.79,9.3,2.4,2.2,0.3,31,14,0.2,136,17.3,74.1
1.1,7.5,24.1,16.6,0.61,14.9,2.4,2.1,0.3,29,14,0.6,14,17.9,72.1
0.9,6.9,23.4,16.6,0.63,19.6,2.3,2.1,0.2,25,15,0.5,9,18.8,68.8
1.1,7.2,21.5,14.3,0.62,22.4,2.3,2.1,0.2,27,11,0.4,344,20.4,65.2
1.1,8.4,25.7,17.3,0.75,21.7,2.3,2.1,0.3,26,14,0.7,294,20.7,64.4
1.2,8.3,25.3,17,0.71,22.6,2.3,2.1,0.2,24,8,0.5,314,21.5,62.5
1.1,4.8,19.9,15.1,0.59,23,2.3,2,0.2,26,10,0.5,21,21.3,64.5
1.1,6.2,27.6,21.4,0.78,17.5,2.4,2.1,0.3,36,14,0.2,330,20.9,66
1.2,6.3,31.2,24.9,0.86,11.3,2.4,2.1,0.3,28,14,0.3,356,20.2,68.6
1.2,4.8,33,28.1,0.96,6.2,2.4,2.1,0.3,35,19,0.2,34,19.4,72.1
1.1,23.2,50.5,27.3,1.08,3.1,2.7,2.2,0.5,41,16,0.1,217,17.7,81.3
1,21.4,48.5,27.1,1.03,3.5,2.7,2.2,0.6,47,23,0.3,178,17,83.6
1,19.3,44.1,24.8,1.01,3.7,2.8,2.2,0.6,39,23,0.2,190,16.6,85.3
```

Sample Submission

- 預測500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 (string, double)，以逗號隔開
- 範例格式：

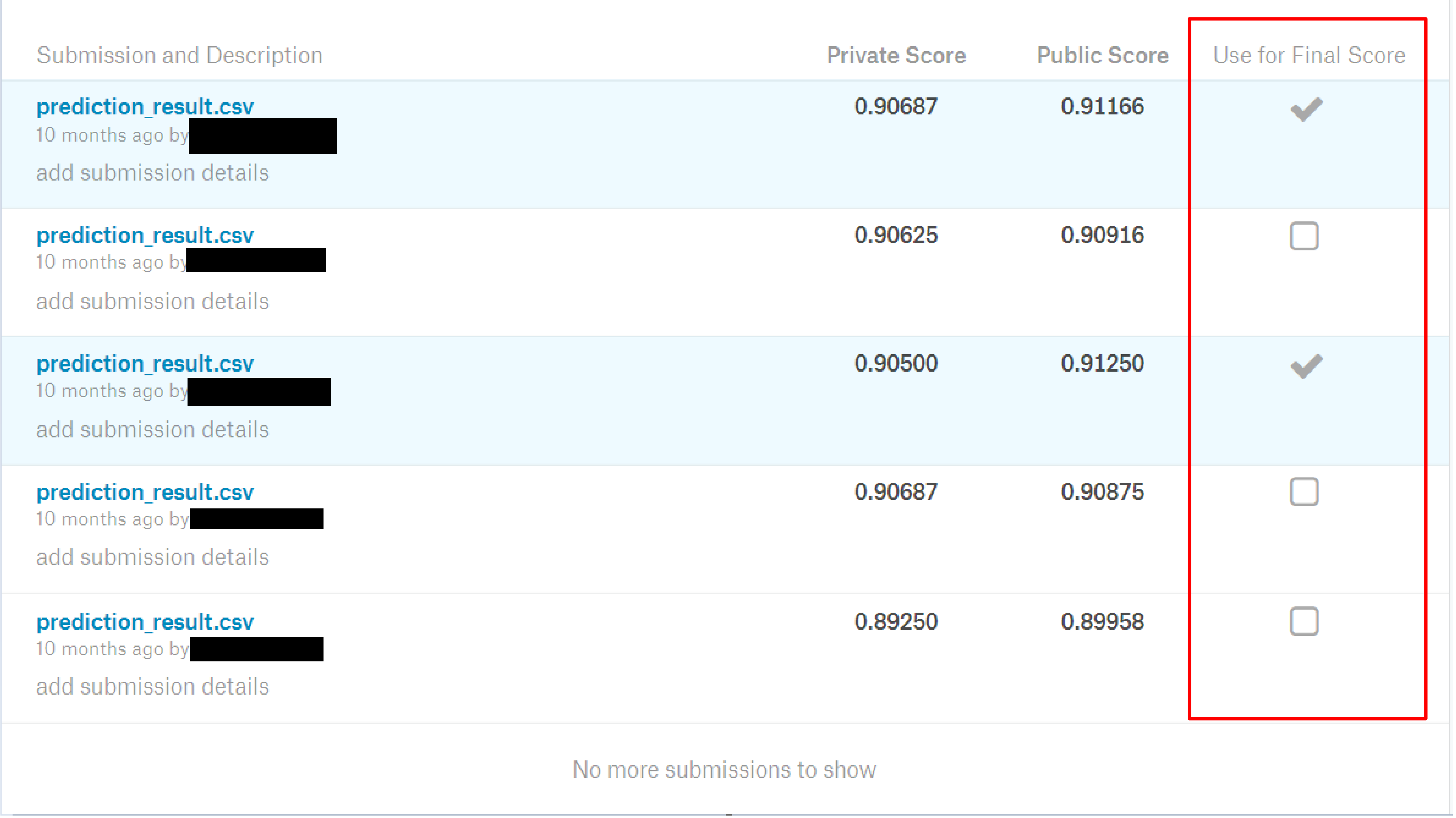


```
id,value
id_0,0
id_1,0
id_2,0
id_3,0
id_4,0
id_5,0
id_6,0
id_7,0
id_8,0
id_9,0
id_10,0
id_11,0
id_12,0
id_13,0
id_14,0
id_15,0
id_16,0
id_17,0
id_18,0
id_19,0
id_20,0
id_21,0
"sampleSubmission.csv" [dos] 241L, 2300C
```

Kaggle Info

- 請自行到kaggle創建帳號（務必使用ntu信箱）
- Link: [Machine Learning (2020, FALL) 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: b09901666_只會tune參數）**
  - 旁聽：旁聽_任意名稱
- Maximum Daily Submission: 5 times
- Simple Baseline Deadline: 10/03/2020 23:59:59 (GMT+8)
- Kaggle Deadline: 10/16/2020 23:59:59 (GMT+8)
- Github Deadline: 10/18/2020 23:59:59 (GMT+8)
- test_data.csv的500筆資料分為：250筆public、250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 Info



Submission and Description	Private Score	Public Score	Use for Final Score
prediction_result.csv 10 months ago by [redacted] add submission details	0.90687	0.91166	✓
prediction_result.csv 10 months ago by [redacted] add submission details	0.90625	0.90916	□
prediction_result.csv 10 months ago by [redacted] add submission details	0.90500	0.91250	✓
prediction_result.csv 10 months ago by [redacted] add submission details	0.90687	0.90875	□
prediction_result.csv 10 months ago by [redacted] add submission details	0.89250	0.89958	□

No more submissions to show

Kaggle Baselines

- Public Leaderboard
  - 250 out of 500 from the testing dataset
  - Participants receive instant feedback about their performance.
  - Be sure not to overfit on the public leaderboard.
- Private Leaderboard
  - 250 out of 500 from the testing dataset
  - Remain unknown until the end of the competition.
- 請不要壓死線上傳，預留時間!

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

- Kaggle Deadline : 10/16/2020 23:59:59 (GMT+8)
- Early Baseline Point - 1%
