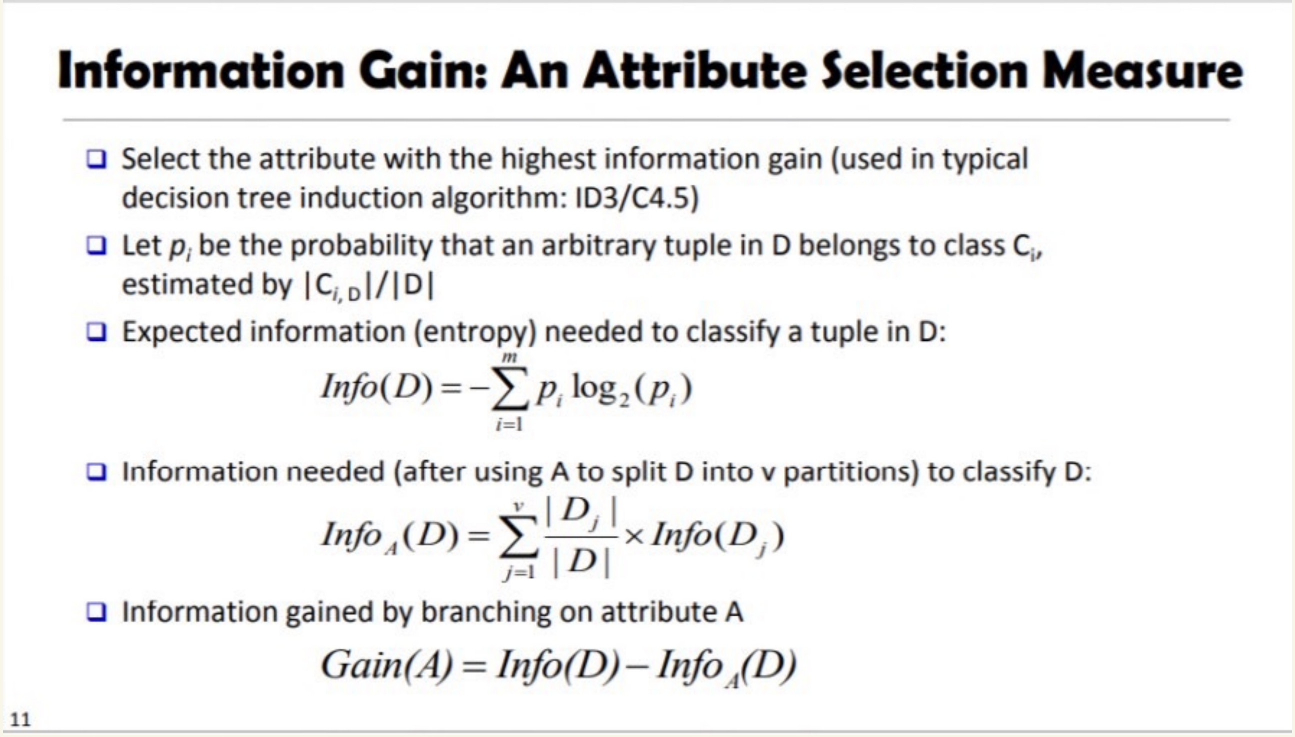
Decision Tree Induction



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:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

