CS 412 Intro. to Data Mining

Chapter 8. Classification: Basic Concepts

Jiawei Han, Computer Science, Univ. Illinois at Urbana-Champaign, 2017

Chapter 8. Classification: Basic Concepts

- Classification: Basic Concepts
- Decision Tree Induction
- Bayes Classification Methods
- Linear Classifier
- Model Evaluation and Selection
- Techniques to Improve Classification Accuracy: Ensemble Methods
- Additional Concepts on Classification
- Summary

สร้าง Model แบบมีคนสอน, มีจุดมุ่งหมาย ไม่มีคนสอน

Supervised vs. Unsupervised Learning (1)

Supervised learning (classification)

- Supervision: The training data such as observations or measurements are accompanied by **labels** indicating the classes which they belong to

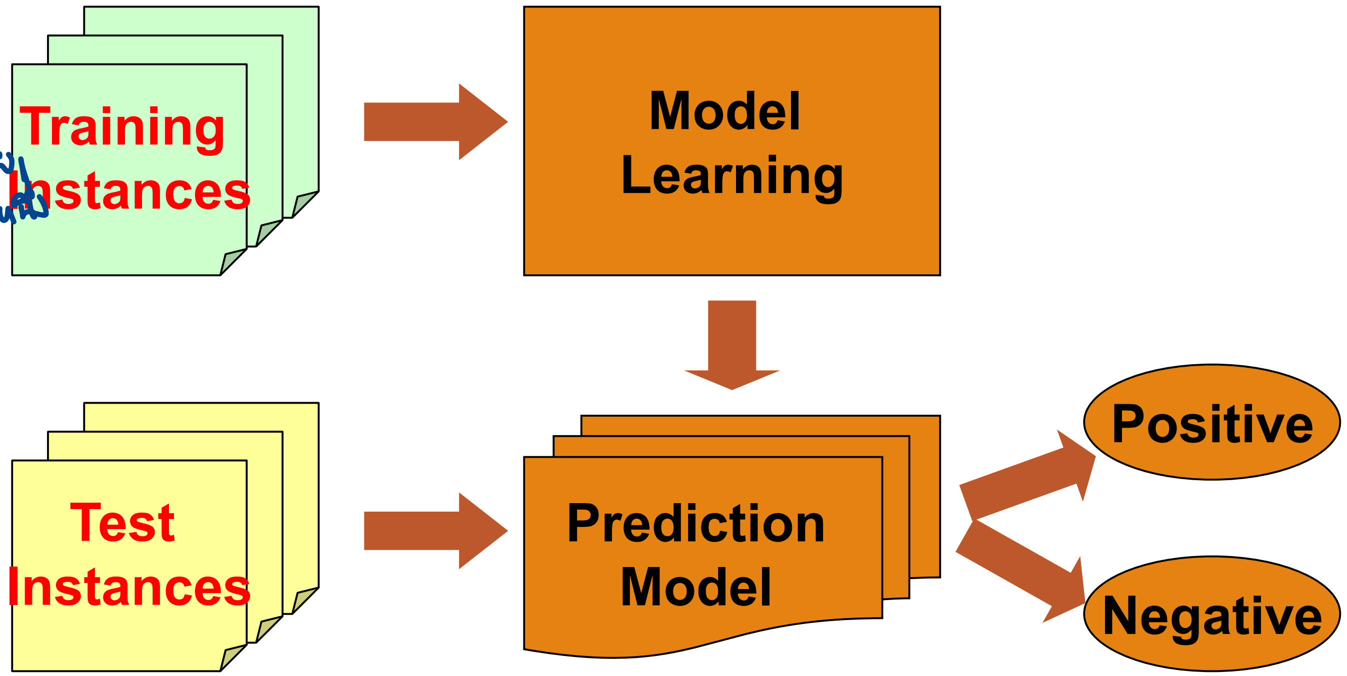
มีการทำ 2 ขั้นตอน

- New data is classified based on the models built from the training set

Training Data with class label:

