Assignment 2
Decision Tree

Yi Shiuan Tseng

Po-Chih Kuo

Goal

- Implement a **binary** decision tree with the *restaurant waiting* dataset
- To build a machine learning model to predict the patients' death (‘hospDIED’) from **real** data
- Preprocess the data or fine-tune the model for better performance

Implementation (70%)

Implement the binary decision tree in 3 steps with Restaurant dataset:

- Step 1: calculate the entropy (20%)
- Step 2: search for the best split (20%)
- Step 3: build the decision tree (30%)

	Unnamed: 0	Alternate	Bar	Friday	Hungry	Patrons	Price	Raining	Reservation	Type	WaitEstimate	Wait
0	X1	T	F	F	T	Some	High	F	T	French	8	T
1	X2	T	F	F	T	Full	Low	F	F	Thai	40	F
2	X3	F	T	F	F	Some	Low	F	F	Burger	8	T
3	X4	T	F	T	T	Full	Low	F	F	Thai	12	T
4	X5	T	F	T	F	Full	High	F	T	French	70	F
5	X6	F	T	F	T	Some	Medium	T	T	Italian	3	T
6	X7	F	T	F	F	None	Low	T	F	Burger	7	F
7	X8	F	F	F	T	Some	Medium	T	T	Thai	6	T
8	X9	F	T	T	F	Full	Low	T	F	Burger	80	F
9	X10	T	T	T	T	Full	High	F	T	Italian	20	F
10	X11	F	F	F	F	None	Low	F	F	Thai	8	F
11	X12	T	T	T	T	Full	Low	F	F	Burger	40	T

Prediction in Real Case (20%)

To classify **death('hospDIED')** in the *MIMIC* dataset.

x_train

	subject_id	age	los	CA	DNR	CMO	DNI	indextime	female	first_careunit
0	10246985	48	1.826250	0	0	0	0	30JUN38:20:25:48	1	10
1	14538806	59	10.581123	0	0	0	0	03SEP59:14:15:14	1	10
2	12850130	73	2.065255	0	0	0	0	30SEP87:20:50:56	1	4
3	11810761	84	1.435428	0	0	0	0	06JUN22:11:36:44	1	10
4	13109236	66	1.443414	0	0	0	0	25MAR31:21:09:15	0	1
...	...	...	...	...	...	...	...	...	...	...
24636	10192748	82	6.488113	0	0	0	0	07OCT39:06:40:24	1	11
24637	16510199	52	2.485833	0	0	0	0	22MAY53:18:51:15	0	11
24638	16753060	62	2.914444	0	0	0	0	01DEC76:21:46:37	0	10
24639	18223630	30	9.803310	0	0	0	0	12NOV77:14:57:21	1	2
24640	18101124	72	6.168958	0	0	0	0	27FEB83:17:38:45	1	1

24641 rows × 84 columns

...

y_train

hospDIED
0
0
0
0
0
...
0
0
0
0
0

Note:

- Decision tree is recommended but not mandatory.

The MIMIC Database

- Medical Information Mart for Intensive Care
- A large, freely-available database
- Over 40,000 patients who stayed in critical care units

- We extract 27379 cases with 84 attributes and 1 label(hospDIED)
- 2738 cases split to the test set
- Data Description: https://docs.google.com/spreadsheets/d/1pxqxQFhIcv_hrgWEtwhXE6zBVQ5ISa-13PIhvXMtWCY/edit#gid=0

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035.

Data

- The Restaurant *Waiting* data(data.csv) for implementation
- The *MIMIC* dataset(x_train, y_train, x_test.csv) for real prediction
- Both are included in the template already.

