

Hw.

# Decision Tree

Training data set: Who buys computer?

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

Class

X: feature

age:

	yes	no
<= 30	2	3
31... 40	4	0
> 40	3	2
income:		
high	2	2
med	4	2
Low	3	1
student:		
yes	6	1
No	3	5
credit:		
ex	3	3
fair	6	2

income:

student:

credit:

1) <sup>Yes No</sup>  $q_1$  class  $(9, 5)$  \*

$$\begin{aligned} \text{sums } \text{Info}(D) &= - \sum_{i=1}^m p_i \log_2(p_i) \\ &= I(9, 5) \\ &= -\frac{9}{14} \log_2\left(\frac{9}{14}\right) - \frac{5}{14} \log_2\left(\frac{5}{14}\right) \\ &= 0.41 + 0.53 \end{aligned}$$

$$\therefore \text{Info}(D) = 0.94$$

2)  $q_1$  Feature \*

$$\begin{aligned} \text{Info}_{age}(D) &= \frac{5}{14} I(2, 3) + \frac{4}{14} I(4, 0) + \frac{5}{14} I(3, 2) \\ &= \frac{5}{14} \left[ -\frac{2}{5} \log_2\left(\frac{2}{5}\right) - \frac{3}{5} \log_2\left(\frac{3}{5}\right) \right] + \frac{4}{14} \left[ -\frac{4}{4} \log_2\left(\frac{4}{4}\right) - \frac{0}{4} \log_2\left(\frac{0}{4}\right) \right] + \frac{5}{14} \left[ -\frac{3}{5} \log_2\left(\frac{3}{5}\right) - \frac{2}{5} \log_2\left(\frac{2}{5}\right) \right] \\ &= \frac{5}{14} [0.5287 + 0.44217] + \frac{5}{14} [0.44217 + 0.52877] \end{aligned}$$

$$\text{Info}_{age}(D) = 0.694 \quad (\text{age})$$

$$\begin{aligned} \text{Info}_{income}(D) &= \frac{4}{14} I(2, 2) + \frac{6}{14} I(4, 2) + \frac{4}{14} I(3, 1) \\ &= \frac{4}{14} \left[ -\frac{2}{4} \log_2\left(\frac{2}{4}\right) - \frac{2}{4} \log_2\left(\frac{2}{4}\right) \right] + \frac{6}{14} \left[ -\frac{4}{6} \log_2\left(\frac{4}{6}\right) - \frac{2}{6} \log_2\left(\frac{2}{6}\right) \right] + \frac{4}{14} \left[ -\frac{3}{4} \log_2\left(\frac{3}{4}\right) - \frac{1}{4} \log_2\left(\frac{1}{4}\right) \right] \end{aligned}$$

$$\therefore \text{Info}_{income}(D) = 0.911 \quad (\text{income})$$

$$\begin{aligned} \text{Info}_{student}(D) &= \frac{7}{14} I(6, 1) + \frac{7}{14} I(3, 4) \\ &= \frac{7}{14} \left[ -\frac{6}{7} \log_2\left(\frac{6}{7}\right) - \frac{1}{7} \log_2\left(\frac{1}{7}\right) \right] + \frac{7}{14} \left[ -\frac{3}{7} \log_2\left(\frac{3}{7}\right) - \frac{4}{7} \log_2\left(\frac{4}{7}\right) \right] \end{aligned}$$

$$\therefore \text{Info}_{student}(D) = 0.7883 \quad (\text{student})$$

$$\begin{aligned} \text{Info}_{\text{credit}}(D) &= \frac{6}{14} I(3,3) + \frac{8}{14} I(6,2) \\ &= \frac{6}{14} \left[ -\frac{3}{6} \log_2 \left( \frac{3}{6} \right) - \frac{3}{6} \log_2 \left( \frac{3}{6} \right) \right] + \frac{8}{14} \left[ -\frac{6}{8} \log_2 \left( \frac{6}{8} \right) - \frac{2}{8} \log_2 \left( \frac{2}{8} \right) \right] \end{aligned}$$

$$\text{Info}_{\text{credit}}(D) = 0.892 \quad (\text{credit})$$

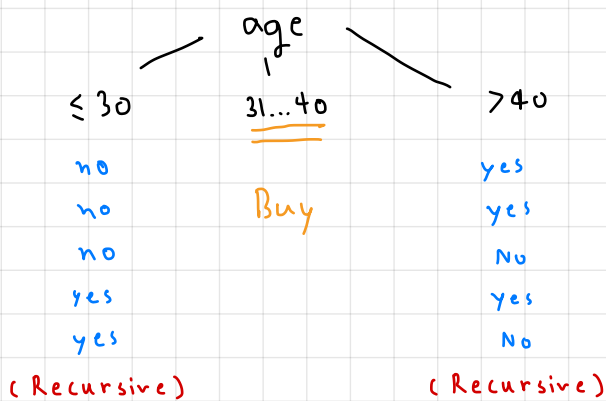
$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D)$$

$$\text{Gain}(\text{age}) = 0.94 - 0.694 = 0.246 \quad (\text{age is chosen as root node})$$

$$\text{Gain}(\text{income}) = 0.94 - 0.9111 = 0.029$$

$$\text{Gain}(\text{student}) = 0.94 - 0.788 = 0.152$$

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



Training data set: Who buys computer?

