บทที่ 3

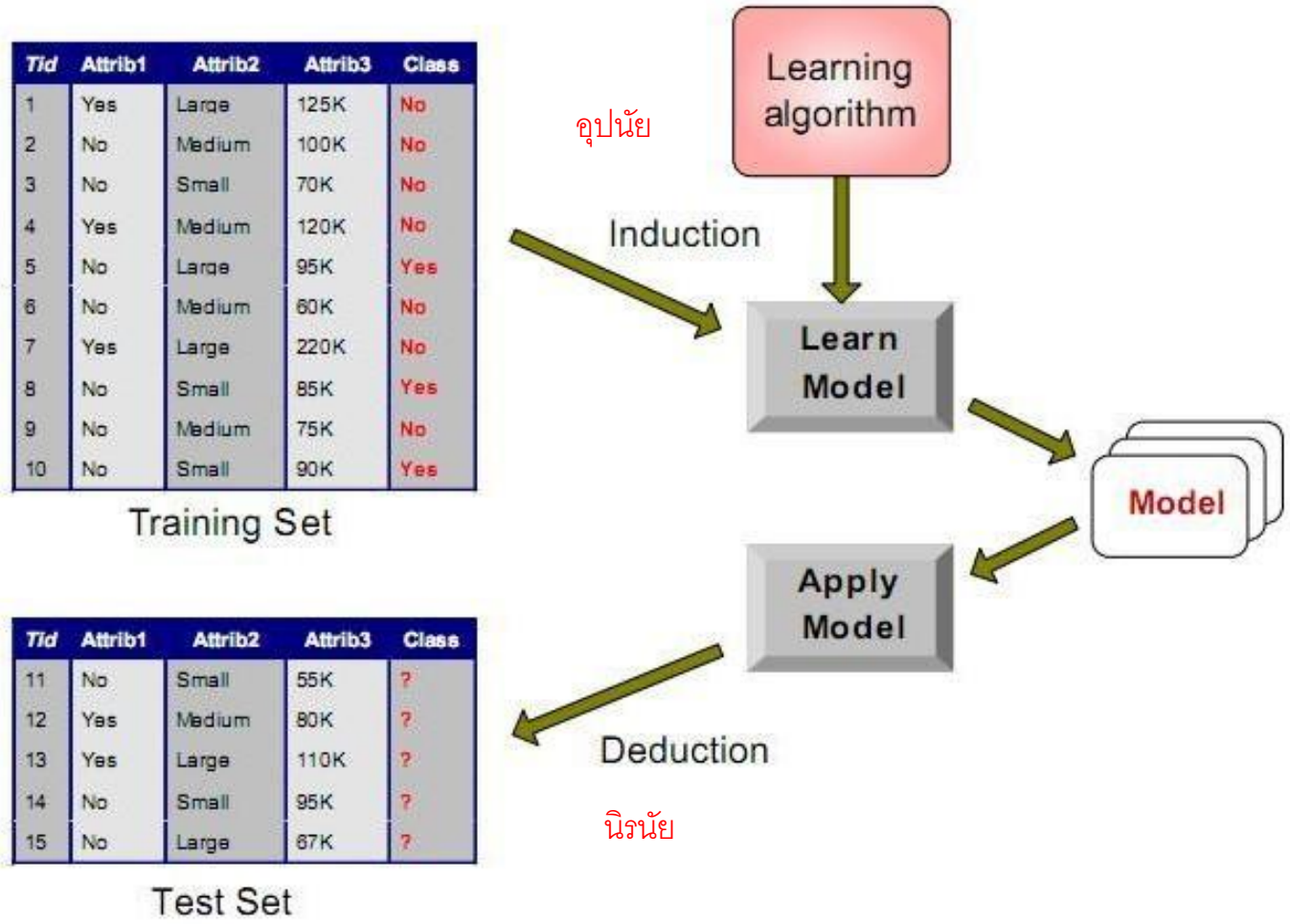
หลักการจำแนกข้อมูลเบื้องต้นการจำแนกข้อมูลด้วยโมเดลต้นไม้ตัดสินใจ

Classification & Prediction

Classification

- เป็นกระบวนการสร้างโมเดล จัดการข้อมูลให้ อยู่กลุ่มที่กำหนดมาให้
- เช่น จัดกลุ่มนักเรียนว่า ดีมาก ดี ปานกลาง ไม่ดี
- โดยพิจารณาจากประวัติและผลการเรียน หรือแบ่งประเภทของลูกค้าว่าเชื่อถือได้หรือเชื่อถือไม่ได้ โดยพิจารณาจากข้อมูลที่มีอยู่

Process of Classification



Classification

1. แบ่งข้อมูลตัวอย่าง (Samples Data) ออกเป็น 3 ส่วนได้แก่
   - Training Datasets
   - Validation Datasets
   - Test Datasets
2. นำ Training Datasets มาสร้าง Decision Tree
3. ใช้ Validation Datasets วัดความถูกต้องในการจำแนกของ Tree ที่สร้าง
4. ทำซ้ำข้อ 2,3 เพื่อให้ได้ความถูกต้องสูงสุด
5. ใช้ Testing Datasets มาทดสอบกับ Tree ที่ได้เพื่อวัดความถูกต้อง

