Week 3: Decision Trees (1)

Step 1: Entropy for the Class

- Class label: {yes, no}
- 9 instances of yes and 5 instances of no

$$Ent(D) = -\frac{9}{14}log_2\frac{9}{14} - \frac{5}{14}log_2\frac{5}{14}$$

$$= -0.642(-0.639) - 0.357(-1.486)$$

$$= 0.410 + 0.531$$

$$= 0.941\ bits$$

Step 2: Entropy for the Attributes

Age:

- Youth: 2 yes; 3 no
- Middle_aged: 4 yes; 0 no
- Senior: 3 yes; 2 no2

$$Ent_{(age)} = \frac{5}{14}\left(-\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5}\right) + \frac{4}{14}\left(-\frac{4}{4}log_2\frac{4}{4} - \frac{0}{4}log_2\frac{0}{4}\right) + \frac{5}{14}\left(-\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5}\right)$$

$$= 0.357\ (-0.4(-1.322) - 0.6(-0.737)) + 0.286(0) + 0.357(-0.6(-0.737) - 0.4(-1.322))$$

$$= 0.357\ (0.971) + 0 + 0.357(0.971) = 0.694\ bits$$

Income:

- Low: 3 yes; 1 no
- Medium: 4 yes; 2 no
- High: 2 yes, 2 no

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

$$= 0.285\big(-0.75(-0.415) - 0.25(-2)\big) + 0.429\big(-0.667(-0.584) - 0.333(-1.586)\big) + 0.285(-0.5(-1) \\ - 0.5(-1))$$

$$= 0.285(0.81125) +\ 0.429(0.9177) + 0.285(1) = 0.9099\ bits$$

Student:

- Yes: 6 yes; 1 no
- No: 3 yes; 4 no

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

$$= 0.5\big(-0.857(-0.223) - 0.143(-2.806)\big) + 0.5(-0.429(-1.221) - 0.571(-0.808))$$

$$= 0.5(0.5924) + 0.5(0.985) = 0.7887\ bits$$

Credit_Rating:

- Fair: 6 yes; 2 no
- Excellent: 3 yes, 3 no

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

$$= 0.571\big(-0.75(-0.415)-0.25(-2)\big)+0.429(-0.5(-1)-0.5(-1))$$

$$= 0.571(0.8112)+0.429(1) = 0.892\ bits$$

Step 3: Calculate Gain

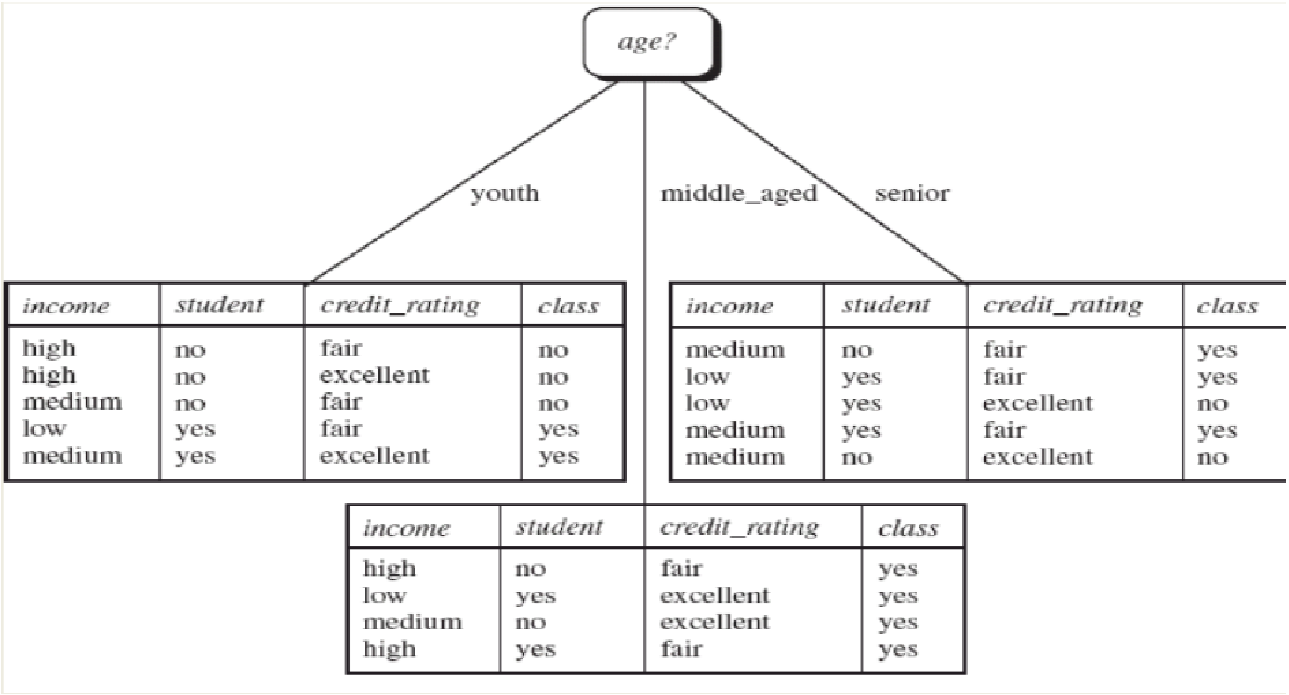
$Gain(Age) = 0.940 - 0.694 = 0.246\ bits$