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

$$\text{Recursive } \leq 30 \quad (\text{yes, no})$$

$$1) \text{ or } \text{Class} * \leq 30$$

$$\text{Info}(D) = I(2,3)$$

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

$$\text{Info}(D) = 0.971$$

$$2.) \text{ or } \text{feature} *$$

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

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

$$\therefore \text{Info}_{\text{income}}(D) = 0.4 \quad (\text{income})$$

$$\begin{aligned} \text{Info (D)}_{\text{student}} &= \frac{2}{5} I(2,0) + \frac{3}{5} I(0,3) \\ &= \frac{2}{5} \left[ -\frac{2}{2} \log_2 \left( \frac{2}{2} \right) - \frac{0}{2} \log_2 \left( \frac{0}{2} \right) \right] + \frac{3}{5} \left[ -\frac{0}{3} \log_2 \left( \frac{0}{3} \right) - \frac{3}{3} \log_2 \left( \frac{3}{3} \right) \right] \end{aligned}$$

$$\text{Info (D)}_{\text{student}} = 0 \quad (\text{student})$$

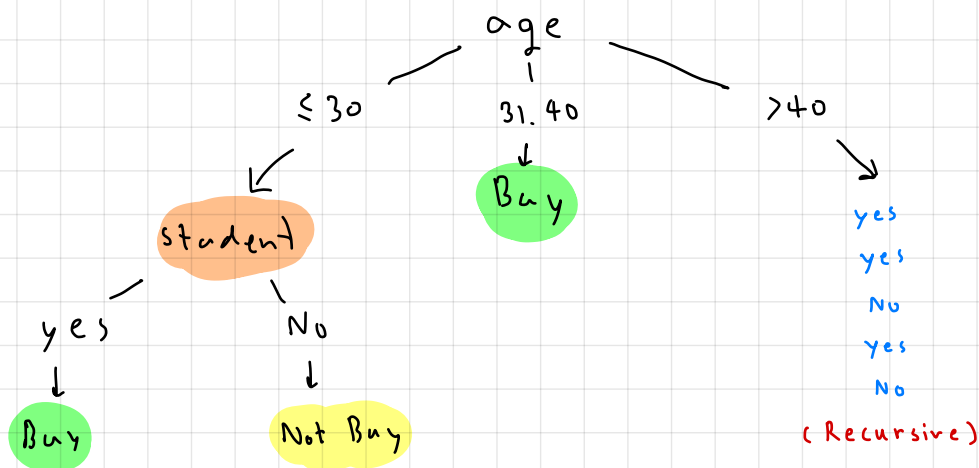
$$\begin{aligned} \text{Info (D)}_{\text{credit}} &= \frac{2}{5} I(1,1) + \frac{3}{5} I(1,2) \\ &= \frac{2}{5} \left[ -\frac{1}{2} \log_2 \left( \frac{1}{2} \right) - \frac{1}{2} \log_2 \left( \frac{1}{2} \right) \right] + \frac{3}{5} \left[ -\frac{1}{3} \log_2 \left( \frac{1}{3} \right) - \frac{2}{3} \log_2 \left( \frac{2}{3} \right) \right] \end{aligned}$$

$$\text{Info (D)}_{\text{credit}} = 0.9509 \quad (\text{credit})$$

$$\text{Gain (income)} = 0.971 - 0.4 = 0.571$$

$$\text{Gain (student)} = 0.971 - 0 = 0.971 \quad (\text{Gain student}) \text{ is } 1$$

$$\text{Gain (credit-rating)} = 0.971 - 0.9509 = 0.0201$$



Recursive > 40

1) 47 Class \*  $\begin{matrix} \text{yes} & \text{No} \\ (3, 2) \end{matrix}$

$$\text{Info (D)} = I(3,2)$$

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

$$\therefore \text{Info (D)} = 0.971$$

an feature

$$\begin{aligned}\text{Info (D)}_{\text{income}} &= \frac{3}{5} I(2,1) + \frac{2}{5} I(1,1) \\ &= \frac{3}{5} \left[ -\frac{2}{3} \log_2 \left( \frac{2}{3} \right) - \frac{1}{3} \log_2 \left( \frac{1}{3} \right) \right]\end{aligned}$$

$$\text{Info (D)}_{\text{income}} = 0.9509$$

$$\begin{aligned}\text{Info (D)}_{\text{student}} &= \frac{3}{5} I(1,1) + \frac{2}{5} I(1,1) \\ &= \frac{3}{5} \left[ -\frac{2}{3} \log_2 \left( \frac{2}{3} \right) - \frac{1}{3} \log_2 \left( \frac{1}{3} \right) \right] + \frac{2}{5} \left[ -\frac{1}{2} \log_2 \left( \frac{1}{2} \right) - \frac{1}{2} \log_2 \left( \frac{1}{2} \right) \right]\end{aligned}$$

$$\text{Info (D)}_{\text{student}} = 0.9509$$

$$\begin{aligned}\text{Info (D)}_{\text{credit}} &= \frac{2}{5} I(0,2) + \frac{3}{5} I(3,0) \\ &= \frac{2}{5} \left[ -\frac{0}{2} \log_2 \left( \frac{0}{2} \right) - \frac{2}{2} \log_2 \left( \frac{2}{2} \right) \right] + \frac{3}{5} \left[ -\frac{3}{3} \log_2 \left( \frac{3}{3} \right) - \frac{0}{3} \log_2 \left( \frac{0}{3} \right) \right]\end{aligned}$$

$$\text{Info (D)}_{\text{credit}} = 0$$

$$\text{Gain}(\text{income}) = 0.971 - 0.9509 = 0.0201$$

$$\text{Gain}(\text{student}) = 0.9710 - 0.9509 = 0.0201$$

$$\text{Gain}(\text{credit\_rating}) = 0.971 - 0 = 0.971$$

