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)

- Supervised learning (classification)
  - 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

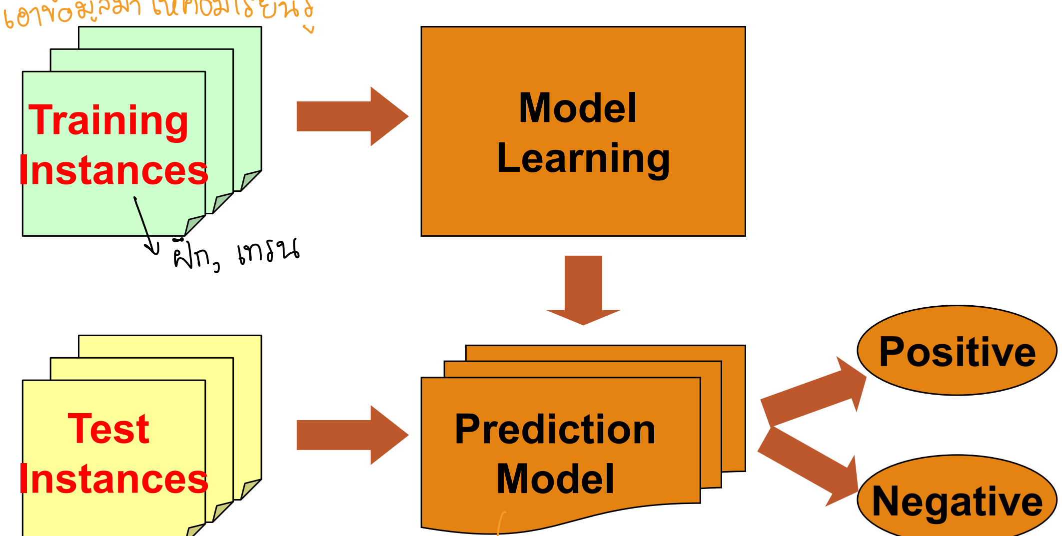
Training Data with class label:

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

มีคำตอบ, ทำนายว่าซื้อหรือไม่ซื้อ

binary

เอาข้อมูลมาให้คอมเรียนรู้



ฝึก, เทรน

ใช่ได้หรือไม่

เอา model มาใช้งาน

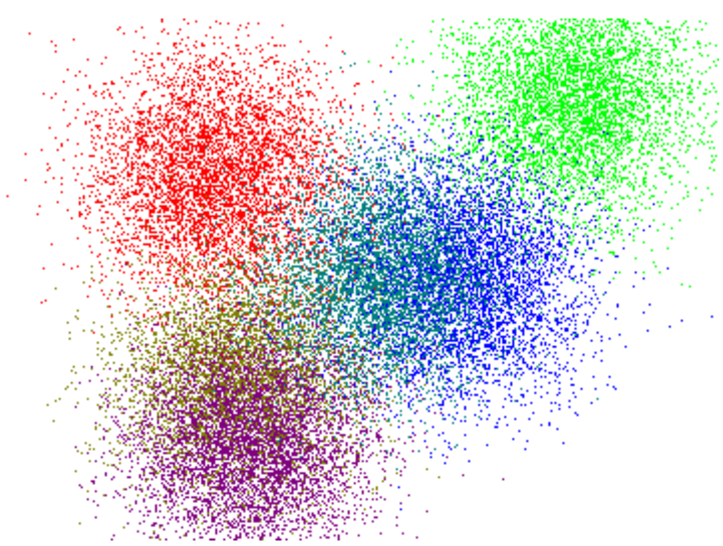
Supervised vs. Unsupervised Learning (2)

ไม่มีผู้สอน , ไม่มีจุดมุ่งหมายในการเรียน

- Unsupervised learning (clustering)

ไม่มีคำตอบ

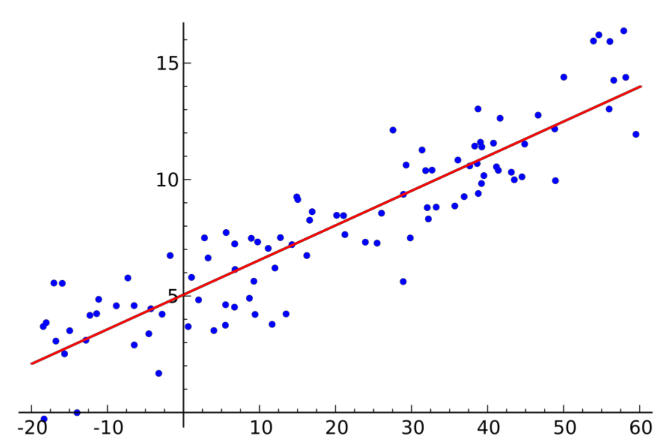
  - The class labels of training data are unknown
  - Given a set of observations or measurements, establish the possible existence of classes or clusters in the data