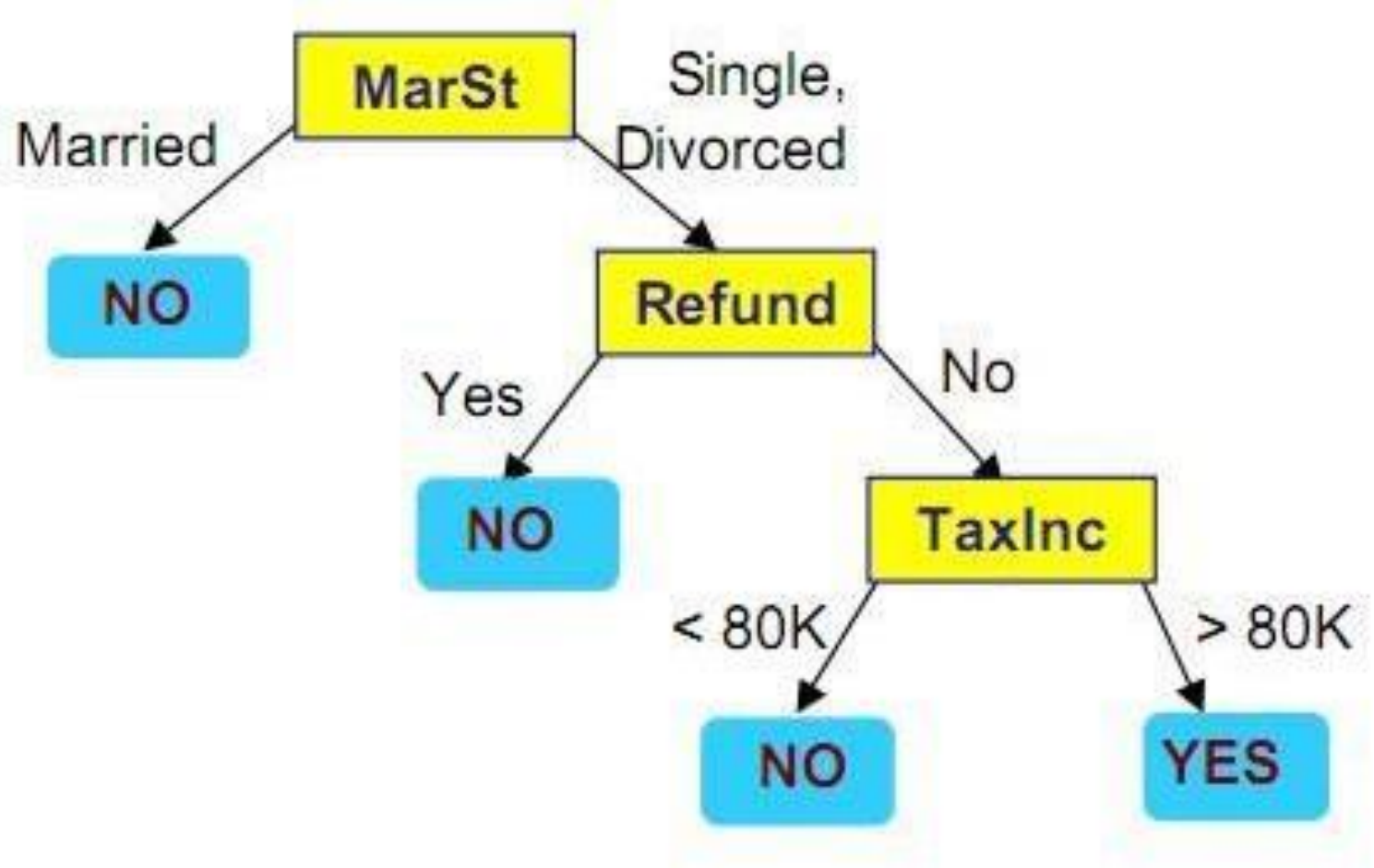
Classification : Definition

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

อัลกอริทึมที่ใช้ในการเหมืองข้อมูลแบบจำแนก ได้แก่

- **Decision Tree induction**
- **Naïve Bayes method**
- **K-nearest neighbor (K-NN)**
- **Neural Network**

Decision Tree Example



Decision Tree Learning Algorithm

- **Decision Tree** เป็นการนำข้อมูลมาสร้างแบบจำลองการพยากรณ์ในรูปแบบโครงสร้างต้นไม้ และมีการทำงานแบบ Supervised Learning (คือการเรียนรู้ของโมเดลแบบมีครูสอน) สามารถสร้างแบบจำลองการจัดหมวดหมู่ได้จากกลุ่มตัวอย่างข้อมูลที่กำหนดไว้ล่วงหน้า และพยากรณ์กลุ่มของรายการที่ยังไม่เคยนำมาจัดหมวดหมู่ ได้ด้วยรูปแบบของ Tree โครงสร้างประกอบด้วย Root Node, Child และ Leaf Node อัลกอริทึม ที่ใช้ในการสร้าง Decision Tree ได้แก่
  - ID3 Algorithm
  - C4.5 Algorithm (มีกระบวนการไม่ยุ่งยากไม่ซับซ้อนใช้ Math ไม่มากนัก)
  - C5.0 Algorithm
  - CART Algorithm

