Machine Learning HW1

MLTAs
ntueemlta2022@gmail.com

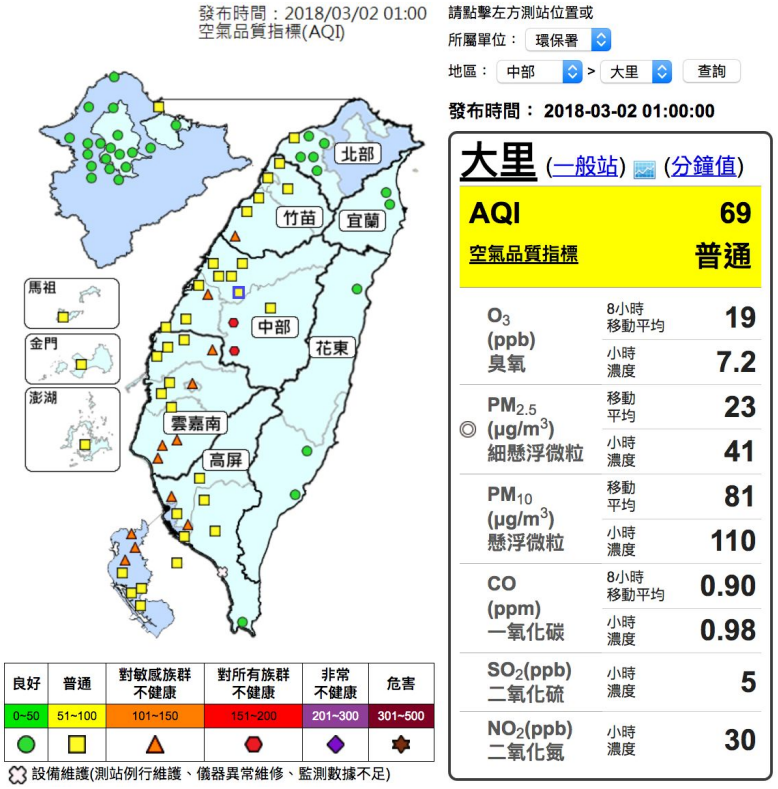
Outline

- HW1 Intro - PM2.5 Prediction
  - Tasks/Data Description
  - Training/Testing Data
  - Sample Submission
- Kaggle
  - Kaggle Info
  - Kaggle Submission
  - Special Regulation
- Grading / Assignment Regulation
  - Deadline
  - Grading Criteria
  - Hand-in Format
- Other Policy
  - TA hour, Hint, etc

HW1 Intro - PM 2.5 Prediction

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

- 本次作業使用的觀測記錄為某一年某地區的觀測資料,
  - training data: 同一年某地區的資料當中取樣出數天, 以連續的24小時為一組數據, 第k~k+8小時的觀測數據當作 train_X, 第 k+9 小時的 PM2.5 當作 train_y。
  - testing data : 同一年某地區的資料當中取樣出數天, 以連續的9小時為一組數據, 前8小時的觀測數據當作 test_X, **請預測第九小時的 PM2.5 當作 train_y。**
  - 一共預測 90 筆第九小時的 PM2.5。
- Data含有 **15** 項數據可作為特徵:

  AMB_TEMP, CO, NO, NO2, NOx, O3, PM10, WS_HR, RAINFALL, RH, SO2, WD_HR, WIND_DIREC, WIND_SPEED, PM2.5。

  ### 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0~7筆去預測第8筆

示意圖(數字僅供參考)

