明理 精工

笃学 致远

第7章 决策树

Decision Tree

电子工程学院、人工智能学院

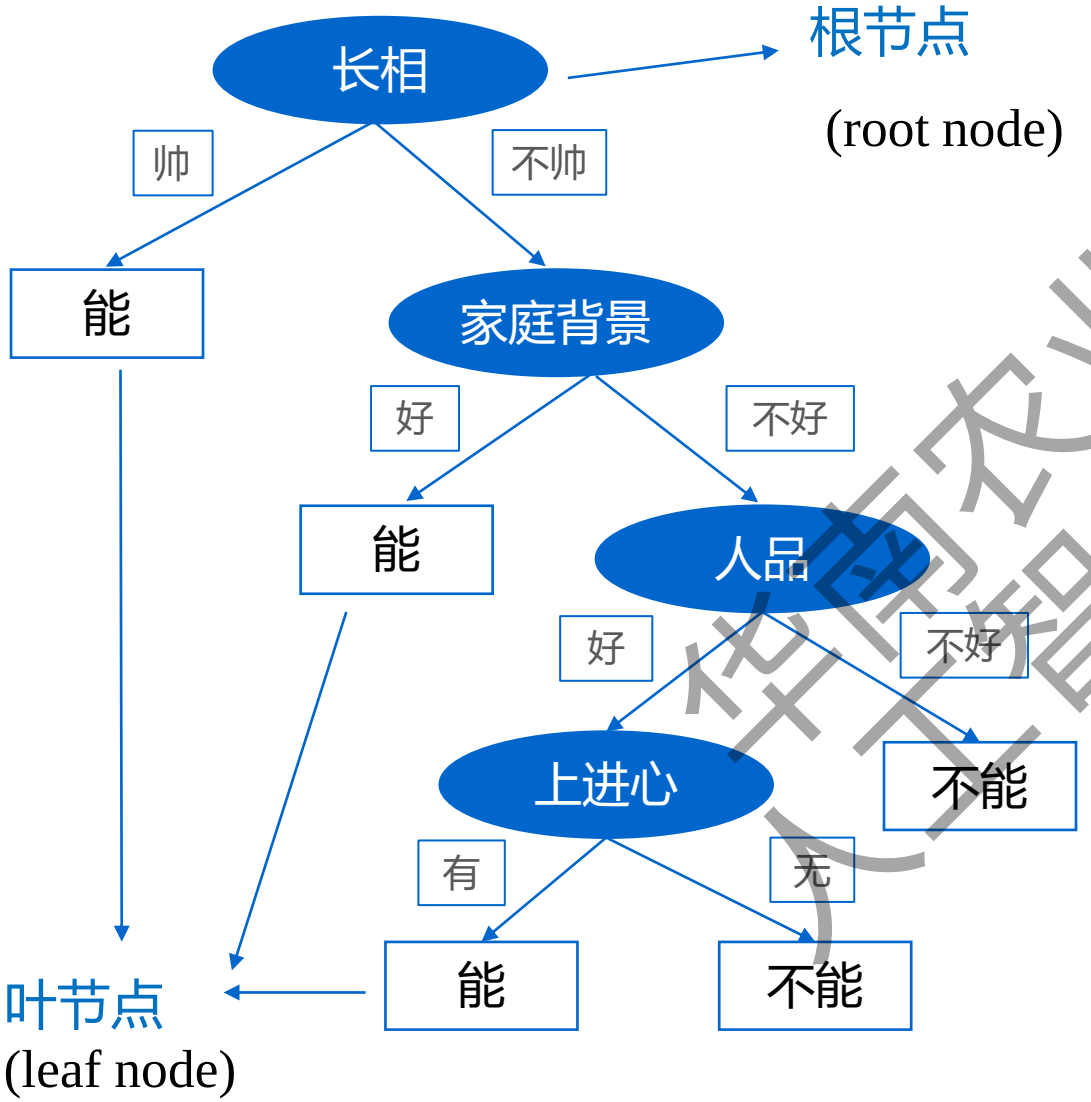
college of Electronic Engineering、college of Artificial Intelligence

- 基本原理
- 建树
- 剪枝

算法的发展过程

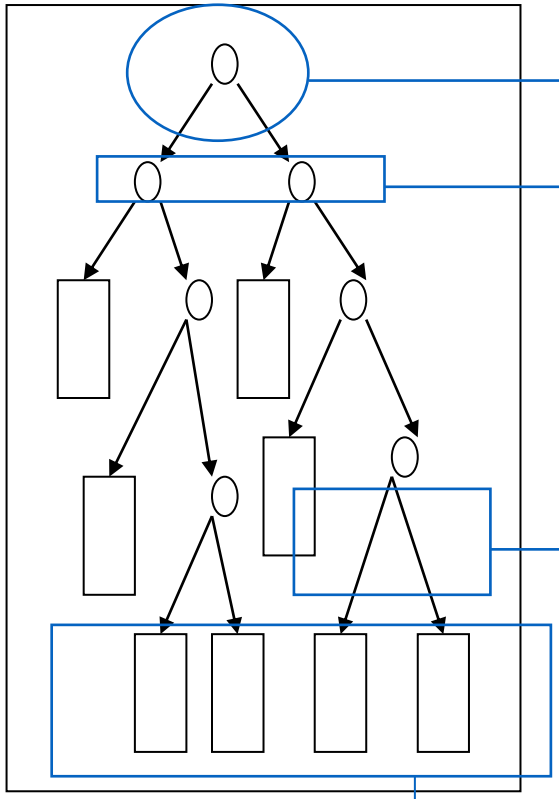
- 1979年, J.R. Quinlan 给出ID3算法，并在1983年和1986年对ID3 进行了总结和简化，使其成为决策树学习算法的典型。
- 1993年，Quinlan 进一步发展了ID3算法，改进成C4.5算法。
- 另一类决策树算法为CART，CART的决策树由二元逻辑问题生成，每个树节点只有两个分枝，分别包括学习实例的正例与反例。

1. 决策树原理（Decision Tree）



- 决策树属于判别模型。
- 决策树是一种树状结构。
- 决策树的决策过程从根节点开始。

✓ 决策树的表示



根节点 (root node)

非叶子节点 (non-leaf node)

(代表测试条件，对数据属性的测试)