  ในแต่ละวิธีมีข้อดีข้อด้อยต่างกันไป ในที่นี้จะขอกล่าวถึงเฉพาะ C4.5

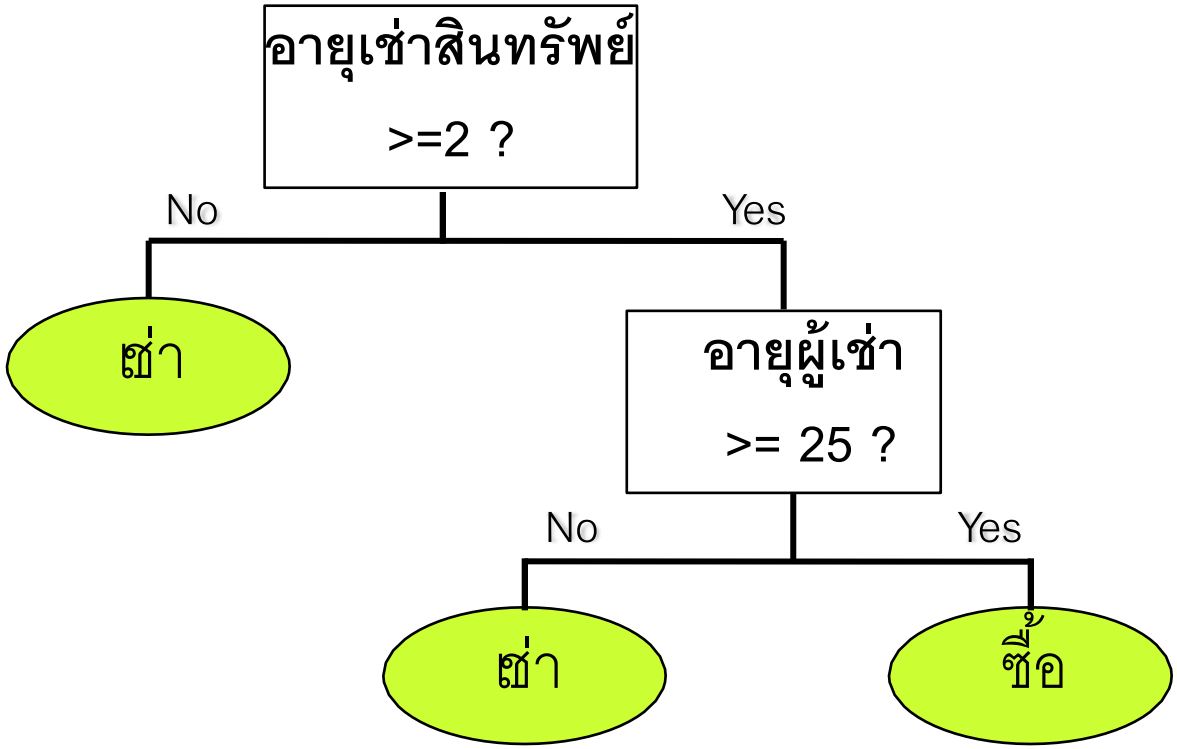
เทคนิคในการทำเหมืองข้อมูลแบบ Decision Tree

- สมมติว่าต้องการพยากรณ์ว่าลูกค้าที่ปัจจุบันได้เช่าสินทรัพย์อย่างใดอย่างหนึ่งไปแล้วจะมีโอกาสตัดสินใจซื้อสินทรัพย์ชนิดนั้นไปเป็นของตนเองหรือไม่ ? โดยใช้ปัจจัยในการวิเคราะห์คือ ระยะเวลาที่ลูกค้าได้เช่าสินทรัพย์มาและอายุของลูกค้า

อายุเช่าสินทรัพย์	อายุผู้เช่า	ซื้อสินทรัพย์
2	25	YES
3	22	YES
1	28	NO
5	30	YES
4	27	NO
2	21	NO
3	23	NO
…	…	…

เทคนิคในการทำเหมืองข้อมูลแบบ Decision Tree

- จากข้อมูลพบว่า “ลูกค้าที่มีอายุเช่าสินทรัพย์ตั้งแต่ 2 ปีขึ้นไป และอายุของผู้เช่าตั้งแต่ 25 ปีขึ้นไป มักจะตกลงซื้อสินทรัพย์เป็นของตนเอง”



Classification Sample Data

Outlook	Temperature	Humidity	Windy	Play(?)
sunny	hot	high	false	no
sunny	hot	high	true	no
cloudy	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
cloudy	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
cloudy	mild	high	true	yes
cloudy	hot	normal	false	yes
rainy	mild	high	true	no

Classification Sample Data (cons't)

- ข้อมูลที่กำหนดในตารางเป็นข้อมูลสภาพอากาศที่ใช้ประกอบการตัดสินใจในการเล่นกีฬาชนิดหนึ่งว่า

  มีสภาพอากาศอย่างไรจึงจะเล่น (play = yes)

  มีสภาพอากาศอย่างไรจึงไม่เล่น (play = no)

- ในงานจำแนกข้อมูล (Classification) ข้อมูลที่เป็นจุดมุ่งหมายใน การจำแนกคือ **แอททรบิวต์** play
- ขณะที่แอททริบิวต์ outlook , temperature , humidity , windy ทำหน้าที่เป็น predicting attributes

Classification Sample Data (cons't)

- ปัญหาที่ต้องพิจารณาคือ จะเลือก Attributes ใดทำหน้าที่เป็น root node ในแต่ละขั้นตอนของการสร้าง tree และ subtree
- เกณฑ์ที่ช่วยตัดสินใจ ในการเลือก root node คือ ทดลองเลือก Attribute แต่ละตัวมาทำหน้าที่เป็น root node แล้วหาค่า Gain ซึ่งเป็นค่าที่ใช้บอกว่า attribute ที่ทำหน้าที่เป็น root node สามารถจำแนกข้อมูลได้ดีมากน้อยเพียงใด

จะเลือก attribute **ที่ให้ค่า** Gain **สูงสุดเป็น** root node

การหาค่า Gain

- Gain เป็นค่าที่บอกระดับความสามารถของการจำแนกคลาสของ attribute หน่วยของการวัดเป็น bits (มาจาก Information Theory)

ถ้าให้

- T แทน เซตของ Training Set
- X แทน แอททริบิวต์ ที่ถูกเลือกให้เป็นตัวจำแนกข้อมูล

$$Gain(x) = info(T) - info_X(T)$$

การหาค่า Gain (ต่อ)

- Info(T) เป็นฟังก์ชัน ที่ระบุปริมาณข้อมูลที่ต้องการเพื่อให้สามารถจำแนกคลาสที่ต้องการได้
- info(T) = $- \sum_{j=1 \text{ to } k} [\text{freq}(C_j, T) / |T|] \times \log_2 [\text{freq}(C_j, T) / |T|]$ bits

