

# Supervised Learning



Dr. Wedad Hussein  
wedad.hussein@cis.asu.edu.eg

Dr. Mahmoud Mounir  
mahmoud.mounir@cis.asu.edu.eg



# **Data Mining:**

---

## **Concepts and Techniques**

**(3<sup>rd</sup> ed.)**


### **— Chapter 8 —**

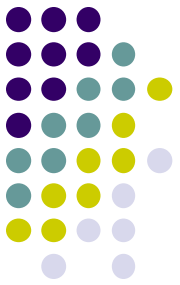
Jiawei Han, Micheline Kamber, and Jian Pei  
University of Illinois at Urbana-Champaign &  
Simon Fraser University

©2011 Han, Kamber & Pei. All rights reserved.

# Chapter 8. Classification: Basic Concepts

---

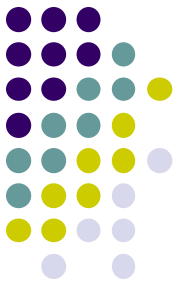
- Classification: Basic Concepts 
- Decision Tree Induction
- Bayes Classification Methods
- Rule-Based Classification
- Model Evaluation and Selection
- Techniques to Improve Classification Accuracy:  
Ensemble Methods
- Summary



# INTRODUCTION-

- Given the following dataset of objects

	Attribute 1	Attribute 2	Attribute 3
Objects	X	Y	Z
OB-1	1	4	1
OB-2	1	2	2
OB-3	1	4	2
OB-4	2	1	2
OB-5	1	1	1
OB-6	2	4	2
OB-7	1	1	2
OB-8	2	1	1



# INTRODUCTION-

- Given the following dataset of objects

	Attribute 1	Attribute 2	Attribute 3	
Objects	X	Y	Z	Class
OB-1	1	4	1	A
OB-2	1	2	2	B
OB-3	1	4	2	B
OB-4	2	1	2	A
OB-5	1	1	1	A
OB-6	2	4	2	B
OB-7	1	1	2	A
OB-8	2	1	1	A

# Supervised vs. Unsupervised Learning

---

- Supervised learning (classification)
  - Supervision: The training data (observations, measurements, etc.) are accompanied by **labels** indicating the class of the observations
  - New data is classified based on the training set
- Unsupervised learning (clustering)
  - The class labels of training data is unknown
  - Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

# Prediction Problems: Classification vs. Numeric Prediction

---

- **Classification**
  - predicts categorical class labels (discrete or nominal)
  - classifies data (constructs a model) based on the training set and the values (**class labels**) in a classifying attribute and uses it in classifying new data
- **Numeric Prediction**
  - models continuous-valued functions, i.e., predicts unknown or missing values
- Typical applications
  - 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