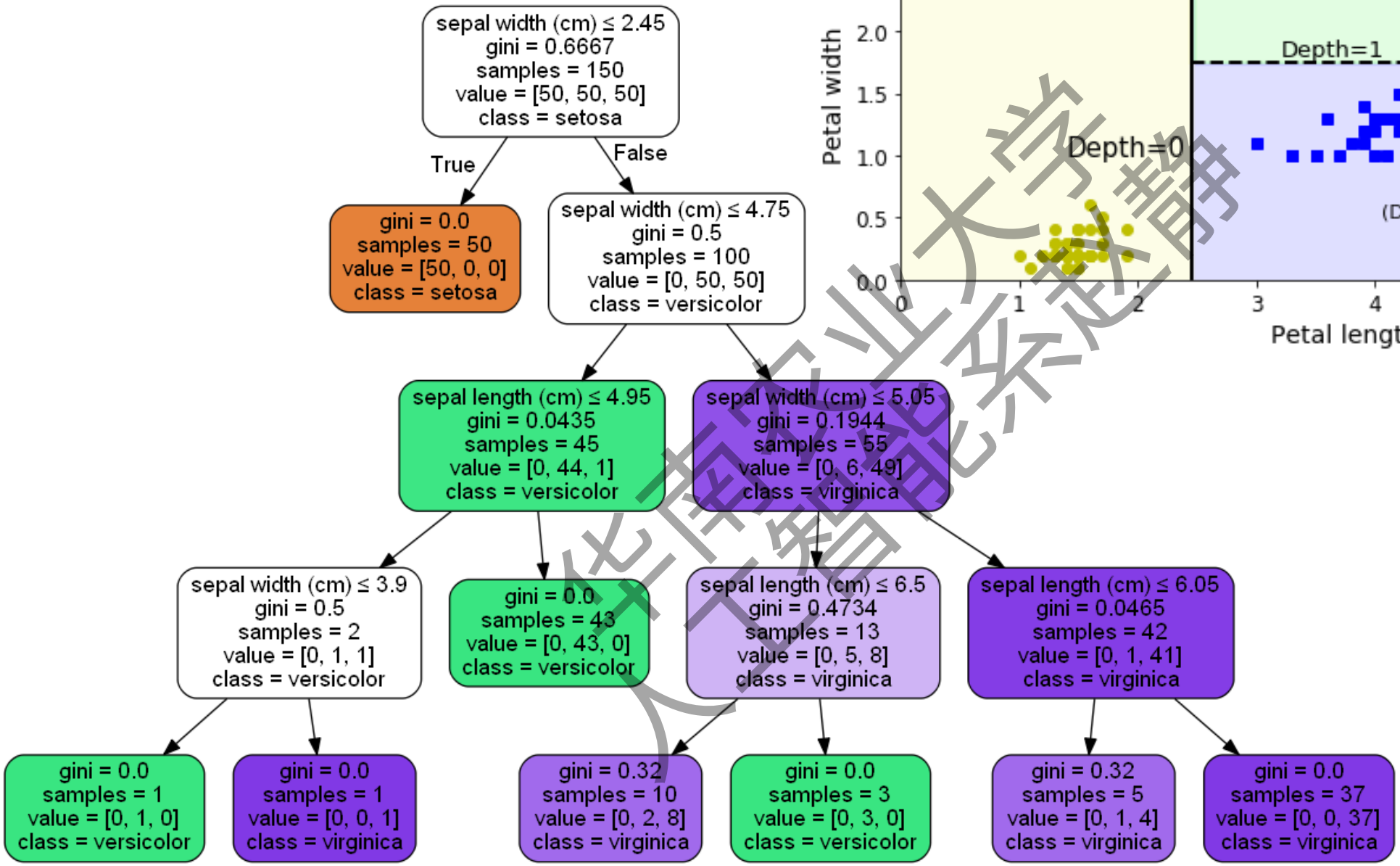
分支 (branches) (代表测试结果)

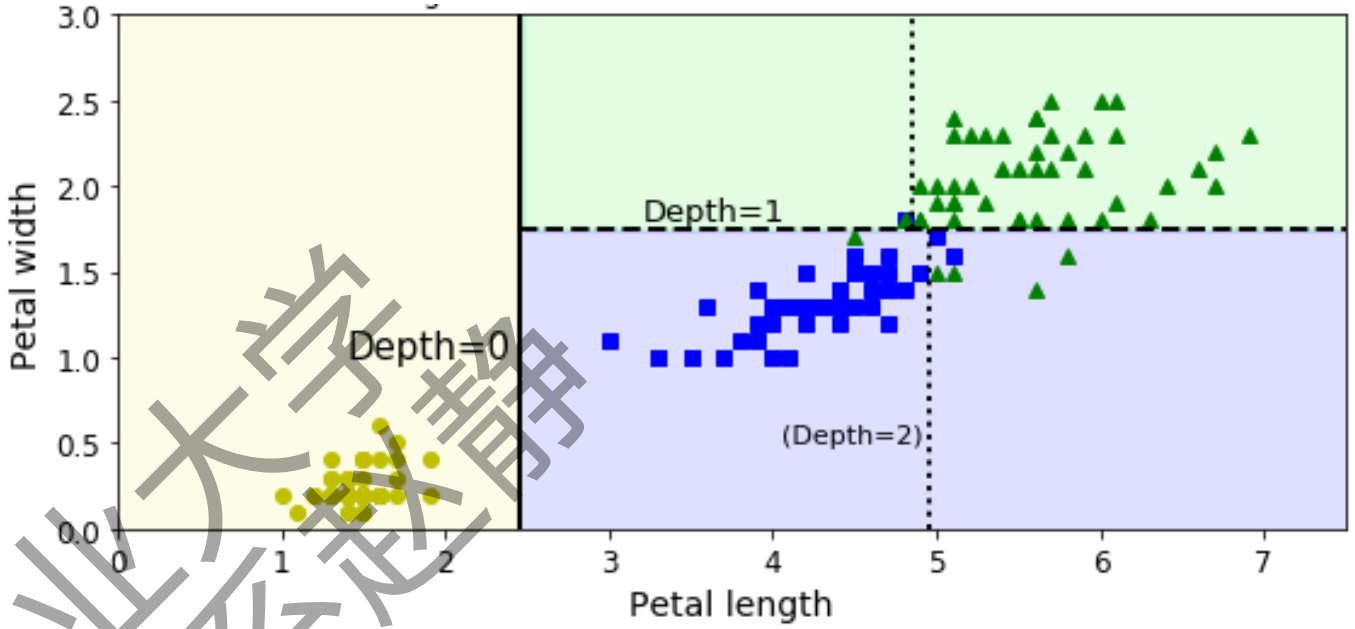
叶节点 (leaf node)

(代表分类后所获得的分类标记)

