Pattern Recognition
Homework 3 announcement

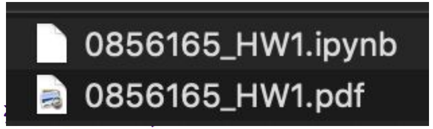
TA: 楊証琨 Jimmy

Ph.D. student at National Taiwan Universitiy

d08922002@csie.ntu.edu.tw

Homework 3

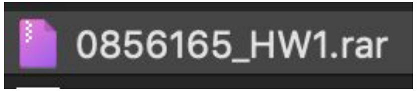
- **Deadline: May. 4, 23:59**
  1. Code assignment (80%): Implementing decision tree, adaboost and random forest by only **NumPy**
  2. Short answer questions (20%)
- **Submit your code (.py/.ipynb) and reports (.pdf) on E3**
  - Sample Code
  - HW3 questions
- Please follow the **file naming rules <STUDENT ID>_HW3.pdf,** otherwise, you will get penalty of your scores



0856165_HW1.ipynb

0856165_HW1.pdf

Compress →

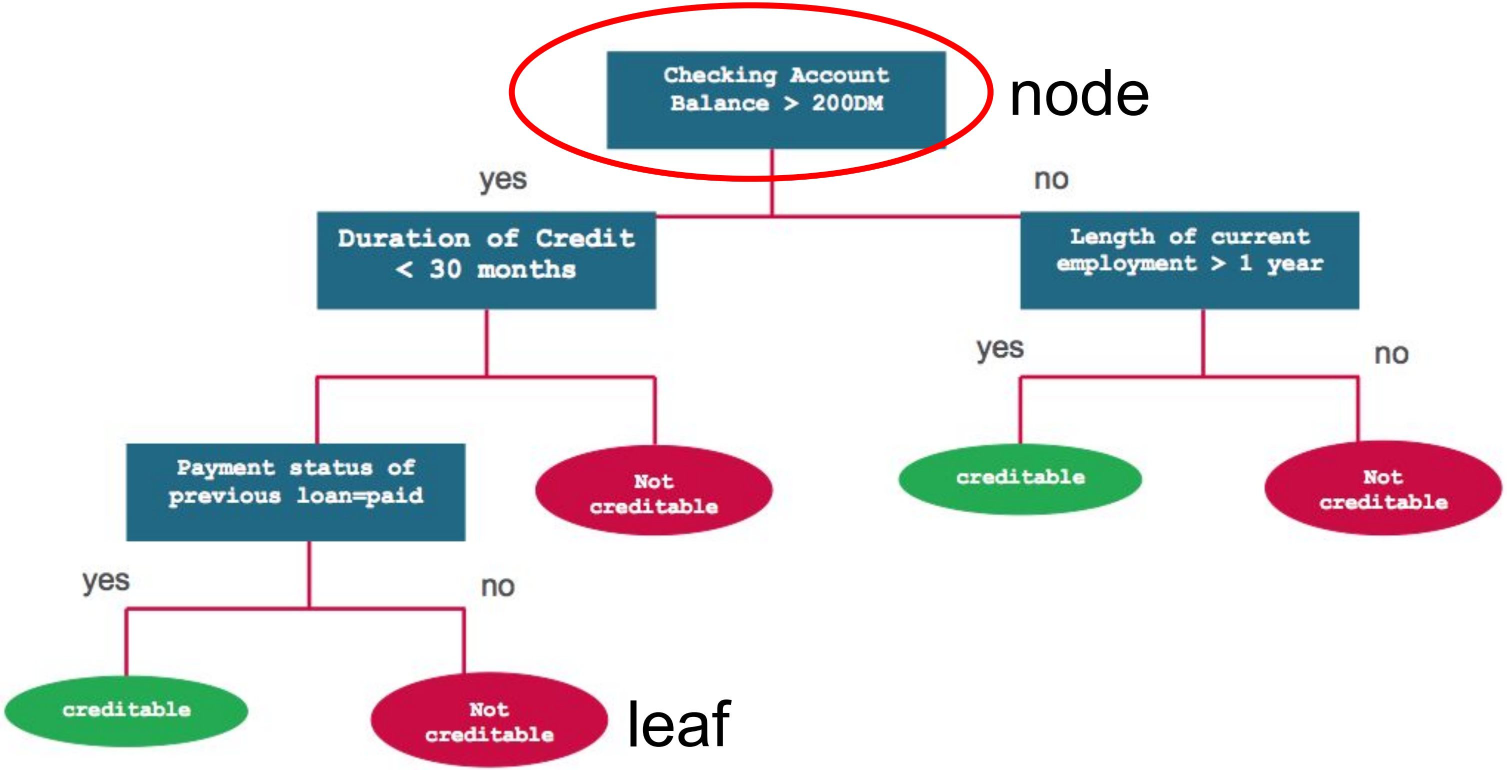


0856165_HW1.rar

submit → **E3**

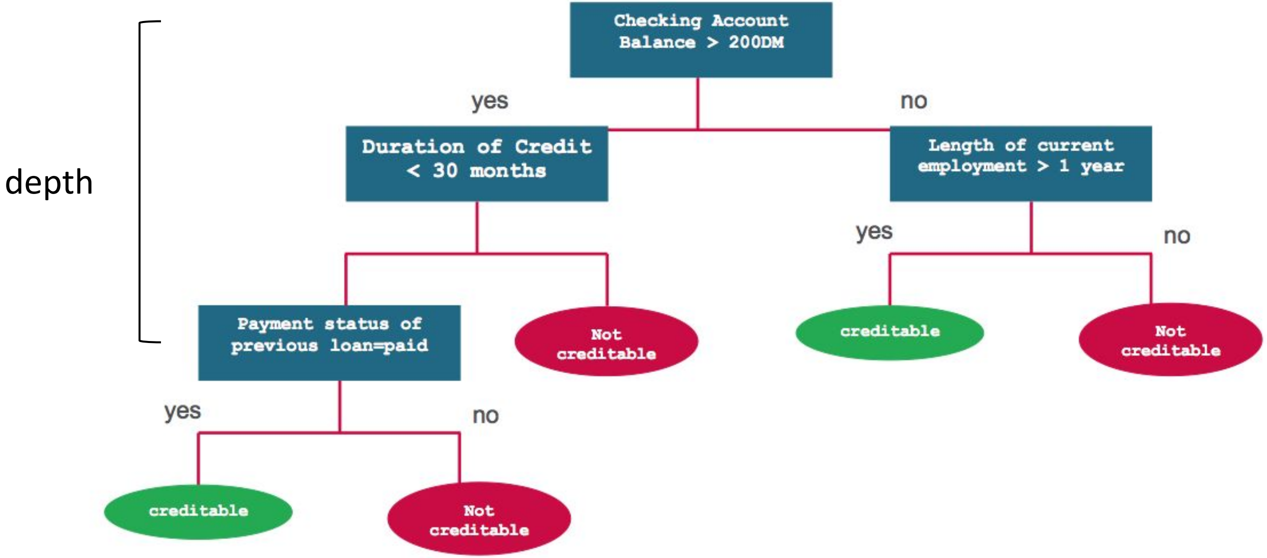
Decision tree algorithm

- Whether to approve the loan for a customer?


Checking Account Balance > 200DM
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
leaf

Decision tree algorithm

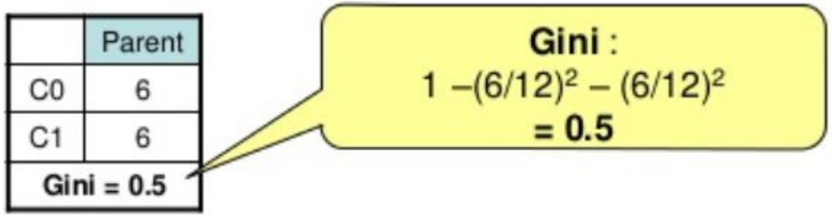
- How to find the feature to make the decisions?
- Find the feature to split data that the class at the resulting nodes are as **pure** as possible


