CS 412 Intro. to Data Mining

การทำนาย

Chapter 8. Classification: Basic Concepts

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

การสร้างโมเดลแบบใช้มีผู้สอน

Supervised vs. Unsupervised Learning (1)

- Supervised learning (classification) → เก็บข้อมูลว่ามีอะไรบ้าง เช่น รายได้
  - 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

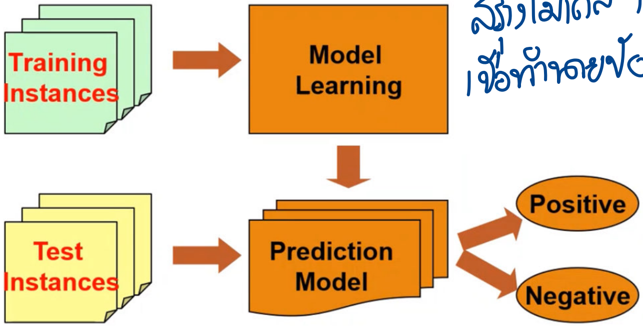
ไบนารี

ทำนายว่าจะซื้อหรือไม่ซื้อ

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

1



สร้างโมเดล , ML
เพื่อทำนายข้อมูล

Supervised vs. Unsupervised Learning (2)

ต่างกันคือ Data แบ่งเป็น 2 ส่วน

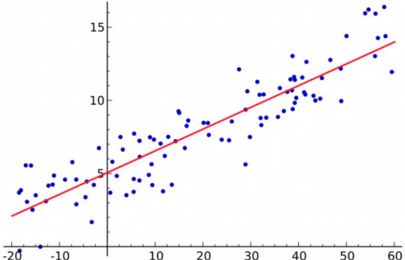
- Unsupervised learning (clustering)

มีแค่ X ตัวเดียว

  - 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

- Classification → ทำนาย class → ซื้อไม่ซื้อ ค่าทำนายออกมาเป็นตัวเลขจะเรียกว่า Regretion
  - 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 ที่มีคำตอบมาสร้าง Model แล้วไปทำนายคำตอบว่าถูกมากน้อยแค่ไหน
  - 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

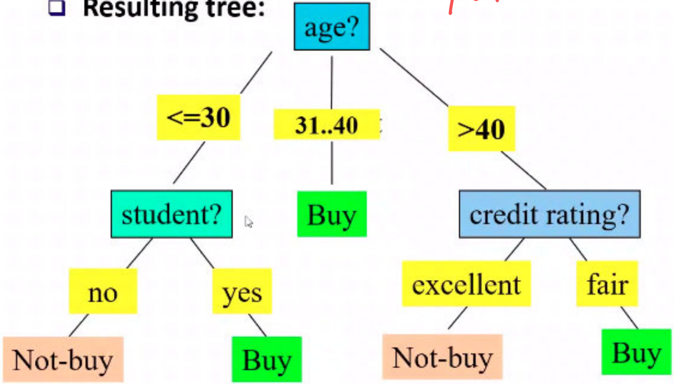
- Classification: Basic Concepts
- Decision Tree Induction ต้นไม้ตัดสินใจ
- Bayes Classification Methods
- Linear Classifier
- Model Evaluation and Selection
- Techniques to Improve Classification Accuracy: Ensemble Methods
- Additional Concepts on Classification
- Summary

Decision Tree Induction: An Example

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

y = f(x)

- **Resulting tree:**



X (feature) y (labal)

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

Example: Attribute Selection with Information Gain

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

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

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

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

$\frac{5}{14}I(2,3)$ 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

$$Gain(income) = 0.029$$

$$Gain(student) = 0.151$$

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

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:

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

- 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:

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

หา Info(D)

$$\mathrm{Info}(D) = I(\overset{Y}{9},\overset{N}{5}) = \overset{Y}{\boxed{-\frac{9}{14}\log_{(2)}\left(\frac{9}{14}\right)}} - \overset{N}{\boxed{\frac{5}{14}\log_{(2)}\left(\frac{5}{14}\right)}}$$

$$= 0.94$$

หา $\mathrm{Info}_{age}(D)$

$$\mathrm{Info}_{age}(D) = \overset{<=30}{\boxed{\frac{5}{14} I(\overset{Y}{2},\overset{N}{3})}} + \overset{31-40}{\boxed{\frac{4}{14} I(\overset{Y}{4},\overset{N}{0})}}$$

$$I(\overset{Y}{2},\overset{N}{3}) = -\frac{2}{5}\log_{(2)}\left(\frac{2}{5}\right) - \frac{3}{5}\log_{(2)}\left(\frac{3}{5}\right) = 0.971$$

$$I(\overset{Y}{4},\overset{N}{0}) = -\frac{4}{4}\log_{(2)}\left(\frac{4}{4}\right) - \frac{0}{4}\log_{(2)}\left(\frac{0}{4}\right) = 0$$

$$I(\overset{Y}{3},\overset{N}{2}) = -\frac{3}{5}\log_{(2)}\left(\frac{3}{5}\right) - \frac{2}{5}\log_{(2)}\left(\frac{2}{5}\right) = 0.971$$

