สรุป Chapter 8

Classification เป็นกระบวนการสร้างโมเดล จัดการข้อมูลให้ อยู่กลุ่มที่กำหนดมาให้ เช่น จัดกลุ่มนักเรียนว่า ดีมาก ดี ปานกลาง ไม่ดี โดยพิจารณาจากประวัติและผลการเรียน หรือแบ่ง ประเภทของลูกค้าว่าเชื่อถือได้หรือเชื่อถือไม่ได้โดยพิจารณาจากข้อมูลที่มีอยู่

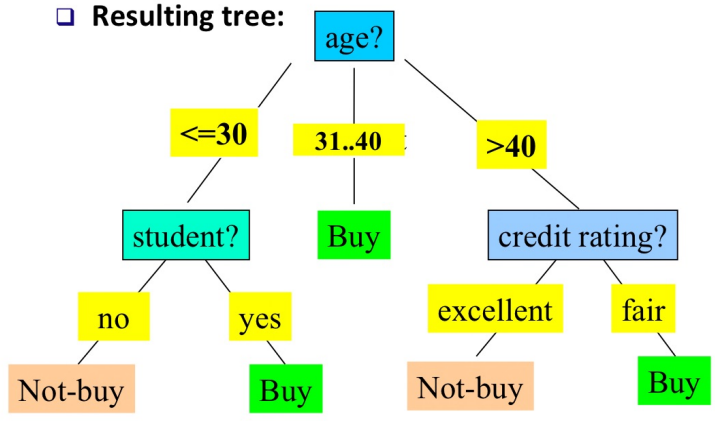
แบ่งข้อมูลตัวอย่าง (Samples Data) ออกเป็น 3 ส่วนได้แก่
- Training Datasets
- Validation Datasets
- Test Datasets

นำ Training Datasets มาสร้าง Decision Tree
ใช้ Validation Datasets วัดความถูกต้องในการจำแนกของ Tree ที่สร้าง
ทำซ้ำข้อ 2,3 เพื่อให้ได้ความถูกต้องสูงสุด
ใช้ Testing Datasets มาทดสอบกับ Tree ที่ได้เพื่อวัดความ ถูกต้อง

Decision Tree เป็นการนำข้อมูลมาสร้างแบบจำลองการพยากรณ์ใน รูปแบบโครงสร้างต้นไม้ และมีการทำงานแบบ Supervised Learning (คือการเรียนรู้ของโมเดลแบบมีครูสอน) สามารถสร้างแบบจาลองการจัด หมวดหมู่ได้จากกลุ่มตัวอย่างข้อมูลที่กาหนดไว้ล่วงหน้า และพยากรณ์กลุ่ม ของรายการที่ยังไม่เคยนามาจัดหมวดหมู่ ได้ด้วยรูปแบบของ Tree โครงสร้างประกอบด้วย Root Node, Child และ Leaf Node อัลกอริทึม ที่ใช้ในการสร้าง Decision Tree

Decision Tree Induction: An Example

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



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

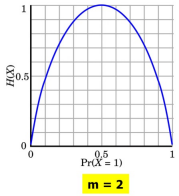
Note: The data set is adapted from "Playing Tennis" example of R. Quinlan

Gain เป็นค่าที่บอกระดับความสามารถของการจำแนกคลาสของ attribute หน่วยของการวัดเป็น bits (มาจาก Information Theory)

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, …, 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

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

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

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

ตัวอย่างการหาค่า Gain

Example: Attribute Selection with Information Gain

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

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

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

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