ใส่ข้อมูลแล้วให้มันแบ่งกลุ่ม

Prediction Problems: Classification vs. Numeric Prediction

- Classification → ทำนาย class , กลุ่มคน
  - Predict categorical class labels (discrete or nominal) ทำนายว่าอยู่กลุ่มไหน
  - 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**
  - 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

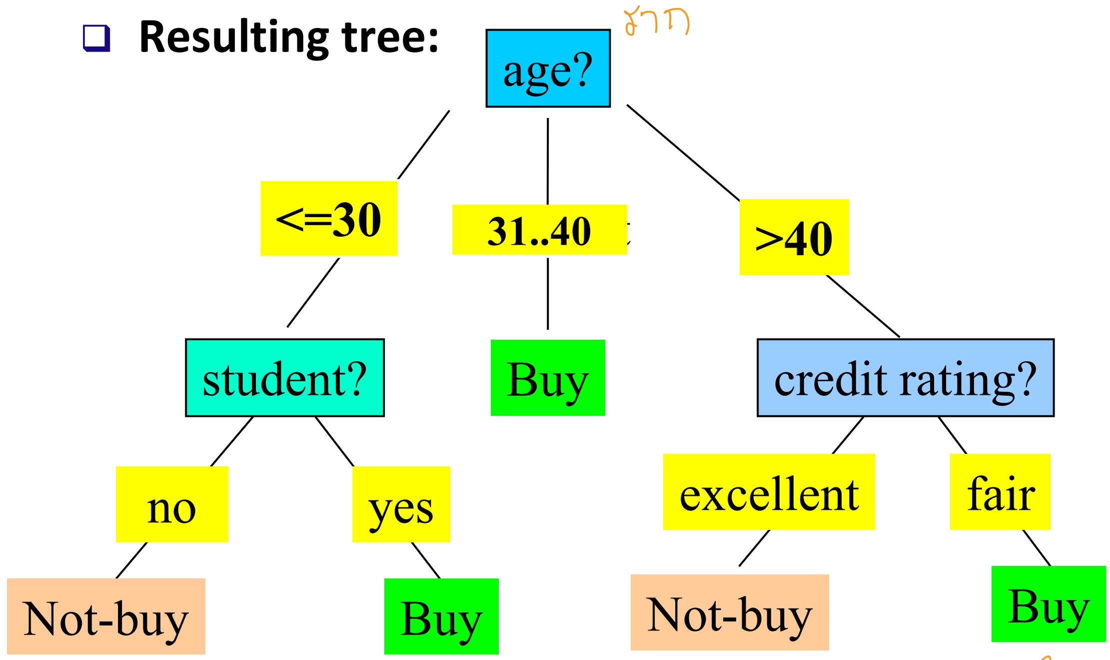
- Classification: Basic Concepts
- Decision Tree Induction
- Bayes Classification Methods
- Linear Classifier
- Model Evaluation and Selection
- Techniques to Improve Classification Accuracy: Ensemble Methods
- Additional Concepts on Classification
- Summary

ทำนายว่าใครจะซื้อไม่ซื้อ

ต้นไม้ตัดสินใจ

Decision Tree Induction: An Example

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



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

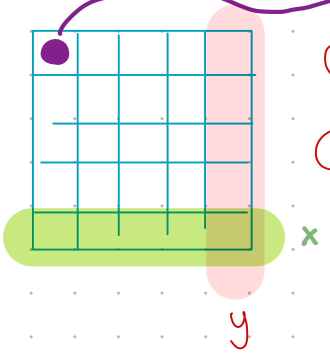
Note: The data set is adapted from "Playing Tennis" example of R. Quinlan

