

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
<=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

$$Gain(income) = 0.029$$

$$Gain(student) = 0.151$$

$$Gain(credit_rating) = 0.048$$

คำนวณ Entropy ของข้อมูลทั้งหมด $Info(D)$

มีทั้งหมด 14 ตัวอย่าง แบ่งเป็น:

9 ตัวอย่างที่ "yes" (ซื้อคอมพิวเตอร์)

5 ตัวอย่างที่ "no" (ไม่ซื้อคอมพิวเตอร์)

คำนวณ Entropy ตามสูตร:

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

จากตาราง เราได้ค่า:

$$Info(D) = 0.940$$

คำนวณค่า Entropy สำหรับ Attribute "age"

แบ่งข้อมูลตามค่า age:

$age \leq 30$: 5 ตัวอย่าง → (2 yes, 3 no)

$31 \leq age \leq 40$: 4 ตัวอย่าง → (4 no, 0 yes)

$age > 40$: 5 ตัวอย่าง → (3 yes, 2 no)

คำนวณ Entropy ของแต่ละกลุ่ม:

$$I(2,3) = - \left(\frac{2}{5} \log_2 \frac{2}{5} + \frac{3}{5} \log_2 \frac{3}{5} \right) = 0.971$$

$$I(4,0) = - \left(\frac{4}{4} \log_2 \frac{4}{4} + \frac{0}{4} \log_2 \frac{0}{4} \right) = 0$$

$$I(3,2) = - \left(\frac{3}{5} \log_2 \frac{3}{5} + \frac{2}{5} \log_2 \frac{2}{5} \right) = 0.971$$

จากนั้น คำนวณค่า $Info_{age}(D)$:

$$\begin{aligned} Info_{age}(D) &= \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2) \\ &= \frac{5}{14} (0.971) + \frac{4}{14} (0) + \frac{5}{14} (0.971) \\ &= 0.694 \end{aligned}$$

คำนวณค่า Information Gain

ใช้สูตร:

$$Gain(Attribute) = Info(D) - Info_{Attribute}(D)$$

สำหรับ age:

$$Gain(age) = 0.940 - 0.694 = 0.246$$

เปรียบเทียบกับ Attribute อื่น

ค่าที่ได้จากการคำนวณอื่น ๆ (จากตาราง):

$$Gain(income) = 0.029$$

$$Gain(student) = 0.151$$

$$Gain(credit_rating) = 0.048$$

เนื่องจาก $Gain(age) = 0.246$ สูงสุด

ควรเลือก age เป็น Attribute แรกในการแบ่งข้อมูล

สรุป

ค่า Information Gain ของ age คือ 0.246 ซึ่งเป็นค่าที่มากที่สุด ดังนั้น ควรเลือก age เป็น Attribute แรกในการแบ่งข้อมูล