แทนค่า $\mathrm{Info}_{age}(D) = \frac{5}{14}(0.971) + \frac{4}{14}(0) + \frac{5}{14}(0.971) = 0.694$

หาค่า Gain(age)

$$\mathrm{Gain}(age) = 0.94 - 0.694 = 0.246$$

หา $\mathrm{Info}_{income}(D)$

$$\mathrm{Info}_{income}(D) = \overset{high}{\boxed{\frac{4}{14} I(\overset{Y}{2},\overset{N}{2})}} + \overset{medium}{\boxed{\frac{6}{14} I(\overset{Y}{4},\overset{N}{2})}} + \overset{low}{\boxed{\frac{4}{14} I(\overset{Y}{3},\overset{N}{1})}}$$

$$I(2,2) = -\frac{2}{4}\log_{(2)}\left(\frac{2}{4}\right) - \frac{2}{4}\log_{(2)}\left(\frac{2}{4}\right) = 1$$

$$I(4,2) = -\frac{4}{6}\log_{(2)}\left(\frac{4}{6}\right) - \frac{2}{6}\log_{(2)}\left(\frac{2}{6}\right) = 0.918$$

$$I(3,1) = -\frac{3}{4}\log_{(2)}\left(\frac{3}{4}\right) - \frac{1}{4}\log_{(2)}\left(\frac{1}{4}\right) = 0.811$$

แทนค่า $\mathrm{Info}_{income}(D) = \frac{4}{14}(1) + \frac{6}{14}(0.918) + \frac{4}{14}(0.811) = 0.911$

หาค่า Gain(income)

$$\mathrm{Gain}(income) = 0.94 - 0.911 = 0.029$$

หา $Info_{student}(D)$

$$Info_{student}(D) = \overset{\text{yes}}{\boxed{\frac{7}{14} I(\overset{Y}{6},\overset{N}{1})}} + \overset{\text{NO}}{\boxed{\frac{7}{14} I(\overset{Y}{3},\overset{N}{4})}}$$

$$I(\overset{Y}{6},\overset{N}{1}) = -\frac{6}{7}\log_{(2)}\left(\frac{6}{7}\right) - \frac{1}{7}\log_{(2)}\left(\frac{1}{7}\right) = 0.592$$

$$I(\overset{Y}{3},\overset{N}{4}) = -\frac{3}{7}\log_{(2)}\left(\frac{3}{7}\right) - \frac{4}{7}\log_{(2)}\left(\frac{4}{7}\right) = 0.985$$

แทนค่า $Info_{student}(D) = \frac{7}{14}(0.592) + \frac{7}{14}(0.985) = 0.789$

หา Gain(student)

$$Gain(student) = 0.94 - 0.789 = 0.151$$

หา $Info_{credit\text{-}rating}(D)$

$$Info_{credit\text{-}rating}(D) = \overset{\text{fair}}{\boxed{\frac{8}{14} I(\overset{Y}{6},\overset{N}{2})}} + \overset{\text{excellent}}{\boxed{\frac{6}{14} I(\overset{Y}{3},\overset{N}{3})}}$$

$$I(\overset{Y}{6},\overset{N}{2}) = -\frac{6}{8}\log_{(2)}\left(\frac{6}{8}\right) - \frac{2}{8}\log_{(2)}\left(\frac{2}{8}\right) = 0.811$$

$$I(\overset{Y}{3},\overset{N}{3}) = -\frac{3}{6}\log_{(2)}\left(\frac{3}{6}\right) - \frac{3}{6}\log_{(2)}\left(\frac{3}{6}\right) = 1$$

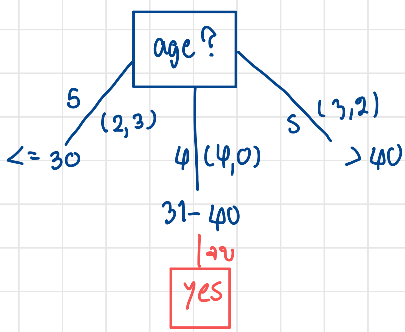
แทนค่า $Info_{credit\text{-}rating}(D) = \frac{8}{14}(0.811) + \frac{6}{14}(1) = 0.892$

หา Gain(credit-rating)

$$Gain(credit\text{-}rating) = 0.94 - 0.892 = 0.048$$

จาก Gain

Gain(age)	=	0.246
Gain(income)	=	0.029
Gain(student)	=	0.151
Gain(credit-rating)	=	0.048



เลือก Gain ที่มีค่ามากที่สุด มาเป็นราก นั่นคือ Gain(age)

age (<= 30)