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
depth

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

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

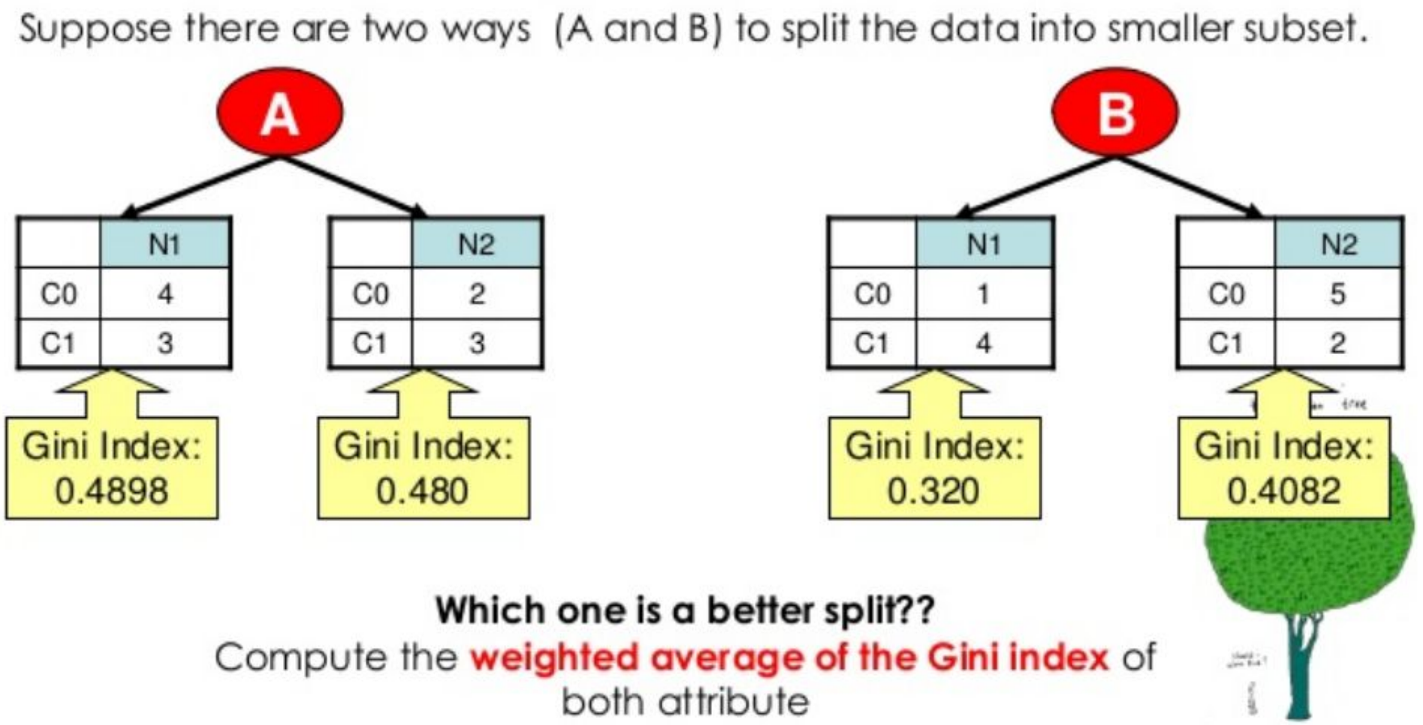
- If all classes are the same in one node

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

- 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.

A:

	N1
C0	4
C1	3

Gini Index: 0.4898

	N2
C0	2
C1	3

Gini Index: 0.480

B:

	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

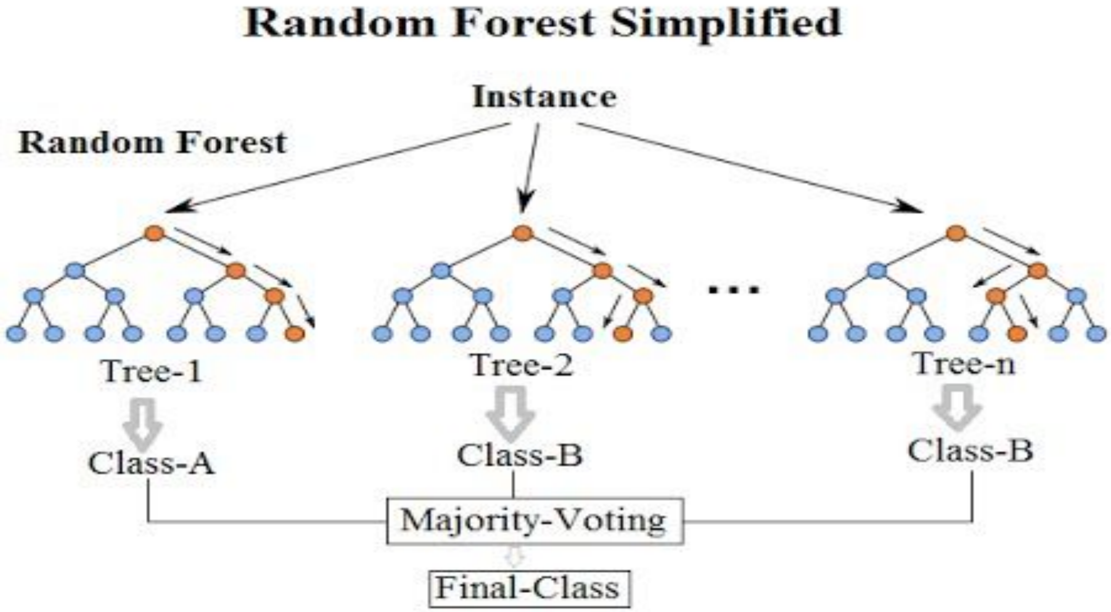
- Until stopped
  a. Select a node
  b. loop all values of all features
     - partition the node and calculate the purity of data
     - find the value of feature can yield lowest value of gini or entropy
  c. Split the node using the feature value found in step b.
  d. Go to next node and repeat step a to c.
- Stopping criteria
  - The data in each leaf-node belongs to the same class
  - **Depth of the tree is equal to 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**