- 决策树算法属于监督学习方法。
- 决策树算法采用贪心算法。
- 在决策树的生成过程中，分割方法即属性选择的度量是关键。

✓ 决策树原理

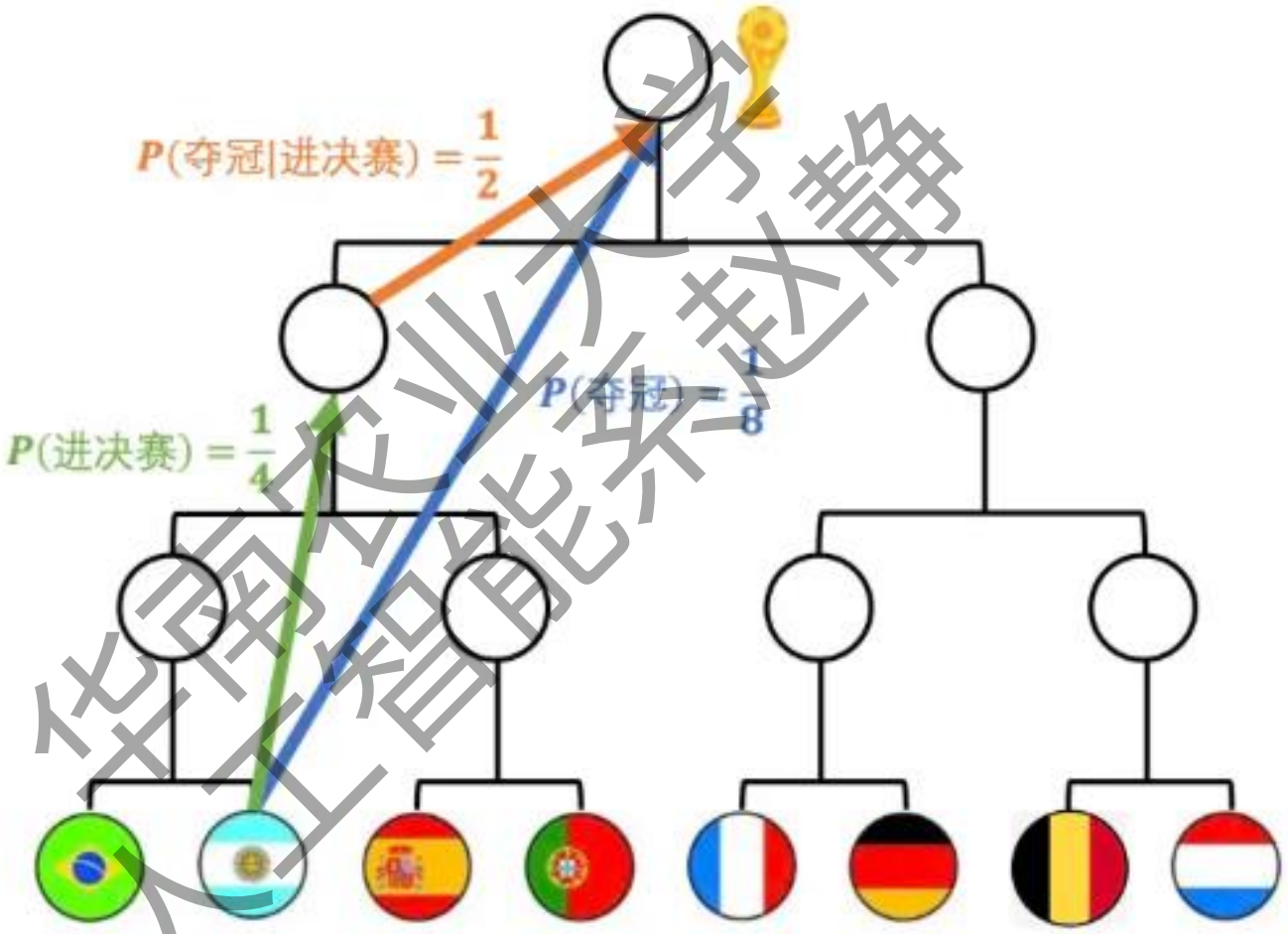




- ✓ 属性选择的度量

熵(entropy)

- **信息量**

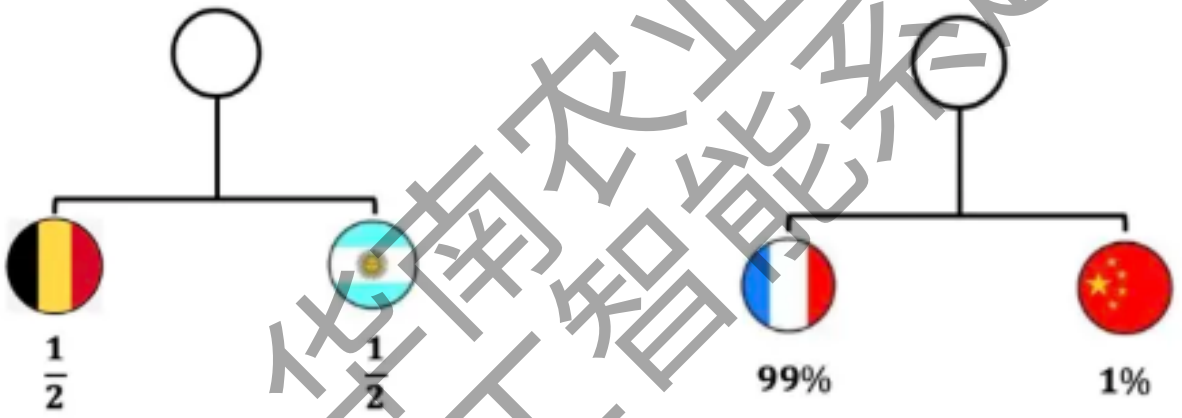


信息量(1/8)=信息量(1/4)+信息量(1/2)

信息量(1/4 × 1/2)=信息量(1/4)+信息量(1/2)

- **信息熵（Information Entropy）**

$$\mathrm{Ent}(\mathrm{D}) = -\sum p_i \log p_i, i = 1, 2, \ldots, n$$



信息量：

$-\log_2 \frac{1}{2} = 1$　　$-\log_2 \frac{1}{2} = 1$　　$-\log_2 \frac{99}{100} \approx 0.0145$　　$-\log_2 \frac{1}{100} \approx 6.6439$

对系统贡献的信息量：

$\frac{1}{2}(-\log_2 \frac{1}{2}) = 0.5$　　$\frac{1}{2}(-\log_2 \frac{1}{2}) = 0.5$　　$\frac{99}{100}(-\log_2 \frac{99}{100}) \approx 0.014355$　　$\frac{1}{100}(-\log_2 \frac{1}{100}) \approx 0.066439$

$$Ent(D) = -\sum p_i \log p_i, i = 1, 2, \ldots, n$$

例:

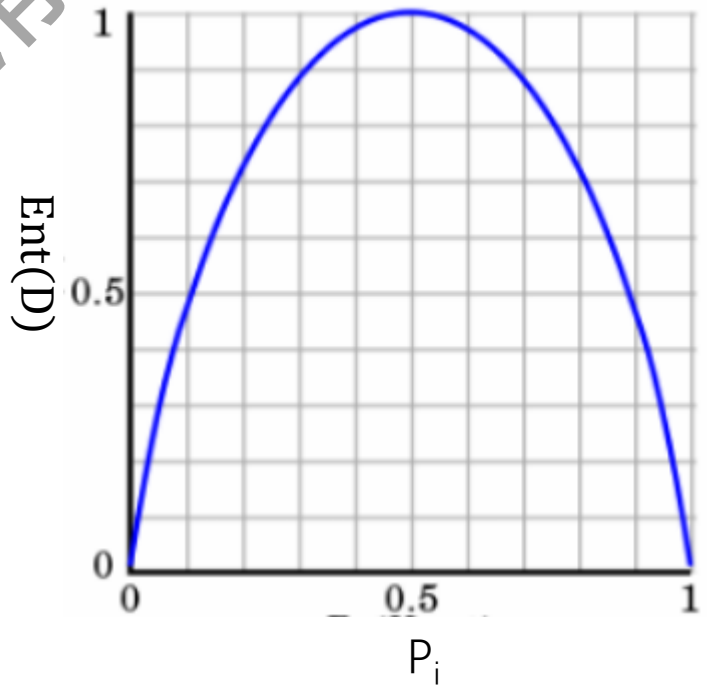
A集合y=[1,1,1,1,1,1,1,1,2,2]

B集合y=[1,2,3,4,5,6,6,5,3,1]

二分类问题中：

$$P(X=1)=p, \quad P(X=0)=1-p, \quad 0 \leqslant p \leqslant 1$$

$$Ent(D) = -\ p\log_2 p - (1-p)\log_2(1-p)$$



二分类熵值与概率的关系

- 基本原理
- 建树
  - ID3
  - C4.5
  - CART
- 剪枝

2.1 ID3建树——信息增益

令当前节点的样本集合为$\mathcal{D}$,分裂后第v个子集样本集合为$\mathcal{D}_v$

- 用样本的比例估计概率分布：$p(Y=\mathrm{c})=\frac{|N_c|}{|N|}$
- 经验熵（分裂之前的熵）：$H(\mathcal{D})=-\sum_{\mathrm{c}=1}^{\mathrm{C}} p(Y=\mathrm{c})\log p(Y=\mathrm{c})$
- 经验条件熵（分裂成V个子集后的熵）：$H(\mathcal{D}|X)=\sum_{v=1}^{V}\frac{|\mathcal{D}_v|}{|\mathcal{D}|}H(\mathcal{D}_v)$
- **信息增益**：$gain_X(\mathcal{D})=H(\mathcal{D})-H(\mathcal{D}|X)$