เพื่อทำนาย Column นี้

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no



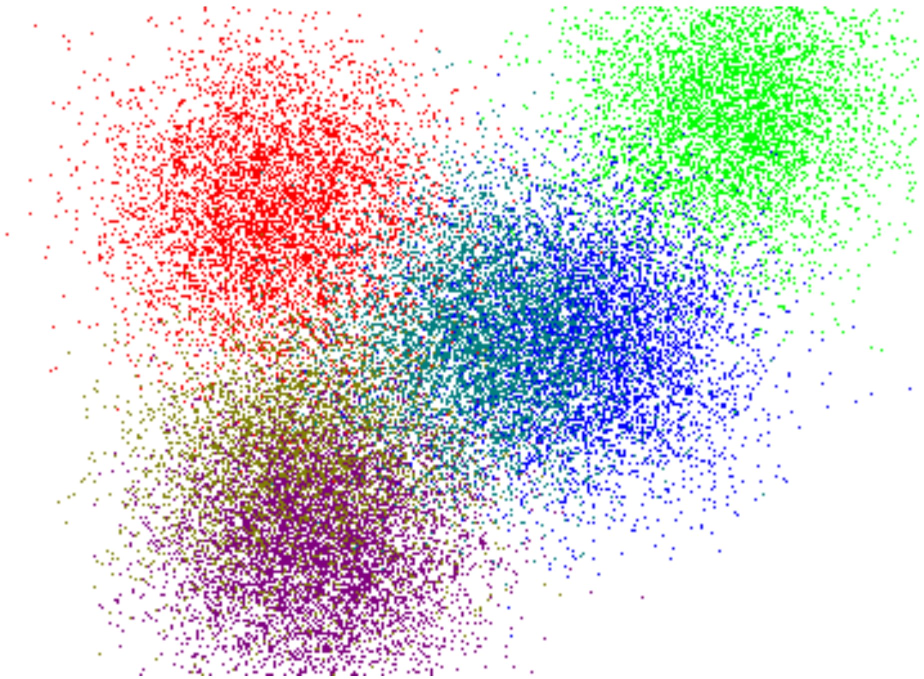
Supervised vs. Unsupervised Learning (2)

- Unsupervised learning (clustering)
  - The class labels of training data are unknown ไม่รู้จัก
  - Given a set of observations or measurements, establish the possible existence of classes or clusters in the data

ให้มันแบ่งกลุ่ม ตัวใกล้กันจับกลุ่มกัน

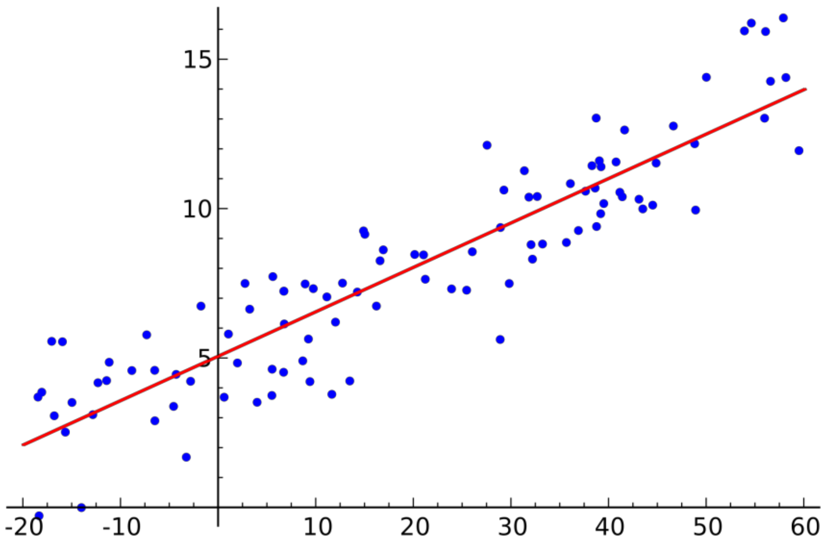
ไม่เหมือนกัน

ก็แยกกัน



Prediction Problems: Classification vs. Numeric Prediction

- Classification
  - Predict categorical class labels (discrete or nominal)
  - Construct a model based on the training set and the **class labels** (the values in a classifying attribute) and use it in classifying new data
- Numeric prediction
  - Model continuous-valued functions (i.e., predict unknown or missing values)
- Typical applications of classification
  - Credit/loan approval
  - Medical diagnosis: if a tumor is cancerous or benign
  - Fraud detection: if a transaction is fraudulent
  - Web page categorization: which category it is



