Introduction to Pattern Recognition Homework 3 announcement

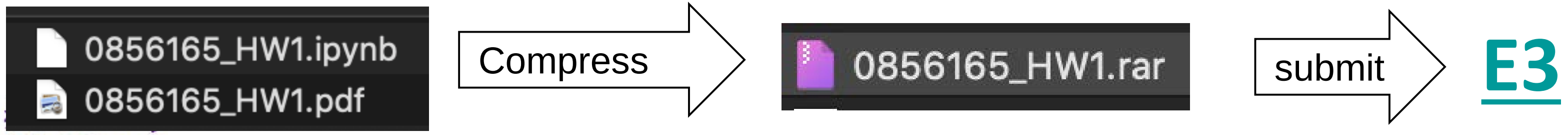
TA: 楊証琨, Jimmy

Ph.D. student at National Taiwan Universitiy

d08922002@ntu.edu.tw

Homework 3

- **Deadline: May. 22, Fri at 23:59.**
  1. Code assignment (80%): Implementing Decision Tree & Random Forest
  2. Short answer questions (20%)
- Submit your **1) code (.py/.ipynb)** and **2) reports (.pdf)** on E3
  - Sample Code
  - HW3 questions
- Please follow the **file naming rules <STUDENT ID>_HW3.pdf,** otherwise, you will get penalty of your scores

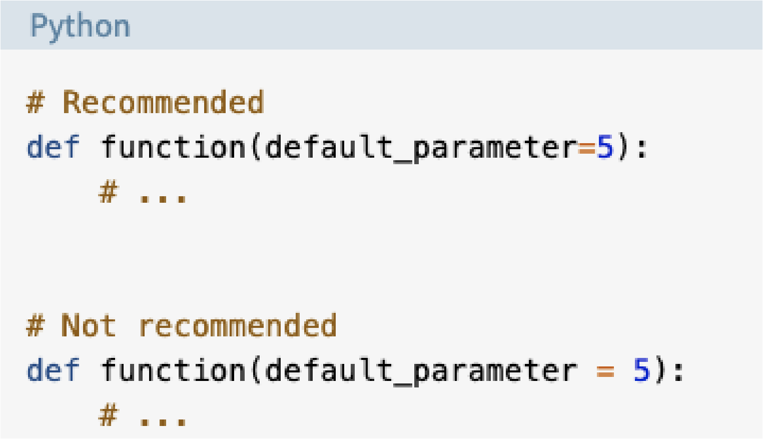


0856165_HW1.ipynb
0856165_HW1.pdf

0856165_HW1.rar

Coding

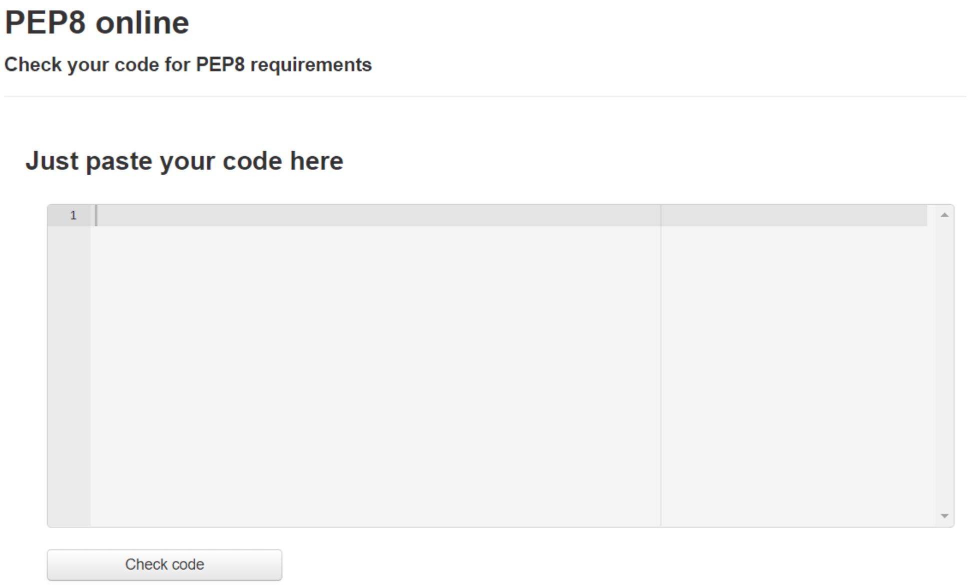
- Write beautiful Python codes with PEP8 guidelines for readability. Basic requirement: use whitespace correctly!
- PEP8 online checker