train_x1
某天觀測的
第 0~7 小時

train_y1
某天觀測的
第 8 小時之
pm2.5

AMB_TEMP	CO	NO	NO2	NOx	O3	PM10	WS_HR	RAINFALL	RH	SO2	WD_HR	WIND_DIREC	WIND_SPEED	PM2.5
10.8	0.32	1.7	8.6	10.3	22.9	21	0.6	0	71	1.9	172	171	0.6	15
10.8	0.27	1.6	6.2	7.8	23.8	20	1.4	0	71	1.7	161	129	1.8	13
11	0.25	0.9	5.4	6.3	27.4	21	0.8	0	68	1.6	152	147	1.5	12
11	0.23	0.7	3.1	3.8	29.5	21	1.8	0	68	1.6	138	145	1.7	9
11.3	0.22	0.8	2.9	3.8	30.7	16	1.9	0	67	1.6	140	139	1.7	9
11.6	0.23	0.9	2.9	3.8	31.2	21	1.8	0	67	1.5	141	145	2.2	9
11.7	0.25	0.9	3.9	4.8	30	23	2.1	0	67	1.5	142	141	2.4	10
12	0.27	1.3	5.1	6.4	28.7	22	2.3	0	67	1.5	141	142	2.5	9
12.4	0.28	1.5	5	6.6	28.3	21	2.5	0	66	1.4	142	144	2.4	10
12.7	0.3	1.8	5.8	7.6	27.5	25	1.9	0	66	1.5	139	140	1.7	11
13.8	0.35	3.4	9.9	13.4	25.5	24	0.5	0	63	1.7	128	111	1.3	10
14.8	0.33	3.4	7.6	11	31.7	34	1.3	0	57	1.7	84	91	1.7	8
15.3	0.32	2.6	6.5	9.2	34.7	30	1.1	0	55	1.7	97	43	1	11
15.7	0.32	2.8	6.8	9.7	36.5	28	0.9	0	57	1.8	65	47	1.3	11
15.6	0.29	2.2	6.4	8.7	37.7	23	1.5	0	59	1.8	77	93	2	9
15.2	0.32	2.2	9.6	11.8	34.9	21	1.2	0	62	1.8	63	62	2	10

Training Data

每一筆資料都是相鄰的
以0~7筆去預測第8筆
以1~8筆去預測第9筆

示意圖(數字僅供參考)

train_x2
某天觀測的
第 1~8 小時

train_y2
某天觀測的
第 9 小時之
pm2.5

AMB_TEMP	CO	NO	NO2	NOx	O3	PM10	WS_HR	RAINFALL	RH	SO2	WD_HR	WIND_DIREC	WIND_SPEED	PM2.5
10.8	0.32	1.7	8.6	10.3	22.9	21	0.6	0	71	1.9	172	171	0.6	15
10.8	0.27	1.6	6.2	7.8	23.8	20	1.4	0	71	1.7	161	129	1.8	13
11	0.25	0.9	5.4	6.3	27.4	21	0.8	0	68	1.6	152	147	1.5	12
11	0.23	0.7	3.1	3.8	29.5	21	1.8	0	68	1.6	138	145	1.7	9
11.3	0.22	0.8	2.9	3.8	30.7	16	1.9	0	67	1.6	140	139	1.7	9
11.6	0.23	0.9	2.9	3.8	31.2	21	1.8	0	67	1.5	141	145	2.2	9
11.7	0.25	0.9	3.9	4.8	30	23	2.1	0	67	1.5	142	141	2.4	10
12	0.27	1.3	5.1	6.4	28.7	22	2.3	0	67	1.5	141	142	2.5	9
12.4	0.28	1.5	5	6.6	28.3	21	2.5	0	66	1.4	142	144	2.4	10
12.7	0.3	1.8	5.8	7.6	27.5	25	1.9	0	66	1.5	139	140	1.7	11
13.8	0.35	3.4	9.9	13.4	25.5	24	0.5	0	63	1.7	128	111	1.3	10
14.8	0.33	3.4	7.6	11	31.7	34	1.3	0	57	1.7	84	91	1.7	8
15.3	0.32	2.6	6.5	9.2	34.7	30	1.1	0	55	1.7	97	43	1	11
15.7	0.32	2.8	6.8	9.7	36.5	28	0.9	0	57	1.8	65	47	1.3	11
15.6	0.29	2.2	6.4	8.7	37.7	23	1.5	0	59	1.8	77	93	2	9
15.2	0.32	2.2	9.6	11.8	34.9	21	1.2	0	62	1.8	63	62	2	10

