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

Supervised vs. Unsupervised Learning (1)

สร้าง model หรือใช้ฟีเจอร์มาทำนายคำตอบ

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

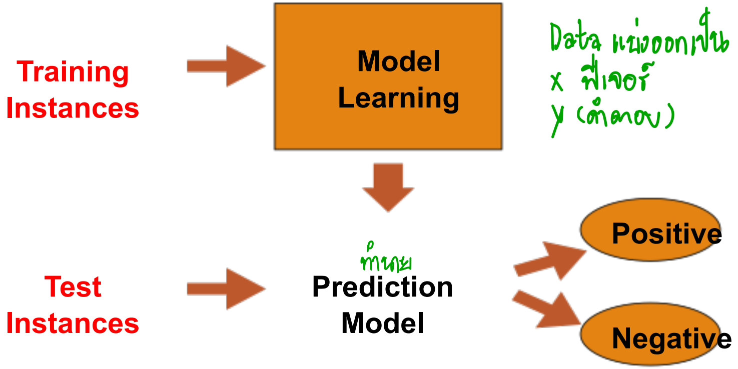
    มีคนสอน, มีจุดมุ่งหมาย / มีการสังเกต / มีการวัดผล

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

    สร้างที่เก็บข้อมูลมาทำนาย column ต่อไป

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)

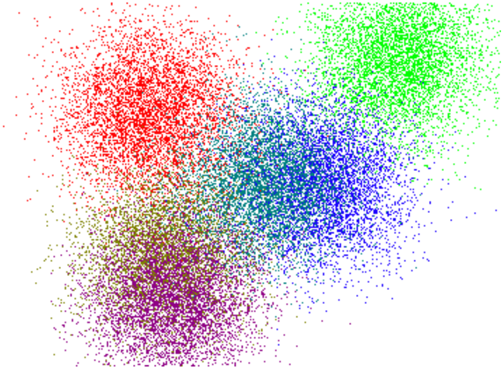
มีแค่ x

ไม่มีผู้สอน, แบ่งกลุ่มเฉยๆ

- Unsupervised learning (clustering)

ฟีเจอร์ → เหมือนกัน → ใกล้กัน
ไม่เหมือนกัน → ไกลกัน

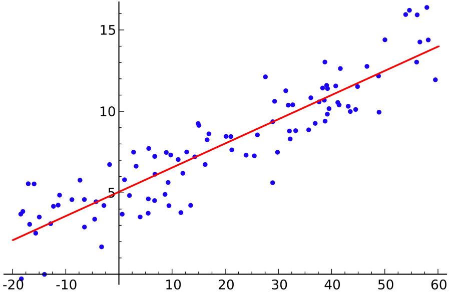
  - The class labels of training data are unknown
  - Given a set of observations or measurements, establish the possible existence of classes or clusters in the data



Prediction Problems: Classification vs. Numeric Prediction

ทำนาย Class แบ่งกลุ่ม

- 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

สร้าง model → วัดผล, ทำนายคำตอบ $\xrightarrow{\text{ผ่าน}}$ นำไปใช้งาน

- **Model construction**
  - 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

