Attribute selection technique

data set

Age	Income	student	CR	class Buy computer
youth	High	NO	Fair	N
Y...	H...	NO	Excellent	N
middle age	H..	NO	F	Y
Snr	medium	NO	F	Y
Snr	low	yes	F	Y
m..	low	yes	F	N
Y..	m...	yes	E	Y
Y..	low	no	F	N
S..	m..	yes	E	Y
Y..	m...	yes	E	Y
m	m...	no	E	Y
m	H..	yes	F	Y
m	m..	no	E	N
S	Low	yes	E	Y

1) Information gain / Entropy measure

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

P_i - Probability that arbitrary tuple in 'D'

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

Gain = $info(D) - info_A(D)$

of "no" in class = 5 $\rightarrow$ 4 = 5 + 9 = 14

of "yes" in class = 9

$$\text{Info}(D) = \frac{-9}{14}\log_2(9/14) - (5/14)\log_2(5/14)$$

$$= -(-0.4098) - (-0.5305)$$

$$\text{Info}(D) = 0.9403$$

say now we select "Age" as the factor to decide

$$\text{Info}_{\text{age}}(D) = 5/14\left[-2/5\log(2/5) - 3/5\log(3/5)\right] +$$

$$4/14\left[-4/4\log_2(4/4) - 0/4\log_2(0/4)\right] +$$

$$5/14\left[-3/5\log_2(3/5) - 2/5\log(2/5)\right]$$

$$= 0.694$$

$$\text{gain}(\text{age}) = 0.9403 - 0.694$$

$$= 0.2463$$

Illy

$$\text{gain(income)} = 0.029$$

$$\text{gain(student)} = 0.151$$

$$\text{gain(credit rating)} = 0.048$$

$$0.2463_{\text{(gain)}} > 0.151_{\text{(student)}} > 0.048_{\text{credit}} > 0.029_{\text{income}}$$