案例：账号真实性判断

日志密度L	好友密度F	是否使用真实头像H	账号是否真实R
s	s	no	no
s	l	yes	yes
l	m	yes	yes
m	m	yes	yes
l	m	yes	yes
m	l	no	yes
m	s	no	no
l	m	no	yes
m	s	no	yes
s	s	yes	no

$$H(\mathcal{D}) = -0.7\log_2 0.7 - 0.3\log_2 0.3 = 0.879$$

日志密度L:

$$H_L(\mathcal{D}) = 0.3 \times \left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) +$$
$$0.4 \times \left(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right) +$$
$$0.3 \times \left(-\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3}\right) +$$
$$= 0.603$$

$$gain_L(\mathcal{D}) = H(\mathcal{D}) - H_L(\mathcal{D}) = 0.276$$

$$gain_F(\mathcal{D}) = 0.553$$

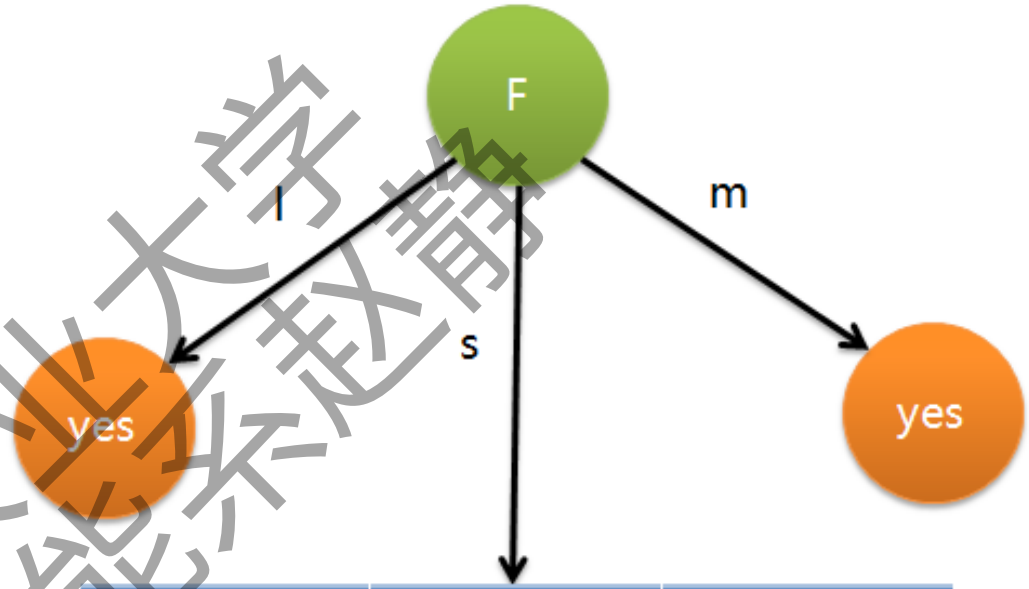
$$gain_H(\mathcal{D}) = 0.033$$

日志密度L	好友密度F	是否使用真实头像H	账号是否真实R
s	s	no	no
s	l	yes	yes
l	m	yes	yes
m	m	yes	yes
l	m	yes	yes
m	l	no	yes
m	s	no	no
l	m	no	yes
m	s	no	yes
s	s	yes	no

$$gain_L(\mathcal{D}) = 0.276$$

$$gain_F(\mathcal{D}) = 0.553$$

$$gain_H(\mathcal{D}) = 0.033$$



日志密度	是否使用真实头像	账号是否真实
s	no	no
m	no	no
m	no	yes
s	yes	no

2.2 C4.5建树——信息增益比

信息增益 $gain_X(\mathcal{D}) = H(\mathcal{D}) - H(\mathcal{D}|X)$

经验条件熵（分裂成V个子集后的熵）：$H(\mathcal{D}|X) = \sum_{v=1}^{V} \frac{|\mathcal{D}_v|}{|\mathcal{D}|} H(\mathcal{D}_v)$

信息增益率 $gain\_ratio_X(\mathcal{D}) = \frac{gain_X(\mathcal{D})}{split\_info_X(\mathcal{D})}$

分裂信息 $split\_info_X(\mathcal{D}) = HX(\mathcal{D}) = -\sum_{v=1}^{V} \frac{|\mathcal{D}_v|}{|\mathcal{D}|} \log_2 \frac{|\mathcal{D}_v|}{|\mathcal{D}|}$

例:就业因素调查

学号	性别	学生干部	综合成绩	毕业论文	就业情况
1	男	是	70-79	优	已
2	女	是	80-89	中	已
3	男	不是	60-69	不及格	未
4	男	是	60-69	良	已
5	男	是	70-79	中	已
6	男	不是	70-79	良	未
7	女	是	60-69	良	已
8	男	是	60-69	良	已
9	女	是	70-79	中	未
10	男	不是	60-69	及格	已
11	男	是	80-89	及格	已
12	男	是	70-79	良	已
13	男	不是	70-79	及格	未
14	男	不是	60-69	及格	已
15	男	是	70-79	良	已
16	男	不是	70-79	良	未
17	男	不是	80-89	良	未
18	女	是	70-79	良	已
19	男	不是	70-79	不及格	未
20	男	不是	70-79	良	未
21	女	是	60-69	优	已
22	男	是	60-69	良	已

学号	性别	学生干部	综合成绩	毕业论文	就业情况
3	男	不是	60-69	不及格	未
6	男	不是	70-79	良	未
9	女	是	70-79	中	未
13	男	不是	70-79	及格	未
16	男	不是	70-79	良	未
17	男	不是	80-89	良	未
19	男	不是	70-79	不及格	未
20	男	不是	70-79	良	未
1	男	是	70-79	优	已
2	女	是	80-89	中	已
4	男	是	60-69	良	已
5	男	是	70-79	中	已
7	女	是	60-69	良	已
8	男	是	60-69	良	已
10	男	不是	60-69	及格	已
11	男	是	80-89	及格	已
12	男	是	70-79	良	已
14	男	不是	60-69	及格	已
15	男	是	70-79	良	已
18	女	是	70-79	良	已
21	女	是	60-69	优	已
22	男	是	60-69	良	已

$$\text{Entropy}(\text{就业情况}) = -\frac{14}{22}\log_2\frac{14}{22} - \frac{8}{22}\log_2\frac{8}{22} = 0.94566$$

$$\text{Entropy}(\text{就业情况}) = -\frac{14}{22}\log_2\frac{14}{22} - \frac{8}{22}\log_2\frac{8}{22} = 0.94566$$