Python

```python
# Recommended
def function(default_parameter=5):
    # ...


# Not recommended
def function(default_parameter = 5):
    # ...
```



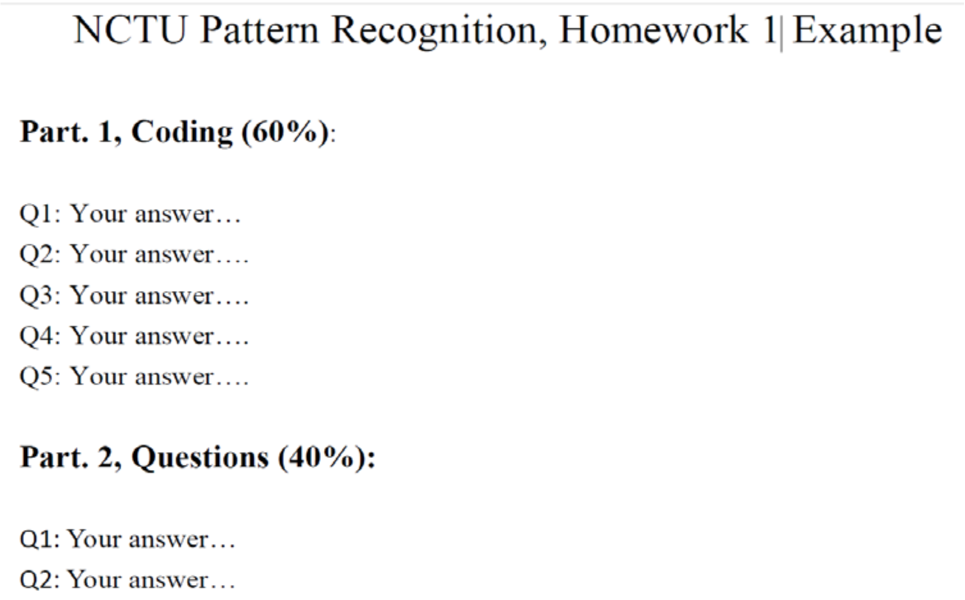
PEP8 online

Check your code for PEP8 requirements

Just paste your code here

Reports

- Submit in PDF format
- Include the answers of coding part in the reports!
- Please see the sample submission file on E3



NCTU Pattern Recognition, Homework 1| Example

Part. 1, Coding (60%):

Q1: Your answer…

Q2: Your answer….

Q3: Your answer….

Q4: Your answer….

Q5: Your answer….

