数据挖掘
HW3
计11 周韦平
2021010699

清華大學 Tsinghua University

一、①

采用cart法找根结点

职业类型

	是	否
职工	3	0
管理	2	1

$$gini(职业类型) = \frac{3}{10}\times\left(1-\left(\frac{3}{3}\right)^2-\left(\frac{0}{3}\right)^2\right)+\frac{3}{10}\times(1-1^2-1^2)+\frac{4}{10}\times\left(1-\left(\frac{2}{4}\right)^2-\left(\frac{2}{4}\right)^2\right)$$
$$=\frac{17}{60}$$

婚姻状况

	是	否
已婚	3	1
单身	2	2
离异	1	1

$$gini_{婚姻状况} = \frac{4}{10}\left(1-\left(\frac{3}{4}\right)^2-\left(\frac{1}{4}\right)^2\right)+\frac{4}{10}\left(1-\left(\frac{2}{4}\right)^2-\left(\frac{2}{4}\right)^2\right)+\frac{2}{10}\left(1-\left(\frac{1}{2}\right)^2-\left(\frac{1}{2}\right)^2\right)$$
$$=\frac{9}{20}$$

教育水平

	是	否
非高等	1	4
高等	4	0

$$gini_{教育水平} = \frac{5}{10}\times(1-1)+\frac{5}{10}\times\left(1-\left(\frac{1}{5}\right)^2-\left(\frac{4}{5}\right)^2\right)=\frac{8}{25}$$

$gini_{教育水平} < gini_{职业类型} < gini_{婚姻状况}$ 选教育水平为根

教育水平为高，一定是，教育水平为低，进一步计算下一个结点

职业类型

	是	否
职工	0	3
管理	1	0
警察	0	1

$$gini_{职业类型} = \frac{3}{5}\times(1-1)+\frac{1}{5}\times(1-1)+\frac{1}{5}\times(1-1)=0$$

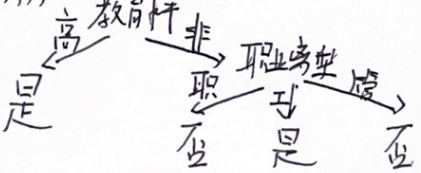
婚姻状况

	是	否
已婚	0	1
单身	1	2
离异	0	1

$$gini_{婚姻状况} = \frac{1}{5}\times(1-1)+\frac{2}{5}\times\left(1-\left(\frac{1}{2}\right)^2-\left(\frac{1}{2}\right)^2\right)+\frac{1}{5}(1-1)=\frac{4}{15}$$

$gini_{婚姻状况} > gini_{职业类型}$ 选职业类型

决策树为



给定 非高等 → 职员 → 单身

对基金不感兴趣

② 职业类型

$P(职|是)=\frac{1}{6}$ $P(职|否)=\frac{3}{4}$

$P(工|是)=\frac{3}{6}$ $P(工|否)=0$

$P(管|是)=\frac{2}{6}$ $P(管|否)=\frac{1}{4}$

婚姻

$P(已|是)=\frac{3}{6}$ $P(已|否)=\frac{1}{4}$

$P(单|是)=\frac{2}{6}$ $P(单|否)=\frac{2}{4}$

$P(合|是)=\frac{1}{6}$ $P(离|否)=\frac{1}{4}$

教育程

$P(非|是)=\frac{1}{6}$ $P(非|否)=1$

$P(高|是)=\frac{5}{6}$ $P(高|否)=0$

$P(是,职,单,非)=P(职|是)P(单|是)P(非|是)P(是)=\frac{1}{6}\times\frac{2}{6}\times\frac{1}{6}\times\frac{6}{10}=\frac{1}{180}$

$P(否,职,单,非)=P(管|否)P(单|否)P(非|否)P(否)=\frac{1}{4}\times\frac{2}{4}\times1\times\frac{6}{10}=\frac{3}{20}$

$\therefore P(否,职,单非)>P(是,职,单非)$ $P(否|职,单非)>P(是|单,职,非)$

∴ 不感兴趣

	(1,1)	(7,5)
(2,1)	✓	
(1,2)	✓	
(0,1)	✓	
(3,3)	✓	
(2,2)	✓	
(4,4)	✓	
(5,4)		✓
(6,3)		✓
(5,6)		✓
(6,5)		✓
(7,4)		✓
(7,6)		✓

更新中心 → $C_1 = (2.5, \frac{11}{6})$

$C_2 = (6, \frac{14}{5})$

再次迭代，发现分组仍是

C_1: {(2,1), (1,2), (3,1), (3,3), (2,2), (4,2)}

C_2: {(5,4), (6,3), (5,6), (6,5), (7,4), (7,6)} 不变

∴ 中心为 $C_1(2.5, \frac{11}{6})$, $C_2(6, \frac{14}{5})$, 分组为 ↑