学号	性别	学生干部	综合成绩	毕业论文	就业情况
3	男	不是	60-69	不及格	未
6	男	不是	70-79	良	未
13	男	不是	70-79	及格	未
16	男	不是	70-79	良	未
17	男	不是	80-89	良	未
19	男	不是	70-79	不及格	未
20	男	不是	70-79	良	未
1	男	是	70-79	优	已
4	男	是	60-69	良	已
5	男	是	70-79	中	已
8	男	是	60-69	良	已
10	男	不是	60-69	及格	已
11	男	是	80-89	及格	已
12	男	是	70-79	良	已
14	男	不是	60-69	及格	已
15	男	是	70-79	良	已
22	男	是	60-69	良	已
9	女	是	70-79	中	未
2	女	是	80-89	中	已
7	女	是	60-69	良	已
18	女	是	70-79	良	已
21	女	是	60-69	优	已

$$\text{Entropy}(\text{男}) = -\frac{10}{17}\log_2\frac{10}{17} - \frac{7}{17}\log_2\frac{7}{17} = 0.97742$$

$$\text{Entropy}(\text{女}) = -\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5} = 0.72193$$

$$\text{Entropy}(\text{性别}) = \frac{17}{22} * 0.97742 + \frac{5}{22} * 0.72193 = 0.91935$$

$$\text{Gain}(\text{性别}) = 0.94566 - 0.91935 = 0.02631$$

$$Ent_A(\text{性别}) = -\frac{17}{22}\log_2\frac{17}{22} - \frac{5}{22}\log_2\frac{5}{22} = 0.77323$$

$$\text{Gain\_Ratio}(\text{性别}) = 0.02631/0.77323 = 0.03403$$

$$\text{Entropy}(\text{就业情况}) = -\frac{14}{22}\log_2\frac{14}{22} - \frac{8}{22}\log_2\frac{8}{22} = 0.94566$$

学号	性别	学生干部	综合成绩	毕业论文	就业情况
3	男	不是	60-69	不及格	未
6	男	不是	70-79	良	未
13	男	不是	70-79	及格	未
16	男	不是	70-79	良	未
17	男	不是	80-89	良	未
19	男	不是	70-79	不及格	未
20	男	不是	70-79	良	未
10	男	不是	60-69	及格	已
14	男	不是	60-69	及格	已
1	男	是	70-79	优	已
4	男	是	60-69	良	已
5	男	是	70-79	中	已
8	男	是	60-69	良	已
11	男	是	80-89	及格	已
12	男	是	70-79	良	已
15	男	是	70-79	良	已
22	男	是	60-69	良	已
9	女	是	70-79	中	未
2	女	是	80-89	中	已
7	女	是	60-69	良	已
18	女	是	70-79	良	已
21	女	是	60-69	优	已

$$\text{Gain}(\text{学生干部}) = 0.94566 - 0.54382 = 0.40184$$

$$Ent_A(\text{学生干部}) = -\frac{13}{22}\log_2\frac{13}{22} - \frac{9}{22}\log_2\frac{9}{22} = 0.97602$$

$$\text{Gain\_Ratio}(\text{学生干部}) = 0.40184/0.97602 = 0.41171$$

$$\text{Entropy}(\text{就业情况}) = -\frac{14}{22}\log_2\frac{14}{22} - \frac{8}{22}\log_2\frac{8}{22} = 0.94566$$

学号	性别	学生干部	综合成绩	毕业论文	就业情况
3	男	不是	60-69	不及格	未
10	男	不是	60-69	及格	已
14	男	不是	60-69	及格	已
4	男	是	60-69	良	已
8	男	是	60-69	良	已
22	男	是	60-69	良	已
7	女	是	60-69	良	已
21	女	是	60-69	优	已
6	男	不是	70-79	良	未
13	男	不是	70-79	及格	未
16	男	不是	70-79	良	未
19	男	不是	70-79	不及格	未
20	男	不是	70-79	良	未
1	男	是	70-79	优	已
5	男	是	70-79	中	已
12	男	是	70-79	良	已
15	男	是	70-79	良	已
9	女	是	70-79	中	未
18	女	是	70-79	良	已
17	男	不是	80-89	良	未
11	男	是	80-89	及格	已
2	女	是	80-89	中	已

$$\text{Gain}(\text{综合成绩}) = 0.94566 - 0.819897 = 0.125763$$

$$Ent_A(\text{综合成绩}) = -\frac{8}{22}\log_2\frac{8}{22} - \frac{11}{22}\log_2\frac{11}{22} - \frac{3}{22}\log_2\frac{3}{22} = 1.422675$$

$$\text{Gain\_Ratio}(\text{综合成绩}) = 0.125763/1.422675 = 0.088391$$

$$Entropy(就业情况) = -\frac{14}{22}\log_2\frac{14}{22} - \frac{8}{22}\log_2\frac{8}{22} = 0.94566$$

学号	性别	学生干部	综合成绩	毕业论文	就业情况
3	男	不是	60-69	不及格	未
19	男	不是	70-79	不及格	未
10	男	不是	60-69	及格	已
14	男	不是	60-69	及格	已
13	男	不是	70-79	及格	未
11	男	是	80-89	及格	已
4	男	是	60-69	良	已
8	男	是	60-69	良	已
22	男	是	60-69	良	已
7	女	是	60-69	良	已
6	男	不是	70-79	良	未
16	男	不是	70-79	良	未
20	男	不是	70-79	良	未
12	男	是	70-79	良	已
15	男	是	70-79	良	已
18	女	是	70-79	良	已
17	男	不是	80-89	良	未
21	女	是	60-69	优	已
1	男	是	70-79	优	已
5	男	是	70-79	中	已
9	女	是	70-79	中	未
2	女	是	80-89	中	已

$$Gain(毕业论文) = 0.94566 - 0.745557 = 0.200103$$

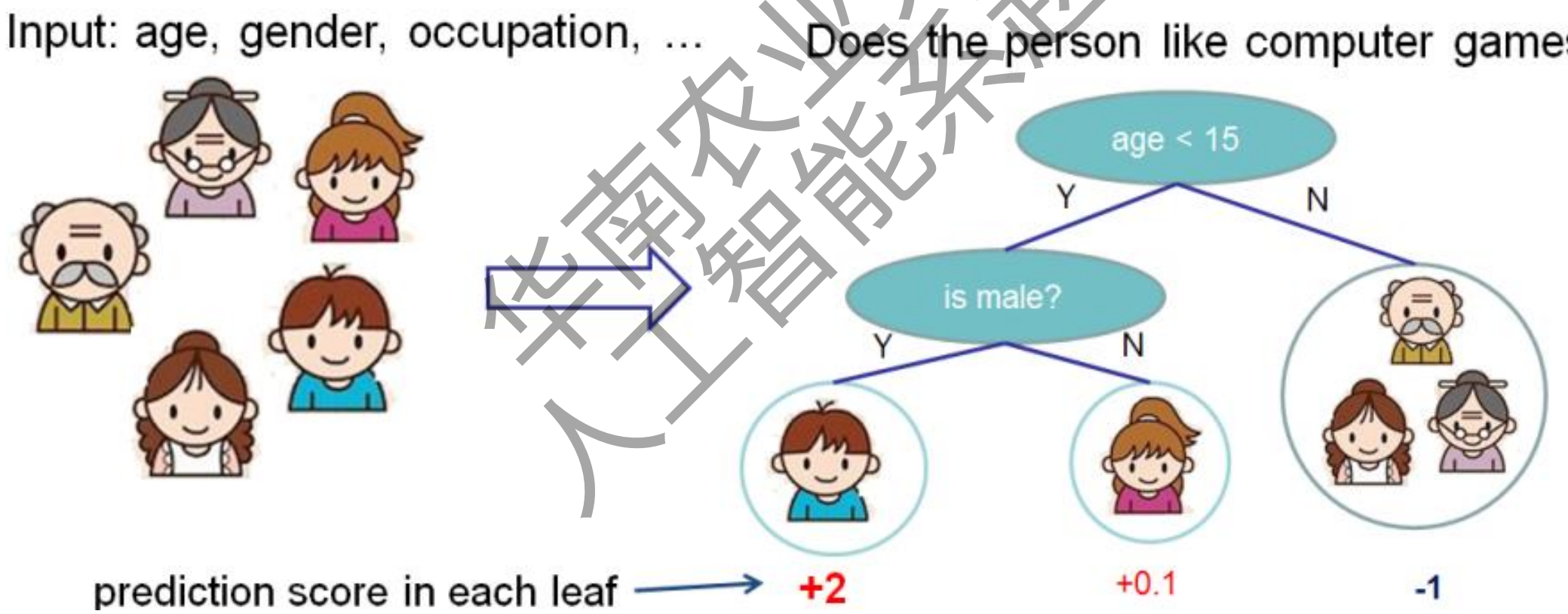
$$Ent_A(毕业论文) = -\frac{2}{22}\log_2\frac{2}{22} - \frac{4}{22}\log_2\frac{4}{22} - \frac{11}{22}\log_2\frac{11}{22} - \frac{2}{22}\log_2\frac{2}{22} - \frac{3}{22}\log_2\frac{3}{22} = 2.100806$$

