Hw2: Attribute Selection with Information Gain

643020513-9 ภัทรธร ก้อนผงศ์

$Info(D) = I(9,5) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) = 0.940$ → Expected information (entropy) ของทั้งหมด

$Info_{age}(D) = \frac{5}{14}I(2,3) + \frac{4}{14}I(4,0) + \frac{5}{14}I(3,2) = 0.694$ → Expected information แต่ละตัวเพื่อหา Root node

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

$= \frac{4}{14}\left(-\frac{2}{4}\log_2(\frac{2}{4}) - \frac{2}{4}\log_2(\frac{2}{4})\right) + \frac{6}{14}\left(-\frac{4}{6}\log_2(\frac{4}{6}) - \frac{2}{6}\log_2(\frac{2}{6})\right) + \frac{4}{14}\left(-\frac{3}{4}\log_2(\frac{3}{4}) - \frac{1}{4}\log_2(\frac{1}{4})\right)$

$= 0.911$

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

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

$= 0.788$

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

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

$= 0.892$

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

$Gain(income) = Info(D) - Info_{income}(D) = 0.910 - 0.911 = 0.029$

$Gain(student) = Info(D) - Info_{student}(D) = 0.910 - 0.788 = 0.152$

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

สรุป ดังนั้น Root node คือ ค่า Gain ที่เยอะที่สุดก็คือ age มีค่า Gain = 0.246

ถ้า age: <= 30

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

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

$= \frac{1}{5}(-1\log_2(1) - 0) + \frac{2}{5}\left(-\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2})\right) + \frac{2}{5}\left(-\frac{2}{2}\log_2(\frac{2}{2}) - 0\right) = 0.4$

$Info_{student}(D) = \frac{3}{5}\overset{No}{I(0,3)} + \frac{2}{5}\overset{Yes}{I(2,0)} = \frac{3}{5}\left(-\frac{3}{3}\log_2(\frac{3}{3}) - 0\right) + \frac{2}{5}\left(-\frac{2}{2}\log_2(\frac{2}{2}) - 0\right) = 0$

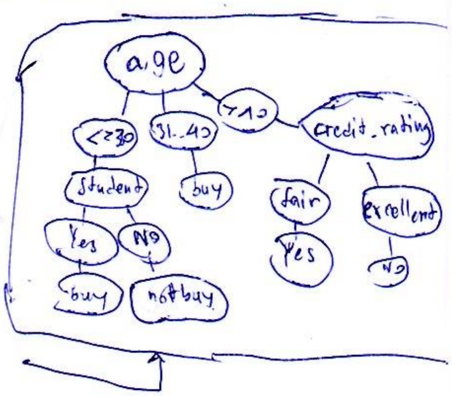
$Info_{credit\_rating}(D) = \frac{3}{5}\overset{fair}{I(1,2)} + \frac{2}{5}\overset{excellent}{(1,1)} = \frac{3}{5}\left(-\frac{1}{3}\log_2(\frac{1}{3}) - \frac{2}{3}\log_2(\frac{2}{3})\right) + \frac{2}{5}\left(-\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2})\right) = 0.951$

$Gain(income) = Info_{age:<=30}(D) - Info_{income}(D) = 0.971 - 0.4 = 0.571$

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

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

ดังนั้น Decision node แรก คือ student เพราะมีค่า Gain สูงสุด →



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

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

$= 0$

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

ถ้า >40

Info age:>40 (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) = $\frac{2}{5}I(1,1) + \frac{3}{5}I(2,1) = \frac{2}{5}(-\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2})) + \frac{3}{5}(-\frac{2}{3}\log_2(\frac{2}{3}) - \frac{1}{3}\log_2(\frac{1}{3})) = 0.951$

Info (student) (D) = $\frac{3}{5}I(2,1) + \frac{2}{5}I(1,1) = \frac{3}{5}(-\frac{2}{3}\log_2(\frac{2}{3}) - \frac{1}{3}\log_2(\frac{1}{3})) + \frac{2}{5}(-\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2})) = 0.951$

Info (credit-rating) (D) = $\frac{3}{5}I(3,0) + \frac{2}{5}I(0,2) = \frac{3}{5}(-\frac{3}{3}\log_2(\frac{3}{3})) + \frac{2}{5}(-\frac{2}{2}\log_2(\frac{2}{2})) = 0$

Gain(income) = Info age:>40(D) − Info income (D) = 0.971 − 0.951 = 0.02

Gain (student) = Info age:>40 (D) − Info student (D) = 0.971 − 0.951 = 0.02

Gain (credit-rating) = Info age:>40(D) − Info credit-rating (D) = 0.971 − 0 = 0.971

หรือนั่นคือ credit-rating เป็น decision node ของ (เพราะมีค่า Gain สูงสุดและ Info(D) = 0 นั่นหมายถึง แน่ว่าซื้อ ไม่ซื้อ ได้เลย)