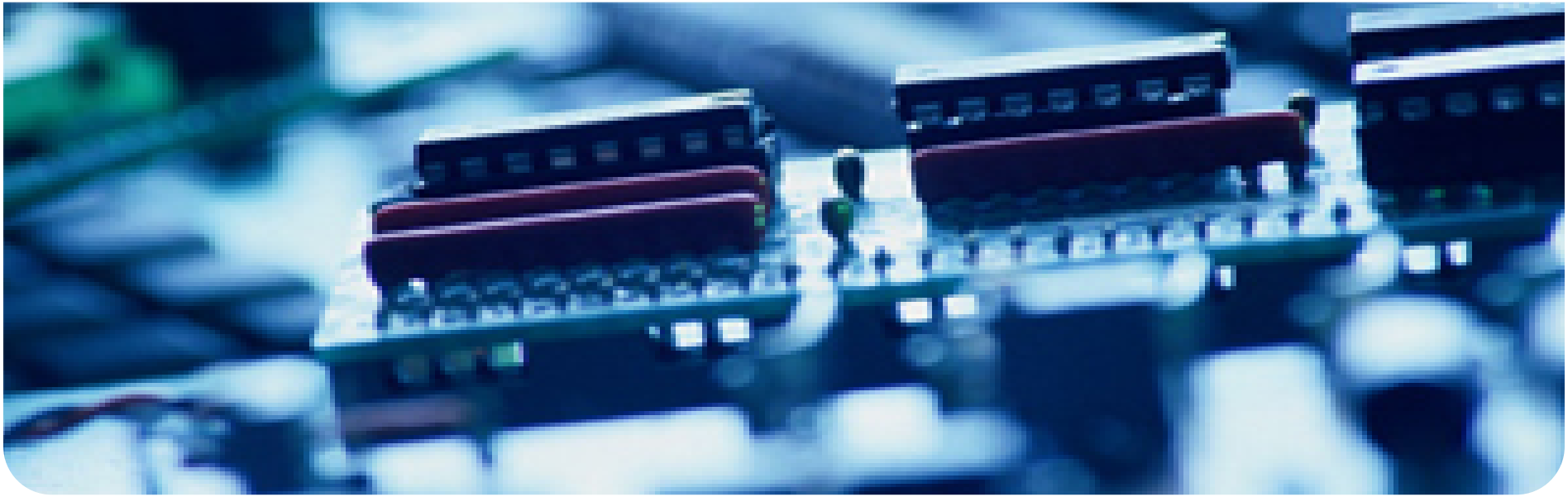
数据挖掘和大数据分析



Outline

① Decision Tree Algorithm （C）

② Random Forest

③ Quiz 2

DATA ANALYTICS:

DATA MINING AND BIG DATA

—— Machine Learning 5

Information Entropy

信息熵

$$H(X) = -\sum_{x \in X} P(x) \log P(x)$$

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2 p_i$$

$$Info_A(D) = \sum_{j=1}^{v} \left[\left(\frac{|D_j|}{|D|} \right) * Info(D_j) \right]$$

Decision Tree Algorithm

New Concept: **Pure** （纯度）

（1）DataSets：D【50% “+” & 50% “-”】

$H(D) = -0.5 * \log_2 0.5 - 0.5 * \log_2 0.5 = 1$

（2）DataSets：D【20% “+” & 80% “-”】

$H(D) = -0.2 * \log_2 0.2 - 0.8 * \log_2 0.8 = 0.722$

（3）DataSets：D【100% “+” & 0% “-”】

$H(D) = -1 * \log_2 1 - 0 * \log_2 0 = 0$

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2 p_i$$

Decision Tree Algorithm

Information Entropy → **Conditional Entropy**

信息熵	条件熵
$$Info(D) = -\sum_{i=1}^{m} p_i \log_2 p_i$$	$$Info_A(D) = \sum_{j=1}^{v} \left[\left(\frac{\lvert D_j \rvert}{\lvert D \rvert} \right) * Info(D_j) \right]$$
DT Label Information Entropy	DT Feature Value Conditional Entropy

$$Gain(A) = Info(D) - Info_A(D)$$

Age	Salary	STU	Credit	Buy Computer
<30	H	No	OK	No
<30	H	No	Good	No
30-40	H	No	OK	Yes
>40	M	No	OK	Yes
>40	L	Yes	OK	Yes
>40	L	Yes	Good	No
30-40	L	Yes	Good	Yes
<30	M	No	OK	No
<30	L	Yes	OK	Yes
>40	M	Yes	OK	Yes
<30	M	Yes	Good	Yes
30-40	M	No	Good	Yes
30-40	H	Yes	OK	Yes
>40	M	No	Good	No