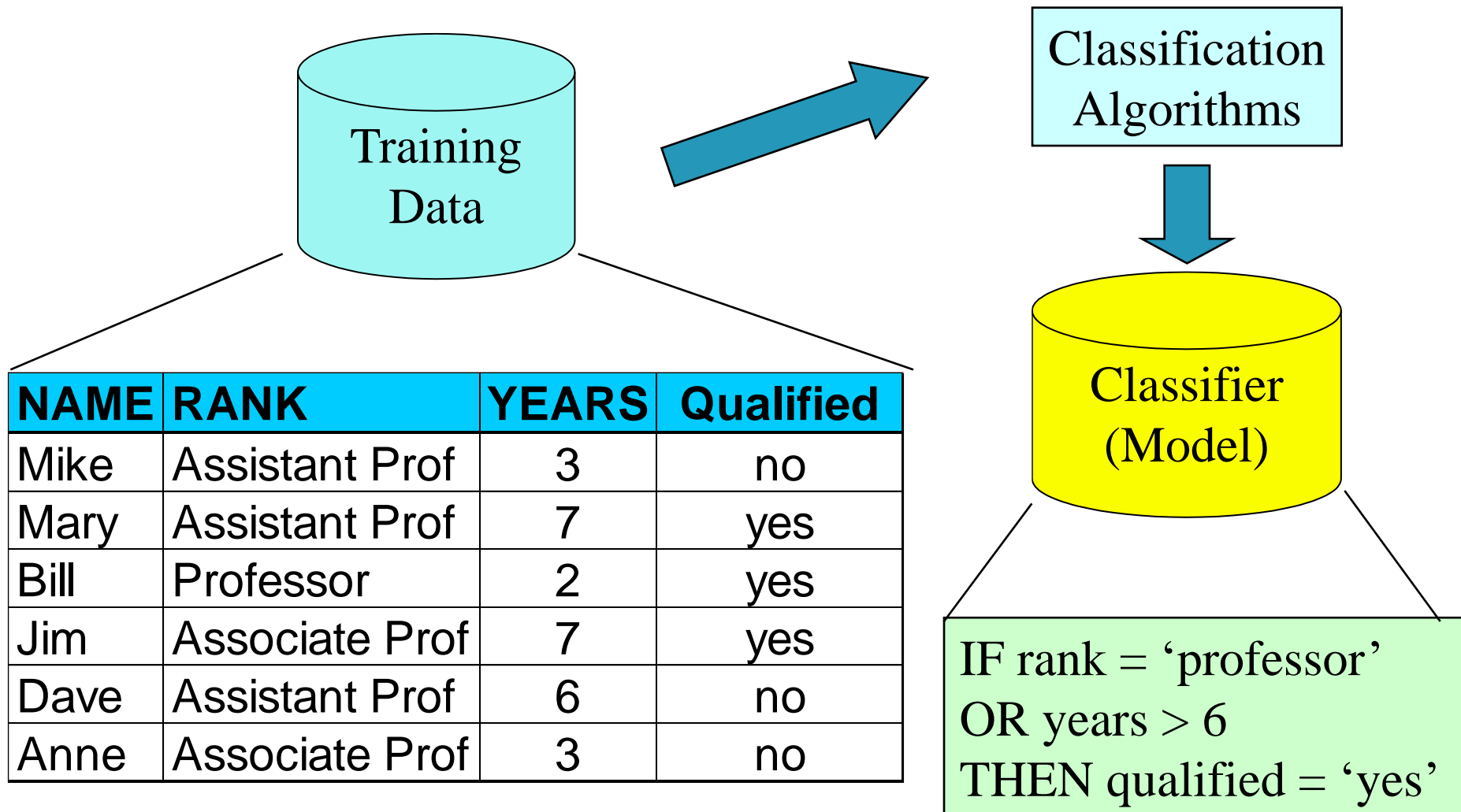
# Classification—A Two-Step Process

---

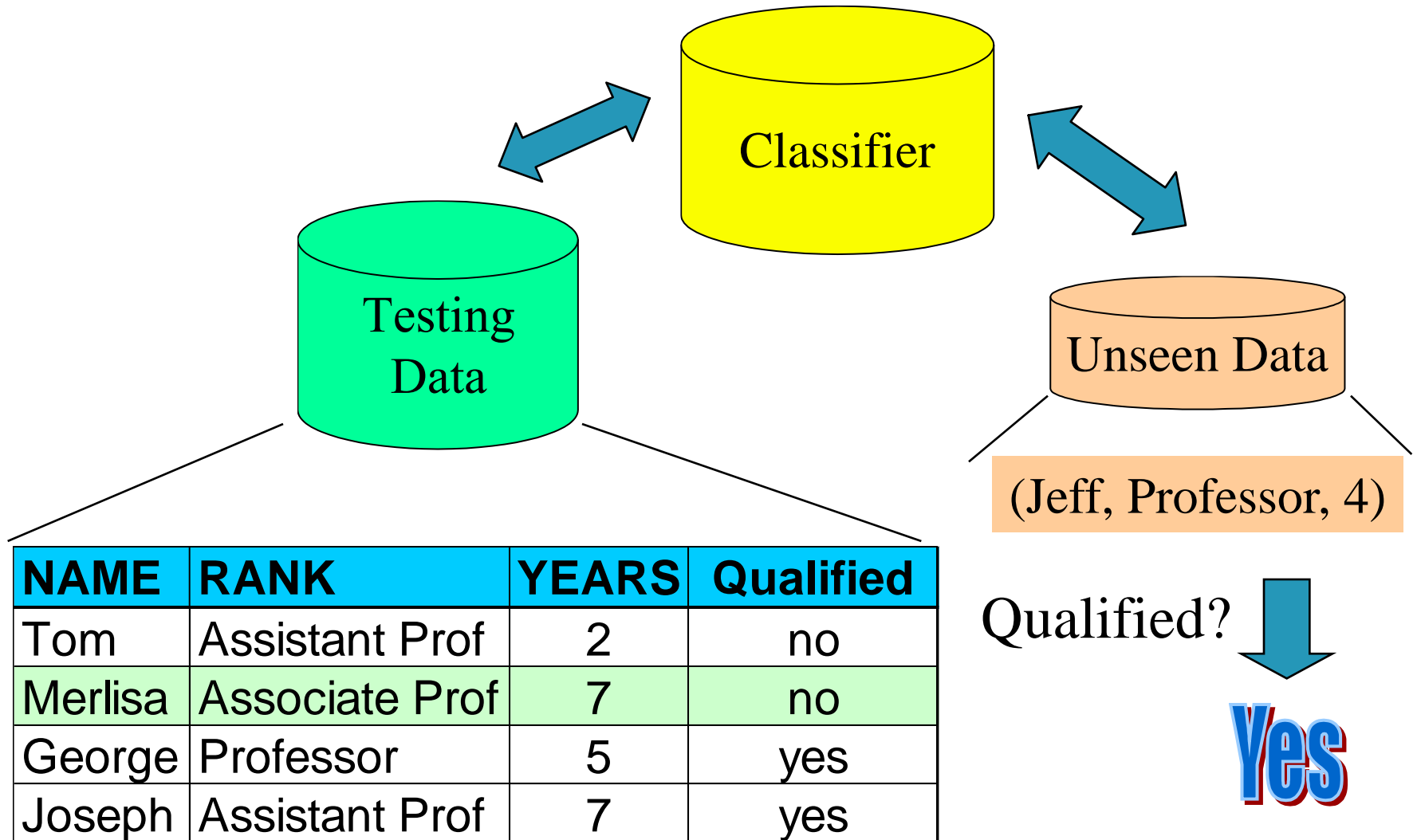
- **Model construction**: describing a set of predetermined classes
  - Each tuple/sample is assumed to belong to a predefined class, as determined by the **class label attribute**
  - The set of tuples used for model construction is **training set**
  - The model is represented as classification rules, decision trees, or mathematical formulae
- **Model usage**: for classifying future or unknown objects
  - **Estimate accuracy** of the model
    - The known label of test sample is compared with the classified result from the model
    - **Accuracy** rate is the percentage of test set samples that are correctly classified by the model
    - **Test set** is independent of training set (otherwise overfitting)
  - If the accuracy is acceptable, use the model to **classify new data**
- Note: If *the test set* is used to select models, it is called **validation (test) set**