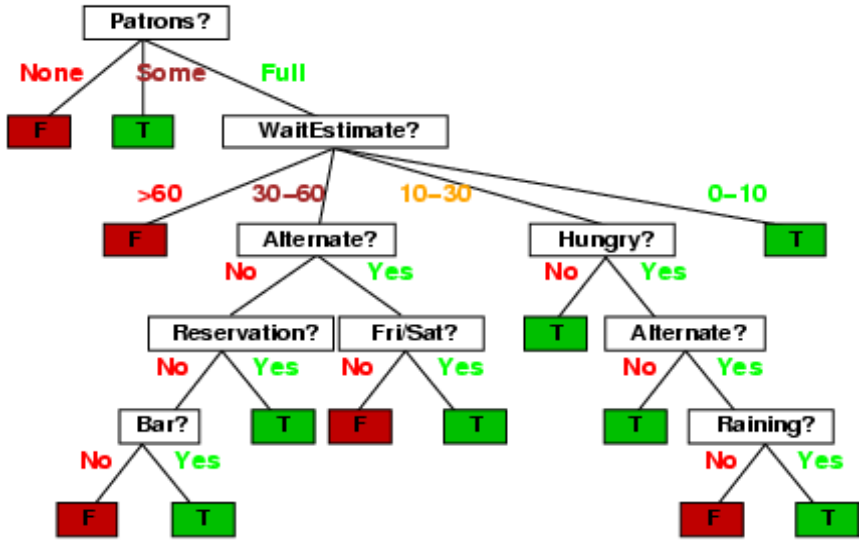
```
#Read  data
x_train = pd.read_csv('https://raw.githubusercontent.com/aubreyyy24/HW2_data/main/x_train.csv')
y_train = pd.read_csv('https://raw.githubusercontent.com/aubreyyy24/HW2_data/main/y_train.csv')
x_train.head()
```

	subject_id	age	los	CA	DNR	CMO	DNI	indextime	female	first_careunit	insurance	ethnicity	admission_type	smoking	congestive_heart_failure
0	10246985	48	1.826250	0	0	0	0	30JUN38:20:25:48	1	10	1	6	6	1	0
1	14538806	59	10.581123	0	0	0	0	03SEP59:14:15:14	1	10	1	4	9	1	0
2	12850130	73	2.065255	0	0	0	0	30SEP87:20:50:56	1	4	1	4	6	0	0
3	11810761	84	1.435428	0	0	0	0	06JUN22:11:36:44	1	10	3	4	6	1	0
4	13109236	66	1.443414	0	0	0	0	25MAR31:21:09:15	0	1	1	4	8	1	0

5 rows × 84 columns

Bonus (extra 10%)

- Visualize your decision tree of the classification part (with MIMIC data)
- Your visualization image of the decision tree can contain **five** levels at most
- Save your visualization image as **[STUDENT_ID]_visualization.png**!



Report (10%)

- List the top 3 splitting features and their thresholds of your model (in the MIMIC dataset)
- Briefly describe how you build the decision tree
- Describe if you pay extra effort to improve your model
- If you preprocess the MIMIC data in the second part(selecting feature s...), describe the work and reasons
- Summarize your work
- Do not exceed 2 pages!
- Name your report file as "**[STUDENT_ID]_report.pdf**"

Grading Policy

Item	Score
Implementation	70%
Classification (Performance)	20%
Report	10%
Decision Tree Visualization (bonus)	10%

You will have the following items

- Template: **HW2_Decision_Tree.ipynb** (input data inside)
- Sample output: **sample_implementation.csv**
  sample_prediction.csv

Template

- Except for the imported packages in the template, you cannot use any other packages in the first(implementation) part
- Remember to save the code file to **[STUDENT_ID]_hw2.ipynb**

HW2_Decision_Tree.ipynb

檔案 編輯 檢視畫面 插入 執行階段 工具 說明 已儲存所有變更

+ 程式碼 + 文字

HW2: Decision Tree

In assignment 2, you need to:

1. Implement the decision tree in 3 steps (with the *breast cancer* dataset)

- Step 1: calculate the entropy
- Step 2: search for the best split
- Step 3: build the decision tree

2. Predict the patients' death (hospDIED) in the *MIMIC* dataset

Please fill in your **studentID** here.

1. Implement the Decision Tree

In the first part, you need to implement the decision tree by completing the given funcitions.

Also, you need to run those functions with given input variables and save the output to the implementation csv file **[STUDENT_ID]_implementation.csv**.

2. Classification with the MIMIC Dataset

In the second part, you need to classify **'hospDIED'(death)** in the MIMIC dataset.