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

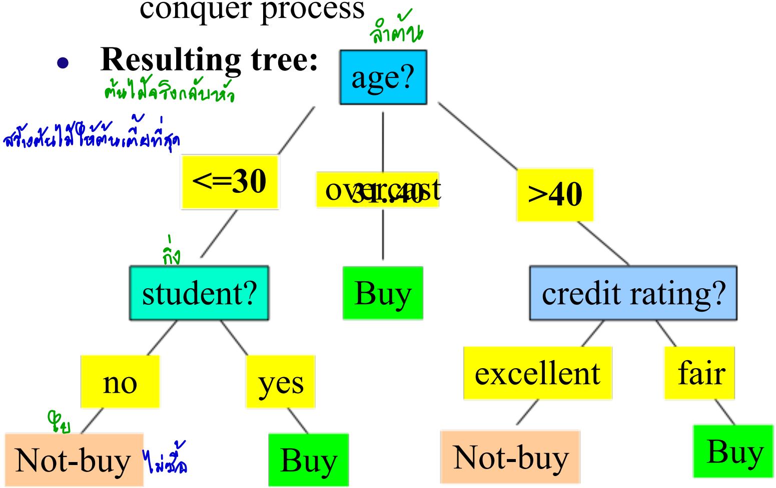
ต้นไม้ตัดสินใจ

Decision Tree Induction: An Example

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

ต้นไม้จริงกลับหัว

•สร้างต้นไม้ให้ต้นเตี้ยที่สุด



Training data set: Who buys computer?

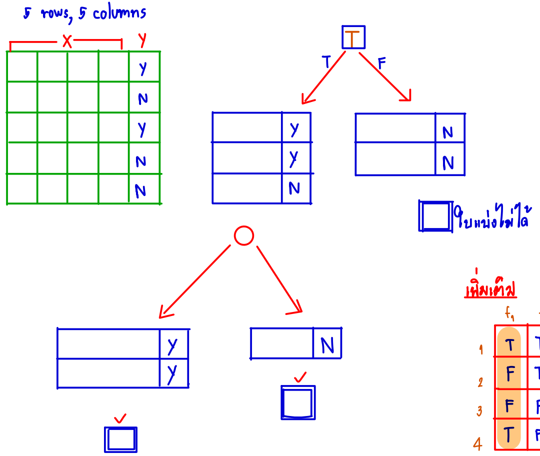
x y

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

วิธีคิด

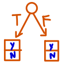
- สร้างจาก root ก่อนค่อยมาใบ
- มี Data 2 ส่วน → x, y
- เอา Data 5 ตัวมาแบ่ง root node (ตัวที่แบ่งได้ดีที่สุด)



เพิ่มเติม

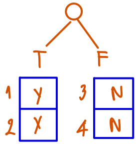
	f_1	f_2	f_3	y
1	T	T	F	y
2	F	T	F	y
3	F	F	F	N
4	T	F	T	N

ดู f_1



T → y ไปใส่ช่อง T
F → y ไปใส่ช่อง F

f_2

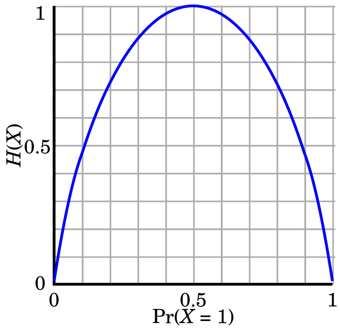


• ไม่ต้องแบ่งต่อ

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 {y1, y2, …,

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



m = 2

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

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

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 *pi* be the probability that an arbitrary tuple in D belongs to class Ci, estimated by |C*i*, D|/|D|
- Expected information (entropy) needed to classify a tuple in D:

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

แต่ละชิ้นของต้นไม้
Info(0) จะเหมือนกันหมด

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

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

คำนวณจากจำนวนฟีเจอร์

- Information gained by branching on attribute A

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

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

Example: Attribute Selection with Information Gain

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

$$Info(D) = I(\overset{\text{yes, No}}{9,5}) = \boxed{-\frac{9}{14}\log_2(\frac{9}{14})} \boxed{-\frac{5}{14}\log_2(\frac{5}{14})} = 0.940$$

แบ่ง yes กับ no ว่ามีกี่ตัว

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

$$Info_{age}(D) = \boxed{\frac{5}{14}I(2,3)}^{<=30} + \boxed{\frac{4}{14}I(4,0)}^{31-40} + \boxed{\frac{5}{14}I(3,2)}_{>41} = 0.694$$