Classification—Model Construction, Validation and Testing

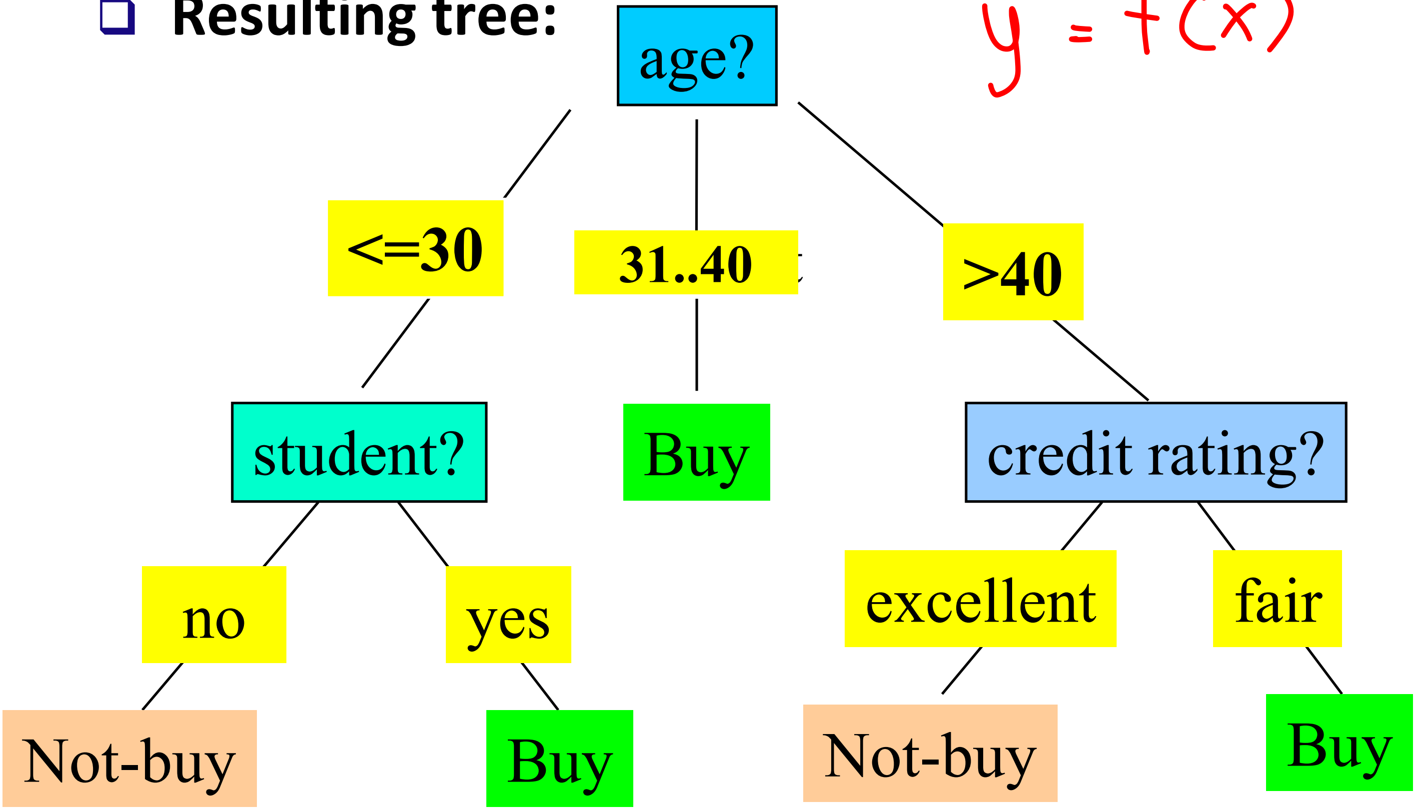
- **Model construction** เอา Data มา train โดย Algorithm ต่างๆ
  - Each sample is assumed to belong to a predefined class (shown by the **class label**)
  - The set of samples used for model construction is **training set**
  - Model: Represented as decision trees, rules, mathematical formulas, or other forms
- **Model Validation and Testing**:
  - **Test:** Estimate accuracy of the model
    - The known label of test sample is compared with the classified result from the model
    - *Accuracy:* % of test set samples that are correctly classified by the model
    - Test set is independent of training set
  - **Validation**: If *the test set* is used to select or refine models, it is called **validation** (or development) **(test) set**
- **Model Deployment:** If the accuracy is acceptable, use the model to classify new data

