四川大學
SICHUAN UNIVERSITY

机器学习-第六章 决策树

教师：胡俊杰 副教授

邮箱：hujunjie@scu.edu.cn

1.数据集划分

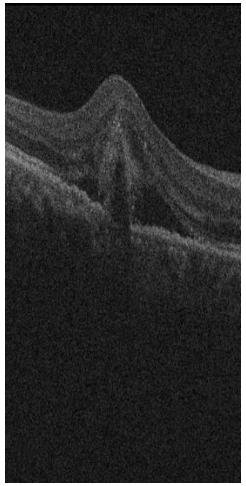
训练集（Training Set）：训练模型所使用到的数据，通过该部分数据确定模型所包含的各学习参数。

验证集（Validation Set）：有时也叫做开发集（Dev Set），用来做模型选择（model selection），评价模型的训练效果

测试集（Test Set）： 测试已经训练好的模型的性能。

- 三者划分：训练集（80%）、验证集（10%）、测试集（10%）
- 实际应用中，训练集/验证集/测试集的具体比例可调整
- 如果只划分训练集和验证集，通常训练集（80%），验证集（20%）

不平衡数据的处理

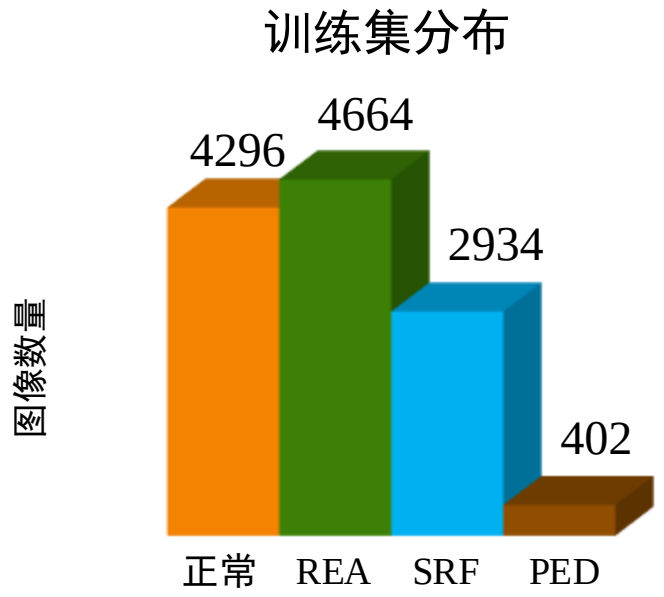


输入图像（OCT）

任务I：黄斑水肿病变分类 → 是否正常、REA、SRF、PED？

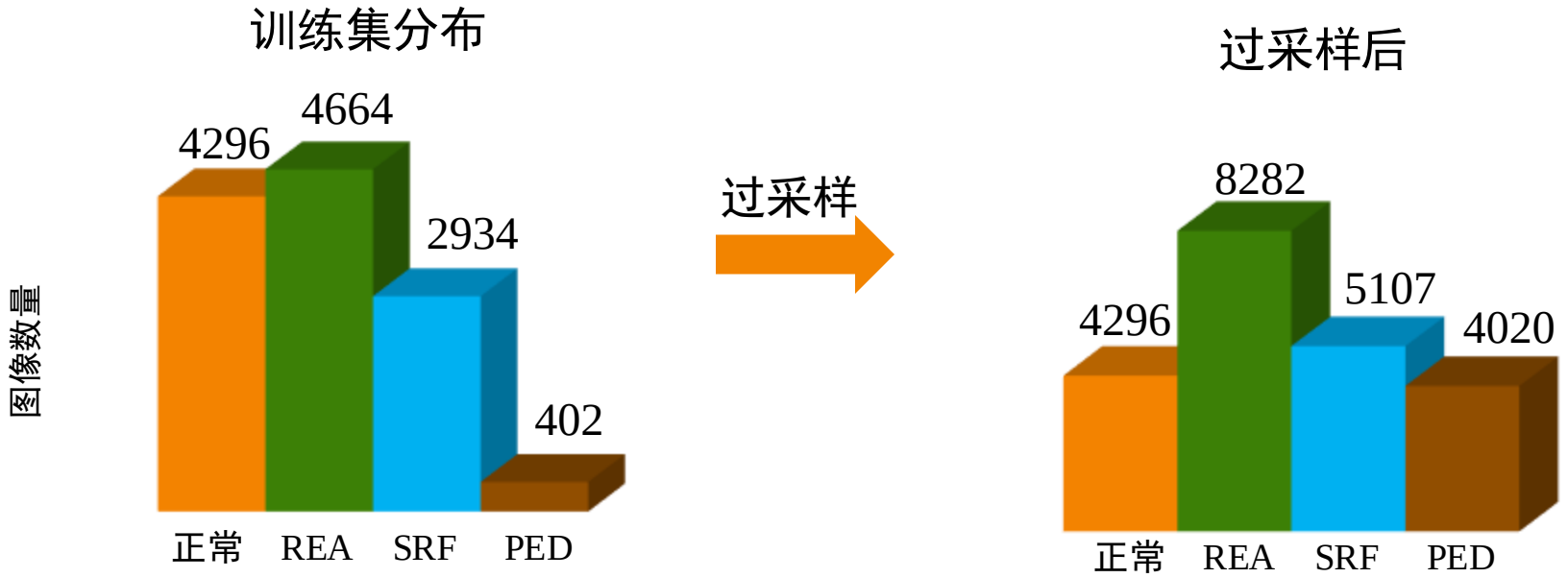
任务II：黄斑水肿病变分割 → 

识别各病灶区域



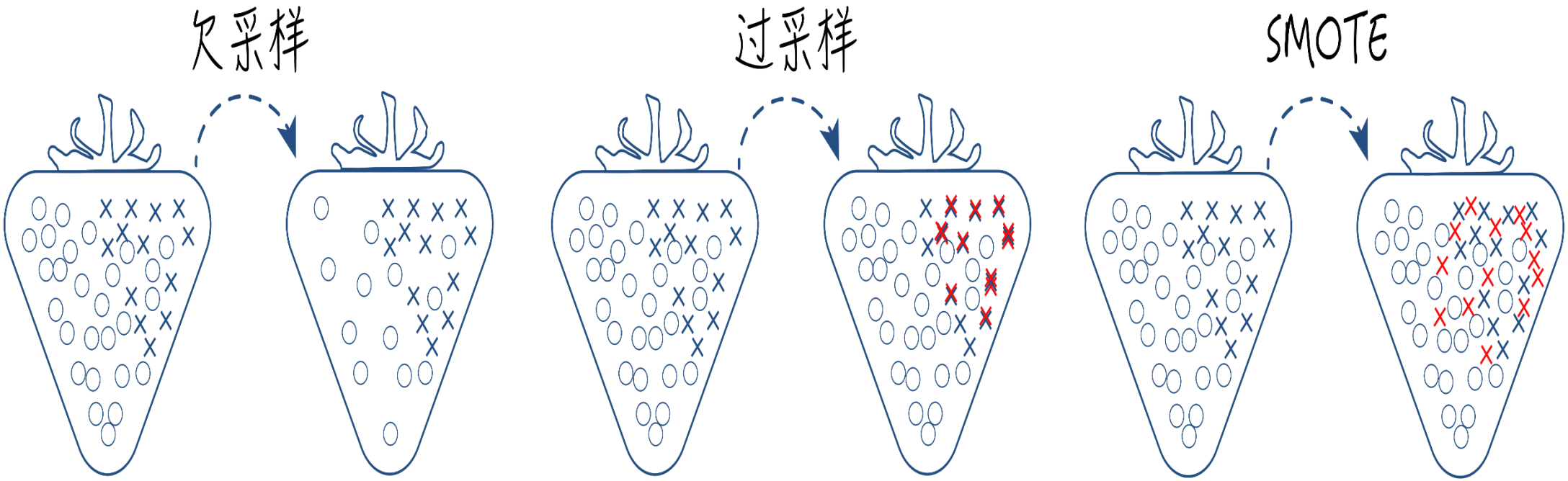
- 数据不平衡是指数据集中各类样本数量不均衡的情况
- 不加处理的话，模型会倾向于预测正常和REA类别，忽略PED类别