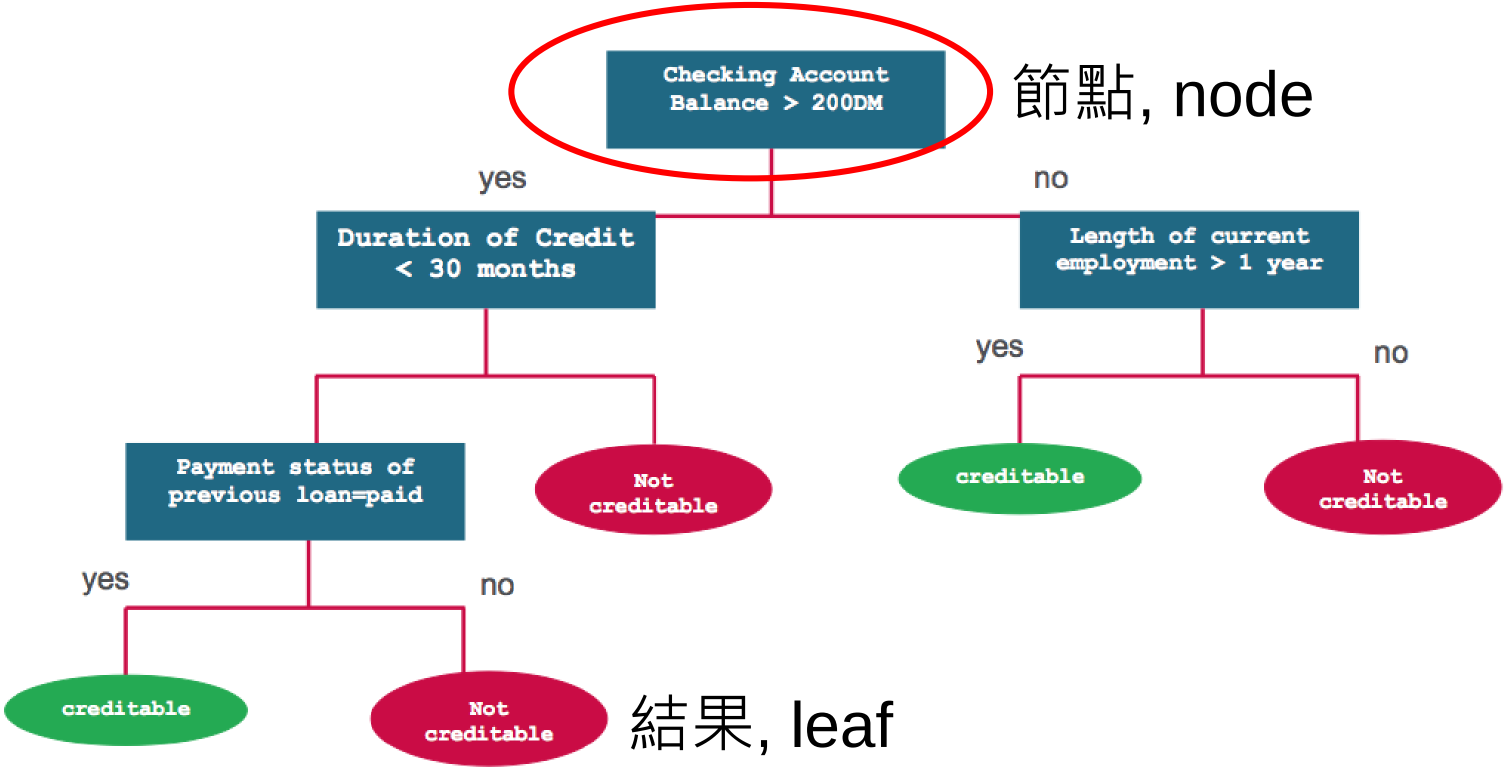
Part. 2, Questions (40%):

Q1: Your answer…

Q2: Your answer…

Decision Tree Algorithm

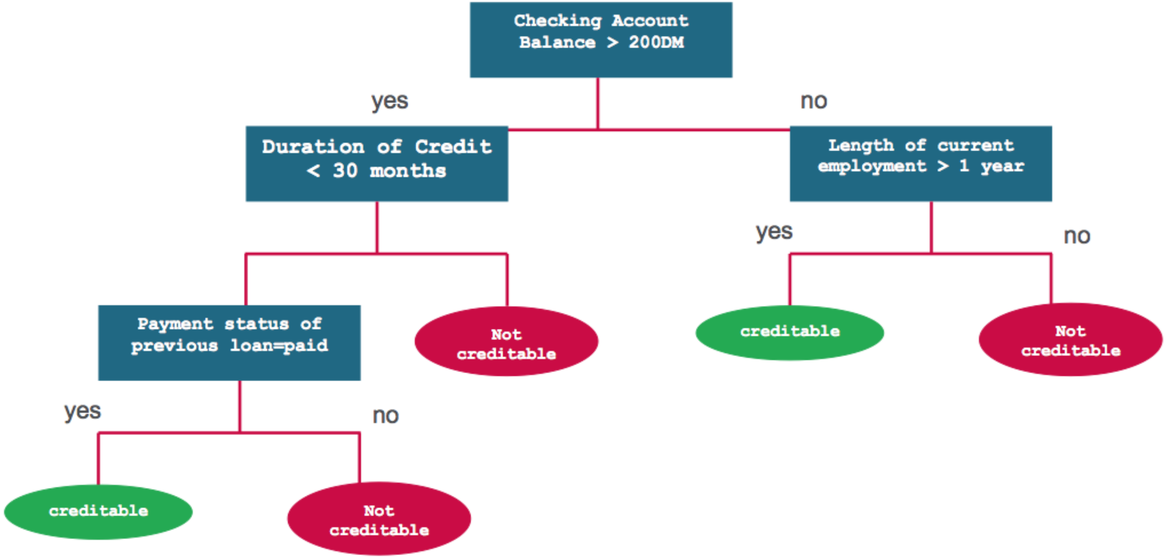
- Whether to approve the loan for customer?