# Process (1): Model Construction




# Process (2): Using the Model in Prediction



# Chapter 8. Classification: Basic Concepts

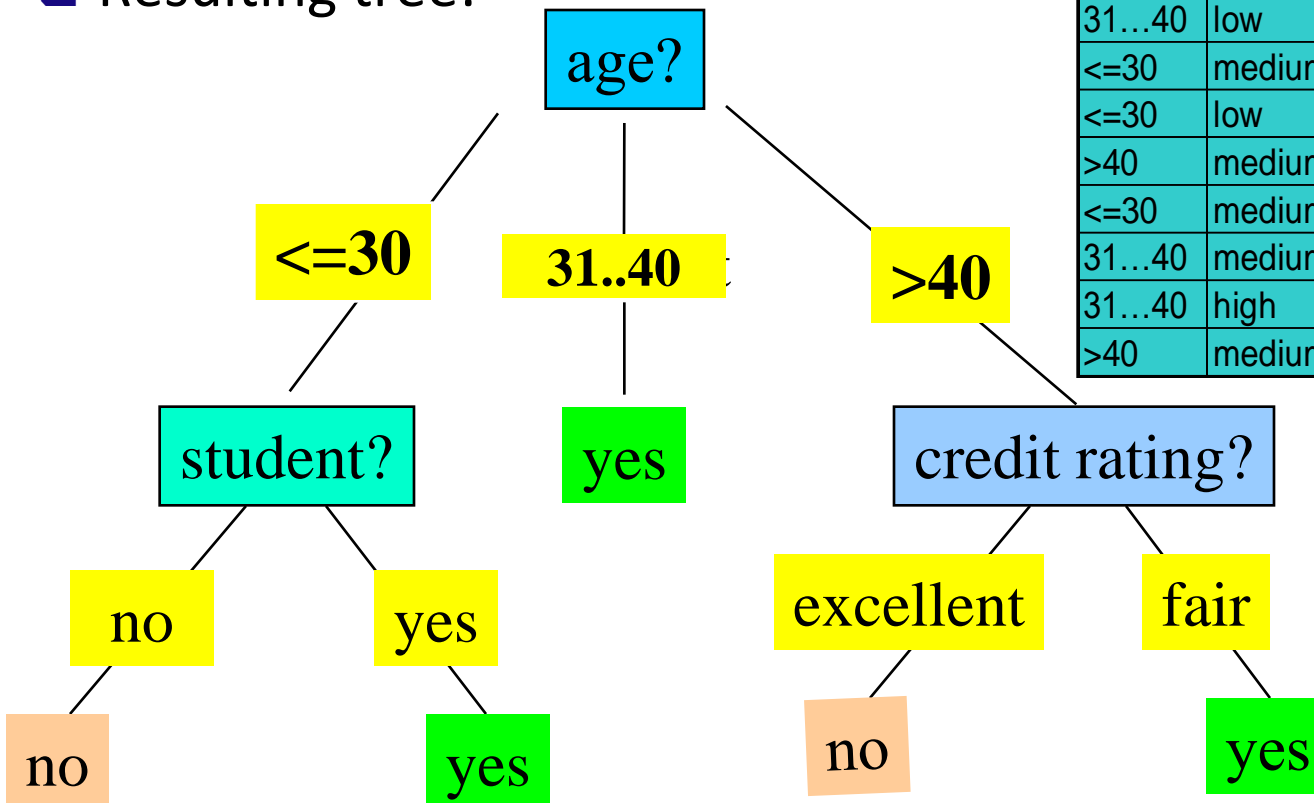
---

- Classification: Basic Concepts
- Decision Tree Induction 
- Bayes Classification Methods
- Rule-Based Classification
- Model Evaluation and Selection
- Techniques to Improve Classification Accuracy:  
Ensemble Methods
- Summary