Note: You can now import the packages you need here!

Note: Decision tree is recommended but not mandatory.

Report *(10%)*

Report should be submitted as a pdf file! (**[STUDENT_ID]_report.pdf**)

- List the top 3 splitting features and their thresholds of your model
- Briefly describe how you build the decision tree
- Summarize your work
- If you preprocess the MIMIC data in the second part (selecting features...), describe the work and reasons
- Do not exceed 2 pages!

Save the Code File

Please save your code and submit it as an ipynb file! (**[STUDENT_ID]_hw2.ipynb**)

Output CSV File Format - Implementation

- Named as "**[StudentID]_implementation.csv**"
- There should be 3+2n rows in your csv file:
  - Entropy : 1
  - BestSplit column 2, BestSplit value 3
  - Tree features 4~4+(n-1), Tree thresholds 4+n~4+2n-1
  - n is the number of the features you used
- Please make sure that your model can correctly output this format of csv file

example: if n=4

		A
Entropy	1	0
BestSplit	2	Patrons
	3	0
Tree	4	Patrons
	5	Price
	6	Raining
	7	Hungry
	8	0
	9	0
	10	0
	11	0

column

threshold

features

thresholds

Output CSV File Format - Prediction

- Named as "**[StudentID]_prediction.csv**"
- y_test contains 2738 cases
- Each row represents "subject_id, hospDIED (Prediction)"
- Please make sure your model can correctly output this format of csv file

subject_id	hospDIED
10246985	0
14538806	0
12850130	0
11810761	0
13109236	0
12601474	1
19738421	0
15051600	0
19734681	1
11236474	0
16968810	0
11382142	0
16093826	1
10625523	0
15676460	0
10761467	0
12084606	0
18908038	0
18910094	0
10337761	0
15566609	1
18122436	0
11639209	0
16962073	0
11779110	0

Assignment 2 Requirement

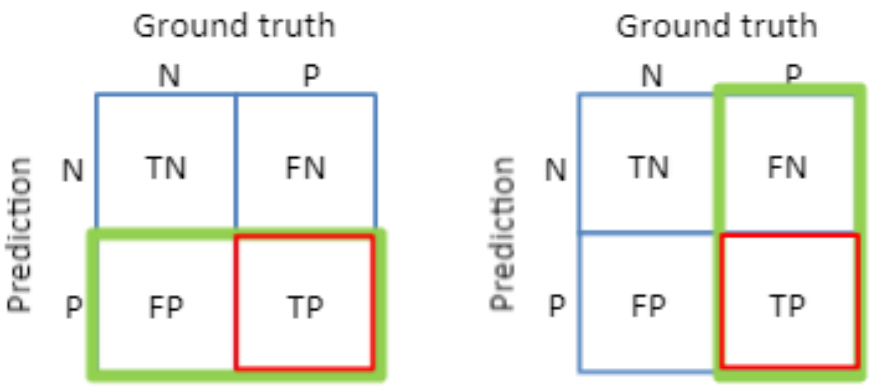
- Do it individually! Not as a team! (team is for final project)
- Announce date: 2021/10/21
- Deadline: 2021/11/4 23:59 (Late submission is not allowed!)
- Hand in your files in the following format
    - [StudentID]_hw2.ipynb
    - [StudentID]_implementation.csv
    - [StudentID]_prediction.csv
    - [StudentID]_visualization.png
    - [StudentID]_report.pdf
    - Compress all files into [StudentID]_HW2.zip

The Evaluation Metric

- F1 score: **F1 Score = 2*(Recall * Precision) / (Recall + Precision)**

- For example:
- The class you predicted:
- $\hat{y}$ = [1, 1, 0, 0, 0, 0, 1]
- Ground Truth :
- y = [0, 0, 0, 0, 0, 1, 1]

- F1 score = 0.4



Penalty

- 0 points if any of the following conditions

- Plagiarism
- Late submission
- Not using template or import any other packages in implementation part
- Incorrect input/output format
- Incorrect submission format

Questions?

- TA: Yi Shiuan Tseng (aubreytys@gapp.nthu.edu.tw)