$$Gain\_Ratio(毕业论文) = 0.200103/2.00103 = 0.10167158$$

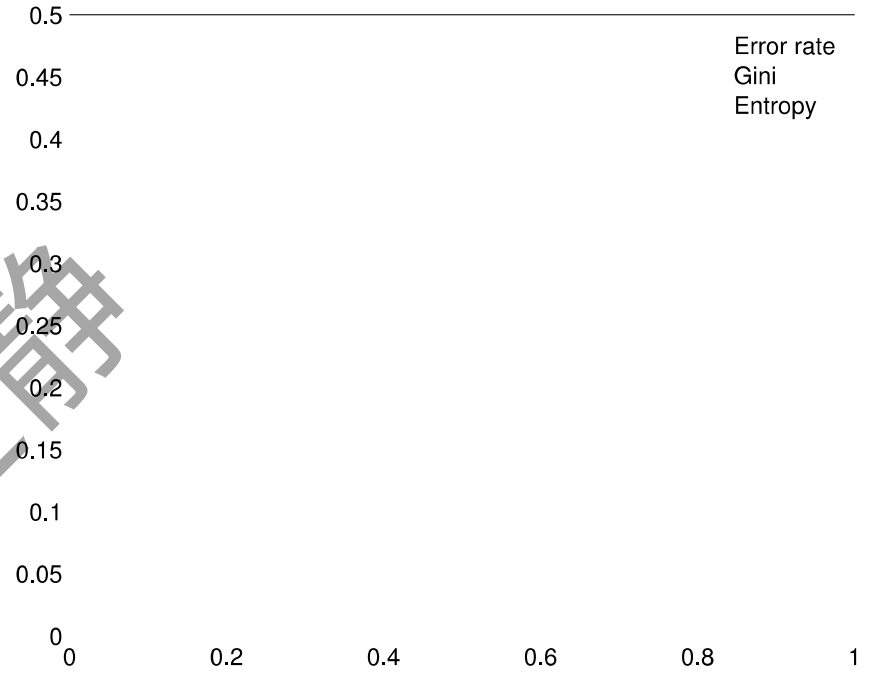
	性别	学生干部	综合成绩	论文
Gain	0.02631	0.40184	0.12576	0.200103
Gain_ratio	0.03403	0.41171	0.08839	0.10167158

2.3 CART树

二叉树：二分递归划分，将当前样本集合划分为两个子集为两个子节点，使得生成的每个非叶子结点都有两个分支



➢ CART分类——Gini Index



两类分类的不纯度度量

✓ 常用的不纯净度量:

- 错误率： $H(\mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}}\mathbb{I}(y_i \neq \hat{y}) = 1 - \hat{\pi}_{\hat{y}}$
- 熵（ID3/C4.5）： $H(\mathcal{D}) = -\sum_{c=1}^{C}\hat{\pi}_c \log \hat{\pi}_c$
- Gini指数（无需log，计算更快）：

$$Gini(\mathcal{D}) = \sum_{c=1}^{C}\hat{\pi}_c(1-\hat{\pi}_c) = \sum_{c}\hat{\pi}_c - \sum_{c}\hat{\pi}_c^2 = 1 - \sum_{c}\hat{\pi}_c^2$$

$$\hat{\pi}_c = \frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}}\mathbb{I}(y_i = c) = N_c/N$$

✓ 建树

- 令节点的样本集合为$\mathcal{D}$，对候选分裂$\boldsymbol{\theta} = (j, t_m)$，选择特征$j$，分裂阈值为$t_m$，将样本分裂成左右两个分支$\mathcal{D}_L$和$\mathcal{D}_R$

$$\mathcal{D}_L(\boldsymbol{\theta}) = \{(\mathbf{x}_i, y_i) | x_{ij} \leq t_m\}$$

$$\mathcal{D}_R(\boldsymbol{\theta}) = \{(\mathbf{x}_i, y_i) | x_{ij} > t_m\}$$

- 分裂原则：分裂后两个分支的样本越纯净越好

$$G(\mathcal{D}, \boldsymbol{\theta}) = \frac{N_{left}}{N_m} H\big(\mathcal{D}_L(\boldsymbol{\theta})\big) + \frac{N_{right}}{N_m} H\big(\mathcal{D}_R(\boldsymbol{\theta})\big)$$

不纯净性$G(\mathcal{D}, \boldsymbol{\theta})$最小: $\boldsymbol{\theta}^* = argmin_{\boldsymbol{\theta}} G(\mathcal{D}, \boldsymbol{\theta})$

例：贷款审批

$$Gini(D,A)=\frac{|D_1|}{|D|}Gini(D_1)+\frac{|D_2|}{|D|}Gini(D_2)$$

$$Gini(p)=\sum_{k=1}^{K}p_k(1-p_k)$$

$$Gini(D,A_1=\text{青年})=\frac{5}{15}\times\left(2\times\frac{2}{5}\times\left(1-\frac{2}{5}\right)\right)+\frac{10}{15}\times\left(2\times\frac{7}{10}\times\left(1-\frac{7}{10}\right)\right)=0.44$$