หา Info (D) ของ age (<=30)

$$Info(D) = I(\overset{Y}{2}, \overset{N}{3}) = 0.971$$

หา Info (D) ของ age (<=30)

$$Info(D) \text{ ของ } age(<=30) = \frac{2}{5} I(\overset{Y}{0}, \overset{N}{2}) + \frac{2}{5} I(\overset{Y}{1}, \overset{N}{1}) + \frac{1}{5} I(\overset{Y}{1}, \overset{N}{0})$$

$$I(\overset{Y}{0}, \overset{N}{2}) = -\frac{0}{2}\log_{(2)}\left(\frac{0}{2}\right) - \frac{2}{2}\log_{(2)}\left(\frac{2}{2}\right) = 0$$

$$I(\overset{Y}{1}, \overset{N}{1}) = -\frac{1}{2}\log_{(2)}\left(\frac{1}{2}\right) - \frac{1}{2}\log_{(2)}\left(\frac{1}{2}\right) = 1$$

$$I(\overset{Y}{1}, \overset{N}{0}) = -\frac{1}{1}\log_{(2)}\left(\frac{1}{1}\right) - \frac{0}{1}\log_{(2)}\left(\frac{0}{2}\right) = 0$$

แทนค่า $Info(D)$ ของ $age(<=30) = \frac{2}{5}(0) + \frac{2}{5}(1) + \frac{1}{5}(0) = 0.4$

หา Gain (income) ของ age (<=30)

$$Gain(income) \text{ ของ } age(<=30) = 0.971 - 0.4 = 0.571$$

หา $Info_{student}(D)$ ของ age (<=30)

$$Info_{student}(D) \text{ ของ } age(<=30) = \overset{yes}{\boxed{\frac{2}{5} I(\overset{Y}{2}, \overset{N}{0})}} + \overset{No}{\boxed{\frac{3}{5} I(\overset{Y}{0}, \overset{N}{3})}}$$

สิ้นสุด yes → yes (buy - Computer) , No → no (buy - Computer)

เลือกแบ่งด้วย student เพราะสามารถแบ่งข้อมูลได้แบบสมบูรณ์

age (>40)

หา Info(D) ของ age (>40)

$$\text{Info}(D) \text{ ของ age } (>40) = I(\overset{Y}{3},\overset{N}{2}) = 0.971$$

หา $\text{Info}_{\text{income}}(D)$ ของ age (>40)

$$\text{Info}_{\text{income}}(D) \text{ ของ age } (>40) = \overset{\text{medium}}{\boxed{\frac{3}{5} I(\overset{Y}{2},\overset{N}{1})}} + \overset{\text{low}}{\boxed{\frac{2}{5} I(\overset{Y}{1},\overset{N}{1})}}$$

$$I(\overset{Y}{2},\overset{N}{1}) = -\frac{2}{3}\log_{(2)}\left(\frac{2}{3}\right) - \frac{1}{3}\log_{(2)}\left(\frac{1}{3}\right) = 0.918$$

$$I(\overset{Y}{1},\overset{N}{1}) = 1$$

แทนค่า $\text{Info}(D)$ ของ age $(>40) = \frac{3}{5}(0.918) + \frac{2}{5}(1) = 0.951$

หา Gain (income) ของ age (>40)

$$\text{Gain (income) ของ age } (>40) = 0.971 - 0.951 = 0.02$$

หา $\text{Info}_{\text{student}}(D)$ ของ age (>40)

$$\text{Info}_{\text{student}}(D) \text{ ของ age } (>40) = \overset{\text{Yes}}{\boxed{\frac{3}{5} I(\overset{Y}{2},\overset{N}{1})}} + \overset{\text{No}}{\boxed{\frac{2}{5} I(\overset{Y}{1},\overset{N}{1})}}$$

$$I(\overset{Y}{2},\overset{N}{1}) = -\frac{2}{3}\log_{(2)}\left(\frac{2}{3}\right) - \frac{1}{3}\log_{(2)}\left(\frac{1}{3}\right) = 0.918$$

$$I(\overset{Y}{1},\overset{N}{1}) = 1$$

แทนค่า $\text{Info}_{\text{student}}(D)$ ของ age $(>40) = \frac{3}{5}(0.918) + \frac{2}{5}(1) = 0.951$

หา Gain (student) ของ age (>40)

$$\text{Gain (student) ของ age } (>40) = 0.971 - 0.951 = 0.02$$

หา $\text{Info}_{credit\_rating}(D)$ ของ age (>40)

$$\text{Info}_{credit\text{-}rating}(D) \text{ ของ age } (>40) = \overset{\text{fair}}{\boxed{\frac{3}{5} I(\overset{y}{3},\overset{N}{0})}} + \overset{\text{excellent}}{\boxed{\frac{2}{5} I(\overset{y}{0},\overset{N}{2})}}$$

สิ้นสุด fair → yes (buy-Computer) , excellent → no (buy-Computer)

เลือกแบ่งด้วย credit_rating เพราะสามารถแบ่งข้อมูลได้สมบูรณ์

สรุป