# Decision Tree Induction: An Example

- ❑ Training data set: Buys\_computer
- ❑ The data set follows an example of Quinlan's ID3 (Playing Tennis)
- ❑ Resulting tree:

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



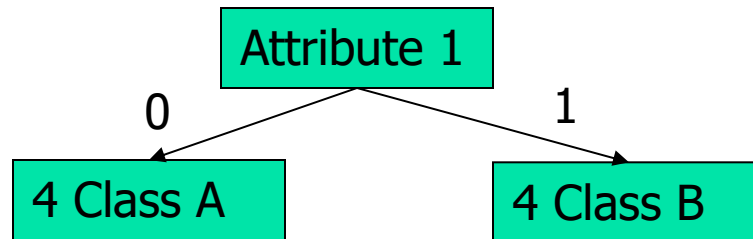
# Algorithm for Decision Tree Induction

---

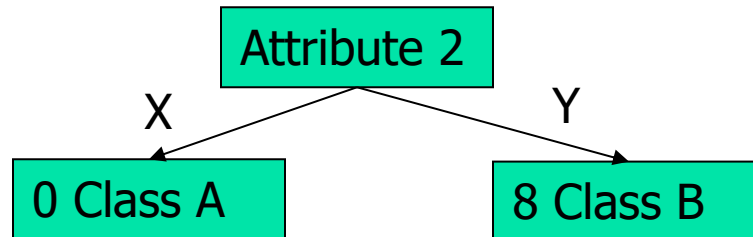
- Basic algorithm (a greedy algorithm)
  - Tree is constructed in a **top-down recursive divide-and-conquer manner**
  - At start, all the training examples are at the root
  - Attributes are categorical (if continuous-valued, they are discretized in advance)
  - Examples are partitioned recursively based on selected attributes
  - Test attributes are selected on the basis of a heuristic or statistical measure (e.g., **information gain**)
- Conditions for stopping partitioning
  - All samples for a given node belong to the same class
  - There are no remaining attributes for further partitioning – **majority voting** is employed for classifying the leaf
  - There are no samples left

# Brief Review of Entropy

- $Entropy = H(D) = -\sum_i P_i \log_2(P_i)$



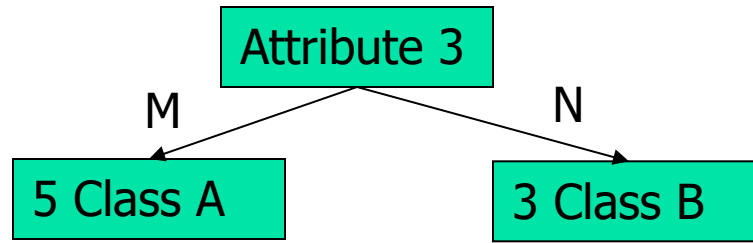
- $Entropy = H(\text{Attribute 1}) = -\frac{4}{8} \log_2\left(\frac{4}{8}\right) - \frac{4}{8} \log_2\left(\frac{4}{8}\right) = \underline{\underline{1}}$



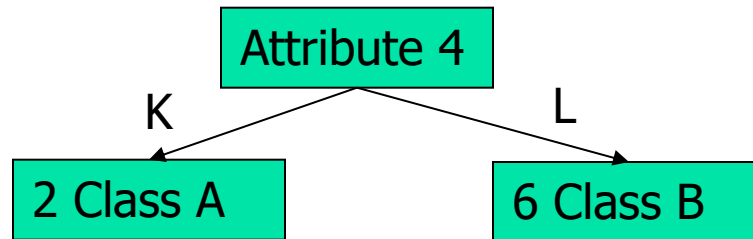
- $Entropy = H(\text{Attribute 2}) = -\frac{0}{8} \log_2\left(\frac{0}{8}\right) - \frac{8}{8} \log_2\left(\frac{8}{8}\right) = \underline{\underline{0}}$

# Brief Review of Entropy

---



- $Entropy = H(\text{Attribute 3}) = -\frac{5}{8} \log_2\left(\frac{5}{8}\right) - \frac{3}{8} \log_2\left(\frac{3}{8}\right)$   
 $= 0.424 + 0.531 = \underline{\underline{0.955}}$



- $Entropy = H(\text{Attribute 4}) = -\frac{2}{8} \log_2\left(\frac{2}{8}\right) - \frac{6}{8} \log_2\left(\frac{6}{8}\right) = \underline{\underline{0.811}}$

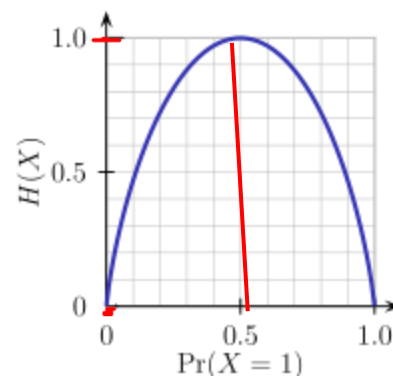
# Brief Review of Entropy

## ■ Entropy (Information Theory)

- A measure of uncertainty associated with a random variable
- Calculation: For a discrete random variable  $Y$  taking  $m$  distinct values  $\{y_1, \dots, y_m\}$ ,
  - $H(Y) = -\sum_{i=1}^m p_i \log(p_i)$ , where  $p_i = P(Y = y_i)$
- Interpretation:
  - Higher entropy => higher uncertainty
  - Lower entropy => lower uncertainty