$Gain(income) = 0.940 - 0.9099 = 0.030\ bits$

$Gain(student) = 0.940 - 0.7887 = 0.151\ bits$

$Gain(credit\_rating) = 0.940 - 0.892 = 0.048\ bits$

Because *age* has the highest information gain among the attributes, it is selected as the splitting attribute



income	*student*	*credit_rating*	*class*
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes

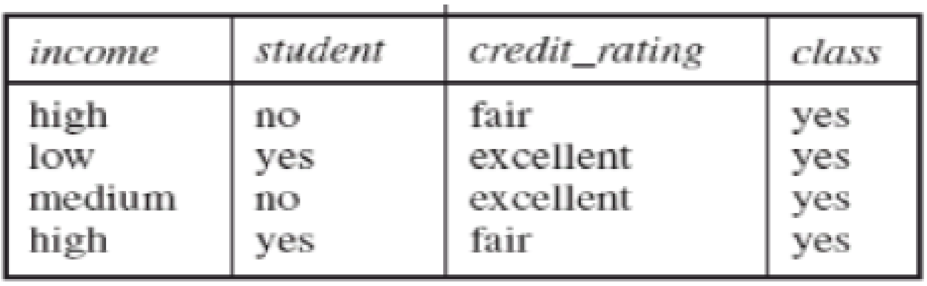
income	*student*	*credit_rating*	*class*
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

income	*student*	*credit_rating*	*class*
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

Let's now consider the 2nd level attribute selection:

Middle_aged branch:

If we examine the table below, we can see that there are no possible splitting. Thus, this will be the final node.



income	*student*	*credit_rating*	*class*
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

Youth branch:

Step 1: Entropy for the Class

- Class label: {yes, no}
- 2 instances of yes and 3 instances of no

$$Ent(D) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5}$$

$$= -0.4(-1.322) - 0.6(-0.737) = 0.971\ bits$$

Step 2: Entropy for the Attributes

Income:

- Low: 1 yes; 0 no
- Medium: 1 yes; 1no
- High: 0 yes, 2 no

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

$$= 0.2\big(-1(0) - 0(-\infty)\big) + 0.4\big(-0.5(-1) - 0.5(-1)\big) + 0.4(-0(-\infty) - 1(0))$$

$$= 0 + 0.4(1) + 0 = 0.40\ bits$$

Student:

- Yes: 2 yes; 0 no
- No: 0 yes; 3 no

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

$$= 0.4\big(-1(0) - 0(-\infty)\big) + 0.6\big(-0(-\infty) - 1(0)\big) = 0\ bits$$

Credit_rating

- Fair: 1 yes; 2 no
- Excellent: 1 yes, 1 no

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

$$= 0.571(-0.75(-0.415) - 0.25(-2)) + 0.429(-0.5(-1) - 0.5(-1))$$

$$= 0.571(0.8113) + 0.429(1) = 0.892 \text{ bits}$$

Step 3: Calculate Gain

$Gain(income) = 0.971 - 0.4 = 0.571\ bits$

$Gain(student) = 0.971$ – 0 = 0.971 bits

$Gain(credit\_rating) = 0.971 - 0.892 = 0.079\ bits$

Because *student* has the highest information gain among the attributes, it is selected as the splitting attribute