Training Data

每一筆資料都是相鄰的
以0~7筆去預測第8筆
以1~8筆去預測第9筆,
以2~9筆去預測第10筆, 依此類推

示意圖(數字僅供參考)

train_x3
某天觀測的
第 2~9 小時

train_y3
某天觀測的
第 10 小時之
pm2.5

AMB_TEMP	CO	NO	NO2	NOx	O3	PM10	WS_HR	RAINFALL	RH	SO2	WD_HR	WIND_DIREC	WIND_SPEED	PM2.5
10.8	0.32	1.7	8.6	10.3	22.9	21	0.6	0	71	1.9	172	171	0.6	15
10.8	0.27	1.6	6.2	7.8	23.8	20	1.4	0	71	1.7	161	129	1.8	13
11	0.25	0.9	5.4	6.3	27.4	21	0.8	0	68	1.6	152	147	1.5	12
11	0.23	0.7	3.1	3.8	29.5	21	1.8	0	68	1.6	138	145	1.7	9
11.3	0.22	0.8	2.9	3.8	30.7	16	1.9	0	67	1.6	140	139	1.7	9
11.6	0.23	0.9	2.9	3.8	31.2	21	1.8	0	67	1.5	141	145	2.2	9
11.7	0.25	0.9	3.9	4.8	30	23	2.1	0	67	1.5	142	141	2.4	10
12	0.27	1.3	5.1	6.4	28.7	22	2.3	0	67	1.5	141	142	2.5	9
12.4	0.28	1.5	5	6.6	28.3	21	2.5	0	66	1.4	142	144	2.4	10
12.7	0.3	1.8	5.8	7.6	27.5	25	1.9	0	66	1.5	139	140	1.7	11
13.8	0.35	3.4	9.9	13.4	25.5	24	0.5	0	63	1.7	128	111	1.3	10
14.8	0.33	3.4	7.6	11	31.7	34	1.3	0	57	1.7	84	91	1.7	8
15.3	0.32	2.6	6.5	9.2	34.7	30	1.1	0	55	1.7	97	43	1	11
15.7	0.32	2.8	6.8	9.7	36.5	28	0.9	0	57	1.8	65	47	1.3	11
15.6	0.29	2.2	6.4	8.7	37.7	23	1.5	0	59	1.8	77	93	2	9
15.2	0.32	2.2	9.6	11.8	34.9	21	1.2	0	62	1.8	63	62	2	10

Testing Data

格式和 training data一樣
但請以0~8筆去預測 test_y1
以9~17筆去預測 test_y2
.....
總共產生90個預測結果

示意圖(數字僅供參考)

AMB_TEMP	CO	NO	NO2	NOx	O3	PM10	WS_HR	RAINFALL	RH	SO2	WD_HR	WIND_DIREC	WIND_SPEED	PM2.5
10.8	0.32	1.7	8.6	10.3	22.9	21	0.6	0	71	1.9	172	171	0.6	15
10.8	0.27	1.6	6.2	7.8	23.8	20	1.4	0	71	1.7	161	129	1.8	13
11	0.25	0.9	5.4	6.3	27.4	21	0.8	0	68	1.6	152	147	1.5	12
11	0.23	0.7	3.1	3.8	29.5	21	1.8	0	68	1.6	138	145	1.7	9
11.3	0.22	0.8	2.9	3.8	30.7	16	1.9	0	67	1.6	140	139	1.7	9
11.6	0.23	0.9	2.9	3.8	31.2	21	1.8	0	67	1.5	141	145	2.2	9
11.7	0.25	0.9	3.9	4.8	30	23	2.1	0	67	1.5	142	141	2.4	10
12	0.27	1.3	5.1	6.4	28.7	22	2.3	0	67	1.5	141	142	2.5	9
12.4	0.28	1.5	5	6.6	28.3	21	2.5	0	66	1.4	142	144	2.4	10
12.7	0.3	1.8	5.8	7.6	27.5	25	1.9	0	66	1.5	139	140	1.7	11
13.8	0.35	3.4	9.9	13.4	25.5	24	0.5	0	63	1.7	128	111	1.3	10
14.8	0.33	3.4	7.6	11	31.7	34	1.3	0	57	1.7	84	91	1.7	8
15.3	0.32	2.6	6.5	9.2	34.7	30	1.1	0	55	1.7	97	43	1	11
15.7	0.32	2.8	6.8	9.7	36.5	28	0.9	0	57	1.8	65	47	1.3	11
15.6	0.29	2.2	6.4	8.7	37.7	23	1.5	0	59	1.8	77	93	2	9
15.2	0.32	2.2	9.6	11.8	34.9	21	1.2	0	62	1.8	63	62	2	10