Chapter 8. Classification: Basic Concepts

- Classification: Basic Concepts
- Decision Tree Induction
- Bayes Classification Methods
- Linear Classifier
- Model Evaluation and Selection
- Techniques to Improve Classification Accuracy: Ensemble Methods
- Additional Concepts on Classification
- Summary

Decision Tree Induction: An Example

- **Decision tree construction**:
  - A top-down, recursive, divide-and-conquer process
- **Resulting tree:**



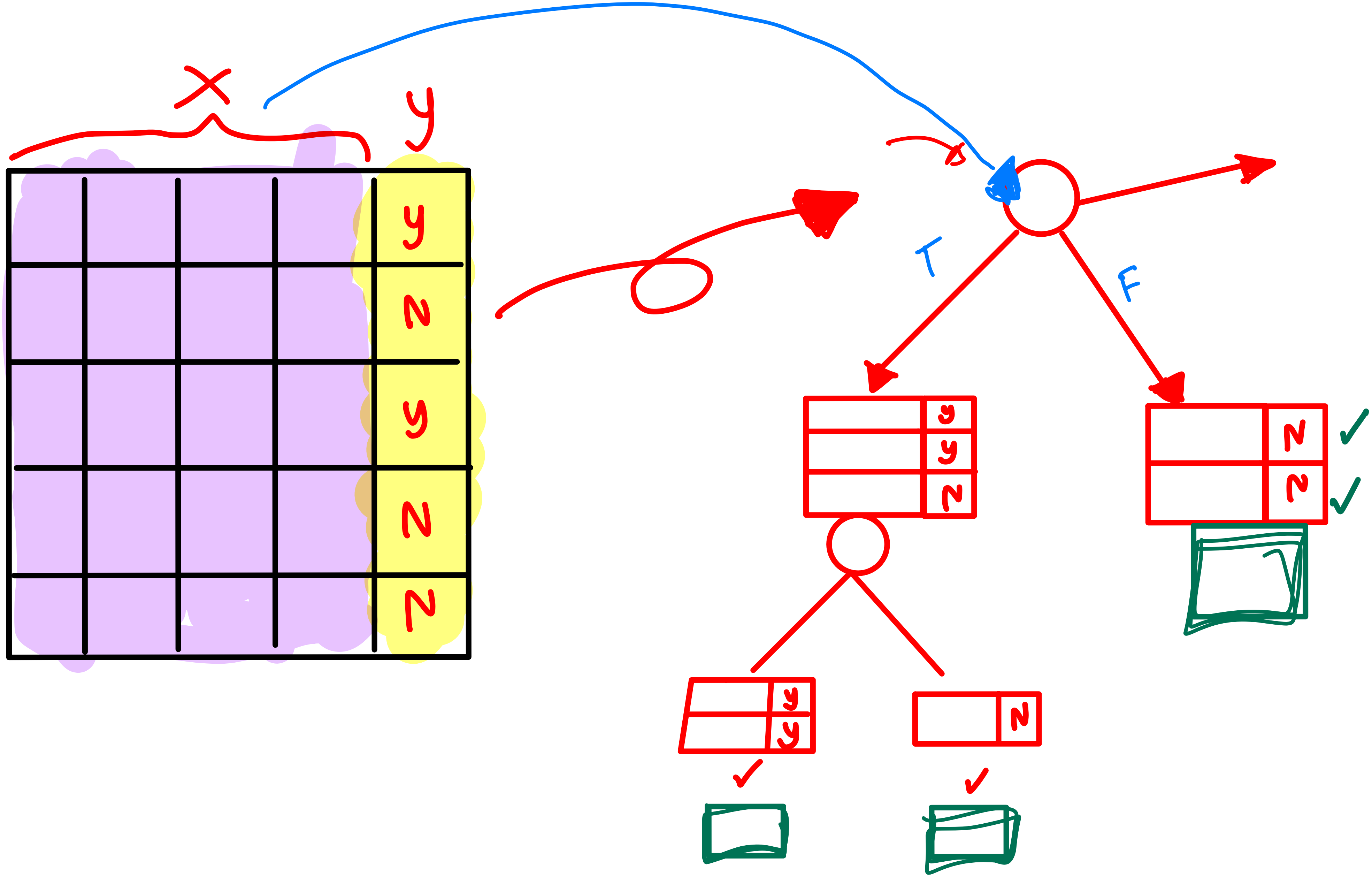
$y = f(x)$

X (feature) Y (label)

Training data set: Who buys computer?

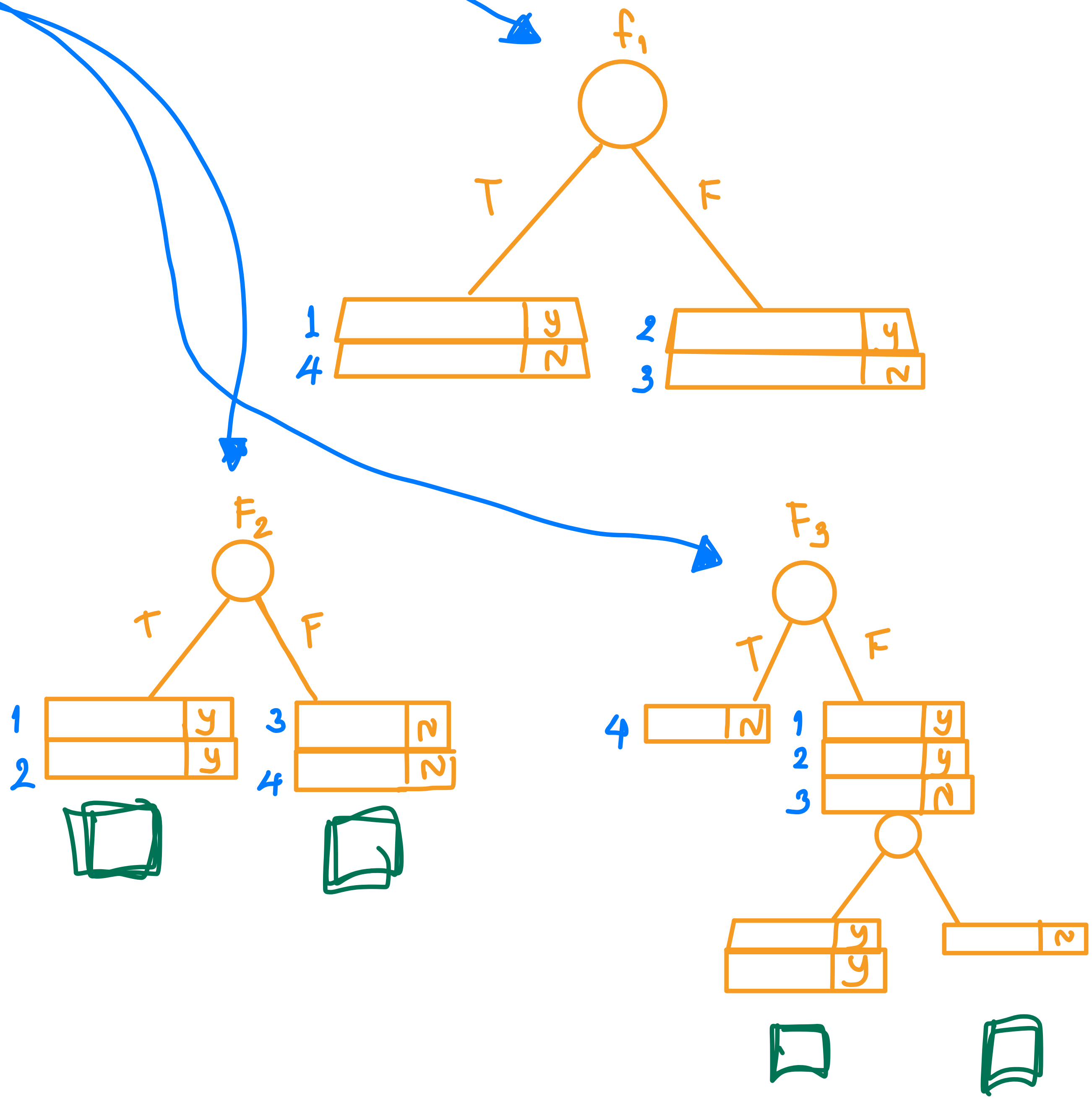
age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Note: The data set is adapted from "Playing Tennis" example of R. Quinlan