不平衡数据的处理



- 对数量较少的类别采用有放回的方式重复采样。由于该任务是一个多标签任务（一张图像对应多个标签），对SRF、PED类别过采样也将增加REA类别的数量
- 实际应用中，应尽可能保证各类别样本数量相近

不平衡数据的处理



- 过采样方法仅是对数量较少类别样本的简单复制，如何增加样本的多样性
- SMOTE: Synthetic Minority Over-sampling TEchnique

[*] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of artificial intelligence research, 2002, 16: 321-357.

评价指标

混淆矩阵（confusion matrix）

		预测	
		Positive	Negative
标签	Positive	TP	FN
	Negative	FP	TN

召回率：$Recall = \frac{TP}{TP+FN}$

- 针对真实的阳性样本，即在全体阳性样本中，模型预测出的比例

精准率：$Precision = \frac{TP}{TP+FP}$

- 针对模型预测为阳性的样本，即在全体模型预测为阳性的样本，真实阳性样本所占的比例

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

- 同时结合了召回率和精准率特点的综合性评价指标

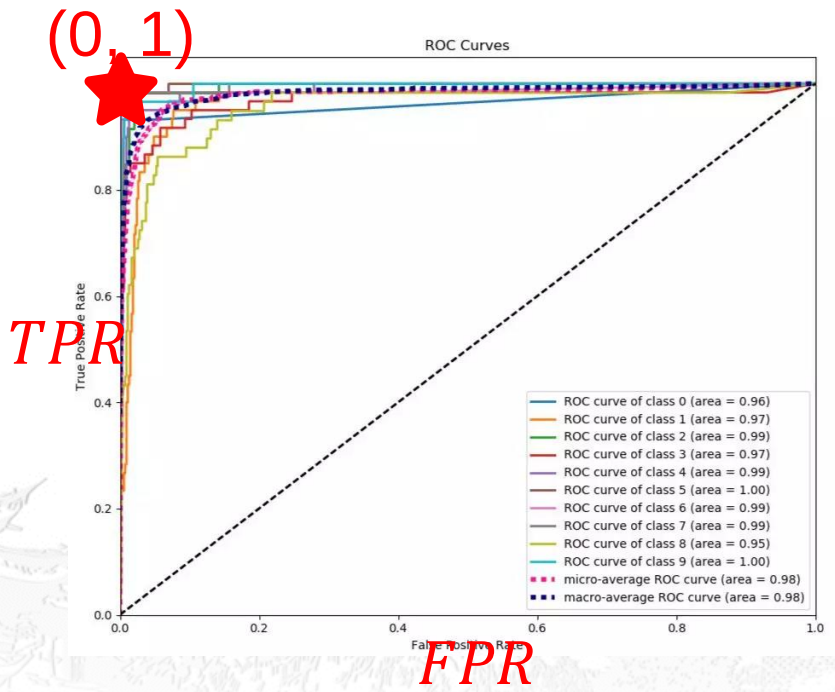
评价指标

ROC曲线， Receiver operating characteristic curve，也称为受试者工作特征曲线

$$FPR = \frac{FP}{FP + TN}$$ 假阳率：针对所有阴性样本，被错误预测为阳性的比例 ↓ 越小越好

$$TPR = \frac{TP}{TP + FN}$$ 真阳率：针对所有阳性样本，被正确预测为阳性的比例 ↑ 越大越好

- 随着阈值的不同，FPR和TPR都在同步变化，(FPR, TPR)所构成的曲线则称为ROC曲线
- ROC曲线与坐标轴围成的面积，称为AUC（Area Under Curve, 曲线下面积），面积越大，则模型性能越好

