HW2 : Attribute Selection with Information Gain

$Info(D) = I(9,5) = -\overset{yes}{\frac{9}{14}\log_2\left(\frac{9}{14}\right)} - \overset{NO}{\frac{5}{14}\log_2\left(\frac{5}{14}\right)} = 0.940$ → Expected information (entropy) ของทั้งหมด

$Info_{age}(D) = \frac{5}{14}\overset{<=30}{I(2,3)} + \frac{4}{14}\overset{31...40}{I(4,0)} + \frac{5}{14}\overset{>40}{I(3,2)} = 0.694$ → Expected information แต่ละตัว เพื่อหา Root node

$Info_{income}(D) = \frac{4}{14}\overset{high}{I(2,2)} + \frac{4}{14}\overset{low}{I(3,1)} + \frac{6}{14}\overset{medium}{I(4,2)}$

$= \frac{4}{14}\left(-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right) + \frac{4}{14}\left(-\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right)$

$+ \frac{6}{14}\left(-\frac{4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right)\right)$

$= 0.911$

$Info_{student}(D) = \frac{7}{14}\overset{no}{I(4,3)} + \frac{7}{14}\overset{yes}{I(6,1)}$

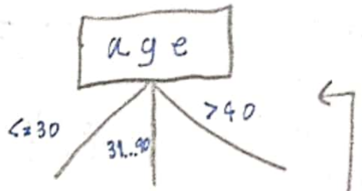
$= \frac{7}{14}\left(-\frac{4}{7}\log_2\left(\frac{4}{7}\right) - \frac{3}{7}\log_2\left(\frac{3}{7}\right)\right) + \frac{7}{14}\left(-\frac{6}{7}\log_2\left(\frac{6}{7}\right) - \frac{1}{7}\log_2\left(\frac{1}{7}\right)\right)$

$= 0.788$

$Info_{credit\_rating}(D) = \frac{8}{14}\left(-\frac{6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right)\right) + \frac{6}{14}\left(-\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)\right)$

$= \frac{8}{14}\overset{fair}{I(6,2)} + \frac{6}{14}\overset{excellent}{I(3,3)}$

$= 0.892$



$Gain(age) = Info(D) - Info_{age}(D) = 0.940 - 0.694 = 0.246$

$Gain(income) = Info(D) - Info_{income}(D) = 0.940 - 0.911 = 0.029$

$Gain(student) = Info(D) - Info_{stadent}(D) = 0.940 - 0.788 = 0.152$

$Gain(credit\_rating) = Info(D) - Info_{credit\_rating}(D) = 0.940 - 0.892 = 0.048$

สรุปดังนั้น Root node ~~[illegible]~~ คือ ค่า Gain ที่เยอะที่สุดก็คือ age มีค่า Gain = 0.246

ถ้า age: <=30

$Info_{age:<=30}(D) = I(2,3) = -\overset{Yes}{\frac{2}{5}\log_2\left(\frac{2}{5}\right)} - \overset{NO}{\frac{3}{5}\log_2\left(\frac{3}{5}\right)} = 0.971$

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

$= \frac{1}{5}\left(-\frac{1}{1}\log_2\left(\frac{1}{1}\right) - 0\right) + \frac{2}{5}\left(-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) + \frac{2}{5}\left(-\frac{2}{2}\log_2\left(\frac{2}{2}\right) - 0\right)$

$= 0.4$

$Info_{student}(D) = \frac{3}{5}\underset{NO}{(0,3)} + \frac{2}{5}\underset{yes}{(2,0)} = \frac{3}{5}\left(-\frac{3}{3}\log_2\left(\frac{3}{3}\right)\right) + \frac{2}{5}\cancel{\log_2\left(\frac{2}{2}\right)}\left(-\frac{2}{2}\log_2\left(\frac{2}{2}\right)\right)$

$= 0$

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

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

$$= 0.951$$

$$Gain(income) = Info_{age:<=30}(D) - Info_{income}(D) = 0.971 - 0.400 = 0.571$$

$$Gain(student) = Info_{age:<=30}(D) - Info_{student}(D) = 0.971 - 0 = 0.971$$

$$Gain(credit\_rating) = Info_{age:<=30}(D) - Info_{credit\_rating}(D) = 0.971 - 0.951 = 0.020$$

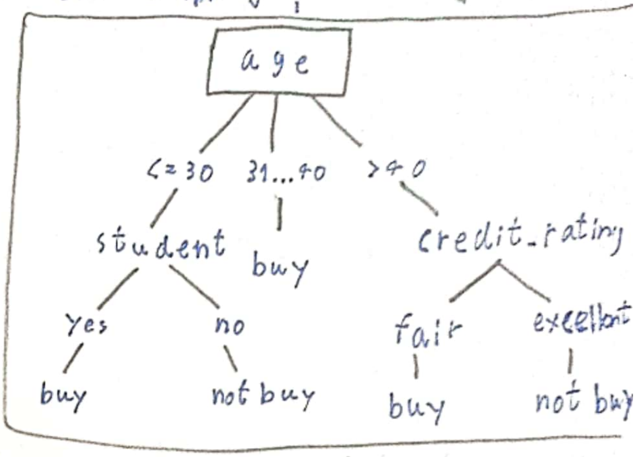
ดังนั้น decision node แรกคือ student เพราะมีค่า Gain สูงสุด

$$Info_{age:31...40}(D) = I(4,0)$$

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

$$= 0$$

ดังนั้น 31...40 จึงไม่มี decision node แต่จากสมการ I(4,0) ทำให้พบว่า age:31...40 จะซื้อทั้งหมด



$$Info_{age:>40}(D) = I(3,2) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) = 0.971$$

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

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

$$= 0.951$$

$$Info_{(student)}(D) = \frac{3}{5}I(2,1) + \frac{2}{5}I(1,1) = \frac{3}{5}\left(-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right) + \frac{2}{5}\left(-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)$$

$$= 0.951$$

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

$$= \frac{3}{5}\left(-\frac{3}{3}\log_2\left(\frac{3}{3}\right)\right) + \frac{2}{5}\left(-\frac{2}{2}\log_2\left(\frac{2}{2}\right)\right) = 0$$

$$Gain(income) = Info_{age:>40}(D) - Info_{income}(D) = 0.971 - 0.951 = 0.02$$

$$Gain(student) = Info_{age:>40}(D) - Info_{\cancel{income}\ student}(D) = 0.971 - 0.951 = 0.02$$

$$Gain(credit\_rating) = Info_{age:>40}(D) - Info_{\cancel{student}\ cr}(D) = 0.971 - 0 = 0.971$$

สรุป ดังนั้น credit_rating เป็น decision node สอง เพราะมีค่า Gain สูงสุดและ Info(D) = 0 จึงสามารถแบ่งว่าซื้อไม่ซื้อได้เลย