    - 在 10/03/2020 23:59:59 (GMT+8) 前於 **public scoreboard** 通過 **early baseline** : **1%**
- Private Score Point - 4%
    - 以 10/16/2019 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%**
    - 以上皆須通過 Reproduce 才給分
- Bonus - 1%
    - (1.0%) private leaderboard 排名前五名且於助教時間上台分享的同學

配分 Grading Criteria - report(5%)

- Programming Report - 2%
  - https://drive.google.com/file/d/1m0LRSjlRlnuFimGVJW6yX1dSwxwF-A0C/view?usp=sharing
- Math Problem - 3%
  - https://hackmd.io/RFiu1FsYR5uQTrrpdxUvlw?view
  - Type in latex(preferable) or take pictures of your handwriting
- Write them in report.pdf

繳交格式 Handin Format

- Kaggle deadline：10/16/2019 11:59:59 (GMT+8)
  Github code & report deadline：10/18/2019 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）：
  - ○ ML2020FALL/hw1/report.pdf
  - ○ ML2020FALL/hw1/hw1.sh
  - ○ ML2020FALL/hw1/hw1_best.sh
  - ○ 請勿上傳 year1-data.csv, testing_data.csv等等dataset!!!
- 你的github上可能還有其他檔案：
  - ○ e.g. ML2020FALL/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

作業規定 Assignment Regulation

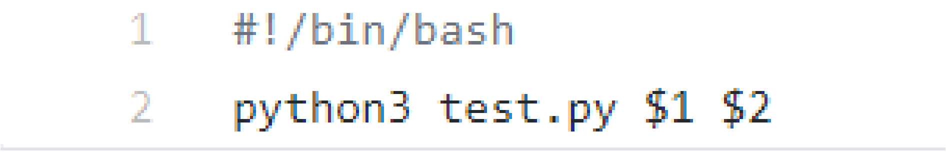
- Only Python 3.6 available !!!!
- 開放使用套件
  - All python standard library
  - numpy ==1.16.5
  - scipy == 1.3.1
  - pandas == 0.25.1
  - python standard library
  - numpy.linalg.lstsq是不可以用的!!!
- 請實作linear regression，方法限定使用Gradient Descent。
- hw1_best.sh不限做法，開放以下套件（但有版本限制請注意）
  - pytorch == 1.2.0
  - tensorflow == 1.14.0
  - keras == 2.2.4
  - scikit-learn == 0.21.3
- 助教 Conda File (同學可自行下載改 prefix 測試)
- 若需使用其他套件，請儘早寄信至助教信箱或到 FB 討論版詢問，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 highest score you choose in kaggle
  - You can use any methods with allowed python packages
- report.pdf
  - Please refer to report template

Shell Script

- 其格式如下(py檔名可自訂):



```
#!/bin/bash
python3 test.py $1 $2
```

- 該script須能執行**testing**的部分，但若是執行結果與kaggle差太多，會執行training的部分，因此也請同學一併傳上**training code**

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/testing_data.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_data.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。

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。

Report 格式

- 限制
  - 檔名必須為 report.pdf !!!
  - 檔名必須為 report.pdf !!!
  - 檔名必須為 report.pdf !!!
  - 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: 10/18/2019 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。
  - 會在社團以討論串形式處理
- 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。

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

Tips for HW1

- What you should learn
    - Familiar with matrix operation
    - Familiar with numpy, pandas
    - Implement gradient descent
    - Data preprocessing

TA Hour

- 9/29, 10/6, 10/13 (Tue) @ BL530
- 14:20~16:10