Checking Account Balance > 200DM
節點, node
yes
no
Duration of Credit < 30 months
Length of current employment > 1 year
yes
no
Payment status of previous loan=paid
Not creditable
creditable
Not creditable
yes
no
creditable
Not creditable
結果, leaf

Decision Tree Algorithm

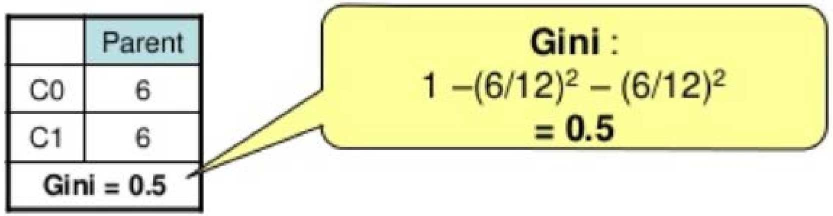
- How to find the feature for making decisions? What’s the value of feature?
- Find the features to separate data that the class at the resulting nodes are as **pure** as possible


Checking Account Balance > 200DM
yes
no
Duration of Credit < 30 months
Length of current employment > 1 year
yes
no
Payment status of previous loan=paid
Not creditable
creditable
Not creditable
yes
no
creditable
Not creditable

How to measure “pure”?

1. Entropy: the smaller, the purer
2. Gini-index: the smaller, the purer

$$Gini = 1 - \sum_j p_j^2$$



	Parent
C0	6
C1	6
Gini = 0.5	

Gini :
1 –(6/12)^2 – (6/12)^2
= 0.5

$$Entropy = -\sum_j p_j \log_2 p_j$$

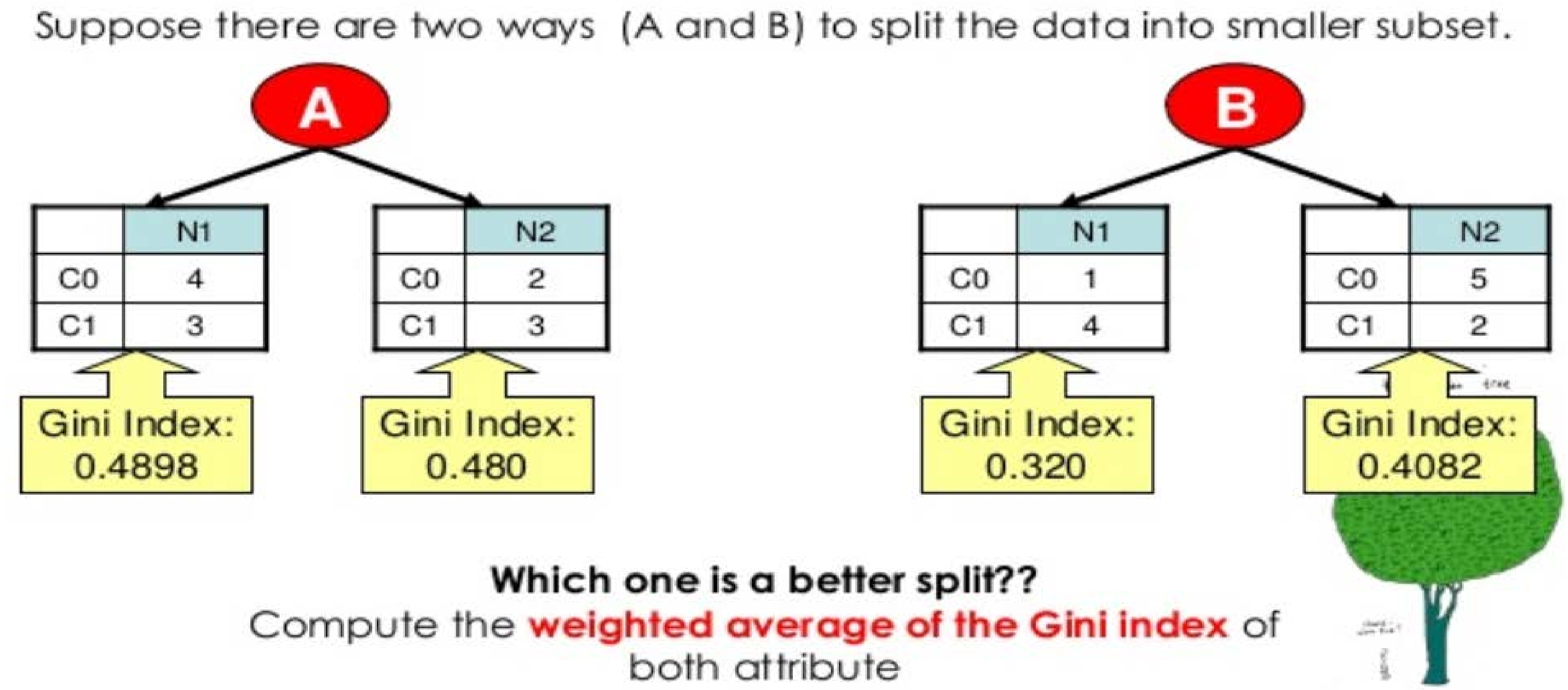
- If all classes are the same in one node