## ■ Conditional Entropy

- $H(Y|X) = \sum_x p(x)H(Y|X = x)$





# Attribute Selection Measure: Information Gain (ID3/C4.5)

---

- Select the attribute with the highest information gain
- 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  $A$

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

# Decision Trees Using ID3 Algorithm

- a. What is the entropy of buys computer?

Buys Computer	
No	Yes
5	9

$$\text{Entropy}_{(\text{Buy Computer})} = H_{(\text{Buy Computer})}$$

$$= -\frac{5}{14} \log_2 \left( \frac{5}{14} \right) - \frac{9}{14} \log_2 \left( \frac{9}{14} \right) = 0.531 + 0.41$$

$$= \underline{0.941}$$

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

# Decision Trees Using ID3 Algorithm

- b. Which attribute should you choose as the root of a decision tree?

Age (14) <u>0.941</u>		
<b>&lt;= 30</b>	<b>31..40</b>	<b>&gt;40</b>
5	4	5
[3 No , 2 Yes]	[0 No , 4 Yes]	[2 No , 3 Yes]

$$\begin{aligned}
 H_{(Age)} &= \frac{5}{14} H[3,2] + \frac{4}{14} H[0,4] + \frac{5}{14} H[2,3] = \\
 &= \frac{5}{14} \left[ -\frac{3}{5} \log_2 \left( \frac{3}{5} \right) - \frac{2}{5} \log_2 \left( \frac{2}{5} \right) \right] + \frac{4}{14} \left[ -\frac{0}{4} \log_2 \left( \frac{0}{4} \right) - \frac{4}{4} \log_2 \left( \frac{4}{4} \right) \right] \\
 &+ \frac{5}{14} \left[ -\frac{2}{5} \log_2 \left( \frac{2}{5} \right) - \frac{3}{5} \log_2 \left( \frac{3}{5} \right) \right] = \\
 &= 0.347 + 0 + 0.347 = \underline{0.694}
 \end{aligned}$$

$$IG_{(Buys\ Computer/Age)} = H_{(Buys\ Computer)} - H_{(Age)} = 0.941 - 0.694 = \underline{0.247}$$

# Decision Trees Using ID3 Algorithm

- b. Which attribute should you choose as the root of a decision tree?

Income (14) <u>0.941</u>		
High	Medium	Low
4	6	4
[2 No , 2 Yes]	[2 No , 4 Yes]	[1 No , 3 Yes]

$$\begin{aligned}H_{(Income)} &= \frac{4}{14} H[2,2] + \frac{6}{14} H[2,4] + \frac{4}{14} H[1,3] = \\&= \frac{4}{14} \left[ -\frac{2}{4} \log_2 \left( \frac{2}{4} \right) - \frac{2}{4} \log_2 \left( \frac{2}{4} \right) \right] + \frac{6}{14} \left[ -\frac{2}{6} \log_2 \left( \frac{2}{6} \right) - \frac{4}{6} \log_2 \left( \frac{4}{6} \right) \right] \\&+ \frac{4}{14} \left[ -\frac{1}{4} \log_2 \left( \frac{1}{4} \right) - \frac{3}{4} \log_2 \left( \frac{3}{4} \right) \right] = \\&= 0.286 + 0.394 + 0.232 = \underline{0.912}\end{aligned}$$

$$IG_{(Buys\ Computer/Income)} = H_{(Buys\ Computer)} - H_{(Income)} = 0.941 - 0.912 = \underline{0.029}$$

# Decision Trees Using ID3 Algorithm

- b. Which attribute should you choose as the root of a decision tree?

Student (14) <u>0.941</u>	
Yes	No
7	7
[1 No , 6 Yes]	[4 No , 3 Yes]

$$\begin{aligned}H_{(Student)} &= \frac{7}{14} H[1,7] + \frac{7}{14} H[4,3] = \\&= \frac{7}{14} \left[ -\frac{1}{7} \log_2 \left( \frac{1}{7} \right) - \frac{6}{7} \log_2 \left( \frac{6}{7} \right) \right] + \frac{7}{14} \left[ -\frac{4}{7} \log_2 \left( \frac{4}{7} \right) - \frac{3}{7} \log_2 \left( \frac{3}{7} \right) \right] \\&= 0.286 + 0.493 = \underline{0.779}\end{aligned}$$

$$IG_{(Buys\ Computer/Student)} = H_{(Buys\ Computer)} - H_{(Student)} = 0.941 - 0.779 = \underline{0.162}$$

# Decision Trees Using ID3 Algorithm

- b. Which attribute should you choose as the root of a decision tree?