test_x1
第n天測資的
第 0~8 小時

test_x2
第n+1天測資的
第 0~8 小時

Sample Submission

- 預測 90 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, Predicted
  - 第二行開始, 每行分別為id值及預測PM2.5數值 (string, double)
- 範例格式:

Id	Predicted
1	27.3085098
2	22.2179518
3	28.1037993
4	36.0934905
5	31.9884843
6	36.9211695
7	35.0285023
8	36.5633157
9	41.9495499
10	39.2167469
11	36.6579451
12	40.4918864
13	44.0729229
14	46.9932295
15	54.4054407
16	32.6512854
17	51.0049883
18	35.5596795

示意圖(數字僅供參考)

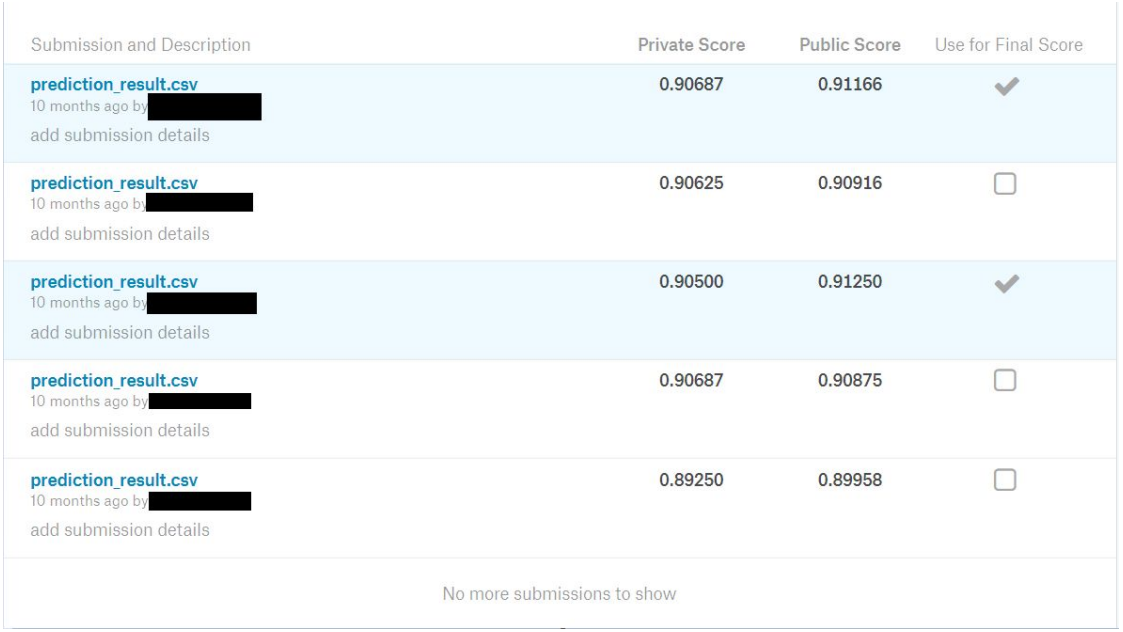
Kaggle

Kaggle Info

- 請自行到kaggle創建帳號(務必使用ntu信箱)
- sample code : code
- Link: https://www.kaggle.com/t/b404886ab374405ba8302aa1add3dab1
- 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

