Attribute Selection: Information Gain

Select the attribute with the highest information gain.

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 Entropy ด้วยสูตร

$$Info(D) = -\sum_{i=1}^{m} p_i \, \boldsymbol{log}_2(p_i)$$

คำนวณความน่าจะเป็นของแต่ละ Features ด้วยสูตร

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

หาค่า Information Gain ด้วยกิ่งของ Features ด้วยสูตร

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

Class P: buys_computer = “yes”

Class N: buys_computer = “no”

ขั้นแรก เราจะหาค่า Entropy ของชุดข้อมูลก่อน โดยที่ข้อมูลจะมีทั้งหมด 14 เรคคอร์ด แบ่งเป็น 2 Class คือ 1.1 Class P คือ ลูกค้าซื้อคอมพิวเตอร์จากร้าน เท่ากับ “yes” มีทั้งหมด 9 เรคคอร์ด 1.2 Class N คือ ลูกค้าไม่ได้ซื้อคอมพิวเตอร์จากร้าน เท่ากับ “no” มีทั้งหมด 5 เรคคอร์ด

จาก

$$Info(D) = -\sum_{i=1}^{m} p_i \boldsymbol{log}_2(p_i)$$

จะได้

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

ต่อไปเราก็จะหา ค่า Gain โดยเริ่มจากการพิจารณาจาก Features age ก่อนเป็นลำดับแรก เราก็จะเห็นว่ามีค่าที่เกิดขึ้นด้วยกันทั้งหมด 3 ค่า คือ <=30, 31…40, >40 ถ้าเราพิจารณาช่วงอายุ <=30 เราจะเห็นว่ามี 2 เรคคอร์ด ที่เป็นลูกค้าที่ซื้อคอมพิวเตอร์จากร้าน และมี 3 เรคคอร์ด ที่เป็นลูกค้าที่ไม่ได้ซื้อคอมพิวเตอร์จากร้าน ในส่วนของช่วงอายุที่เป็น 31…40 จะมี 4 เรคคอร์ดที่อยู่ในช่วงนี้และทั้งหมดเป็นลูกค้าที่ซื้อคอมพิวเตอร์จากร้าน และในส่วนของช่วงอายุที่เป็น >40 จะมี 3 เรคคอร์ด ที่เป็นลูกค้าที่ซื้อคอมพิวเตอร์จากร้าน และมี 2 เรคคอร์ด ที่เป็นลูกค้าที่ไม่ได้ซื้อคอมพิวเตอร์จากร้าน ดังนั้นเราจึงสามารถคำนวณหาความน่าจะเป็นของ Features age ได้เป็น

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

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

$$= \frac{5}{14}\left(-\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right)$$

$$+ \frac{4}{14}\left(-\frac{4}{4}\log_2\left(\frac{4}{4}\right) - \frac{0}{4}\log_2\left(\frac{0}{4}\right)\right)$$

$$+ \frac{5}{14}\left(-\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right)$$