- เมื่อ |T| คือ จำนวนข้อมูลทั้งหมดใน Training Datasets Freq(C_j,T) คือ ความถี่ที่ข้อมูลใน T ปรากฏเป็นคลาส C_j

การหาค่า Gain (ต่อ)

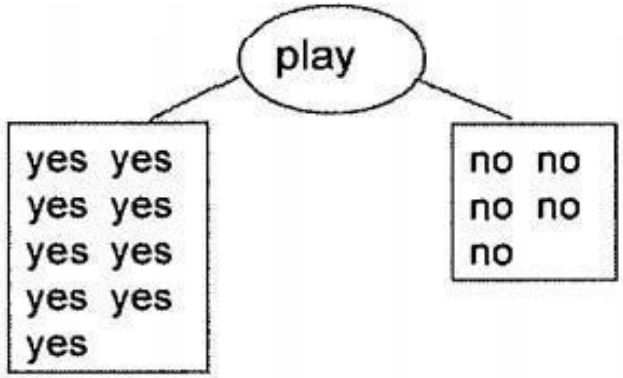
- $Info_X(T)$ หรือ Entropy คือ ฟังก์ชันที่ระบุปริมาณข้อมูลที่ต้องการเพื่อการจำแนกคลาสของข้อมูลโดยใช้ attribute X เป็นตัวตรวจสอบเพื่อแยกข้อมูล
- $Info_X(T)$ =

$$\sum_{i=1 \text{ to } n} \left(|T_i| / |T| \right) \times info(T_i) \quad \text{bits}$$

- เมื่อ i คือ จำนวนค่าที่เป็นไปได้ของแอททริบิวต์ x
- |Ti| คือ จำนวนข้อมูลที่มีค่า x=i

Example

- จากตัวอย่างข้อมูลจะหาค่า Gain ของแต่ละ attribute ที่จะ เลือกเป็น Root node
- 1. จะต้องหาค่า info(T)



$$info(T) = -[9/14 \times \log_2 (9/14)] - [5/14 \times \log_2 (5/14)]$$
$$= 0.940 \text{ bits}$$

Example

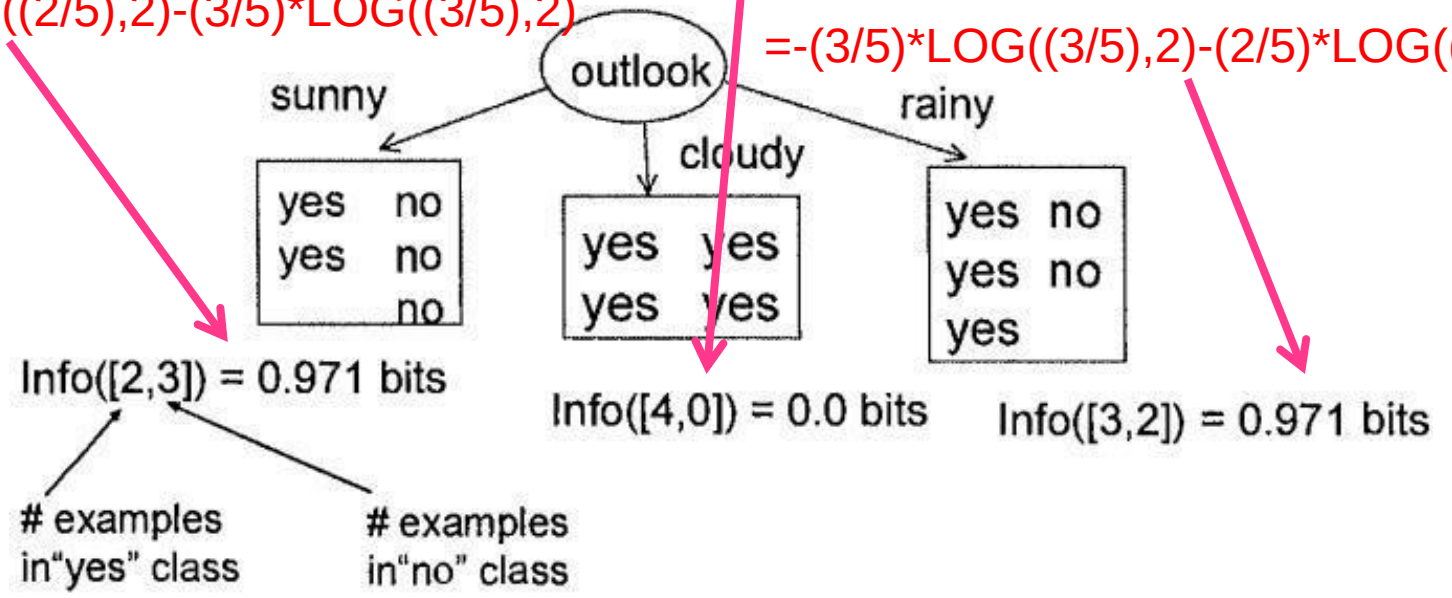
Log_2 0 ไม่นิยาม

- 2. หาค่า $info_X(T)$ ของแต่ละแอททริบิวต์
- ค่า info outlook(T) หาได้ดังนี้

=-(2/5)*LOG((2/5),2)-(3/5)*LOG((3/5),2)

=-(4/4)*LOG(4/4,2)-(0/4)*(LOG(0/4,2)

=-(3/5)*LOG((3/5),2)-(2/5)*LOG((2/5),2)



Average information = info([2,3], [4,0], [3,2])

= 5/14 x 0.971 + 4/14 x 0.0 + 5/14 x 0.971

= 0.693 bits

Example

- การจะจำแนกคลาสของข้อมูลออกเป็น play = yes หรือ play = no ต้องใช้ข้อมูลจากแอททริบิวต์อื่นประกอบการตัดสินใจ ถ้าแอททริบิวต์ outlook จะต้องดูข้อมูลเพื่อ ประกอบการเลือกคลาส ดังนี้