|D|=14

|C1,D|=5

|C2,D|=9

$$Info(D)$$

$$= -\frac{5}{14}\log_2\frac{5}{14} - \frac{9}{14}\log_2\frac{9}{14}$$

$$= 0.940$$

Age	Salary	STU	Credit	Buy Computer
<30	H	No	OK	No
<30	H	No	Good	No
30-40	H	No	OK	Yes
>40	M	No	OK	Yes
>40	L	Yes	OK	Yes
>40	L	Yes	Good	No
30-40	L	Yes	Good	Yes
<30	M	No	OK	No
<30	L	Yes	OK	Yes
>40	M	Yes	OK	Yes
<30	M	Yes	Good	Yes
30-40	M	No	Good	Yes
30-40	H	Yes	OK	Yes
>40	M	No	Good	No

Info年龄(D)

$$= \frac{5}{14}\left(-\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5}\right)$$
$$+ \frac{4}{14}\left(-\frac{4}{4}\log\frac{4}{4} - \frac{0}{4}\log\frac{0}{4}\right)$$
$$+ \frac{5}{14}\left(-\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5}\right)$$
$$= 0.694$$

Gain(年龄)

= Info(D) - Info年龄(D)

= 0.940 - 0.694 = 0.246

Age	Salary	STU	Credit	Buy Computer
<30	H	No	OK	No
<30	H	No	Good	No
30-40	H	No	OK	Yes
>40	M	No	OK	Yes
>40	L	Yes	OK	Yes
>40	L	Yes	Good	No
30-40	L	Yes	Good	Yes
<30	M	No	OK	No
<30	L	Yes	OK	Yes
>40	M	Yes	OK	Yes
<30	M	Yes	Good	Yes
30-40	M	No	Good	Yes
30-40	H	Yes	OK	Yes
>40	M	No	Good	No

Info收入(D)

$$= \frac{4}{14}\left(-\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4}\right)$$

$$+ \frac{6}{14}\left(-\frac{2}{6}\log\frac{2}{6} - \frac{4}{6}\log\frac{4}{6}\right)$$

$$+ \frac{4}{14}\left(-\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4}\right)$$

$$= 0.911$$

Gain(收入)

= Info(D) - Info收入(D)

= 0.940 - 0.911 = 0.029

Age	Salary	STU	Credit	Buy Computer
<30	H	No	OK	No
<30	H	No	Good	No
30-40	H	No	OK	Yes
>40	M	No	OK	Yes
>40	L	Yes	OK	Yes
>40	L	Yes	Good	No
30-40	L	Yes	Good	Yes
<30	M	No	OK	No
<30	L	Yes	OK	Yes
>40	M	Yes	OK	Yes
<30	M	Yes	Good	Yes
30-40	M	No	Good	Yes
30-40	H	Yes	OK	Yes
>40	M	No	Good	No

Info学生(D)

$$= \frac{7}{14}\left(-\frac{1}{7}\log\frac{1}{7} - \frac{6}{7}\log\frac{6}{7}\right)$$
$$+ \frac{7}{14}\left(-\frac{4}{7}\log\frac{4}{7} - \frac{3}{7}\log\frac{3}{7}\right)$$
$$= 0.788$$

Gain(学生)

= Info(D) - Info学生(D)

= 0.940 - 0.788 = 0.152

Age	Salary	STU	Credit	Buy Computer
<30	H	No	OK	No
<30	H	No	Good	No
30-40	H	No	OK	Yes
>40	M	No	OK	Yes
>40	L	Yes	OK	Yes
>40	L	Yes	Good	No
30-40	L	Yes	Good	Yes
<30	M	No	OK	No
<30	L	Yes	OK	Yes
>40	M	Yes	OK	Yes
<30	M	Yes	Good	Yes
30-40	M	No	Good	Yes
30-40	H	Yes	OK	Yes
>40	M	No	Good	No

Info信用(D)

$$= \frac{6}{14}\left(-\frac{3}{6}\log\frac{3}{6} - \frac{3}{6}\log\frac{3}{6}\right)$$

$$+ \frac{8}{14}\left(-\frac{2}{8}\log\frac{2}{8} - \frac{6}{8}\log\frac{6}{8}\right)$$

$$= 0.892$$

Gain(信用)

= Info(D) - Info信用(D)