$$= 0.694$$

เราจึงสามารถคำนวณค่า Information Gain ของอายุ ได้ดังนี้

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

$$= 0.940 - 0.694$$

$$= 0.246$$

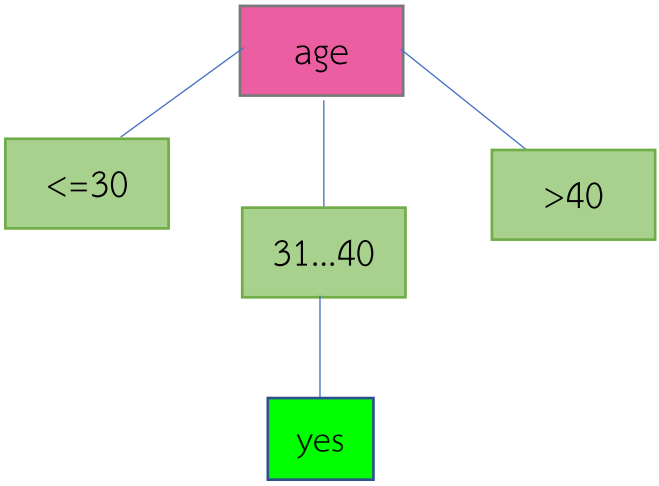
ในทำนองเดียวกัน

$$Gain(income) = 0.029$$

$$Gain(student) = 0.151$$

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

จากการคำนวณค่า Information Gain ของทุก Features พบว่าค่า Information Gain ของ Features age มีค่ามากที่สุด (0.246) ดังนั้นจึงเลือก Features age ขึ้นมาเป็น root node และข้อมูลที่อยู่ในโหนดที่มี Features age = 31...40 มีคลาสเดียวกันหมดคือ buys_computer = “yes” ดังนั้นโหนดนี้ไม่จำเป็นต้องแตกกิ่งออกไปแล้ว แต่โหนดอื่น ๆ จะต้องทำการแตกกิ่งออกไปจนข้อมูลในแต่ละโหนดมีคลาสคำตอบเดียวกันแล้ว



age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
<=30	medium	yes	excellent	yes

เราจะหาค่า Entropy ของชุดข้อมูลที่ age <=30 โดยที่ข้อมูลจะมีทั้งหมด 5 เรคคอร์ด แบ่งเป็น 2 Class คือ 1.1 Class P คือ ลูกค้าซื้อคอมพิวเตอร์จากร้าน เท่ากับ “yes” มีทั้งหมด 2 เรคคอร์ด 1.2 Class N คือ ลูกค้าไม่ได้ซื้อคอมพิวเตอร์จากร้าน เท่ากับ “no” มีทั้งหมด 3 เรคคอร์ด

$$Info(D) = I(2,3) = -\frac{2}{5}log_2(\frac{2}{5}) - \frac{3}{5}log_2(\frac{3}{5}) = 0.971$$

$$Info_{income}(D) = \frac{2}{5}I(0,2) + \frac{2}{5}I(1,1) + \frac{1}{5}I(1,0)$$

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

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

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

$$= 0 + 0.4 + 0 = 0.4$$

$$Info_{student}(D) = \frac{2}{5}I(2,0) + \frac{3}{5}I(0,3)$$

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

$$+ \frac{3}{5}\left(-\frac{0}{3}\log_2\left(\frac{0}{3}\right) - \frac{3}{3}\log_2\left(\frac{3}{3}\right)\right)$$

$$= 0$$

$$Info_{credit\_rating}(D) = \frac{3}{5}I(1,2) + \frac{2}{5}I(1,1)$$

$$= \frac{3}{5}\left(-\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right)$$