$$entropy = -1 \log_2 1 = 0$$

- If the classes are half-and-half

$$entropy = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$$

How to find best split?



Suppose there are two ways (A and B) to split the data into smaller subset.

	N1
C0	4
C1	3

Gini Index: 0.4898

	N2
C0	2
C1	3

Gini Index: 0.480

	N1
C0	1
C1	4

Gini Index: 0.320

	N2
C0	5
C1	2

Gini Index: 0.4082

Which one is a better split??

Compute the **weighted average of the Gini index** of both attribute

Decision Tree pseudo code

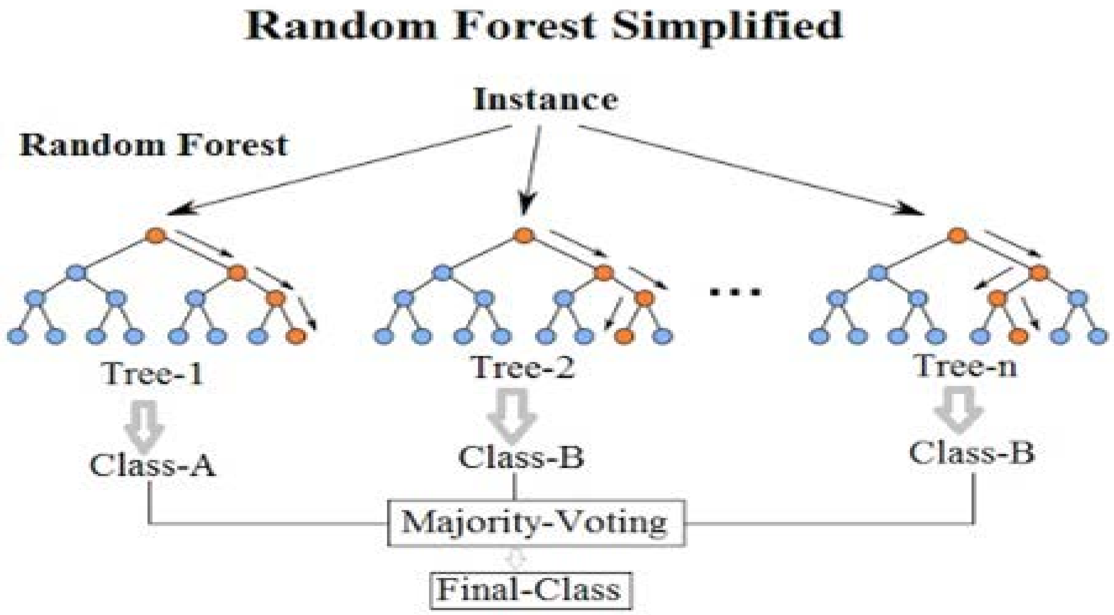
- Until stopped
  a. Select a node
  b. loop all values of all features
     - ◆ partition the node and calculate the pure of data
     - ◆ find the value of feature can yield lowest value of gini or entropy
  c. Split the node using the feature value found in step b.
  d. Go to each node and repeat step a to c.
- Stopping criteria
  - ➢ Each leaf-node contains data of the same class
  - ➢ Depth of the tree is more than some pre-specified limit