Credit Rating (14) <u>0.941</u>	
Fair	Excellent
8	6
[2 No , 6 Yes]	[3 No , 3 Yes]

$$\begin{aligned}H_{(\text{Credit rating})} &= \frac{8}{14} H[2,6] + \frac{6}{14} H[3,3] = \\&= \frac{8}{14} \left[ -\frac{2}{8} \log_2 \left( \frac{2}{8} \right) - \frac{6}{8} \log_2 \left( \frac{6}{8} \right) \right] + \frac{6}{14} \left[ -\frac{3}{6} \log_2 \left( \frac{3}{6} \right) - \frac{3}{6} \log_2 \left( \frac{3}{6} \right) \right] \\&= 0.464 + 0.429 = \underline{0.893}\end{aligned}$$

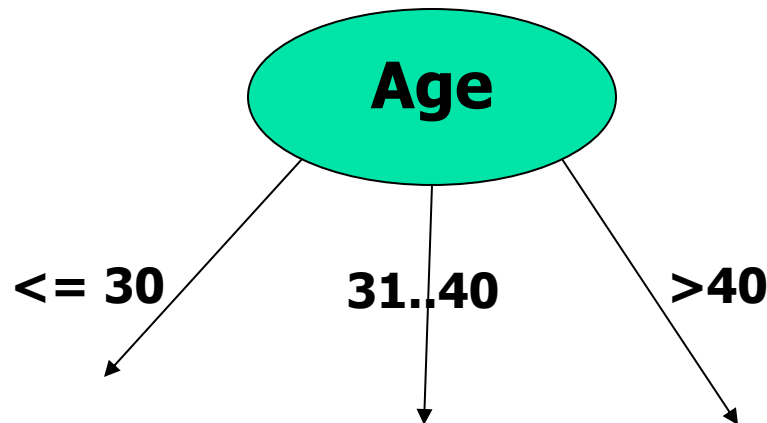
$$IG_{(\text{Buys Computer/Credit Rating})} = H_{(\text{Buys Computer})} - H_{(\text{Credit Rating})} = 0.941 - 0.893 = \underline{0.048}$$

**So, Age is the root of the tree, because  $IG_{(\text{Edible/ Smooth})}$  has the Greatest Information Gain Value**

# Decision Trees Using ID3 Algorithm

- b. Which attribute should you choose as the root of a decision tree?

Age (14) <u>0.941</u>		
<b>&lt;= 30</b>	<b>31..40</b>	<b>&gt;40</b>
5	4	5
[3 No , 2 Yes]	[0 No , 4 Yes]	[2 No , 3 Yes]



# Attribute Selection: 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 <sub>i</sub>	n <sub>i</sub>	I(p <sub>i</sub> , n <sub>i</sub> )
<=30	2	3	0.971
31...40	4	0	0
>40	3	2	0.971

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

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

$$Gain(income) = 0.029$$

$$Gain(student) = 0.151$$

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

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



# Decision Trees Using ID3 Algorithm

---

- You are stranded on a deserted island. Mushrooms of various types grow widely all over the island, but no other food is anywhere to be found. Some of the mushrooms have been determined as poisonous and others as not (determined by your former companions' trial and error). You are the only one remaining on the island. You have the following data to consider:

# Decision Trees Using ID3 Algorithm

---

Example	<i>NotHeavy</i>	<i>Smelly</i>	<i>Spotted</i>	<i>Smooth</i>	Edible
<i>A</i>	1	0	0	0	Yes
<i>B</i>	1	0	1	0	Yes
<i>C</i>	0	1	0	1	Yes
<i>D</i>	0	0	0	1	No
<i>E</i>	1	1	1	0	No
<i>F</i>	1	0	1	1	No
<i>G</i>	1	0	0	1	No
<i>H</i>	0	1	0	0	No

# Decision Trees Using ID3 Algorithm

---

- a. What is the entropy of Edible?
- b. Which attribute should you choose as the root of a decision tree?

# Decision Trees Using ID3 Algorithm

- a. What is the entropy of Edible?

Edible	
No	Yes
5	3

$$\text{Entropy}_{(\text{Edible})} = H_{(\text{Edible})}$$

$$= -\frac{5}{8} \log_2 \left(\frac{5}{8}\right) - \frac{3}{8} \log_2 \left(\frac{3}{8}\right) = 0.4238 + 0.5306$$

$$= \underline{0.9544}$$

Example	<i>NotHeavy</i>	<i>Smelly</i>	<i>Spotted</i>	<i>Smooth</i>	Edible
<i>A</i>	1	0	0	0	Yes
<i>B</i>	1	0	1	0	Yes
<i>C</i>	0	1	0	1	Yes
<i>D</i>	0	0	0	1	No
<i>E</i>	1	1	1	0	No
<i>F</i>	1	0	1	1	No
<i>G</i>	1	0	0	1	No
<i>H</i>	0	1	0	0	No

# Decision Trees Using ID3 Algorithm

- b. Which attribute should you choose as the root of a decision tree?