= 0.940 - 0.892 = 0.048

Decision Tree Algorithm

"Age" Fearure: Max Gain

Age

<30

Sa	Stu	Credit	Computer
H	No	OK	No
H	No	Good	No
M	No	OK	No
L	Yes	OK	Yes
M	Yes	Good	Yes

30-40

Sa	S		uter
H	N		s
L	Y		s
M	N		s
H	Y		s

Yes

>40

Sa	Stu	Credit	Computer
M	No	OK	Yes
L	Yes	OK	Yes
L	Yes	Good	No
M	Yes	OK	Yes
M	No	Good	No

Decision Tree Algorithm

Sa	Stu	Credit	Computer
H	No	OK	No
H	No	Good	No
M	No	OK	No
L	Yes	OK	Yes
M	Yes	Good	Yes

Info收入(D)

= 2/5 * (-2/2 * log2/2 - 0/2 * log0/2)

+ 2/5 * (-1/2 * log1/2 - 1/2 * log1/2)

+ 1/5 * (-1/1 * log1/1 - 0/1 * log0/1) = 0.400

Info学生(D)

= 3/5 * (-3/3 * log3/3 - 0/3 * log0/3)

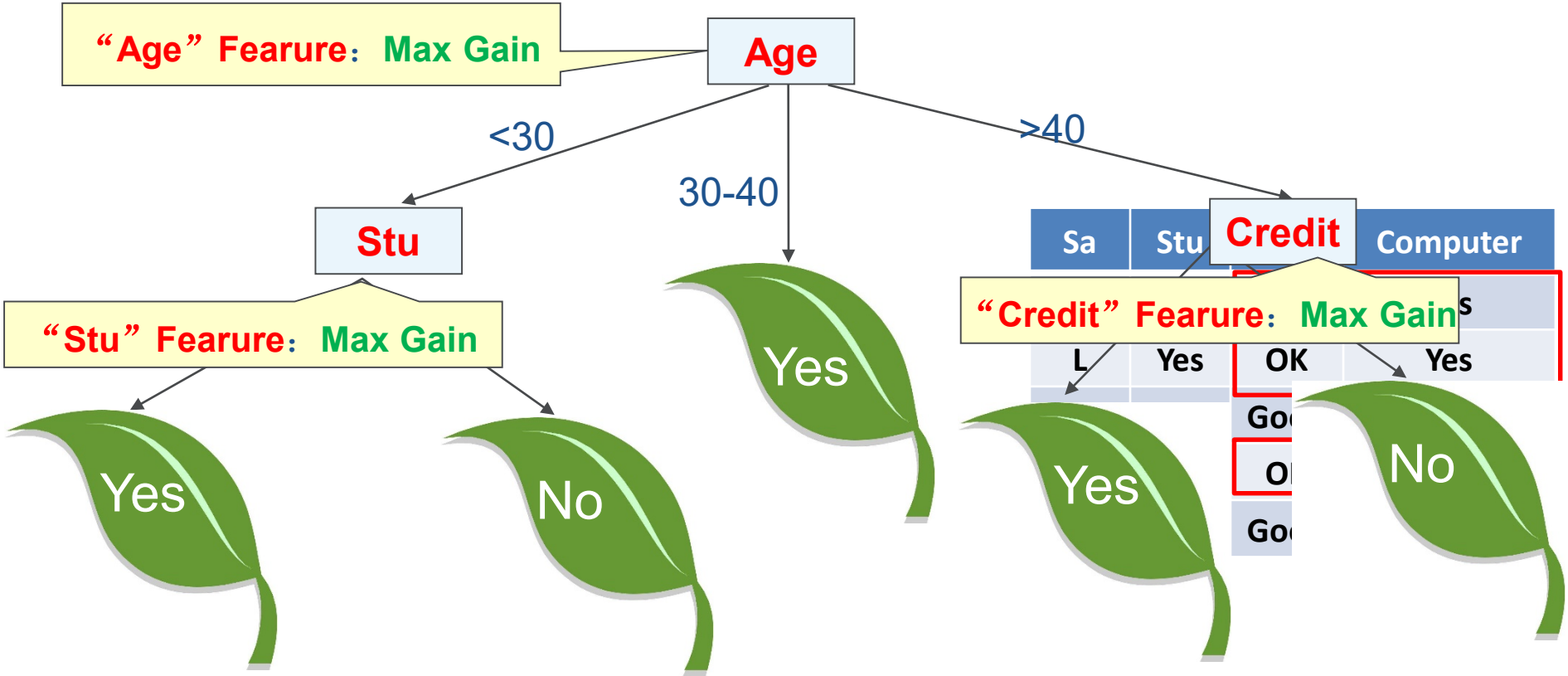
+ 2/5 * (-2/2 * log2/2 - 0/2 * log0/2) = 0