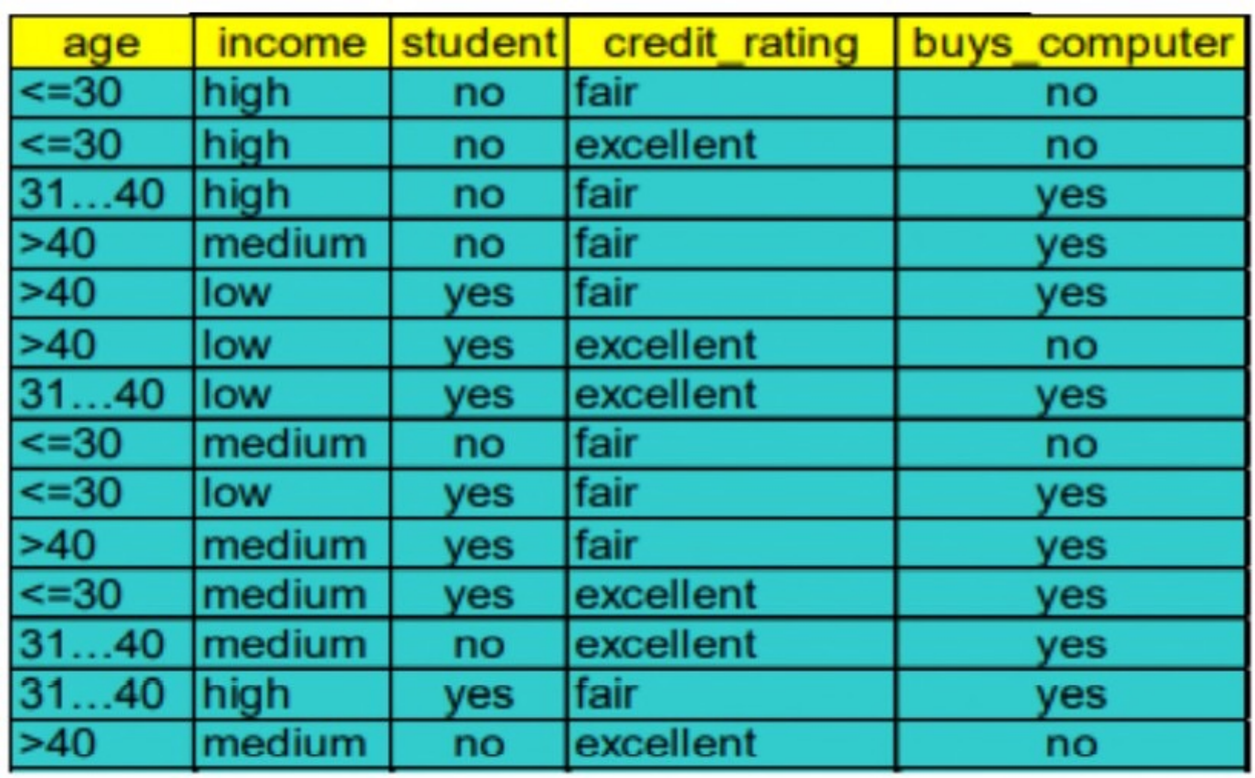
- Information needed (after using A to split D into v partitions) to classify D:

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

- Information gained by branching on attribute A

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

Exaple :



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 Info (D)

$$\text{Info}(D) = I(\overset{Y}{9}, \overset{N}{5}) = \overset{Y}{\left(-\frac{9}{14}\log_{(2)}\left(\frac{9}{14}\right)\right)} - \overset{N}{\left(\frac{5}{14}\log_{(2)}\left(\frac{5}{14}\right)\right)}$$

$$= 0.94$$

หา Info$_{age}$(D)

$$\text{Info}_{age}(D) = \overset{<=30}{\left(\frac{5}{14} I(\overset{Y}{2}, \overset{N}{3})\right)} + \overset{31-40}{\left(\frac{4}{14} I(\overset{Y}{4}, \overset{N}{0})\right)} + \overset{>40}{\left(\frac{5}{14} I(\overset{Y}{3}, \overset{N}{2})\right)}$$

$$I(\overset{Y}{2}, \overset{N}{3}) = -\frac{2}{5}\log_{(2)}\left(\frac{2}{5}\right) - \frac{3}{5}\log_{(2)}\left(\frac{3}{5}\right) = 0.971$$

$$I(\overset{Y}{4}, \overset{N}{0}) = -\frac{4}{4}\log_{(2)}\left(\frac{4}{4}\right) - \frac{0}{4}\log_{(2)}\left(\frac{0}{4}\right) = 0$$

$$I(\overset{Y}{3}, \overset{N}{2}) = -\frac{3}{5}\log_{(2)}\left(\frac{3}{5}\right) - \frac{2}{5}\log_{(2)}\left(\frac{2}{5}\right) = 0.971$$

แทนค่า $\text{Info}_{age}(D) = \frac{5}{14}(0.971) + \frac{4}{14}(0) + \frac{5}{14}(0.971) = 0.694$