X
y
y
N
y
N
N
T
F
y
y
N
N
N
y
y
N

	f_1	f_2	f_3	y
1	T	T	F	y
2	F	T	F	y
3	F	F	F	N
4	T	F	T	N



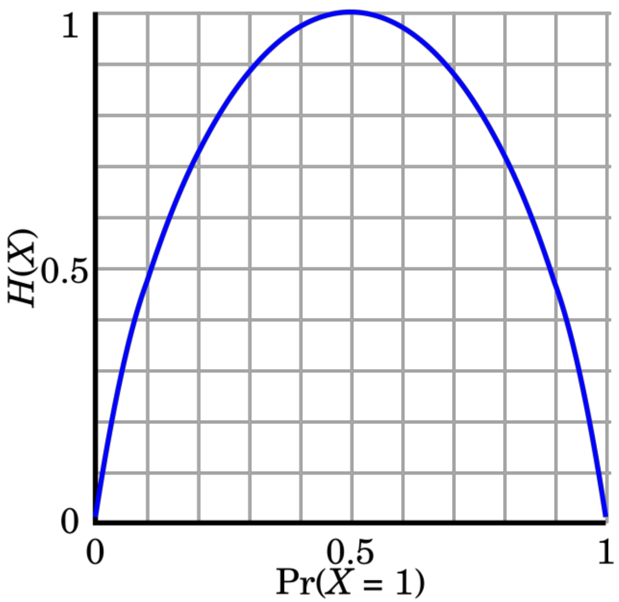
From Entropy to Info Gain: A Brief Review of Entropy

- Entropy (Information Theory)
  - A measure of uncertainty associated with a random number
  - Calculation: For a discrete random variable Y taking m distinct values $\{y_1, y_2, \ldots, y_m\}$

$$H(Y) = -\sum_{i=1}^{m} p_i \log(p_i) \quad where\ p_i = P(Y = y_i)$$

  - Interpretation
    - Higher entropy → higher uncertainty
    - Lower entropy → lower uncertainty
- Conditional entropy

$$H(Y|X) = \sum_{x} p(x) H(Y|X = x)$$



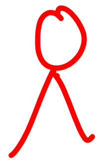
m = 2

Information Gain: An Attribute Selection Measure

- Select the attribute with the highest information gain (used in typical decision tree induction algorithm: ID3/C4.5)
- Let p_i be the probability that an arbitrary tuple in D belongs to class C_i, estimated by $|C_{i,D}|/|D|$

คำนวณ 1 ครั้ง

1



- Expected information (entropy) needed to classify a tuple in D:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

คำนวณตามจำนวน feature

- Information needed (after using A to split D into v partitions) to classify D:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