$Gini(D,A_1=\text{中年})=0.48$

$Gini(D,A_1=\text{老年})=0.44$

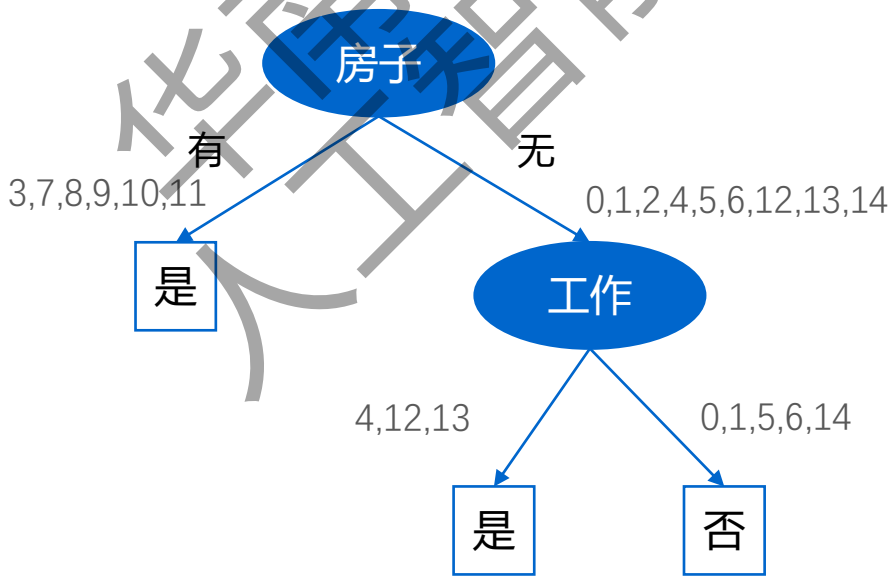
$Gini(D,A_2=\text{是})=0.32$

$Gini(D,A_3=\text{是})=0.27$

$Gini(D,A_4=\text{非常好})=0.36$

$Gini(D,A_4=\text{好})=0.47$

$Gini(D,A_4=\text{一般})=0.32$



	年龄	有工作	有房子	信用	类别
0	青年	否	否	一般	否
1	青年	否	否	好	否
2	青年	是	否	好	是
3	青年	是	是	一般	是
4	青年	否	否	一般	否
5	中年	否	否	一般	否
6	中年	否	否	好	否
7	中年	是	是	好	是
8	中年	否	是	非常好	是
9	中年	否	是	非常好	是
10	老年	否	是	非常好	是
11	老年	否	是	好	是
12	老年	是	否	好	是
13	老年	是	否	非常好	是
14	老年	否	否	一般	否

➢ CART回归

✓ 集合$\mathcal{D}$的不纯净性度量：

$$H(\mathcal{D}) = \sum_{i \in \mathcal{D}} (\bar{y} - y_i)^2$$

$$\bar{y} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} y_i$$

集合中样本的 y 值越接近越纯净

相当于损失函数取L2损失，选择最小L2损失的分裂

L2损失：$L(\hat{y}(\boldsymbol{\theta}), y) = (\hat{y}(\boldsymbol{\theta}) - y)^2 = (\bar{y} - y)^2$

预测值$\hat{y}$为样本均值 $\bar{y} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} y_i$ 时L2损失最小

3.剪枝

剪枝准则：

$$CC(T) = Err(T) + \alpha|T|$$

树的错误率　　正则因子　树的节点数目

形式同机器学习模型的目标函数：$J(\boldsymbol{\theta}, \lambda) = \sum_{i=1}^{N} L(f(\mathbf{x}_i; \boldsymbol{\theta}), y_i) + \lambda R(\boldsymbol{\theta})$