หา Gain (age)

$$\text{Gain}(age) = 0.94 - 0.694 = 0.246$$

หา $\text{Info}_{\text{income}}(D)$

$$\text{Info}_{\text{income}}(D) = \boxed{\overset{\text{high}}{\frac{4}{14} I(\overset{Y}{2}, \overset{N}{2})}} + \boxed{\overset{\text{medium}}{\frac{6}{14} I(\overset{Y}{4}, \overset{N}{2})}} + \boxed{\overset{\text{low}}{\frac{4}{14} I(\overset{Y}{3}, \overset{N}{1})}}$$

$$I(\overset{Y}{2}, \overset{N}{2}) = -\frac{2}{4}\log_{(2)}\left(\frac{2}{4}\right) - \frac{2}{4}\log_{(2)}\left(\frac{2}{4}\right) = 1$$

$$I(\overset{Y}{4}, \overset{N}{2}) = -\frac{4}{6}\log_{(2)}\left(\frac{4}{6}\right) - \frac{2}{6}\log_{(2)}\left(\frac{2}{6}\right) = 0.918$$

$$I(\overset{Y}{3}, \overset{N}{1}) = -\frac{3}{4}\log_{(2)}\left(\frac{3}{4}\right) - \frac{1}{4}\log_{(2)}\left(\frac{1}{4}\right) = 0.811$$

แทนค่า $\text{Info}_{\text{income}}(D) = \frac{4}{14}(1) + \frac{6}{14}(0.918) + \frac{4}{14}(0.811) = 0.911$

หาค่า Gain (income)

$$\text{Gain (income)} = 0.94 - 0.911 = 0.029$$

หา $\text{Info}_{\text{student}}(D)$

$$\text{Info}_{\text{student}}(D) = \boxed{\overset{\text{yes}}{\frac{7}{14} I(\overset{Y}{6}, \overset{N}{1})}} + \boxed{\overset{\text{No}}{\frac{7}{14} (\overset{Y}{3}, \overset{N}{4})}}$$

$$I(\overset{Y}{6}, \overset{N}{1}) = -\frac{6}{7}\log_{(2)}\left(\frac{6}{7}\right) - \frac{1}{7}\log_{(2)}\left(\frac{1}{7}\right) = 0.592$$

$$I(\overset{Y}{3}, \overset{N}{4}) = -\frac{3}{7}\log_{(2)}\left(\frac{3}{7}\right) - \frac{4}{7}\log_{(2)}\left(\frac{4}{7}\right) = 0.985$$

แทนค่า $\text{Info}_{\text{student}}(D) = \frac{7}{14}(0.592) + \frac{7}{14}(0.985) = 0.789$

หา Gain (Student)

$$\text{Gain (Student)} = 0.94 - 0.789 = 0.151$$

หา $\text{Info}_{\text{credit\_rating}}(D)$

$$\text{Info}_{\text{credit\_rating}}(D) = \boxed{\overset{\text{fair}}{\tfrac{8}{14} I(\overset{Y}{6},\overset{N}{2})}} + \boxed{\overset{\text{excellent}}{\tfrac{6}{14} I(\overset{Y}{3},\overset{N}{3})}}$$

$$I(\overset{Y}{6},\overset{N}{2}) = -\frac{6}{8}\log_{(2)}\left(\frac{6}{8}\right) - \frac{2}{8}\log_{(2)}\left(\frac{2}{8}\right) = 0.8111$$

$$I(\overset{Y}{3},\overset{N}{3}) = -\frac{3}{6}\log_{(2)}\left(\frac{3}{6}\right) - \frac{3}{6}\log_{(2)}\left(\frac{3}{6}\right) = 1$$

แทนค่า $\text{Info}_{\text{credit\_rating}}(D) = \frac{8}{14}(0.8111) + \frac{6}{14}(1) = 0.892$

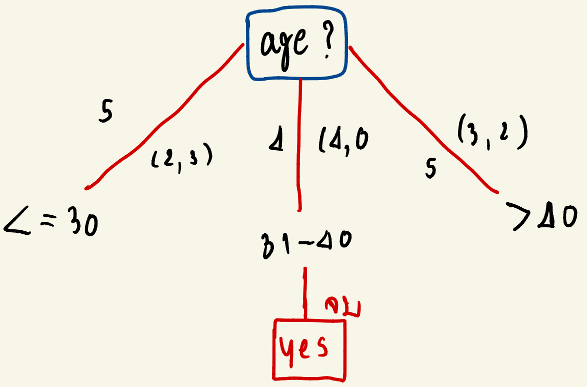
หา Gain (credit_rating)

$$\text{Gain}(\text{credit\_rating}) = 0.94 - 0.892 = 0.048$$

จาก Gain

Gain (age)	=	0.246
Gain (income)	=	0.29
Gain (student)	=	0.151
Gain (credit_rating)	=	0.048

เลือก Gain ที่มีค่ามากที่สุดมาพิจารณาเป็นส่วนแรก ซึ่งในที่นี้คือ Gain (age)



ทองบุญงาม พรโกโพธิ์ 623021051-4