Not Heavy (8) <b><u>0.9544</u></b>	
<b>0</b>	<b>1</b>
3	5
[2 No , 1 Yes]	[3 No , 2 Yes]

$$\begin{aligned}
 H_{(Not\ Heavy)} &= \frac{3}{8} H[2,1] + \frac{5}{8} H[3,2] = \\
 &= \frac{3}{8} \left[ -\frac{2}{3} \log_2 \left( \frac{2}{3} \right) - \frac{1}{3} \log_2 \left( \frac{1}{3} \right) \right] + \frac{5}{8} \left[ -\frac{3}{5} \log_2 \left( \frac{3}{5} \right) - \frac{2}{5} \log_2 \left( \frac{2}{5} \right) \right] = \\
 &= 0.3444 + 0.6068 = \mathbf{\underline{0.9512}}
 \end{aligned}$$

$$IG_{(Edible/Not\ Heavy)} = H_{(Edible)} - H_{(Not\ Heavy)} = 0.9544 - 0.9512 = \mathbf{\underline{0.0032}}$$

Example	NotHeavy	Smelly	Spotted	Smooth	Edible
A	1	0	0	0	Yes
B	1	0	1	0	Yes
C	0	1	0	1	Yes
D	0	0	0	1	No
E	1	1	1	0	No
F	1	0	1	1	No
G	1	0	0	1	No
H	0	1	0	0	No

# Decision Trees Using ID3 Algorithm

- b. Which attribute should you choose as the root of a decision tree?

Smelly (8) <u>0.9544</u>	
0	1
5	3
[3 No , 2 Yes]	[2 No , 1 Yes]

$$\begin{aligned}
 H_{(Smelly)} &= \frac{5}{8} H[3,2] + \frac{3}{8} H[2,1] = \\
 &= \frac{5}{8} \left[ -\frac{3}{5} \log_2 \left( \frac{3}{5} \right) - \frac{2}{5} \log_2 \left( \frac{2}{5} \right) \right] + \frac{3}{8} \left[ -\frac{2}{3} \log_2 \left( \frac{2}{3} \right) - \frac{1}{3} \log_2 \left( \frac{1}{3} \right) \right] = \\
 &= 0.6068 + 0.3444 = \underline{0.9512}
 \end{aligned}$$

$$IG_{(Edible/Smelly)} = H_{(Edible)} - H_{(Smelly)} = 0.9544 - 0.9512 = \underline{0.0032}$$

Example	NotHeavy	Smelly	Spotted	Smooth	Edible
A	1	0	0	0	Yes
B	1	0	1	0	Yes
C	0	1	0	1	Yes
D	0	0	0	1	No
E	1	1	1	0	No
F	1	0	1	1	No
G	1	0	0	1	No
H	0	1	0	0	No

# Decision Trees Using ID3 Algorithm

- b. Which attribute should you choose as the root of a decision tree?

Spotted (8) <u>0.9544</u>	
0	1
5	3
[3 No , 2 Yes]	[2 No , 1 Yes]

$$\begin{aligned}
 H_{(Spotted)} &= \frac{5}{8} H[3,2] + \frac{3}{8} H[2,1] = \\
 &= \frac{5}{8} \left[ -\frac{3}{5} \log_2 \left( \frac{3}{5} \right) - \frac{2}{5} \log_2 \left( \frac{2}{5} \right) \right] + \frac{3}{8} \left[ -\frac{2}{3} \log_2 \left( \frac{2}{3} \right) - \frac{1}{3} \log_2 \left( \frac{1}{3} \right) \right] = \\
 &= 0.6068 + 0.3444 = \underline{0.9512}
 \end{aligned}$$

$$IG_{(Edible/Spotted)} = H_{(Edible)} - H_{(Spotted)} = 0.9544 - 0.9512 = \underline{0.0032}$$

Example	NotHeavy	Smelly	Spotted	Smooth	Edible
A	1	0	0	0	Yes
B	1	0	1	0	Yes
C	0	1	0	1	Yes
D	0	0	0	1	No
E	1	1	1	0	No
F	1	0	1	1	No
G	1	0	0	1	No
H	0	1	0	0	No

# Decision Trees Using ID3 Algorithm

- b. Which attribute should you choose as the root of a decision tree?