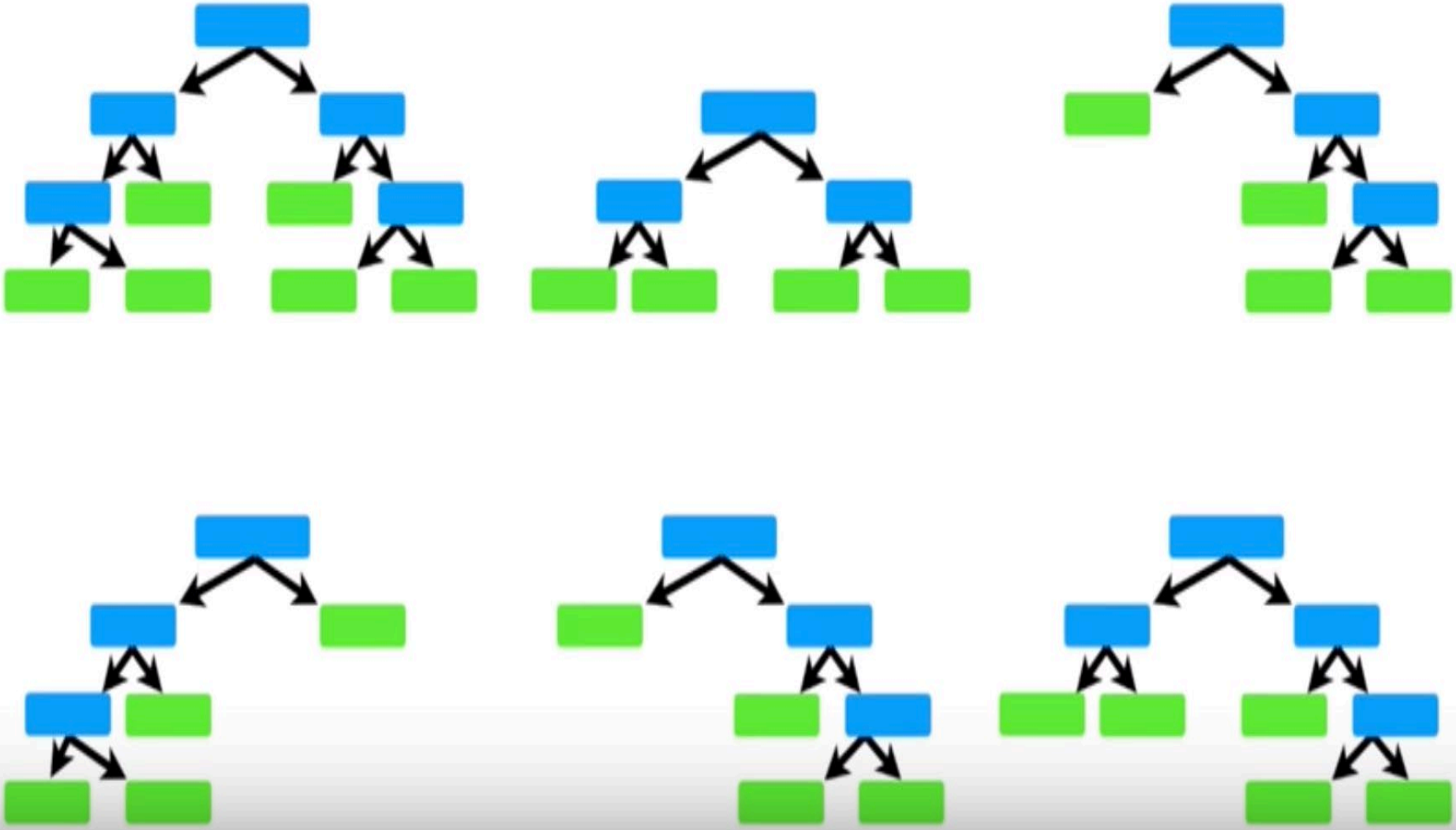
Info信用(D)

= 3/5 * (-2/3 * log2/3 - 1/3 * log1/3)

+ 2/5 * (-1/2 * log1/2 - 1/2 * log1/2) = 0.951

Decision Tree Algorithm





贵在坚持！