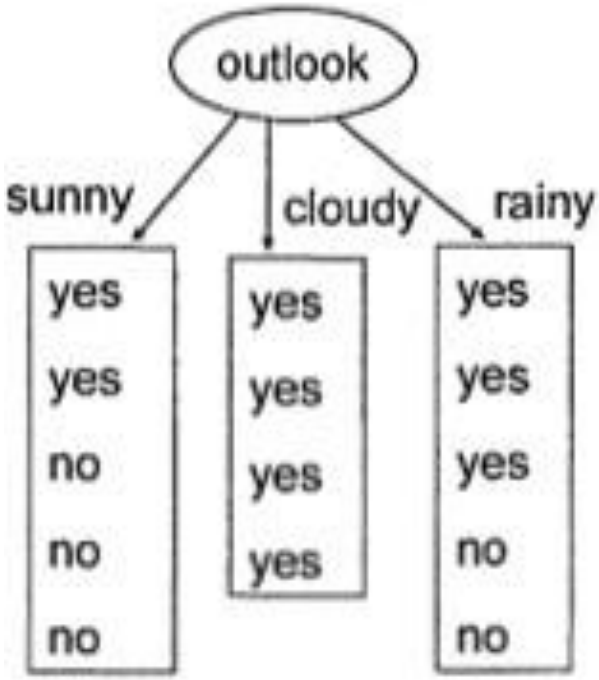
$$\begin{aligned} \text{info}_{\text{outlook}}(T) = \; & (5/14) \times [\, -(2/5) \times \log_2(2/5) - (3/5) \times \log_2(3/5) \,] \\ & + (4/14) \times [\, -(4/4) \times \log_2(4/4) - (0/4) \times \log_2(0/4) \,] \\ & + (5/14) \times [\, -(3/5) \times \log_2(3/5) - (2/5) \times \log_2(2/5) \,] \\ = \; & 0.693 \text{ bits} \end{aligned}$$

Example

$$\text{entropy}(P_1, P_2, \ldots, P_n) = -P_1 \log P_1 - P_2 \log P_2 - \ldots - P_n \log P_n$$

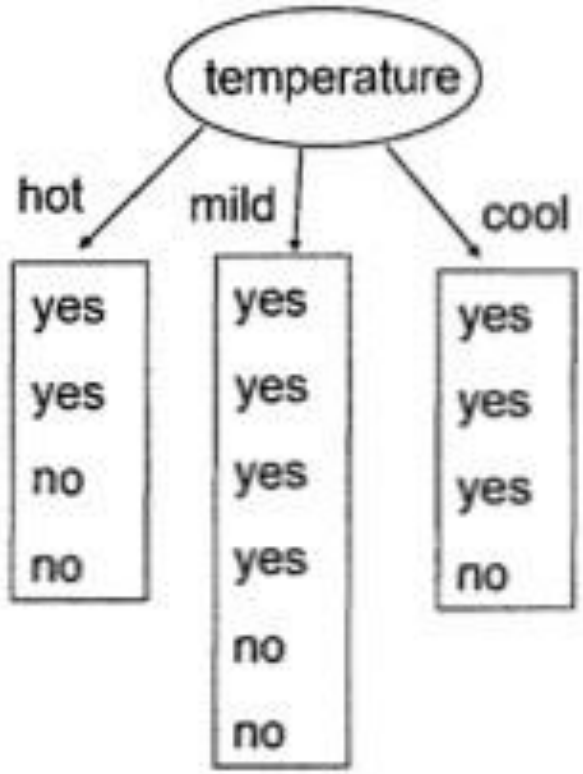
$$\begin{aligned}
\text{info}([2,4,3]) &= \text{entropy}(2/9, 4/9, 3/9) \\
&= -(2/9)\log(2/9) - (4/9)\log(4/9) - (3/9)\log(3/9) \\
&= (-2\log 2 - 4\log 4 - 3\log 3 + 9\log 9)/9
\end{aligned}$$

