CS 412 Intro. to Data Mining

Chapter 8. Classification: Basic Concepts

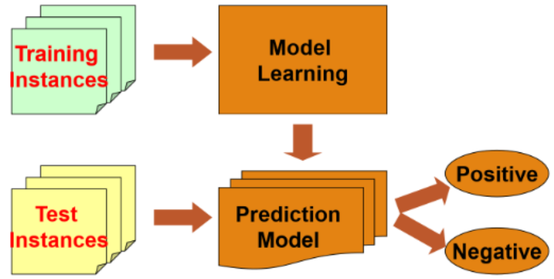
Jiawei Han, Computer Science, Univ. Illinois at Urbana-Champaign, 2017

Supervised vs. Unsupervised Learning (1)

- Supervised learning (classification) → มีคอลัมน์ เป็นจุดมุ่งหมาย → มีผู้สอน
  - Supervision: The training data such as observations or measurements are accompanied by **labels** indicating the classes which they belong to
  - New data is classified based on the models built from the training set

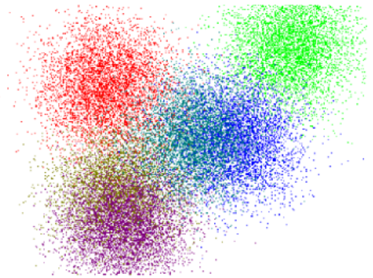
Training Data with class label:

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



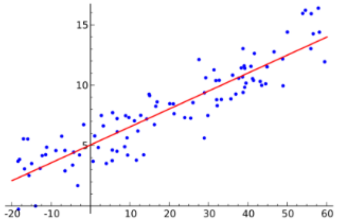
Supervised vs. Unsupervised Learning (2)

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



Prediction Problems: Classification vs. Numeric Prediction

- Classification → เป็น binary ทำนายว่าอยู่กลุ่มไหน
  - 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 → เป็นตัวเลข "Regression"
  - 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

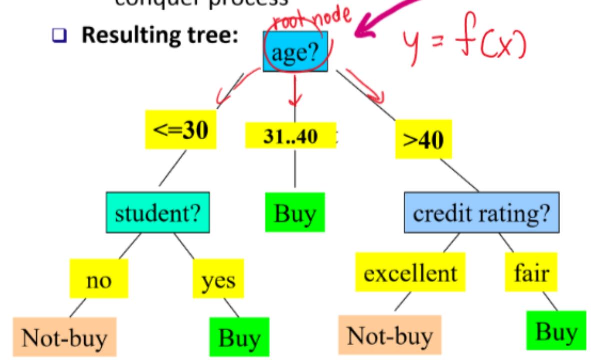
- **Model construction** เอา Data ที่มี featured + คำตอบ มา train
  - Each sample is assumed to belong to a predefined class (shown by the **class label**)
  - The set of samples used for model construction is **training set**
  - Model: Represented as decision trees, rules, mathematical formulas, or other forms
- **Model Validation and Testing**: → เอาไปทำนายเลย
  - **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

Decision Tree Induction: An Example

x : feature

y : label

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



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

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 $|C_{i,D}|/|D|$
- Expected information (entropy) needed to classify a tuple in D:

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

→ คำนวณครั้งเดียว

- 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

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

Example: Attribute Selection with Information Gain

มี data ทั้งหมด 14 ตัว

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

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

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

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

41

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

Hence ← แต่ละ คำถาม

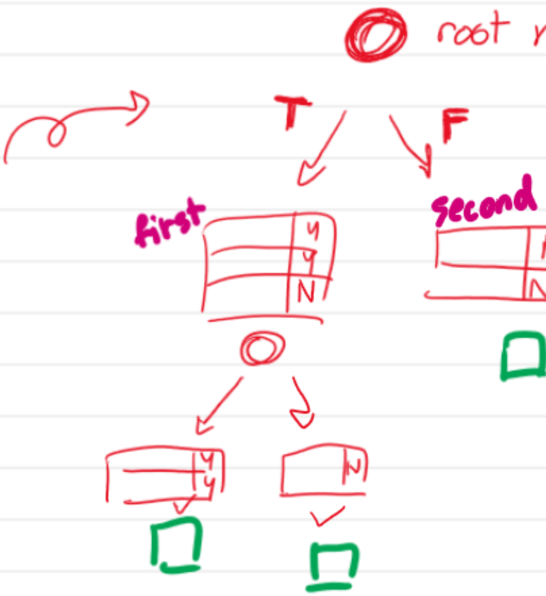
$$Gain(age) = Info(D) - Info_{age}(D) = 0.246$$

Similarly, we can get

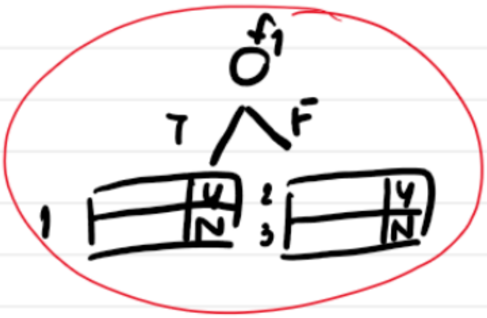
$$Gain(income) = 0.029$$
$$Gain(student) = 0.151$$
$$Gain(credit\_rating) = 0.048$$