Data Structure

$y = f(x)$

x ไปใส่ใน function แล้วเกิด y

หลักการสร้าง Disition tree



1. ต้องเริ่มสร้างจากราก → กิ่ง → ... → ใบ

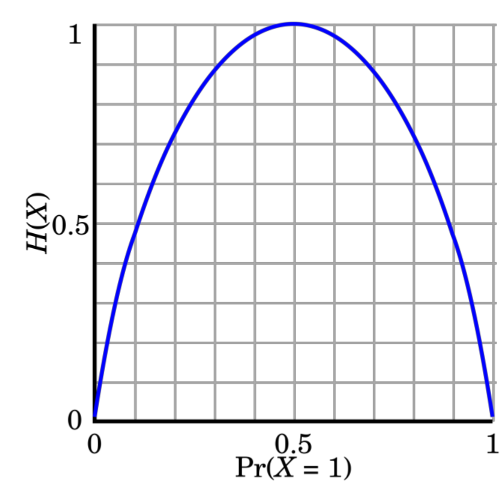
2. อย่าทำต้นไม้ให้สูง ควรทำให้เตี้ยที่สุด

ไม่ได้ใช้หลักการเงื่อนไขประพจน์ label จะไหลไปตามการตอบ

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 $\{y_1, y_2, \ldots, y_m\}$

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



m = 2

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

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

เกณฑ์ของข้อมูล

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:

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

1 ครั้ง

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

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

คำนวณตามจำนวน Freature

- Information gained by branching on attribute A

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

Example: Attribute Selection with Information Gain

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

$$Info(D) = \overset{yes,\ no}{I(9,5)} = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.940$$

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

14 row = S

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

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

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

Hence

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

Similarly, we can get

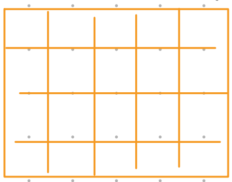
$$Gain(income) = 0.029$$
$$Gain(student) = 0.151$$
$$Gain(credit\_rating) = 0.048$$

เลือกตัวที่มากสุด เป็นราก

$$\underbrace{I(A,B,C)}_{\text{กรณีที่ไม่ใช่ label แค่ yes, no}} = -\frac{A}{S}\log\frac{A}{S} - \dots - \frac{C}{S}\log\frac{C}{S}$$

Homework

ตารางข้อมูล



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

Similarly, we can get

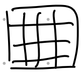
$$Gain(income) = 0.029$$
$$Gain(student) = 0.151$$
$$Gain(credit\_rating) = 0.048$$

1. พิจารณาเลือกราก จากการคำนวณ

เลือกตัวที่มากสุด เป็นราก

2.

ดังนั้นจะได้



age → ซึ่งมี 3 กลุ่ม

- < 30 มี 5 → student → yes / No
- 31 - 40 มี 4 → yes
- > 40 มี 5

ต่อมาคือ student

คำนวณ

<30 ; $Info(D) = I(\overset{yes}{2}, \overset{No}{3}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$

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

$Info_{student}(D) = \frac{2}{5}\overset{yes}{I(2,0)} + \frac{3}{5}\overset{No}{I(0,3)}$

$Info_{credit}(D) = \frac{3}{5}\overset{fair}{I(1,2)} + \frac{2}{5}\overset{exellent}{(1,1)}$

31 → 40 ; $Info(D) = I(\overset{yes}{4}, \overset{No}{0}) = -\frac{4}{4}\cancel{\log_2\frac{4}{4}} - \cancel{\frac{0}{4}\log_2\frac{0}{4}}$

$Info_{income}(D) = \frac{2}{4}\overset{high}{I(1,1)} + \frac{2}{5}\overset{medium}{I(0,1)} + -\overset{low}{I(1,0)}$

$Info_{student}(D) = \frac{2}{4}\overset{yes}{I(2,0)} + \frac{2}{4}\overset{No}{I(2,0)}$

$Info_{credit}(D) = \frac{2}{4}\overset{fair}{I(2,0)} + \frac{2}{4}\overset{exellent}{(2,0)}$

>40 ; $Info(D) = I(\overset{yes}{3}, \overset{No}{2}) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$

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

$Info_{student}(D) = \frac{3}{5}\overset{yes}{I(2,1)} + \frac{2}{5}\overset{No}{I(1,1)}$

$Info_{credit}(D) = \frac{3}{5}\overset{fair}{I(3,0)} + \frac{2}{5}\overset{exellent}{(0,2)}$