Overfitting

- Decision Tree can find a unique path for each data if we don’t pre-specified any limits such as the depth of the node
- It may overfit the training data if there exist some outliers in the data

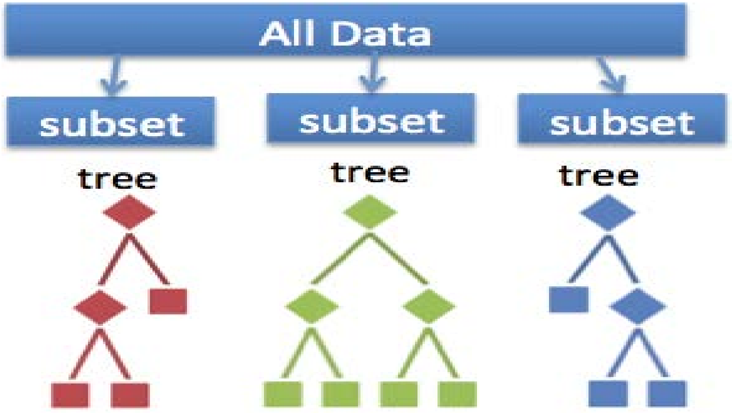
Ensemble method of Decision Trees: Bagging

- **Bagging (Bootstrap aggregating)**: Fit many large trees to bootstrap-resampled versions of the training data, and classify by majority vote



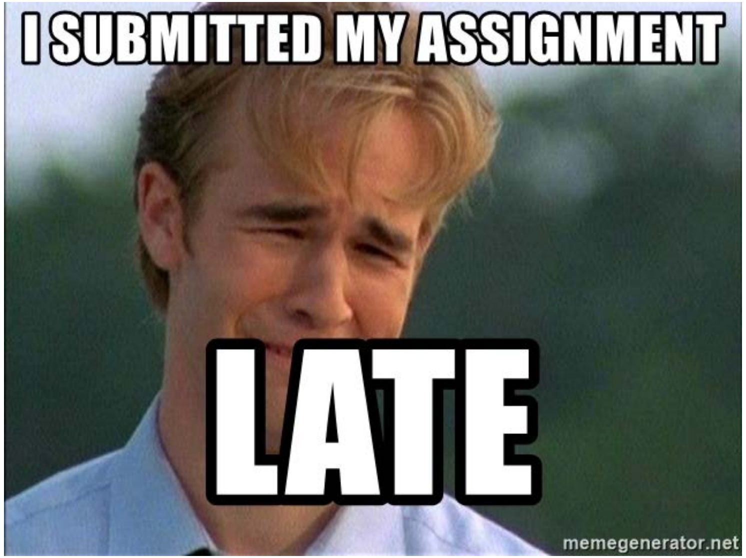
Random Forest: Where is the “Random”?

- Bootstraped dataset
- Each tree in the forest may grow with different data and features
- Which features or data to be used are **randomly** sampled to grow the tree



Late Policy

- We will deduct a late penalty of 20 points per additional late day
- For example, If you get 90 points of this HW but delay for two days, your will get only 90- (20 x 2) = 50 points!



Honor code

- We have found that some students develop their codes based on those by other classmates or on Internet in HW1
  - It is NOT allowed

- You should implement all algorithms by yourself

- If there is any plagiarism in your homework, you will get no points

Notice

- Submit your homework on E3-system !
- Check your email regularly, we will mail you if there are any updates or problems of the homework
- If you have any questions or comments for the homework, please mail Jimmy and Chung-Hsuan and cc Prof. Lin
  - Prof. Lin: **lin@cs.nctu.edu.tw**
  - TA, Jimmy: **d08922002@ntu.edu.tw**
  - TA, Chung-Hsuan: **scott19880525@gmail.com**

Have fun!