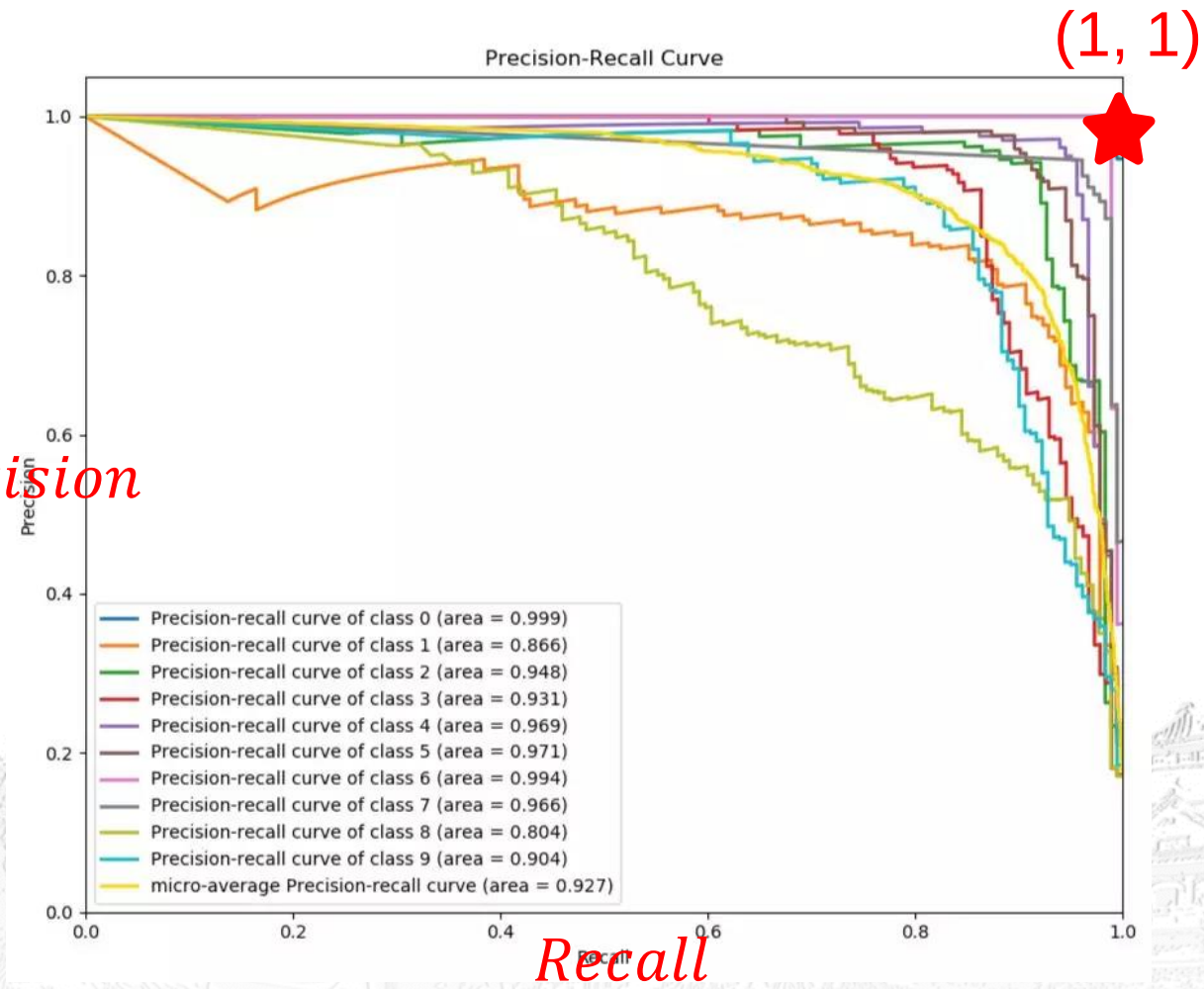


评价指标

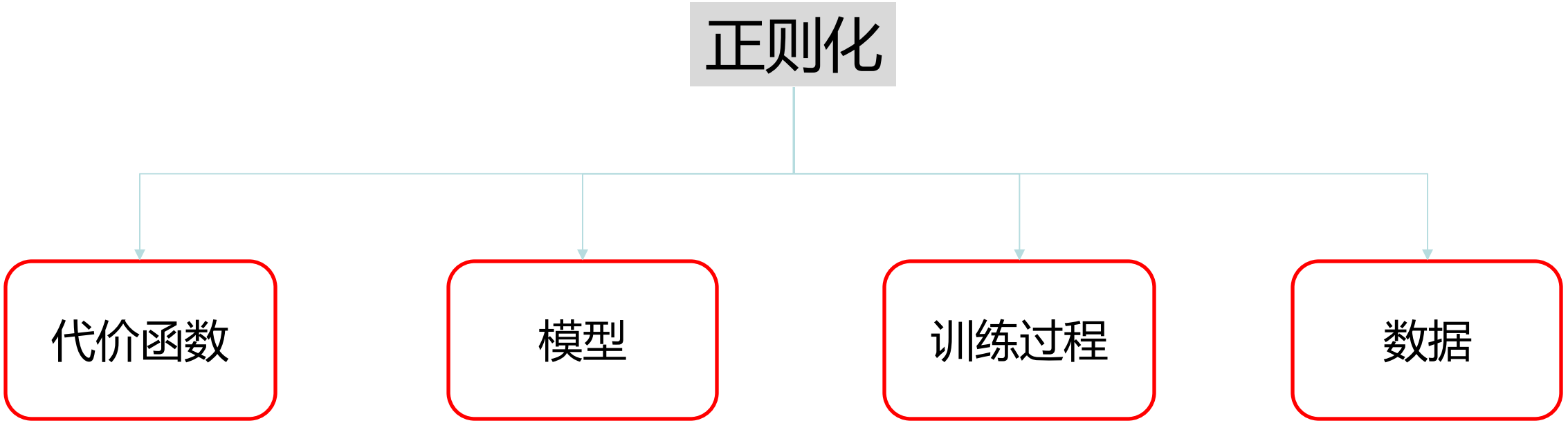
PR（Precision-Recall）曲线

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$



正则化（Regularization）



.....

正则化（Regularization）-代价函数

正则化系数

L_1正则化：$J(w)=\frac{1}{2}\sum_{i=1}^{m}(h(x_i)-y_i)^2+\lambda\sum_{j=1}^{n}|w_j|$，Lasso Regression（Lasso回归）

L_2正则化：$J(w)=\frac{1}{2}\sum_{i=1}^{m}(h(x_i)-y_i)^2+\lambda\sum_{j=1}^{n}w_j^2$，Ridge Regression （岭回归）

Elastic Net：$J(w)=\frac{1}{2}\sum_{i=1}^{m}(h(x_i)-y_i)^2+\lambda(\rho\cdot\sum_{j=1}^{n}|w_j|+(1-\rho)\cdot\sum_{j=1}^{n}w_j^2)$

(弹性网络)

