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

$$H(X) = -\sum_{x \in X} P(x) 1 \text{ o } P(x)$$

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2 p_i \qquad Info_A(D) = \sum_{j=1}^{v} \left[\left(\frac{\left| D_j \right|}{\left| D \right|} \right) * Info(D_j) \right]$$



New Concept: Pure (純麼)

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

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



Information Entropy

信息熵

Conditional Entropy

条件熵

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

DT Label Information Entropy

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

DT Feature Value Conditional Entropy

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

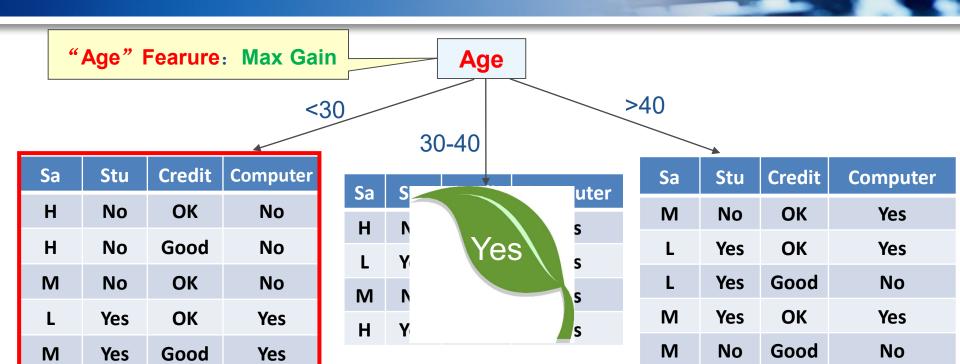
Age	Salary	STU	Credit	Buy Computer	
<30	Н	No	ОК	No	
<30	Н	No	Good	No	
30-40	Н	No	ОК	Yes	D =14
>40	M	No	ОК	Yes	C1,D =5
>40	L	Yes	ОК	Yes	
>40	L	Yes	Good	No	C2,D =9
30-40	L	Yes	Good	Yes	- 4 (-)
<30	M	No	OK	No	Info(D)
<30	L	Yes	OK	Yes	$Info(D)$ $= -\frac{5}{14} \log_2 \frac{5}{14} - \frac{9}{14} \log_2 \frac{9}{14}$
>40	M	Yes	OK	Yes	$= -\frac{5}{14} \log_2 \frac{5}{14} - \frac{9}{14} \log_2 \frac{9}{14}$
<30	M	Yes	Good	Yes	=0.940
30-40	M	No	Good	Yes	
30-40	Н	Yes	OK	Yes	
>40	М	No	Good	No	

Age	Salary	STU	Credit	Buy Computer	
<30	Н	No	ОК	No	
<30	Н	No	Good	No	
30-40	Н	No	ОК	Yes	Info年龄(D)
>40	M	No	OK	Yes	5 (3 1 2 2 2 2)
>40	L	Yes	ОК	Yes	$= \frac{5}{14} \left(-\frac{3}{5} \log \frac{3}{5} - \frac{2}{5} \log \frac{2}{5} \right)$
>40	L	Yes	Good	No	A A A O O
30-40	L	Yes	Good	Yes	$+$ — $(\log\log -)$
<30	M	No	ОК	No	14 4 4 4 4
<30	L	Yes	ОК	Yes	$+\frac{5}{14}\left(-\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5}\right)$
>40	M	Yes	ОК	Yes	14 5 5 5 5 5
<30	M	Yes	Good	Yes	=0.694
30-40	M	No	Good	Yes	
30-40	Н	Yes	ОК	Yes	Gain(年龄)
>40	М	No	Good	No	= Info(D) - Info年龄(D) = 0.940 - 0.694 = 0.246

Age	Salary	STU	Credit	Buy Computer	
<30	Н	No	OK	No	
<30	Н	No	Good	No	
30-40	Н	No	ОК	Yes	Info收入(D)
>40	М	No	OK	Yes	4 2 2 2 2 2 2
>40	L	Yes	ОК	Yes	$= \frac{4}{14} \left(-\frac{2}{4} \log \frac{2}{4} - \frac{2}{4} \log \frac{2}{4} \right)$
>40	L	Yes	Good	No	$\frac{1}{6}$ $\frac{1}{2}$ $\frac{1}{2}$ $\frac{1}{4}$ $\frac{1}{4}$
30-40	L	Yes	Good	Yes	$+-(\log\log -)$
<30	М	No	OK	No	14 6 6 6 6
<30	L	Yes	OK	Yes	$+\frac{4}{14}\left(-\frac{1}{4}\log\frac{1}{4}-\frac{3}{4}\log\frac{3}{4}\right)$
>40	М	Yes	OK	Yes	14 4 4 4 4 4
<30	М	Yes	Good	Yes	= 0.911
30-40	М	No	Good	Yes	Gain(收入)
30-40	Н	Yes	OK	Yes	= Info(D) - Info收入(D)
>40	M	No	Good	No	= 0.940 - 0.911 = 0.029