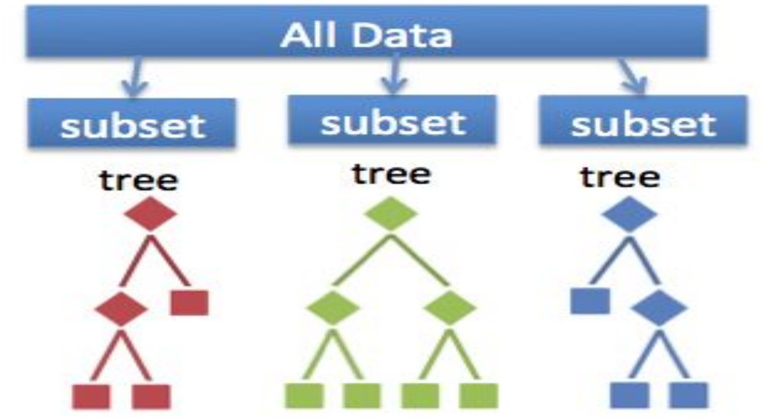
Ensemble method of decision trees: Bagging

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



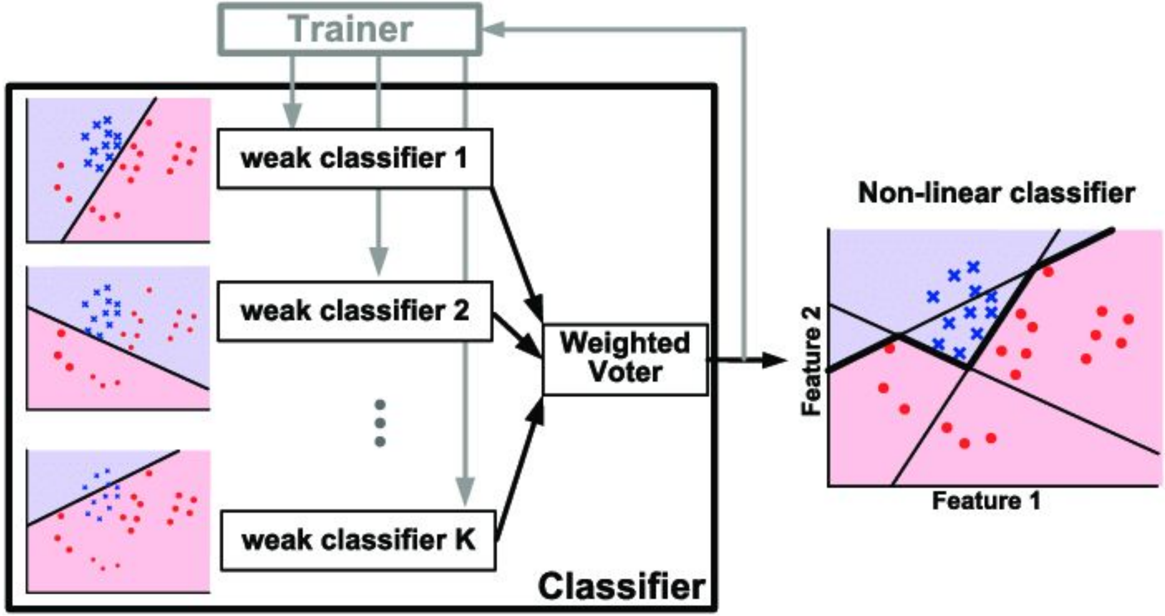
Random forest: Where is the "randomness"?