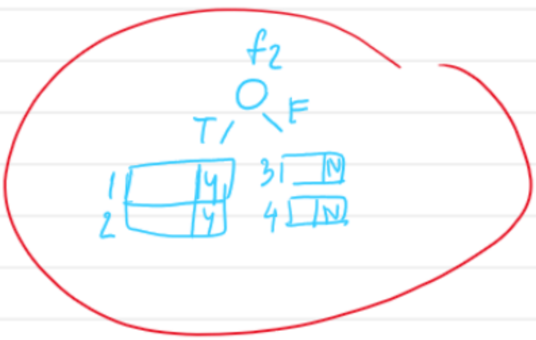
หลักการสร้าง decision tree

root node = ตัวที่แบ่ง data ได้ดีที่สุด



	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





G(age) = 0.246

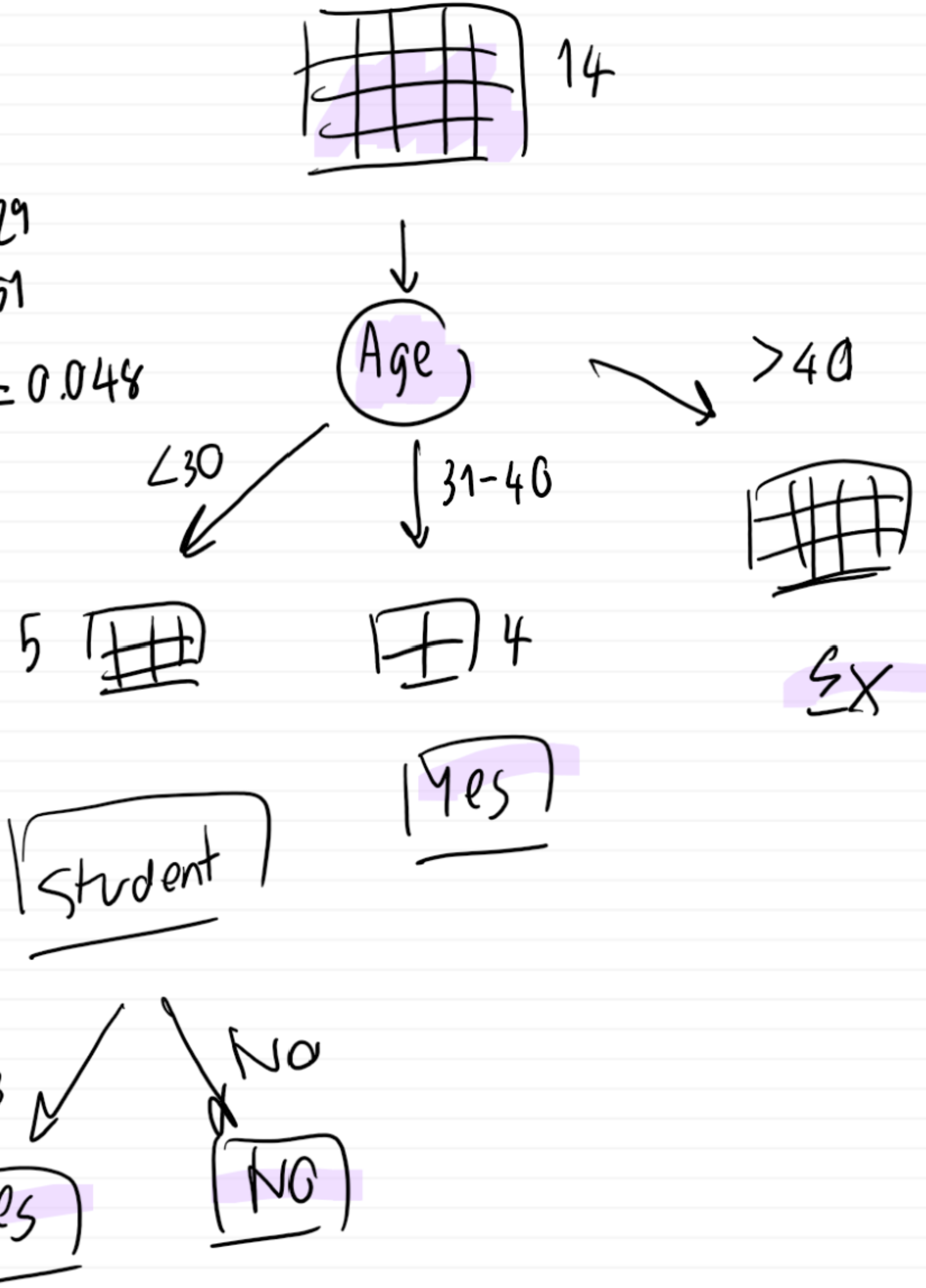
G(income) = 0.029

G(student) = 0.151

G(credit_rating) = 0.048

การบ้าน Hw.

info D
ของแต่ละชิ้น.



<30

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

$$\text{Info}_{\text{income}}(D) = \overset{\text{high}}{\frac{2}{5} I(0,2)} + \overset{\text{medium}}{\frac{2}{5} I(1,1)} + \overset{\text{low}}{\frac{1}{5} I(1,0)}$$

$$\text{Info}_{\text{student}}(D) = \overset{\text{Yes}}{\frac{2}{5} I(2,0)} + \overset{\text{No}}{\frac{3}{5} I(0,3)}$$

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

31 - 40

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

$$\text{Info}_{income}(D) = \frac{2}{4}\overset{high}{I(1,1)} + \frac{1}{4}\overset{medium}{I(0,1)} + \frac{1}{4}\overset{low}{I(1,0)}$$

$$\text{Info}_{student}(D) = \frac{2}{4}\overset{Yes}{I(2,0)} + \frac{2}{4}\overset{NO}{I(2,0)}$$

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

>40

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

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

$$\text{Info}_{\text{student}}(D) = \frac{3}{5}\,\overset{\text{Yes}}{I(2,1)} + \frac{2}{5}\,\overset{\text{NO}}{I(1,1)}$$

$$\text{Info}_{\text{credit}}(D) = \frac{3}{5}\,\overset{\text{fair}}{I(3,0)} + \frac{2}{5}\,\overset{\text{excellent}}{I(0,2)}$$