Age	Salary	STU	Credit	Buy Computer	
<30	Н	No	OK	No	
<30	Н	No	Good	No	
30-40	Н	No	ОК	Yes	
>40	M	No	OK	Yes	Info学生(D)
>40	L	Yes	OK	Yes	7 1 1 6 6
>40	L	Yes	Good	No	$= \frac{7}{14} \left(-\frac{1}{7} \log \frac{1}{7} - \frac{6}{7} \log \frac{6}{7} \right)$
30-40	L	Yes	Good	Yes	
<30	M	No	OK	No	$+\frac{7}{14}\left(-\frac{4}{7}\log\frac{4}{7}-\frac{3}{7}\log\frac{3}{7}\right)$
<30	L	Yes	OK	Yes	$14 \left(7 \right) 7 7 7 7 7 7 7 $
>40	M	Yes	OK	Yes	=0.788
<30	М	Yes	Good	Yes	
30-40	M	No	Good	Yes	Gain(学生)
30-40	Н	Yes	ОК	Yes	= Info(D) - Info学生(D)
>40	M	No	Good	No	= 0.940 - 0.788 = 0.152

Age	Salary	STU	Credit	Buy Computer	
<30	Н	No	ОК	No	
<30	Н	No	Good	No	
30-40	Н	No	ОК	Yes	Info信用(D)
>40	M	No	ОК	Yes	6 3 3 3 3 3
>40	L	Yes	OK	Yes	$= \frac{6}{14} \left(-\frac{3}{6} \log \frac{3}{6} - \frac{3}{6} \log \frac{3}{6} \right)$
>40	L	Yes	Good	No	
30-40	L	Yes	Good	Yes	$+\frac{8}{14}\left(-\frac{2}{8}\log\frac{2}{8} - \frac{6}{8}\log\frac{6}{8}\right)$
<30	M	No	OK	No	
<30	L	Yes	OK	Yes	=0.892
>40	M	Yes	OK	Yes	
<30	M	Yes	Good	Yes	Gain(信用)
30-40	M	No	Good	Yes	= Info(D) - Info信用(D)
30-40	Н	Yes	OK	Yes	= 0.940 - 0.892 = 0.048
>40	M	No	Good	No	



Sa	Stu	Credit	Computer
Н	No	ОК	No
Н	No	Good	No
M	No	ОК	No
L	Yes	ОК	Yes
M	Yes	Good	Yes

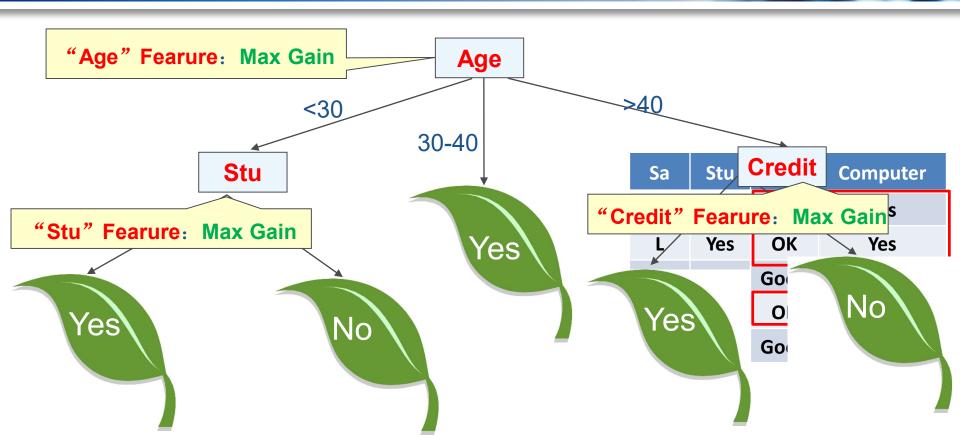
Info收入(D)

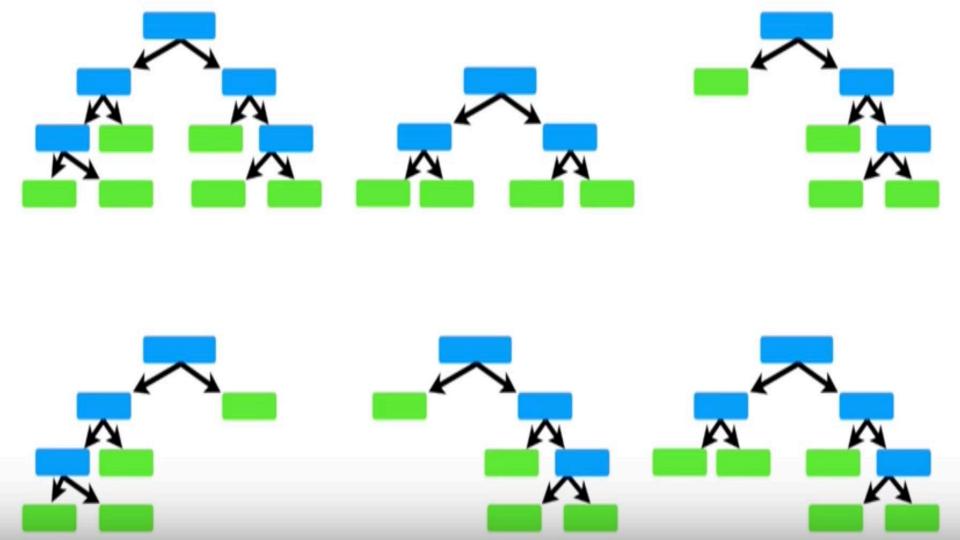
Info学生(D)

$$= 3/5 * (-3/3 * \log 3/3 - 0/3 * \log 0/3) + 2/5 * (-2/2 * \log 2/2 - 0/2 * \log 0/2) = 0$$

Info信用(D)

$$= 3/5 * (-2/3 * log2/3 - 1/3 * log1/3) + 2/5 * (-1/2 * log1/2 - 1/2 * log1/2) = 0.951$$









贵在坚持!