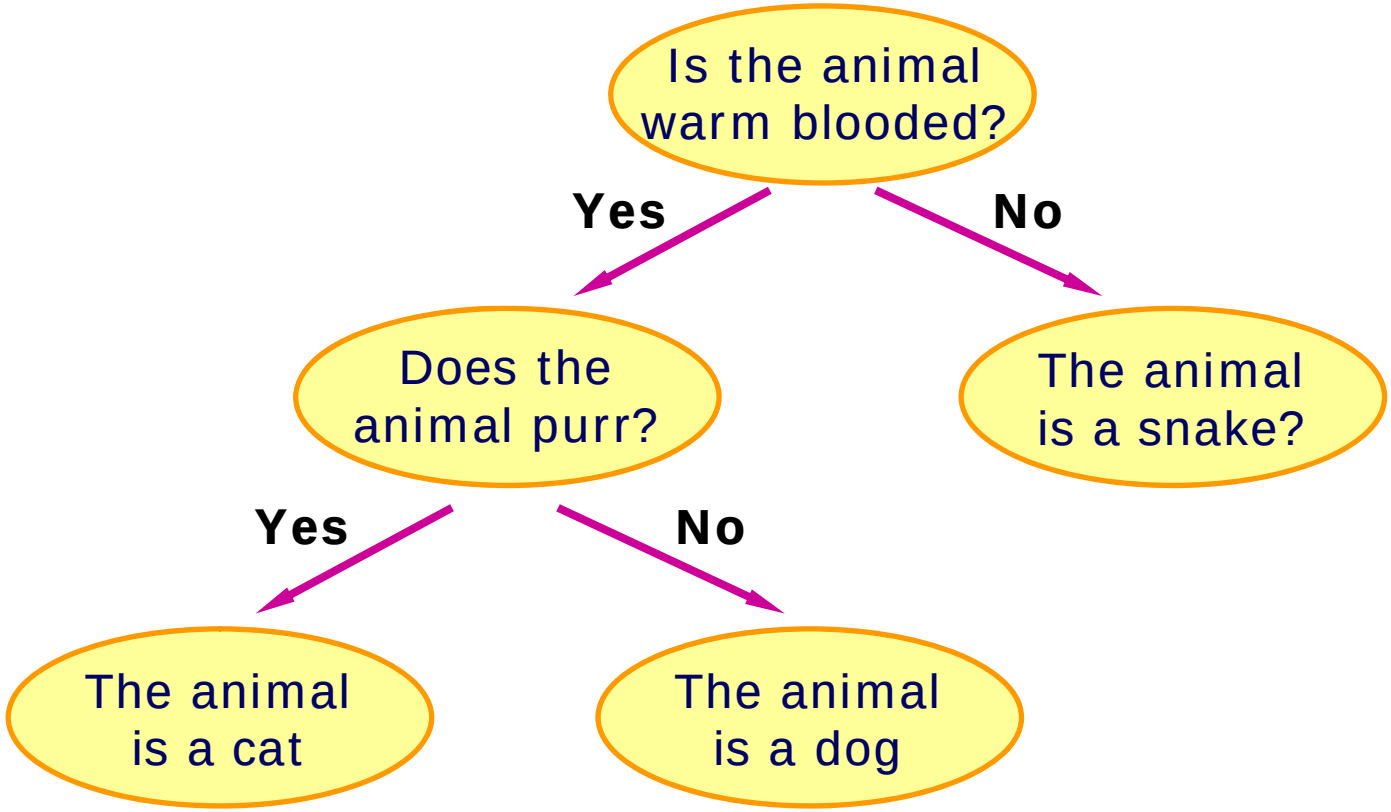
Intelligent Systems Principles and Programming

