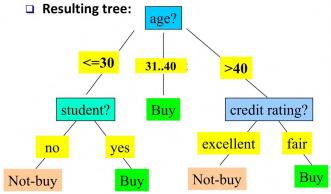
Decision Tree Induction: An Example

- **Decision tree construction**:
  - A top-down, recursive, divide-and-conquer process
- **Resulting 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

Note: The data set is adapted from "Playing Tennis" example of R. Quinlan

9

Class P : buys_computer = "yes" yes : 9

Class N : buys_computer = "no" no : 5

$$Info(D) = I(9,5) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.940$$

age

<=30 yes : 2, no : 3

31-40 yes : 4, no : 0

>40 yes : 3, no : 2

$$Info_{age}(D) = \frac{5}{14}I(2,3) + \frac{4}{14}I(4,0) + \frac{5}{14}I(3,2) = 0.694$$

income

high yes : 2, no : 2

medium yes : 4, no : 2

low yes : 3, no : 1

$$Info_{income}(D) = \frac{4}{14}I(2,2) + \frac{6}{14}I(4,2) + \frac{4}{14}I(3,1)$$
$$= \frac{4}{14}\left[-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right] + \frac{6}{14}\left[-\frac{4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right)\right] + \frac{4}{14}\left[-\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right] = 0.911$$

student

yes yes : 2, no : 2

no yes : 4, no : 2

$$Info_{student}(D) = \frac{7}{14}I(6,1) + \frac{7}{14}I(3,4)$$
$$= \frac{7}{14}\left[-\frac{6}{7}\log_2\left(\frac{6}{7}\right) - \frac{1}{7}\log_2\left(\frac{1}{7}\right)\right] + \frac{7}{14}\left[-\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right)\right] = 0.789$$

credit

fair yes : 2, no : 2

excellent yes : 4, no : 2

$$Info_{credit}(D) = \frac{8}{14}I(6,2) + \frac{6}{14}I(3,3)$$
$$= \frac{8}{14}\left[-\frac{6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right)\right] + \frac{6}{14}\left[-\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)\right] = 0.892$$

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

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

Gain(student) = Info(D) - $Info_{student}(D)$ = 0.940 - 0.789 = 0.151

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

F1

age	income	student	credit_rating	buy_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
<=30	medium	yes	excellent	yes

Class P : buys_computer = "yes" yes : 2 $Info(D) = I(2,3) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right) = 0.971$

Class N : buys_computer = "no" no : 3

income

high yes : 0, no : 2 $Info_{income}(D) = \frac{2}{5}I(0,2) + \frac{2}{5}I(1,1) + \frac{1}{5}I(1,0)$

medium yes : 1, no : 1 $= \frac{2}{5}\left[-\frac{2}{2}\log_2\left(\frac{2}{2}\right)\right] + \frac{2}{5}\left[-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right] + \frac{1}{5}\left[-\frac{1}{1}\log_2\left(\frac{1}{1}\right)\right] = 0.4$

low yes : 1, no : 0

student

yes yes : 2, no : 0 $Info_{student}(D) = \frac{2}{5}I(2,0) + \frac{3}{5}I(0,3)$

no yes : 0, no : 3 $= \frac{2}{5}\left[-\frac{2}{2}\log_2\left(\frac{2}{2}\right)\right] + \frac{3}{5}\left[-\frac{3}{3}\log_2\left(\frac{3}{3}\right)\right] = 0$

credit

fair yes : 2, no : 2 $Info_{credit}(D) = \frac{3}{5}I(1,2) + \frac{2}{5}I(1,1)$

excellent yes : 4, no : 2 $= \frac{3}{5}\left[-\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right] + \frac{2}{5}\left[-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right] = 0.951$

$Gain(income) = Info(D) - Info_{income}(D) = 0.971 - 0.4 = 0.571$

$Gain(student) = Info(D) - Info_{student}(D) = 0.971 - 0.0 = 0.971$ # จะได้ว่าค่า Gain(student) มากที่สุด

$Gain(credit\_rating) = Info(D) - Info_{credit}(D) = 0.971 - 0.951 = 0.02$

age	income	student	credit_rating	buy_computer
31-40	high	no	fair	yes
31-40	low	yes	excellent	yes
31-40	medium	no	excellent	yes
31-40	high	yes	fair	yes

Class P : buys_computer = "yes" yes : 4

Class N : buys_computer = "no" no : 0

income

high yes : 2, no : 0

medium yes : 1, no : 0

low yes : 1, no : 0

student

yes yes : 2, no : 0

no yes : 2, no : 0

credit

fair yes : 2, no : 0

excellent yes : 2, no : 0

age	income	student	credit_rating	buy_computer
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
>40	medium	yes	fair	yes
>40	medium	no	excellent	no

Class P : buys_computer = "yes" yes : 3 $Info(D) = I(3,2) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) = 0.971$

Class N : buys_computer = "no" no : 2

income

high yes : 0, no : 0 $Info_{income}(D) = \frac{3}{5} I(2,1) + \frac{2}{5} I(1,1)$

medium yes : 2, no : 1 $= \frac{3}{5}\left[-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right] + \frac{2}{5}\left[-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right] = 0.951$

low yes : 1, no : 1

student

yes yes : 2, no : 1 $Info_{student}(D) = \frac{3}{5} I(2,1) + \frac{2}{5} I(1,1)$

no yes : 1, no : 1 $= \frac{3}{5}\left[-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right] + \frac{2}{5}\left[-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right] = 0.951$

credit

fair yes : 3, no : 0 $Info_{credit}(D) = \frac{3}{5} I(3,0) + \frac{2}{5} I(0,2)$

excellent yes : 0, no : 2 $= \frac{3}{5}\left[-\frac{3}{3}\log_2\left(\frac{3}{3}\right)\right] + \frac{2}{5}\left[-\frac{2}{2}\log_2\left(\frac{2}{2}\right)\right] = 0$

$Gain(income) = Info(D) - Info_{income}(D) = 0.971 - 0.951 = 0.02$

$Gain(student) = Info(D) - Info_{student}(D) = 0.971 - 0.951 = 0.02$

$Gain(credit\_rating) = Info(D) - Info_{credit}(D) = 0.971 - 0 = 0.971$ # จะได้ว่าค่า Gain(credit_rating) มากที่สุด

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

age	...	buy_computer
<=30		no
<=30		no
<=30		no
<=30		yes
<=30		yes

age	...	buy_computer
31-40		yes
31-40		yes
31-40		yes
31-40		yes

buy

age	...	buy_computer
>40		yes
>40		yes
>40		no
>40		yes
>40		no

Gain(student) = 0.971

Gain(credit_rating) = 0.971

student	...	buy_computer
no		no
no		no
no		no
yes		yes
yes		yes

credit	...	buy_computer
fair		yes
fair		yes
excellent		no
fair		yes
excellent		no

yes — buy

no — not buy

fair — buy

excellent — not buy