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

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

จะเปลี่ยนไปเรื่อยๆ

Similarly, we can get

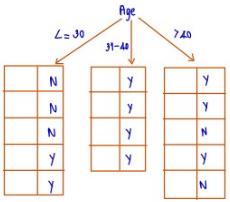
$$Gain(income) = 0.029$$

$$Gain(student) = 0.151$$

$$Gain(credit\_rating) = 0.048$$

Homework

Gain(age)



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

Gain(age)

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

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

$= 0.694$

$Gain(age) = 0.940 - 0.694$

$= 0.246$

Gain(income)

Income	p_i	n_i	$I(p_i, n_i)$
high	2	2	1
medium	4	2	0.918
low	3	1	0.811

$Info_{income}(D) = \frac{4}{14}I(2,2) + \frac{6}{14}I(4,2) + \frac{4}{14}I(3,1)$

$= 0.911$

$Gain(income) = 0.940 - 0.911$

$= 0.029$

Gain(Student)

Student	p_i	n_i	$I(p_i, n_i)$
yes	6	1	0.592
no	3	4	0.985

$Info_{student}(D) = \frac{7}{14}I(6,1) + \frac{7}{14}I(3,4)$

$= 0.789$

$Gain(student) = 0.940 - 0.789$

$= 0.151$

Gain(credit_rating)

credit_rating	p_i	n_i	$I(p_i, n_i)$
fair	6	2	0.8111
excellent	3	3	1

$Info_{credit\_rating}(D) = \frac{7}{14}I(6,2) + \frac{7}{14}I(3,3)$

$= 0.892$

$Gain(credit\_rating) = 0.940 - 0.892$

$= 0.048$

Gain(age) = 0.246
Gain(income) = 0.024
Gain(Student) = 0.151
Gain(credit_rating = 0.048

จะเลือกค่ามาก นั่นคือ Gain(age) เป็น root node

หา age <= 30

$Info(D) = I(2,3) = 0.971$

$Info_{income}(D)$ ของ age(<=30)

Income	p_i	n_i	$I(p_i, n_i)$
high	0	2	0
medium	1	1	1
low	1	0	0

$Info_{income}(D)$ ของ age(<=30) $= \frac{2}{5}I(0,2) + \frac{2}{5}I(1,1) + \frac{1}{5}I(1,0)$

$= 0.4$

$Gain(income)$ ของ age(<=30) $= 0.971 - 0.4$

$= 0.571$

$Info_{student}(D)$

$Info_{student}(D)$ ของ age(<=30) $= \frac{2}{5}I(2,0) + \frac{3}{5}I(2,0) + \frac{3}{5}I(0,3)$

จะได้ yes → yes (buy_computer) → ซื้อ

No → No (buy_computer) → ไม่ซื้อ

age (>40)

$Info(D) = I(3,2) = 0.971$

$Info_{income}(D)$ ของ age(>40)

Income	p_i	n_i	$I(p_i,n_i)$
low	1	1	1
medium	2	1	0.918

$Info_{income}(D)$ ของ age(>40) $= \frac{2}{5}I(1,1) + \frac{-}{5}I(2,1)$

$= 0.951$

Grain(income) ของ age(>40) $= 0.971 - 0.951$

$= 0.020$

$Info_{student}$ ของ age(>40)

Student	p_i	n_i	$I(p_i,n_i)$
yes	2	1	0.916
no	1	1	1

$Info_{student}(D)$ ของ age(>40) $= \frac{3}{5}I(2,1) + \frac{2}{5}I(1,1) = 0.951$

Grain(student) ของ age(>40) $= 0.971 - 0.951 = 0.020$

$Info_{credit\_rating}(D)$ ของ age(>40)

$Info_{credit\_rating}(D)$ ของ age(>40) $= \frac{3}{5}I(3,0) + \frac{2}{5}I(0,2)$

จะได้ fair $\longrightarrow$ yes ซื้อ

excellent $\longrightarrow$ No ไม่ซื้อ

จะได้