Smooth (8) <u>0.9544</u>	
0	1
4	4
[2 No , 2 Yes]	[3 No , 1 Yes]

$$\begin{aligned}
 H_{(Smooth)} &= \frac{4}{8} H[2,2] + \frac{4}{8} H[3,1] = \\
 &= \frac{4}{8} \left[ -\frac{2}{4} \log_2 \left( \frac{2}{4} \right) - \frac{2}{4} \log_2 \left( \frac{2}{4} \right) \right] + \frac{4}{8} \left[ -\frac{3}{4} \log_2 \left( \frac{3}{4} \right) - \frac{1}{4} \log_2 \left( \frac{1}{4} \right) \right] = \\
 &= 0.5 + 0.4056 = \underline{0.9056}
 \end{aligned}$$

$$IG_{(Edible/Smooth)} = H_{(Edible)} - H_{(Smooth)} = 0.9544 - 0.9056 = \underline{0.0488}$$

So, Smooth is the root of the tree, because  $IG_{(Edible/Smooth)}$  has the Greatest Information Gain Value

Example	NotHeavy	Smelly	Spotted	Smooth	Edible
A	1	0	0	0	Yes
B	1	0	1	0	Yes
C	0	1	0	1	Yes
D	0	0	0	1	No
E	1	1	1	0	No
F	1	0	1	1	No
G	1	0	0	1	No
H	0	1	0	0	No



# Computing Information-Gain for Continuous-Valued Attributes

---

- Let attribute A be a continuous-valued attribute
- Must determine the *best split point* for A
  - Sort the value A in increasing order
  - Typically, the midpoint between each pair of adjacent values is considered as a possible *split point*
    - $(a_i + a_{i+1})/2$  is the midpoint between the values of  $a_i$  and  $a_{i+1}$
  - The point with the *minimum expected information requirement* for A is selected as the split-point for A
- Split:
  - D1 is the set of tuples in D satisfying  $A \leq \text{split-point}$ , and D2 is the set of tuples in D satisfying  $A > \text{split-point}$

# Overfitting and Tree Pruning

---

- Overfitting: An induced tree may overfit the training data
  - Too many branches, some may reflect anomalies due to noise or outliers
  - Poor accuracy for unseen samples
- Two approaches to avoid overfitting
  - Prepruning: *Halt tree construction early*-do not split a node if this would result in the goodness measure falling below a threshold
    - Difficult to choose an appropriate threshold
  - Postpruning: *Remove branches* from a “fully grown” tree—get a sequence of progressively pruned trees
    - Use a set of data different from the training data to decide which is the “best pruned tree”

# Enhancements to Basic Decision Tree Induction

---

- Allow for **continuous-valued attributes**
  - Dynamically define new discrete-valued attributes that partition the continuous attribute value into a discrete set of intervals
- Handle **missing attribute values**
  - Assign the most common value of the attribute
  - Assign probability to each of the possible values
- **Attribute construction**
  - Create new attributes based on existing ones that are sparsely represented
  - This reduces fragmentation, repetition, and replication

# Example

- Apply the ID3 algorithm. Suppose we want to train a decision tree using the following instances:

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay in
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

- Which attribute from the previously mentioned table will be the first node in the tree?
- Use your selected root node to partition the previously mentioned instances. Sketch your answer

- 
- (c) Consider the use of real-valued attributes, when learning decision trees, as described in the lecture. The table below shows the relationship between the body height and the gender of a group of persons (the records have been sorted with respect to the value of *height* in cm). Calculate the information gain for potential splitting thresholds and determine the best one.

<i>Height</i>	161	164	169	175	176	179	180	184	185
<i>Gender</i>	F	F	M	M	F	F	M	M	F

- 
- (c) Consider the use of real-valued attributes, when learning decision trees, as described in the lecture. The table below shows the relationship between the body height and the gender of a group of persons (the records have been sorted with respect to the value of *height* in cm). Calculate the information gain for potential splitting thresholds and determine the best one.

<i>Height</i>	161	164	169	175	176	179	180	184	185
<i>Gender</i>	F	F	M	M	F	F	M	M	F

- Potential cut points must lie in the intervals (164, 169), (175, 176), (179, 180), or (184, 185).

- 
- Calculate the information gain for the potential splitting thresholds
  - $C_1 \in (164, 169)$ 
    - resulting class distribution: if  $x < C_1$  then 2 – 0 else 3 – 4
    - conditional entropy: if  $x < C_1$  then  $E = 0$  else  $E = -\frac{3}{7} \log_2 \frac{3}{7} - \frac{4}{7} \log_2 \frac{4}{7} = 0.985$
    - entropy:  $E(C_1|S) = \frac{2}{9} \cdot 0 + \frac{7}{9} \cdot 0.985 = 0.766$
  - $C_2 \in (175, 176)$ 
    - resulting class distribution: if  $x < C_2$  then 2 – 2 else 3 – 2
    - entropy:  $E(C_2|S) = \frac{4}{9} \cdot 1 + \frac{5}{9} \cdot 0.971 = 0.984$
  - $C_3 \in (179, 180)$ 
    - resulting class distribution: if  $x < C_3$  then 4 – 2 else 1 – 2
    - entropy:  $E(C_3|S) = \frac{6}{9} \cdot 0.918 + \frac{3}{9} \cdot 0.918 = 0.918$
  - $C_4 \in (184, 185)$ 
    - resulting class distribution: if  $x < C_4$  then 4 – 4 else 1 – 0
    - entropy:  $E(C_4|S) = \frac{8}{9} \cdot 1 + \frac{1}{9} \cdot 0 = 0.889$