Classification : Constructing Decision Tree



gain = 0.247 bits

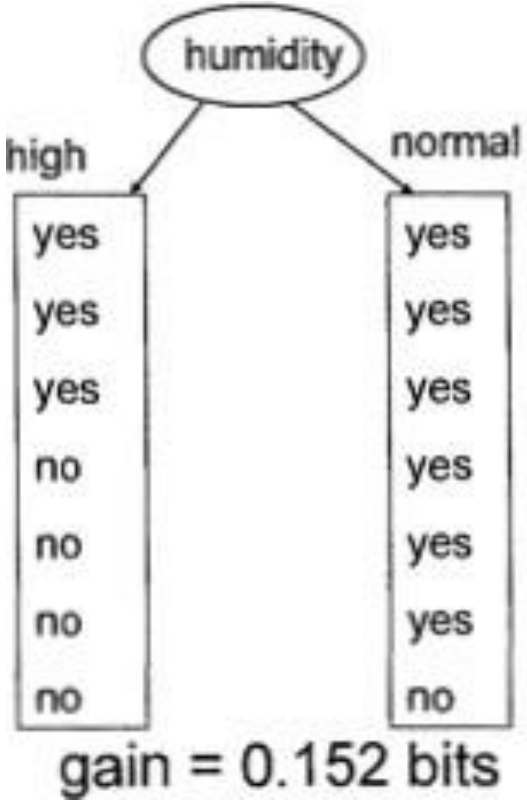
= 0.940-0.693

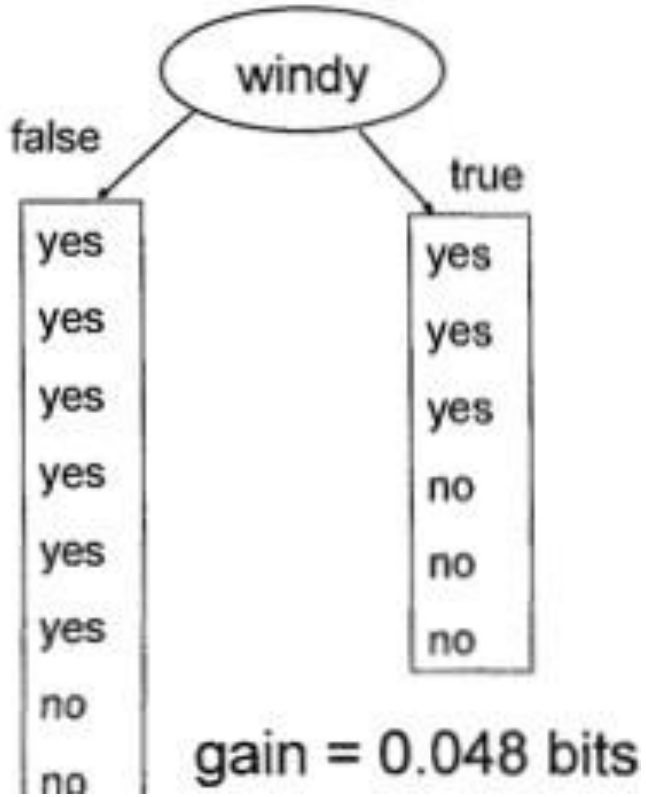


gain = 0.029 bits

= 0.940-0.911

Classification : Constructing Decision Tree (ต่อ)





เปรียบเทียบค่า Gain ที่ได้

$$\text{gain (outlook)} = \text{info (T)} - \text{info}_{\text{outlook}}\ \text{(T)}$$
$$= 0.940 - 0.693$$
$$= 0.247 \quad \text{bits}$$

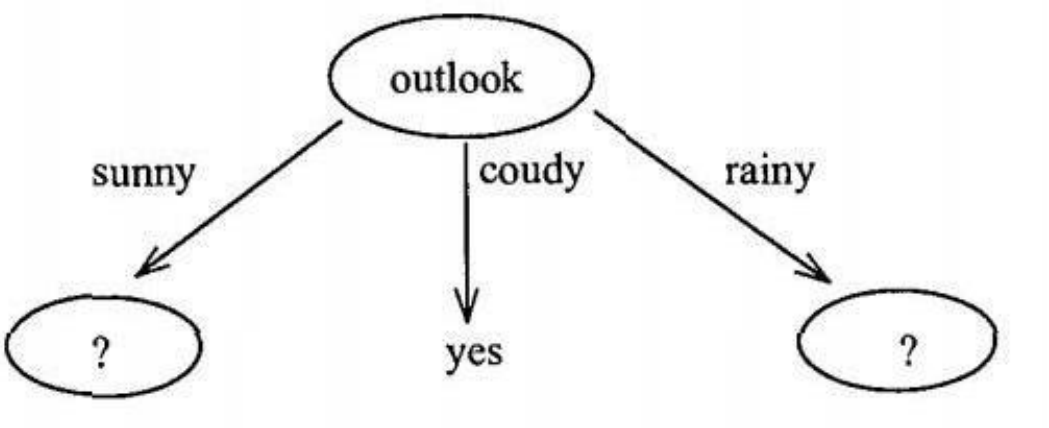
$$\text{gain (temperature)} = \text{info (T)} - \text{info}_{\text{temperature}}\ \text{(T)}$$
$$= 0.940 - 0.911$$
$$= 0.029 \quad \text{bits}$$

$$\text{gain (humidity)} = \text{info (T)} - \text{info}_{\text{humidity}}\ \text{(T)}$$
$$= 0.940 - 0.788$$
$$= 0.152 \quad \text{bits}$$

$$\text{gain (windy)} = \text{info (T)} - \text{info}_{\text{windy}}\text{(T)}$$
$$= 0.940 - 0.892$$
$$= 0.048 \quad \text{bits}$$

Constructing Decision Tree (ต่อ)

- เนื่องจากแอททริบิวต์ outlook ยังไม่สามารถจัดกลุ่มข้อมูลให้เป็นคลาสเดียวกัน ได้ทั้งหมด (outlook = sunny จัดกลุ่มข้อมูลที่เป็นคลาส yes=2 recs, no=3 recs และ outlook = rainy จัดกลุ่มข้อมูลที่เป็นคลาส yes=3 recs, no=2 recs) จึงต้องสร้าง decision tree ต่อ โดยเลือกแอททริบิวต์ที่จะมาเป็น node ในระดับที่ 2 ต่อจาก root node ในกรณี outlook = cloudy ไม่จำเป็นต้องสร้าง node เพิ่มเติม เพราะสามารถจัดกลุ่มข้อมูลที่เป็นคลาส yes ได้ทั้งหมด



Constructing Decision Tree (ต่อ)

- แอททริบิวต์ที่สามารถถูกเลือกเป็น node ในระดับที่ 2 ได้ คือ temperature, humidity และ windy
- พิจารณาการสร้างโหนดลูกทางซ้ายมือ outlook = sunny ถ้าเลือกแอททริบิวต์ temperature จะได้คำนวณค่า gain ได้ดังนี้