Senior branch:

Step 1: Entropy for the Class

- Class label: {yes, no}
- 3 instances of yes and 2 instances of no

$$Ent(D) = -\frac{3}{5}log_2\frac{3}{5}-\frac{2}{5}log_2\frac{2}{5}$$

$$= -0.6(-0.737) - 0.4(-1.322) = 0.971\ bits$$

Step 2: Entropy for the Attributes

Income:

- Low: 1 yes; 0 no
- Medium: 2 yes; 1no
- High: 0 yes, 0 no

$$Ent_{(income)} = \frac{1}{5}\left(-\frac{1}{1}log_2\frac{1}{1}-\frac{0}{1}log_2\frac{0}{1}\right)+\frac{3}{5}\left(-\frac{2}{3}log_2\frac{2}{3}-\frac{1}{3}log_2\frac{1}{3}\right)$$

$$= 0.2\big(-1(0) - 0(-\infty)\big) + 0.6\big(-0.667(-0.584) - 0.333(-1.586)\big)$$

$$= 0 + 0.6(0.917) = 0.551\ bits$$

Student:

- Yes: 2 yes; 1 no
- No: 1 yes; 1 no

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

$$= 0.6\big(-0.667(-0.584) - 0.333(-1.586)\big) + 0.4(-0.5(-1) - 0.5(-1))$$

$$= 0.6(0.918) + 0.4(1) = 0.951\ bits$$

Credit_rating

- Fair: 3 yes; 0 no
- Excellent: 0 yes, 2 no

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

$$= 0.6\big(-1(0) - 0(-\infty)\big) + 0.4\big(-0(-\infty) - 1(0)\big) = 0\ bits$$

Step 3: Calculate Gain

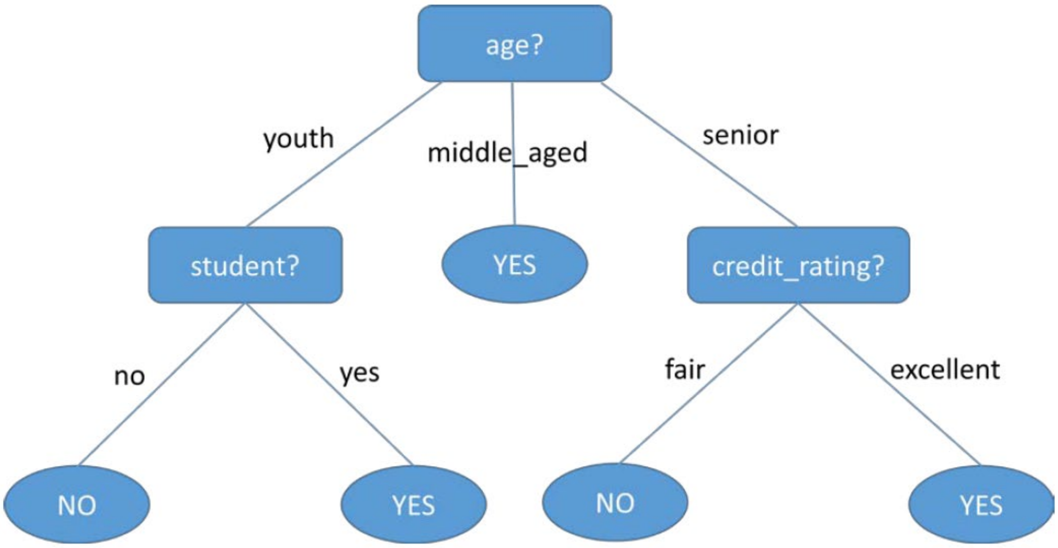
$Gain(income) = 0.971 - 0.551 = 0.42\ bits$

$Gain(student) = 0.971 - 0.951 = 0.02\ bits$

$Gain(credit\_rating) = 0.971 - 0 = 0.971\ bits$

Because *credit_rating* has the highest information gain among the attributes, it is selected as the splitting attribute

The final Decision Tree:


age?
youth
middle_aged
senior
student?
YES
credit_rating?
no
yes
fair
excellent
NO
YES
NO
YES