Xiaoqing Zheng

zhengxq@fudan.edu.cn

Decision tree

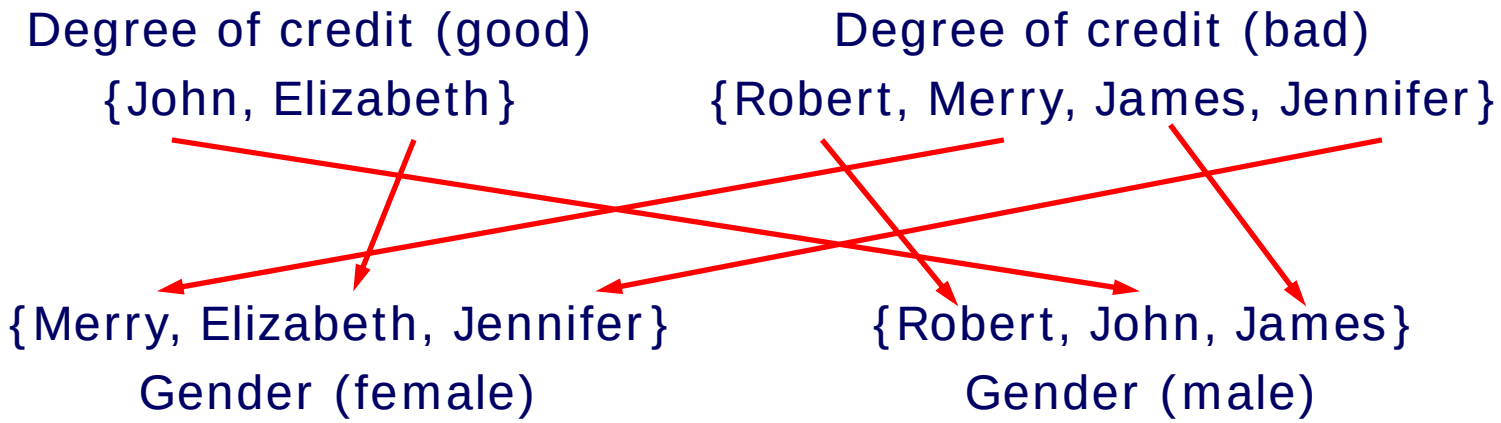


Decision tree representation

- Each internal node tests an attribute
- Each branch corresponds to attribute value
- Each leaf node assigns a classification

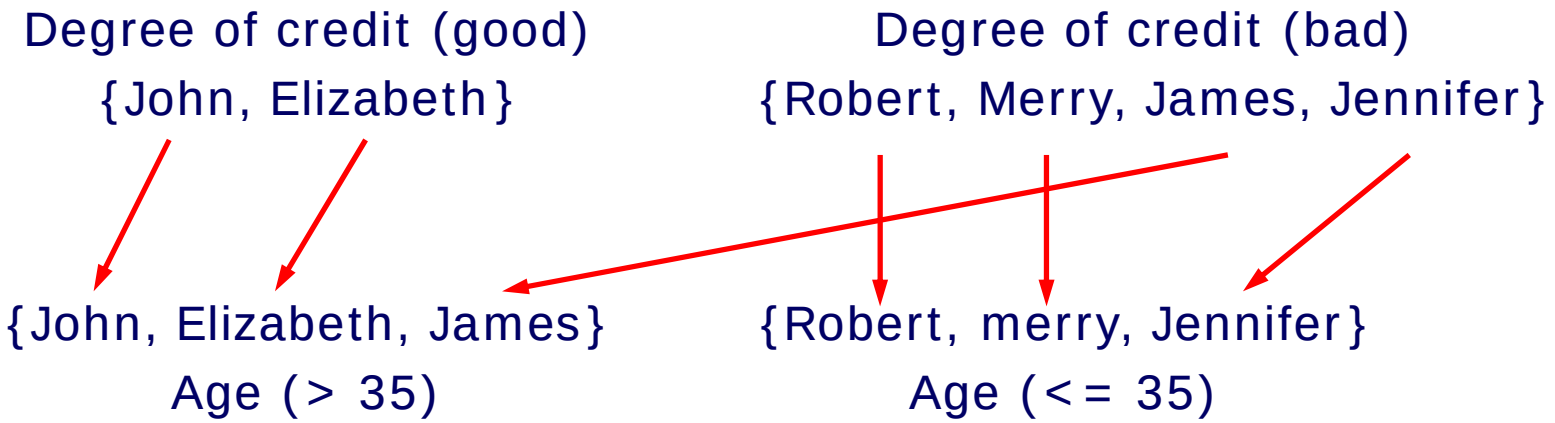
Decision trees

Name	Income	Gender	Age	History record	Credit
Robert	10000	Male	25	Bad	Bad
John	8000	Male	50	Normal	Good
Merry	6000	Female	26	Normal	Bad
Elizabeth	12000	Female	40	Good	Good
James	50000	Male	36	Normal	Bad
Jennifer	3000	Female	18	Good	Bad



Decision trees

Name	Income	Gender	Age	History record	Credit
Robert	10000	Male	25	Bad	Bad
John	8000	Male	50	Normal	Good
Merry	6000	Female	26	Normal	Bad
Elizabeth	12000	Female	40	Good	Good
James	50000	Male	36	Normal	Bad
Jennifer	3000	Female	18	Good	Bad



Discretization

Name	Income	Gender	Age	History record	Credit
Robert	High	Male	Young	Bad	Bad
John	Medium	Male	Elder	Normal	Good
Merry	Medium	Female	Young	Normal	Bad
Elizabeth	High	Female	Elder	Good	Good
James	High	Male	Elder	Normal	Bad
Jennifer	Low	Female	Young	Good	Bad

Income

High >= 10000
Medium >= 5000
Low < 5000

Age

Young <= 35
Elder > 35

Information entropy

s the number of training examples.

m the number of class C_i, $i = \{1, 2, ..., m\}$.

p_i the proportion of the examples in C_i.