Kaggle Submission

- Maximum Daily Submission: 5 times
- test_data.csv的90筆資料分為:45筆public、45筆private
- Leaderboard上所顯示為public score, 在Kaggle Deadline前可以選擇2份submission作為最後的評分依據(public score & private score)。
- 計分排名:會考慮到public以及private的成績



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

Special Regulation

- 限定開放使用套件
  - All python standard library
  - numpy
  - pandas
  - No other packages can be used!!
    - pytorch, sklearn, numpy.linalg.lstsq 是不可以用的, 否則該程式不予計分。
    - 若對可使用套件有疑問, 請聯絡助教。

Grading / Assignment Regulations

Deadline

- Kaggle Deadline: 10/14/2022 23:59:59 (GMT+8)
- Cool Deadline: 10/16/2022 23:59:59 (GMT+8) (晚 Kaggle 兩天)
- 以 kaggle 的上傳時間為準, 請勿壓線上傳！

Grading Criteria - kaggle (4% + Bonus 1%)

- Kaggle Deadline : 10/14/2022 23:59:59 (GMT+8)
- Private Score Point - 4%
  - 以 10/14/2022 23:59:59 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（Optional）- 1%
  - 修課生 private leaderboard 排名前五名可繳交。
  - 繳交投影片描述實作方法，另外需錄製一份講解影片（少於三分鐘）作一個簡單的 presentation，助教將公布給同學們參考

Grading Criteria - report (6%)

- Programming Report - 2%
  - 解釋什麼樣的data preprocessing可以improve你的training/testing accuracy。請提供數據(例如 kaggle public score RMSE)以佐證你的想法。(1%)
  - 請實作 2nd-order polynomial regression model (不用考慮交互項)。(1%)

    (a) 貼上 polynomial regression 版本的 Gradient descent code 內容
    (b) 在只使用 NO 數值作為 feature 的情況下，紀錄該 model 所訓練出的 parameter 數值 (w2, w1, b) 以及 kaggle public score.
  - Template:

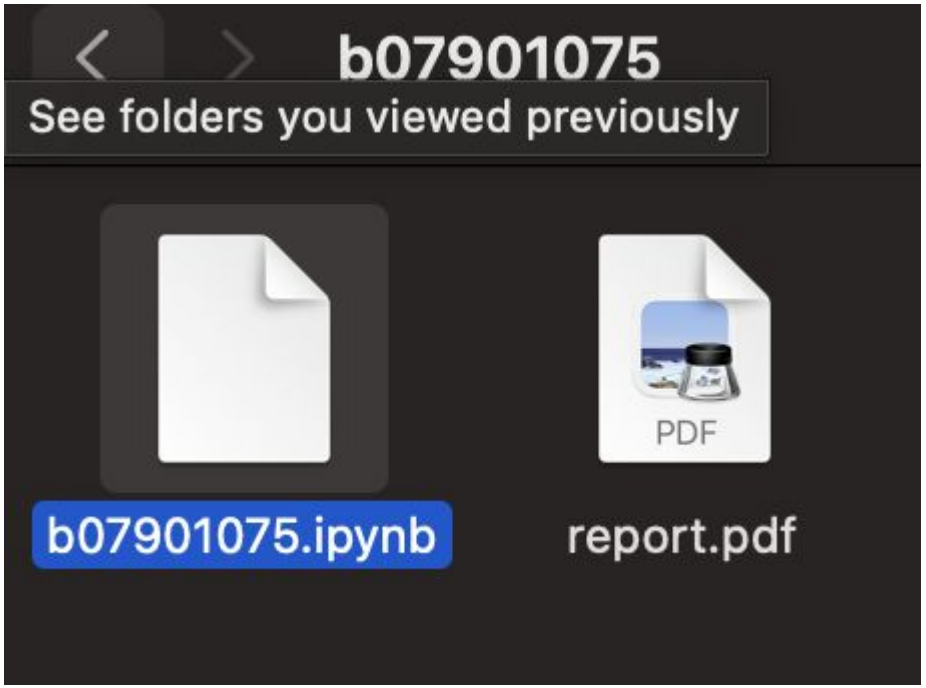
    https://docs.google.com/document/d/1_FLbVbJGv2HwVkUQFerfQ06WbV7teiG_/edit?usp=sharing&ouid=105029010792272788496&rtpof=true&sd=true