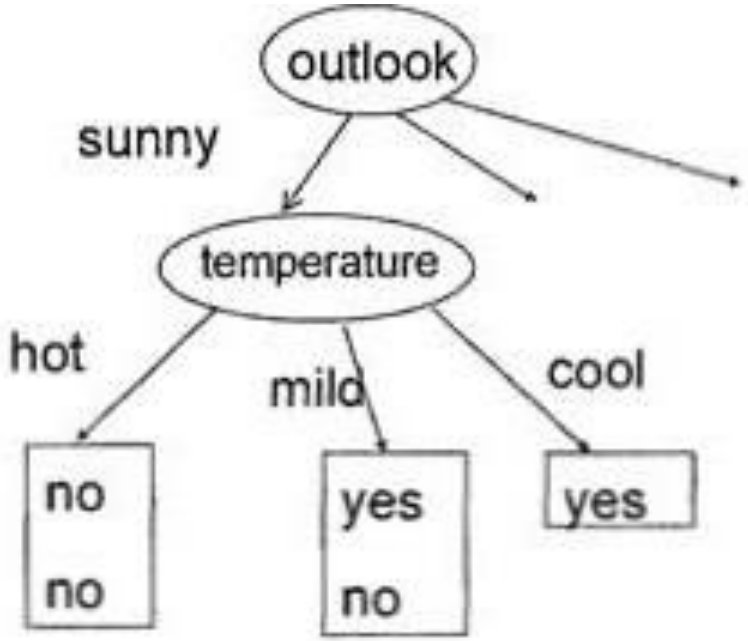
$$\text{gain (temperature)} = \text{info (outlook = sunny)} - \text{info}_{\text{temperature}}\text{ (outlook = sunny)}$$

- เนื่องจาก outlook = sunny จัดกลุ่มข้อมูลที่เป็นคลาส yes=2 recs, no=3 recs ดังนั้น

$$\text{info (outlook = sunny)} = -\frac{2}{5}\times\log_2\left(\frac{2}{5}\right)-\frac{3}{5}\times\log_2\left(\frac{3}{5}\right)$$

$$= 0.971 \text{ bits}$$

Constructing Decision Tree (ต่อ)



gain = 0.571 bits

Constructing Decision Tree (ต่อ)

$$\text{info}_{\text{temperature}}(\text{outlook} = \text{sunny}) = \text{info}([0,2], [1,1], [1,0])$$

$$= \frac{2}{5}\times\left[-\frac{0}{2}\times\log_2\left(\frac{0}{2}\right)-\frac{2}{2}\times\log_2\left(\frac{2}{2}\right)\right]$$

$$+ \frac{2}{5}\times\left[-\frac{1}{2}\times\log_2\left(\frac{1}{2}\right)-\frac{1}{2}\times\log_2\left(\frac{1}{2}\right)\right]$$

$$+ \frac{1}{5}\times\left[-\frac{1}{1}\times\log_2\left(\frac{1}{1}\right)-\frac{0}{1}\times\log_2\left(\frac{0}{1}\right)\right]$$

$$= 0.4 \text{ bits}$$

$$\therefore \text{ gain (temperature)} = 0.971 - 0.4 \text{ bits}$$

$$= 0.571 \text{ bits}$$

Constructing Decision Tree (ต่อ)

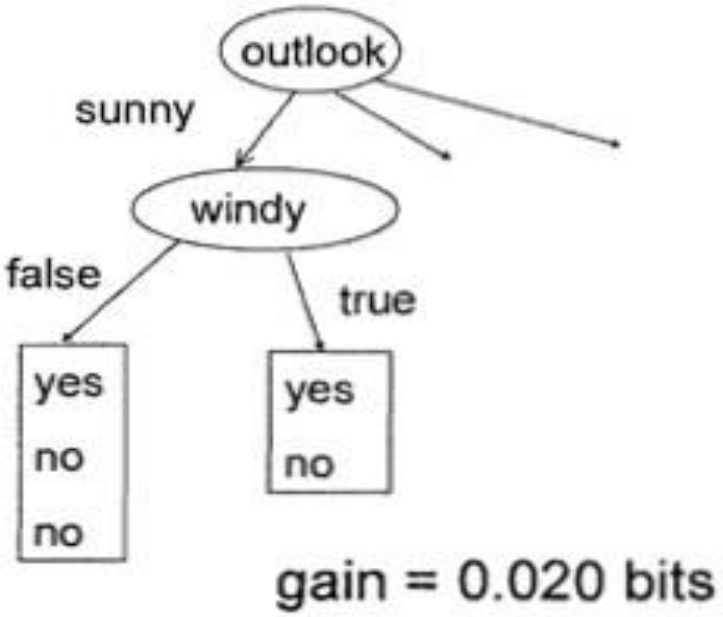
- เมื่อลอง outlook = sunny แล้ว ทดลองจำแนกกลุ่มต่อไปด้วยแอททริบิวต์ cloudy

$$\text{gain (windy)} = \text{info (outlook = sunny)} - \text{info}_{\text{windy}} \text{(outlook = sunny)}$$

$$= 0.971 - \text{info ([1,2] , [1, 1])}$$

$$= 0.971 - 0.951 \text{ bits}$$

$$= 0.020 \text{ bits}$$



Constructing Decision Tree (ต่อ)

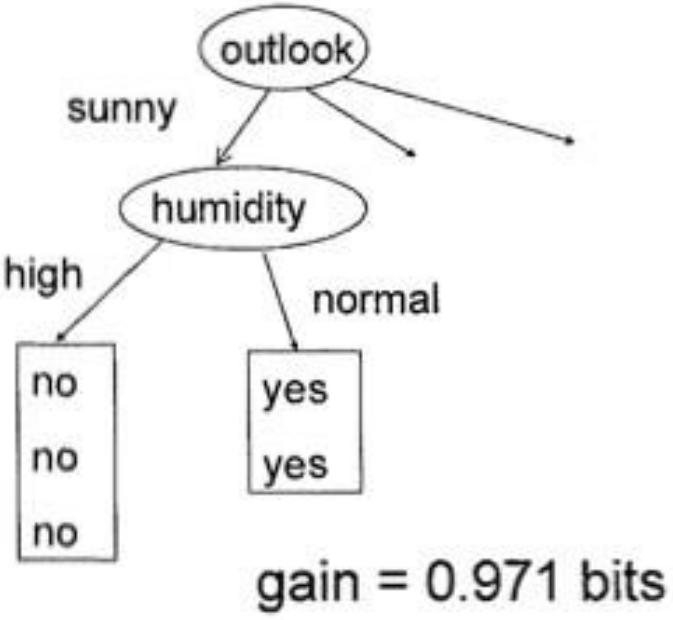
- เมื่อลอง outlook = sunny แล้ว ทดลองจำแนกกลุ่มต่อไปด้วยแอททริบิวต์ humidity