- 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



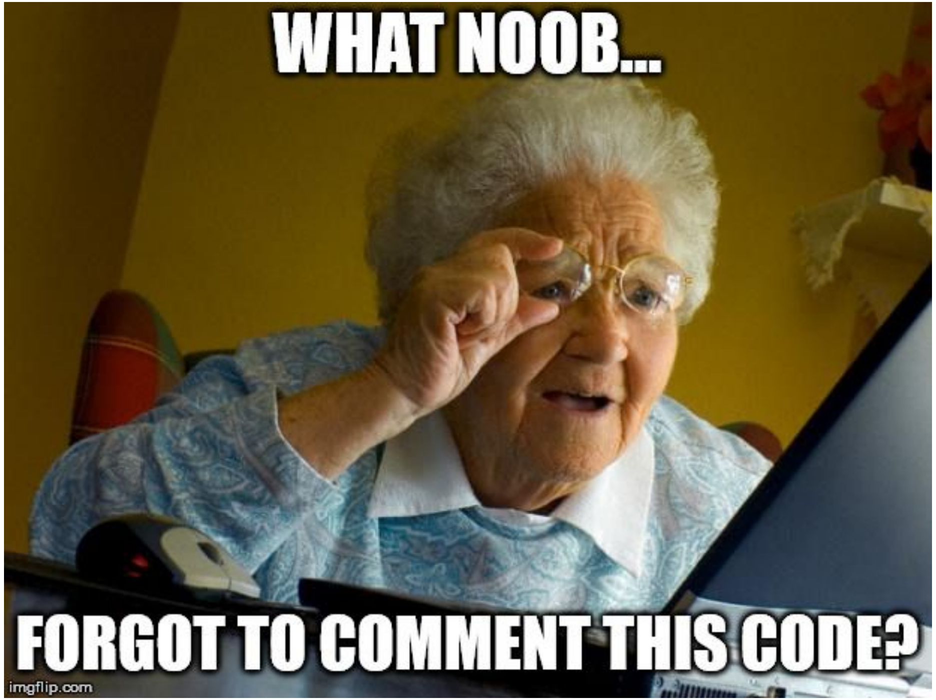
Ensemble method: Boosting

- **Boosting**: Iteratively fit many **shallow** trees and get the results by weighting those classifiers



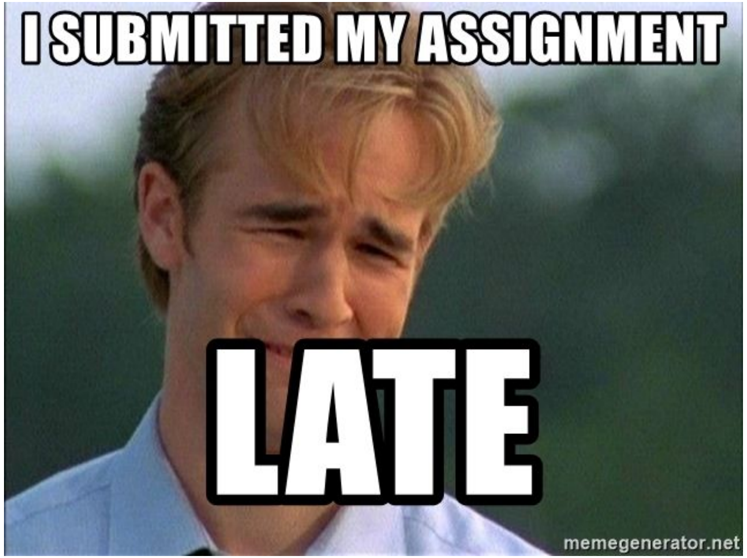
Coding

- **Make sure to comment your code!**
  - Document each step of your model
- [PEP8 online checker]



Late policy

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



Notice

- **All of your model should get the accuracy over 0.8**
- 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 me and cc Prof. Lin
  - Prof. Lin, **lin@cs.nctu.edu.tw**
  - TA Jimmy, **d08922002@csie.ntu.edu.tw**
  - TA 晨軒, **derekt.cs06@nctu.edu.tw**
  - TA 政儒, **ace52751208@gmail.com**

Have fun!