Grading Criteria - report (6%)

- Math Problem - (4+0.2) %
  - https://hackmd.io/@lH2AB7kCSAS3NPw2FffsGg/Sk1n8xPWo?fbclid=IwAR0LiCps2fhIZFJT-gYP8kr7KlvLaRvS9-ftLIaPQY5DVgye1AuHM-RW3Yg
  - 共 5 大題
  - Type in latex (preferable) or take pictures of your handwriting
- Combine programming report and math problem in report.pdf.

Hand-in Format

- 一個 zip 檔案, 檔案名稱為 **學號_hw1.zip**, 需包含
  - 程式碼(任意名稱.ipynb)
  - 程式報告＋數學題(report.pdf)

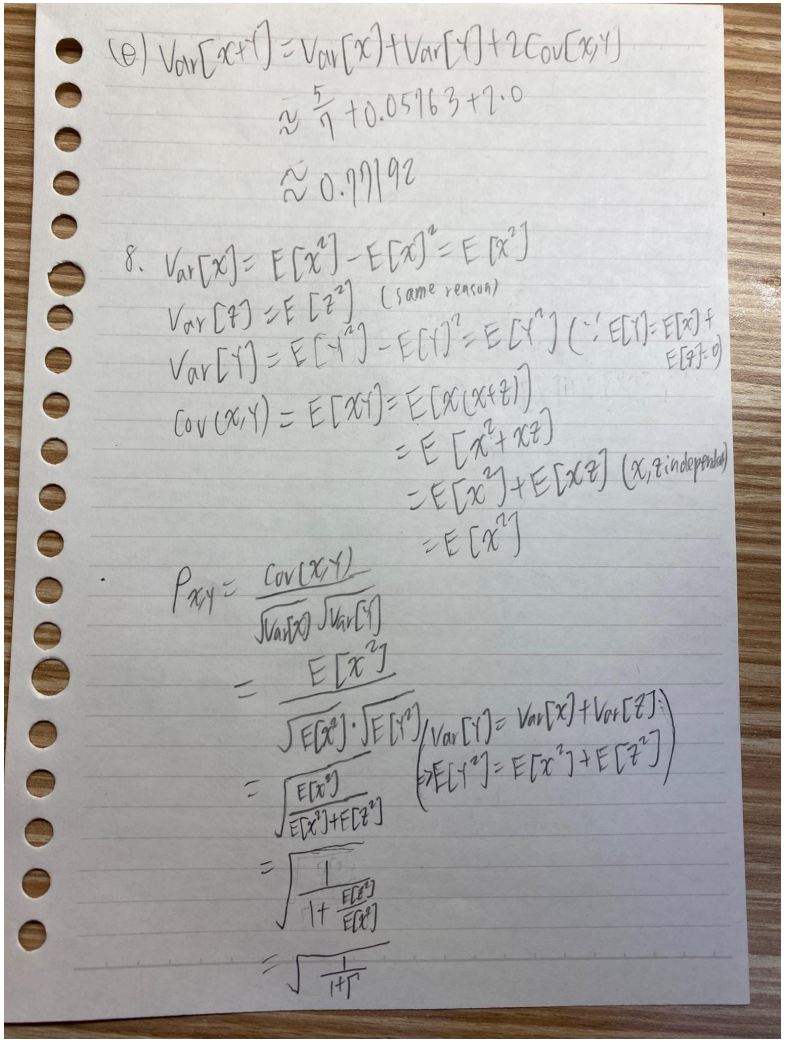


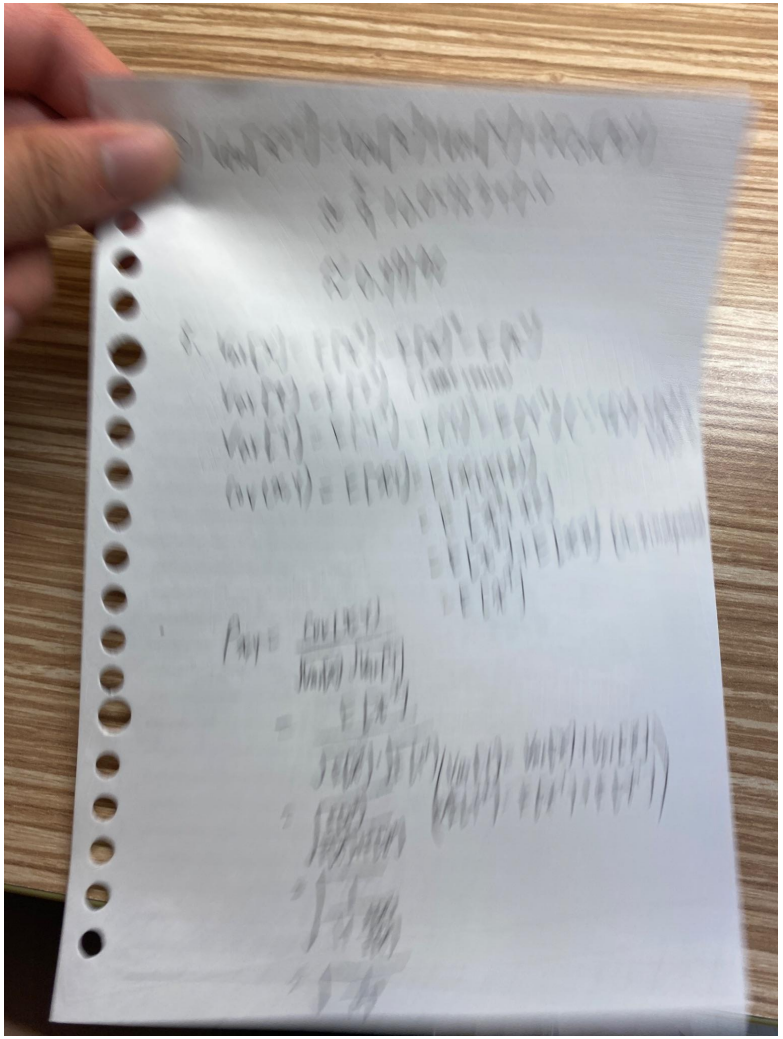
Hand-in Format : Report

- 限制
  - 檔名必須為 report.pdf
  - 請在檔案中標明系級、學號、姓名
  - 按照report模板回答問題, 勿更動題號順序
  - 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)
  - 若有其他參考資料也必須一併附上資料來源、出處

- Cool Deadline: 10/16/2022 23:59:59 (GMT+8) (晚 Kaggle 兩天)

請把作業拍攝清楚(不要興奮到模糊)





其他規定 Other Policy

- Lateness
  - Cool遲交一天(不足一天以一天計算) hw1所得總分將x0.7 (hr)
  - 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
- Reproduce
  - 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。

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
  - 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

TA Hour

- @ 博理530 Tue 13:20-15:20

Hint for HW1 programming

- We provide some suggestion (but not necessary) to pass the baseline.
  - Simple baseline: You might not need so much feature (why?)
  - Strong baseline: Data preprocessing, Training config tuning, Feature selection. (How to define a good feature?)
  - Of course, you can pass the baselines without following the hints. But make sure you don't use the packages which are not allowed!