$$\text{gain (humidity)} = \text{info (outlook = sunny)} - \text{info}_{\text{humidity}}\text{ (outlook = sunny)}$$

$$= 0.971 - \text{info ([0,3] , [2. 0])}$$

$$= 0.971 - 0 \text{ bits}$$

$$= 0.971 \text{ bits}$$



Constructing Decision Tree (ต่อ)

- ในทำนองเดียวกันเมื่อ outlook = sunny แล้ว ทดลองสร้าง decision tree ด้วยแอททริบิวต์ อื่นๆ จะได้ค่า gain ดังต่อไปนี้

  gain(temperature) = 0.571

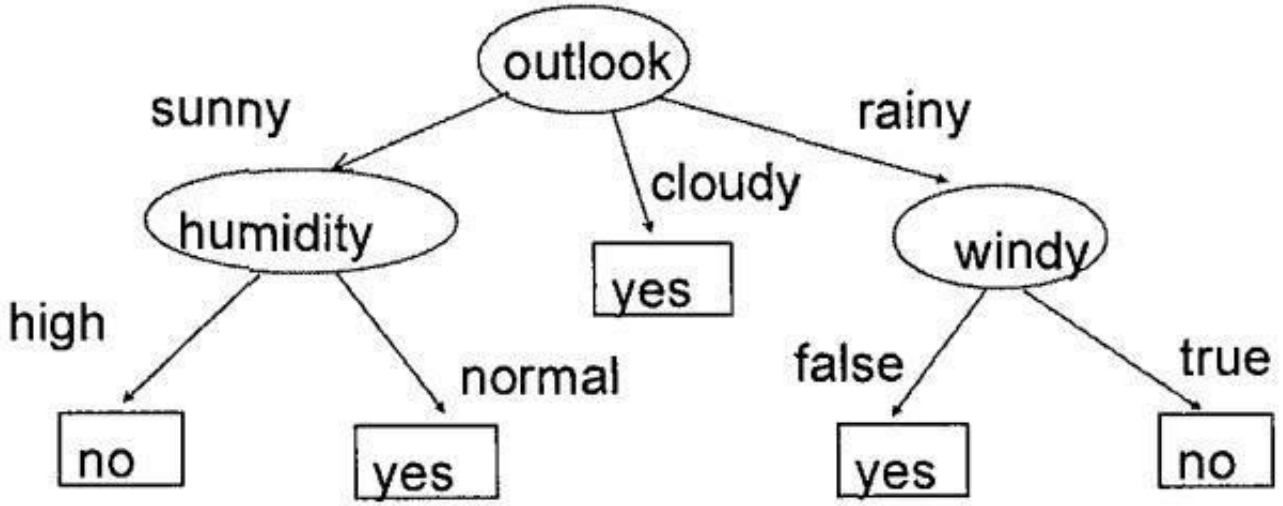
  gain(humidity) = 0.971

  gain(windy) = 0.020

- จึงพิจารณาเลือกแอททริบิวต์ humidity (เพราะมีค่า gain สูงสุด) เป็น node ในระดับที่สองต่อจากโหนด outlook

Constructing Decision Tree (ต่อ)

- decision tree ยังเหลือโหนดลูกทางขวาของโหนด outlook ที่ต้องพิจารณา เลือกแอททริบิวต์และจากวิธีคำนวณค่า gain ที่ผ่านมาก่อนหน้านี้ สามารถเลือกได้ว่าแอททริบิวต์ windy จะให้ค่า gain สูงที่สุด
- กระบวนการสร้าง decision tree จะสิ้นสุด ก็ต่อเมื่อ leaf node เป็นกลุ่มของข้อมูลคลาสเดียวกันทั้งหมด



Classification Rule

- ในกรณีที่มีข้อมูลใหม่ที่ยังไม่ทราบคลาส “outlook = sunny, temperature = cool, humidity = high, windy = true” สามารถใช้ decision tree ทำนายคลาสของข้อมูลนี้ว่า play = no โดยพิจารณาจากเพียง 2 แอททริบิวต์ คือ outlook = sunny และ humidity = high
- Decision tree สามารถแปลงเป็น Classification rule ได้ดังนี้

rule 1 : IF (outlook = sunny) AND (humidity = high) THEN play = no

rule 2 : IF (outlook = sunny) AND (humidity = normal) THEN play = yes

rule 3 : IF (outlook = cloudy) THEN play = yes

rule 4 : IF (outlook = rainy) AND (windy = false) THEN play = yes

rule 5 : IF (outlook = rainy) AND (windy = true) THEN play = no

Exercise

- จากข้อมูล ความคิดเห็นของคน 7 คน ที่ต้องการเลือกผู้สมัครหมายเลข 1 หรือหมายเลข 2 โดยพิจารณาจากอายุ รายได้ และการศึกษาของผู้แสดงความคิดเห็นปรากฏดังตาราง ให้สร้าง Decision Tree พร้อมแสดงกฎการเรียนรู้ที่ได้

No	Age	Income	Education	Candidate
1	>=35	High	High School	1
2	<35	Low	University	1
3	>=35	High	College	2
4	>=35	Low	High School	2
5	>=35	High	University	1
6	<35	High	College	1
7	<35	Low	High School	2