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

$$= 0.551 + 0.4 = 0.951$$

เราจึงสามารถคำนวณค่า Information Gain ได้ดังนี้

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

$$= 0.971 - 0.4$$

$$= 0.246$$

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

$$= 0.971 - 0$$

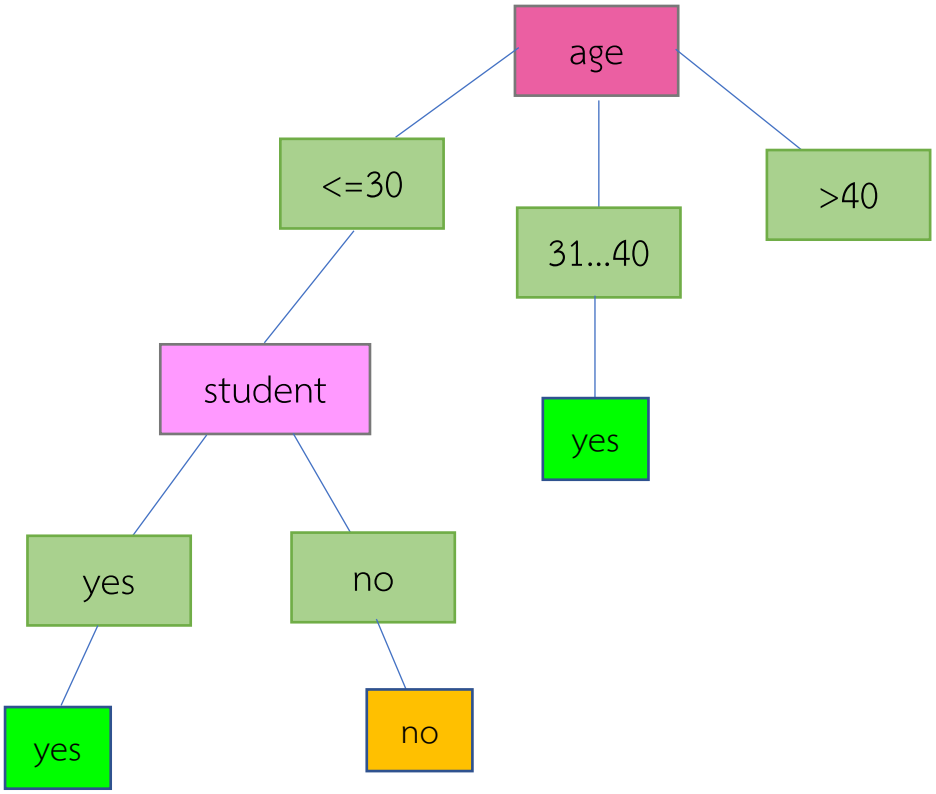
$$= 0.971$$

$$Gain(credit\_rating) = Info(D) - Info_{credit\_ratind}(D)$$

$$= 0.971 - 0.951$$

$$= 0.02$$

จากการคำนวณค่า Information Gain ของทุก Features พบว่าค่า Information Gain ของ Features student มีค่ามากที่สุด (0.971) ดังนั้นจึงเลือก Features student ขึ้นมาเป็น node และข้อมูลที่อยู่ในโหนดที่มี Features student = yes และ Features student = no มีคลาสเดียวกันหมดคือ buys_computer = “yes” และ buys_computer = “no” ตามลำดับ ดังนั้นโหนดนี้ไม่จำเป็นต้องแตกกิ่งออกไปแล้ว



age	income	student	credit_rating	buys_computer
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
>40	medium	yes	fair	yes
>40	medium	no	excellent	no

เราจะหาค่า Entropy ของชุดข้อมูลที่ age >40 โดยที่ข้อมูลจะมีทั้งหมด 5 เรคคอร์ด แบ่งเป็น 2 Class คือ 1.1 Class P คือ ลูกค้าซื้อคอมพิวเตอร์จากร้าน เท่ากับ “yes” มีทั้งหมด 3 เรคคอร์ด 1.2 Class N คือ ลูกค้าไม่ได้ซื้อคอมพิวเตอร์จากร้าน เท่ากับ “no” มีทั้งหมด 2 เรคคอร์ด

$$Info(D) = I(3,2) = -\frac{3}{5}log_2(\frac{3}{5}) - \frac{2}{5}log_2(\frac{2}{5}) = 0.971$$

$$Info_{income}(D) = \frac{0}{5}I(0,0) + \frac{3}{5}I(2,1) + \frac{2}{5}I(1,1)$$

$$= \frac{0}{5}\left(-\frac{0}{0}\log_2\left(\frac{0}{0}\right) - \frac{0}{0}\log_2\left(\frac{0}{0}\right)\right)$$

$$+ \frac{3}{5}\left(-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right)$$

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

$$= 0 + 0.551 + 0.4 = 0.951$$

$$Info_{student}(D) = \frac{3}{5}I(2,1) + \frac{2}{5}I(1,1)$$

$$= \frac{3}{5}\left(-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right)$$

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

$$= 0.551 + 0.4 = 0.951$$

$$Info_{credit\_rating}(D) = \frac{3}{5}I(3,0) + \frac{2}{5}I(0,2)$$

$$= \frac{3}{5}\left(-\frac{3}{3}\log_2\left(\frac{3}{3}\right) - \frac{0}{3}\log_2\left(\frac{0}{3}\right)\right)$$

$$+ \frac{2}{5}\left(-\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right)\right)$$

$$= 0$$

เราจึงสามารถคำนวณค่า Information Gain ได้ดังนี้

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

$$= 0.971 - 0.951$$

$$= 0.02$$

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

$$= 0.971 - 0.951$$

$$= 0.02$$

$$Gain(credit\_rating) = Info(D) - Info_{credit\_ratind}(D)$$

$$= 0.971 - 0$$

$$= 0.971$$

จากการคำนวณค่า Information Gain ของทุก Features พบว่าค่า Information Gain ของ Features credit_rating มีค่ามากที่สุด (0.971) ดังนั้นจึงเลือก Features credit_rating ขึ้นมาเป็น node และข้อมูลที่อยู่ในโหนดที่มี Features credit_rating = fair และ Features student = excellent มีคลาสเดียวกันหมด คือ buys_computer = “yes” และ buys_computer = “no” ตามลำดับ ดังนั้นโหนดนี้ไม่จำเป็นต้องแตกกิ่งออกไปแล้ว