- Information gained by branching on attribute A

ค่า ภามที่ดีที่สุด

$$Gain(A) = Info(D) - Info_A(D)$$

$I(A,B,C) = -\frac{A}{S}\log\frac{A}{S} - \frac{B}{S}\log\frac{B}{S} - \frac{C}{S}\log\frac{C}{S}$

Example: Attribute Selection with Information Gain

- Class P: buys_computer = "yes"
- Class N: buys_computer = "no"

$$Info(D) = I(\overset{y}{9},\overset{N}{5}) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.940$$

age	p_i	n_i	$I(p_i, n_i)$
<=30	2	3	0.971
31…40	4	0	0
>40	3	2	0.971

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

<=30 31…40

$$Info_{age}(D) = \frac{5}{14}I(\overset{y}{2},\overset{N}{3}) + \frac{4}{14}I(\overset{y}{4},\overset{N}{0}) + \frac{5}{14}I(\overset{y}{3},\overset{N}{2}) = 0.694$$

>40

$\frac{5}{14}I(2,3)$ means "age <=30" has 5 out of 14 samples, with 2 yes'es and 3 no's.

Hence

$$Gain(age) = Info(D) - Info_{age}(D) = 0.246$$

Similarly, we can get

$$Gain(income) = 0.029$$
$$Gain(student) = 0.151$$
$$Gain(credit\_rating) = 0.048$$