当α从0开始增大，树的一些分支被剪掉，得到不同α对应的树

训练集

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

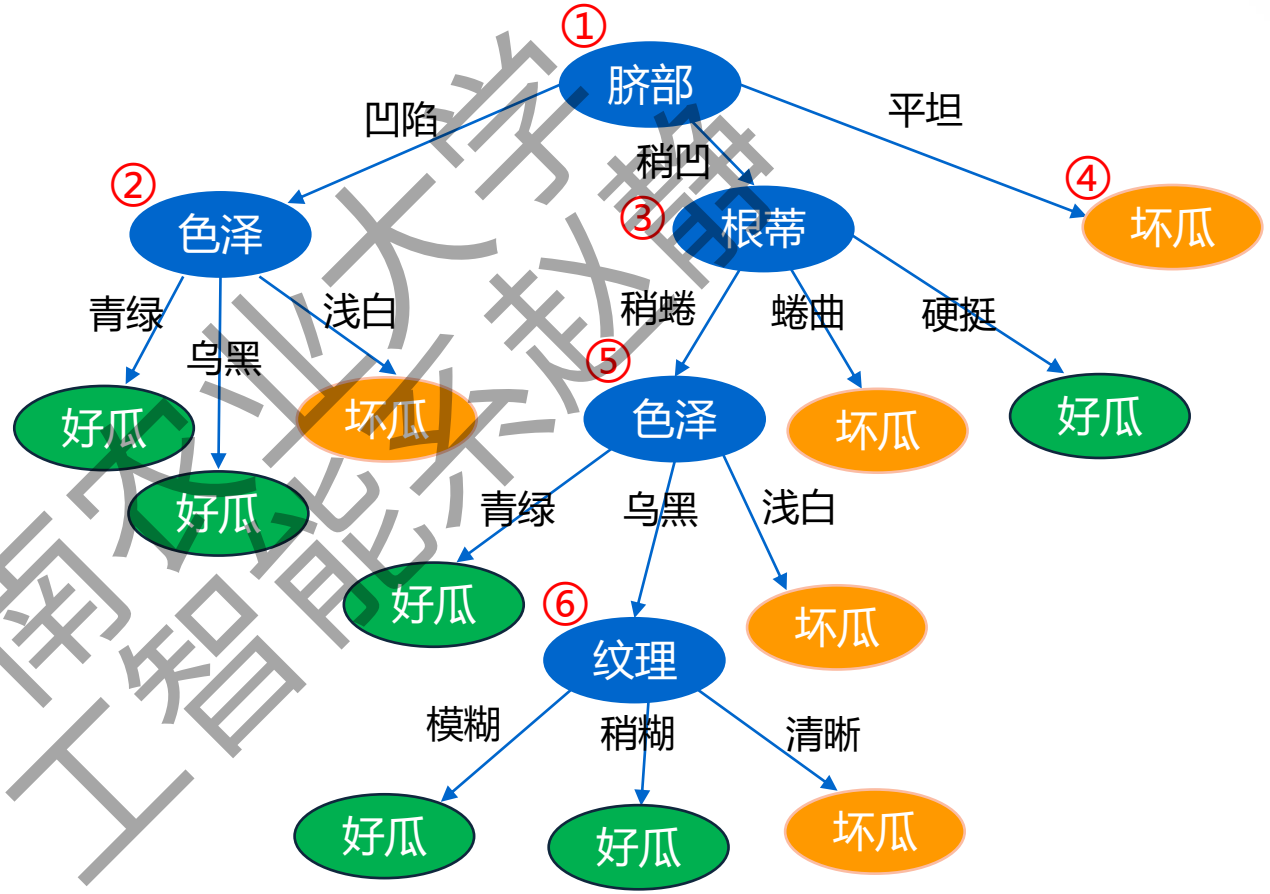
验证集

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否

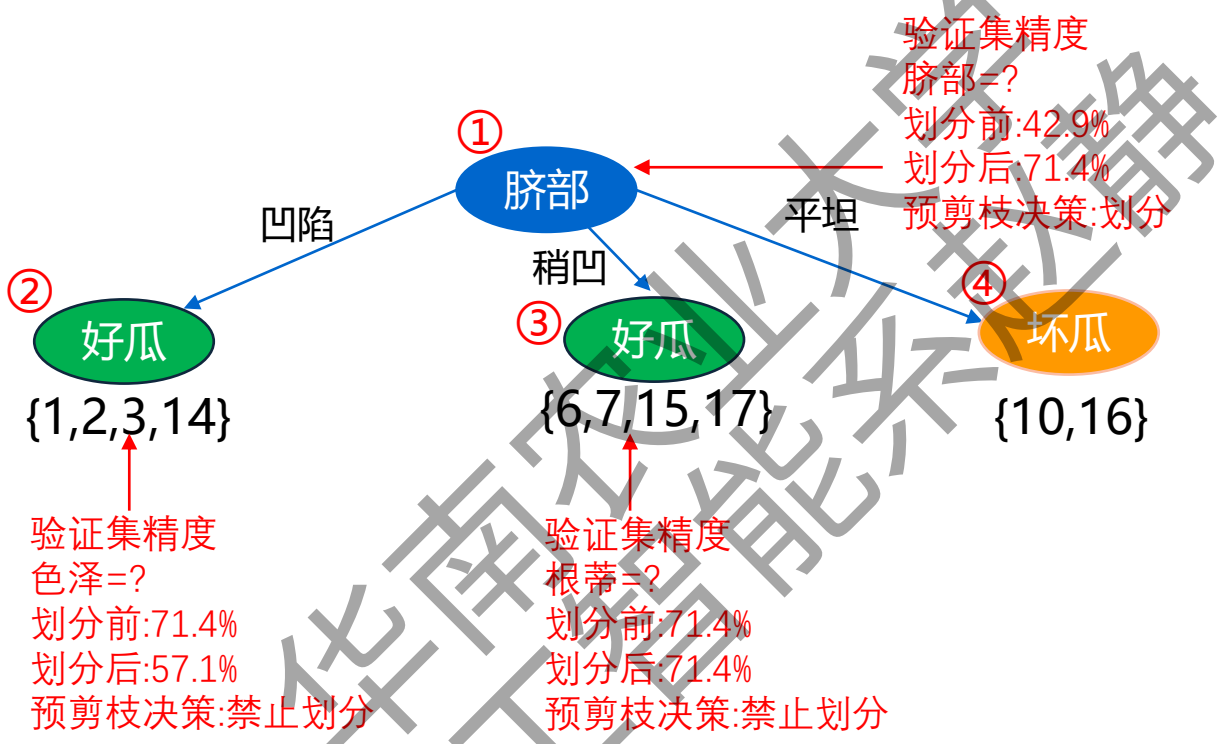
✓ 预剪枝（prepruning）

主要方法有：

- 节点划分前准确率比划分后准确率高
- 限制深度
- 叶子节点个数
- 叶子节点样本数
- 信息增益量等

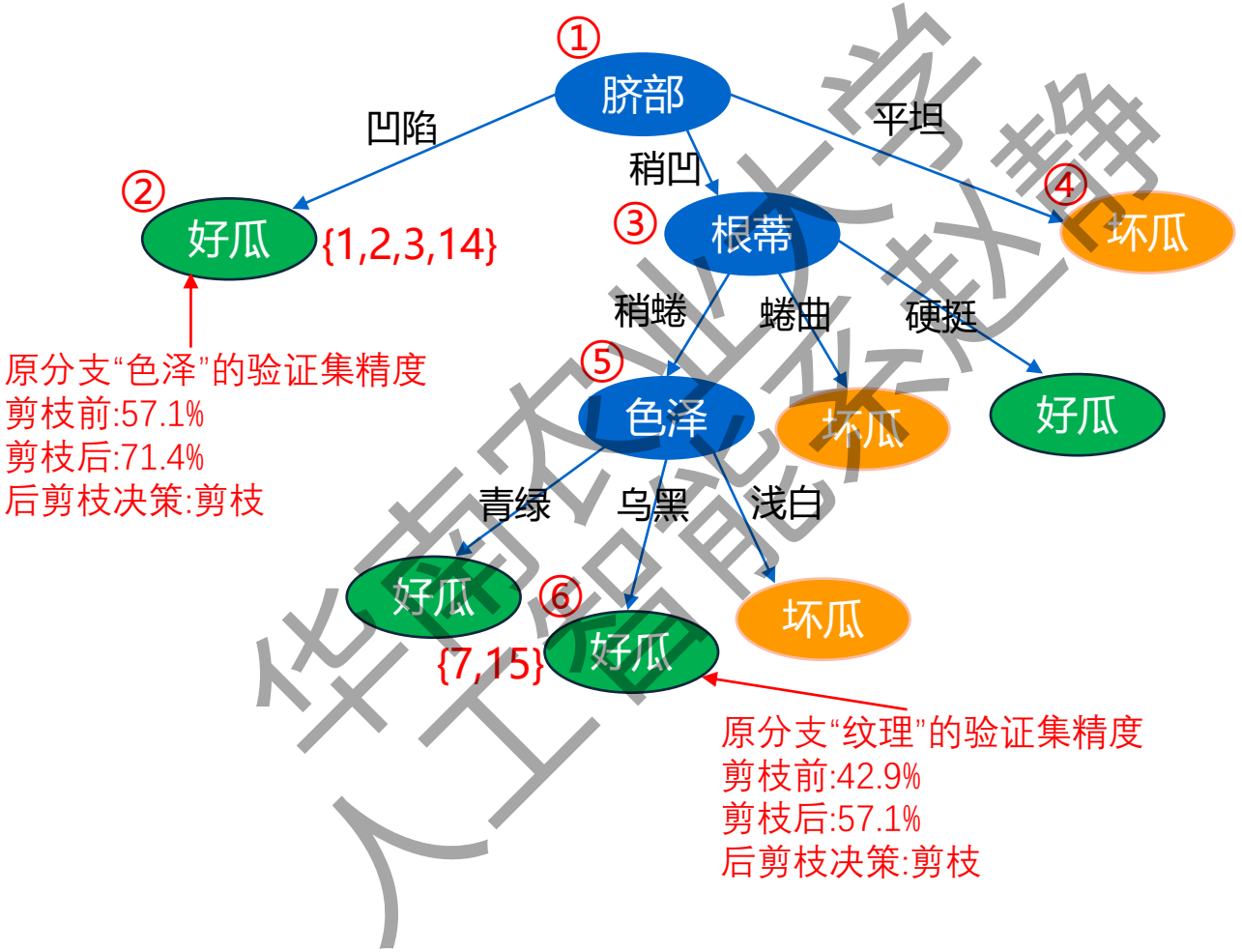


基于表生成未剪枝的决策树



预剪枝的决策树

✓ 后剪枝（ post-pruning ）



后剪枝的决策树

总结

➢ 树模型的优点

- 容易解释
- 对特征预处理要求少
  - 能处理离散值和连续值混合的输入（理论上，实际实现需根据具体工具包的要求）
  - 对特征的单调变换不敏感（只与数据的排序有关）
  - 能自动进行特征选择
  - 可处理缺失数据
- 可扩展到大数据规模

➢ 树模型的缺点

- ■正确率不高：建树过程过于贪心
  可作为Boosting的弱学习器（深度不太深）
- ■模型不稳定（方差大）：输入数据小的变化会带来树结构的变化
  Bagging：随机森林
- ■当特征数目相对样本数目太多时，容易过拟合

➢ 决策树的三种基本类型对比

建立决策树的关键，即在当前状态下选择哪个属性作为分类依据。根据不同的目标函数，决策树主要有三种算法： ID3(Iterative Dichotomiser)、C4.5、CART(Classification And Regression Tree)。

算法	支持模型	树结构	特征选择	连续值处理	缺失值处理	剪枝	特征属性多次使用
ID3	分类	多叉树	信息增益	不支持	不支持	不支持	不支持
C4.5	分类	多叉树	信息增益率	支持	支持	支持	不支持
CART	分类 回归	二叉树	基尼指数 均方差	支持	支持	支持	支持

- ID3——信息增益
- C4.5——信息增益率
- CART——Gini指数