$$I(s_1, s_2, \cdots, s_m) = -\sum_{i=1}^{m} p_i \log(p_i)$$

Information Theory

$h(x, y) = h(x) + h(y)$ if two events x and y are unrelated

$p(x, y) = p(x)p(y)$

$h(x) = -\log_2 p(x)$

Now suppose that a sender wishes to transmit the value of a random variable to a receiver. The average amount of information that they transmit in the process is obtained by taking the expectation of $h(x)$ with respect to the distribution $p(x)$ and is given by

$$\mathrm{H}[x] = -\sum_x p(x) \log_2 p(x)$$

Entropy

Consider a random variable x having 8 possible states, each of which is equally likely. In order to communicate the value of x to a receiver, we would need to transmit a message of length 3 bits.

$$\mathrm{H}[x] = -8 \times \frac{1}{8} \log_2 \frac{1}{8} = 3 \text{ bits}$$

Now consider an example (Cover and Thomas, 1991) of a variable having 8 possible states $\{a, b, c, d, e, f, g, h\}$ for which the respective probabilities are given by $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64})$. The entropy in this case is given by

$$\mathrm{H}[x] = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{8} \log_2 \frac{1}{8} - \frac{1}{16} \log_2 \frac{1}{16} - \frac{4}{64} \log_2 \frac{1}{64} = 2 \text{ bits}.$$

Entropy

we can take advantage of the ***nonuniform distribution*** by using shorter codes for the more probable events, at the expense of longer codes for the less probable events, in the hope of getting a shorter average code length. This can be done by representing the states $\{a, b, c, d, e, f, g, h\}$ using, for instance, the following set of code strings: 0, 10, 110, 1110, 111100, 111101, 111110, 111111. The average length of the code that has to be transmitted is then

$$\text{average code length} = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{16} \times 4 + 4 \times \frac{1}{64} \times 6 = 2 \text{ bits}$$

Information entropy

A Any attribute of the examples that have v different values, $\{a_1, a_2, ..., a_v\}$.

s_{ij} the number of the examples that belong to class C_i and whose value of the attribute A is j.

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + s_{2j} + \cdots + s_{mj}}{s} I(s_{1j}, \cdots, s_{mj})$$

Information gain

Gain(*A*) expected reduction in entropy due to sorting on *A*.

$$Gain(A) = I(s_1, s_2, \cdots, s_m) - E(A)$$

Question:

Which attribute should be tested first?

Example

Credit	Times	Percentage	log(P_i)
Bad	4	0.66667	– 0.58496
Good	2	0.33333	– 1.58496
Total	6	1.00000	

$$I(Credit) = -4/6 \times log_2(4/6) - 2/6 \times log_2(2/6)$$

$$= -0.66667 \times -0.58496$$
$$-0.33333 \times -1.58496$$

$$= \mathbf{0.9183}$$

Example

Gender	Bad	Good	Total
Male	2	1	3
Female	2	1	3
Total	4	2	6

$$I(Gender/Male) = -2/3 \times log_2(2/3) - 1/3 \times log_2(1/3) = 0.9183$$

$$I(Gender/Female) = -2/3 \times log_2(2/3) - 1/3 \times log_2(1/3) = 0.9183$$

$$E(Gender) = 3/6 \times I(Gender/Male) + 3/6 \times I(Gender/Female)$$

$$= 0.5 \times 0.9183 + 0.5 \times 0.9183 = \mathbf{0.9183}$$

$$Gain(Gender) = I(Credit) - E(Gender)$$

$$= 0.9183 - 0.9183$$

$$= \mathbf{0}$$

Example

Age	Bad	Good	Total
Young	3	0	3
Elder	1	2	3
Total	4	2	6

$$I(Age/Young) = -3/3 \times log_2(3/3) - 0/3 \times log_2(0/3) = 0.0000$$

$$I(Age/Elder) = -1/3 \times log_2(1/3) - 2/3 \times log_2(2/3) = 0.9183$$

$$E(Age) = 3/6 \times I(Age/Young) + 3/6 \times I(Age/Elder)$$

$$= 0.5 \times 0.0000 + 0.5 \times 0.9183 = \mathbf{0.4591}$$

$$Gain(Age) = I(Credit) - E(Age)$$

$$= 0.9183 - 0.4591$$

$$= \mathbf{0.4592}$$

Top-down induction

Main loop

1 $A \leftarrow$ the ***best*** decision attribute for next ***node***
2 Assign A as decision attribute for ***node***
3 **for** each value of A
4 **do** create new descendant of ***node***
5 sort training examples to ***leaf nodes***
6 **if** training examples perfectly classified
7 **then** stop
8 **else** iterate over ***new leaf nodes***

When to consider decision trees

- Instances describable by attribute-value pairs
- Target function is discrete valued
- Possibly noisy training data

Examples:

- Equipment or medical diagnosis
- Credit risk analysis
- Modeling calendar scheduling preferences

Any question?

Xiaoqing Zheng
Fundan University