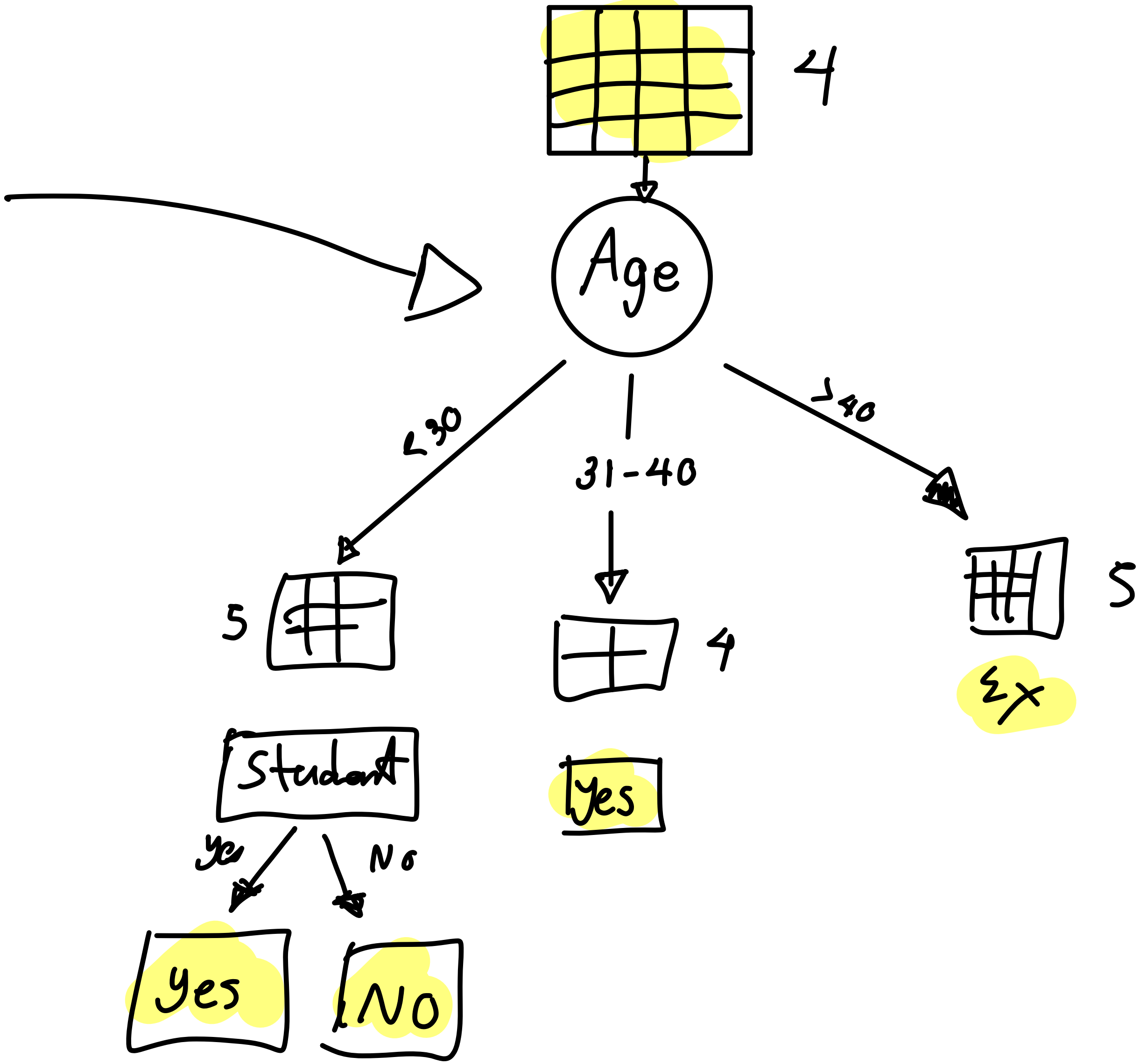
Homework

G(age) = 0.246

G(income) = 0.029

G(student) = 0.151

G(credit) = 0.048



>30

$$\text{Info}(D) = I(\overset{Y}{2},\overset{N}{3}) = \overset{Y}{-\frac{2}{5}\log_2\frac{2}{5}} \overset{N}{-\frac{3}{3}\log_2\frac{3}{5}}$$

$$\text{Info}_{\text{income}}(D) = \frac{2}{5}\overset{h}{I(0,2)} + \frac{2}{5}\overset{m}{I(1,1)} + \frac{1}{5}\overset{l}{I(1,0)}$$

(h = high, m = medium, low)

$$\text{Info}_{\text{student}}(D) = \frac{2}{5}\overset{y}{I(2,0)} + \frac{3}{5}\overset{N}{I(0,3)}$$

$$\text{Info}_{\text{credit\_rating}}(D) = \frac{3}{5}\overset{f}{I(1,2)} + \frac{2}{5}\overset{e}{I(1,1)}$$

(f = fair, e = excellent)

$31 \ldots 40$

$$\text{Info}(D) = I(\overset{Y}{4},\overset{N}{0}) = \overset{Y}{-\frac{4}{4}\log_2\frac{4}{4}} \overset{N}{-\frac{0}{4}\log_2\frac{0}{4}}$$

$$\text{Info}_{\text{income}}(D) = \frac{2}{4}\overset{h}{I(1,1)} + \frac{1}{4}\overset{m}{I(0,1)} + \frac{1}{4}\overset{l}{I(1,0)}$$

(h = high, m = medium, low)

$$\text{Info}_{\text{student}}(D) = \frac{2}{4}\overset{y}{I(2,0)} + \frac{2}{4}\overset{N}{I(2,0)}$$

$$\text{Info}_{\text{credit\_rating}}(D) = \frac{2}{4}\overset{f}{I(2,0)} + \frac{2}{4}\overset{e}{I(2,0)}$$

(f = fair, e = excellent)

> 40

$$\text{Info}(D) = I(\overset{Y}{3}, \overset{N}{2}) = \overset{Y}{-\frac{3}{5}\log_2 \frac{3}{5}} \overset{N}{-\frac{2}{5}\log_2 \frac{2}{5}}$$

(h = high, m = medium, low)

★ $\text{Info}_{income}(D) = \frac{3}{5} I\overset{m}{(2,1)} + \frac{2}{5} I\overset{l}{(1,1)}$

★ $\text{Info}_{student}(D) = \frac{3}{5} I\overset{y}{(2,1)} + \frac{2}{5} I\overset{N}{(1,1)}$

★ $\text{Info}_{credit\_rating}(D) = \frac{3}{5} I\overset{f}{(3,0)} + \frac{2}{5} I\overset{e}{(0,2)}$

(f = fair, e = excellent)