比例系数

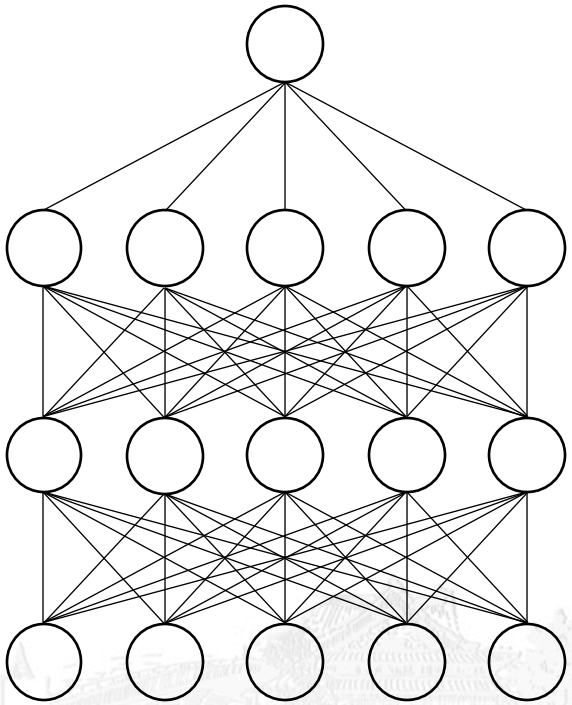
其中：

- λ为正则化系数，调整正则化项与训练误差的比例，λ>0。
- 1≥ρ≥0为比例系数，调整L_1正则化与L_2正则化的比例。

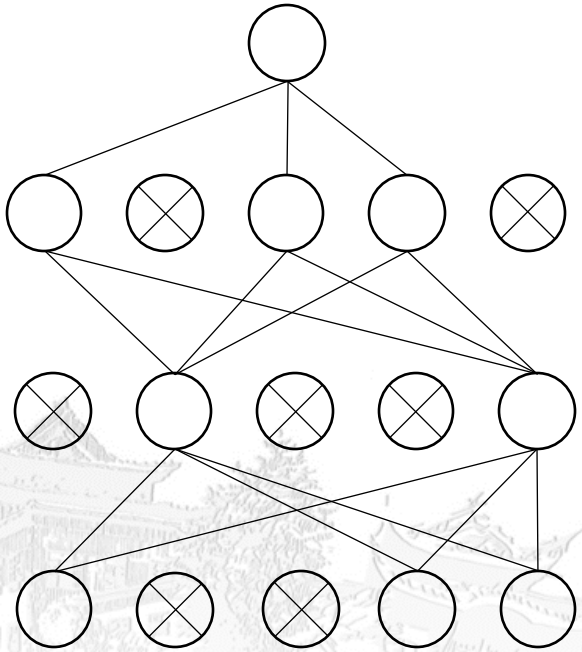
正则化-模型

Dropout（深度神经网络训练过程中常用的正则化）

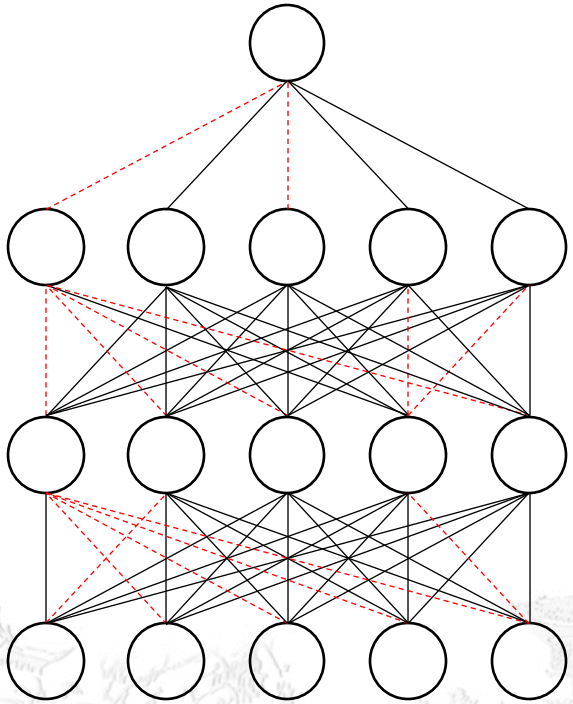
- 训练过程：以概率p随机地禁止/激活每个神经元（伯努利分布）
- 测试过程：保留全体神经元，但激活值强度乘以p



标准的深度神经网络



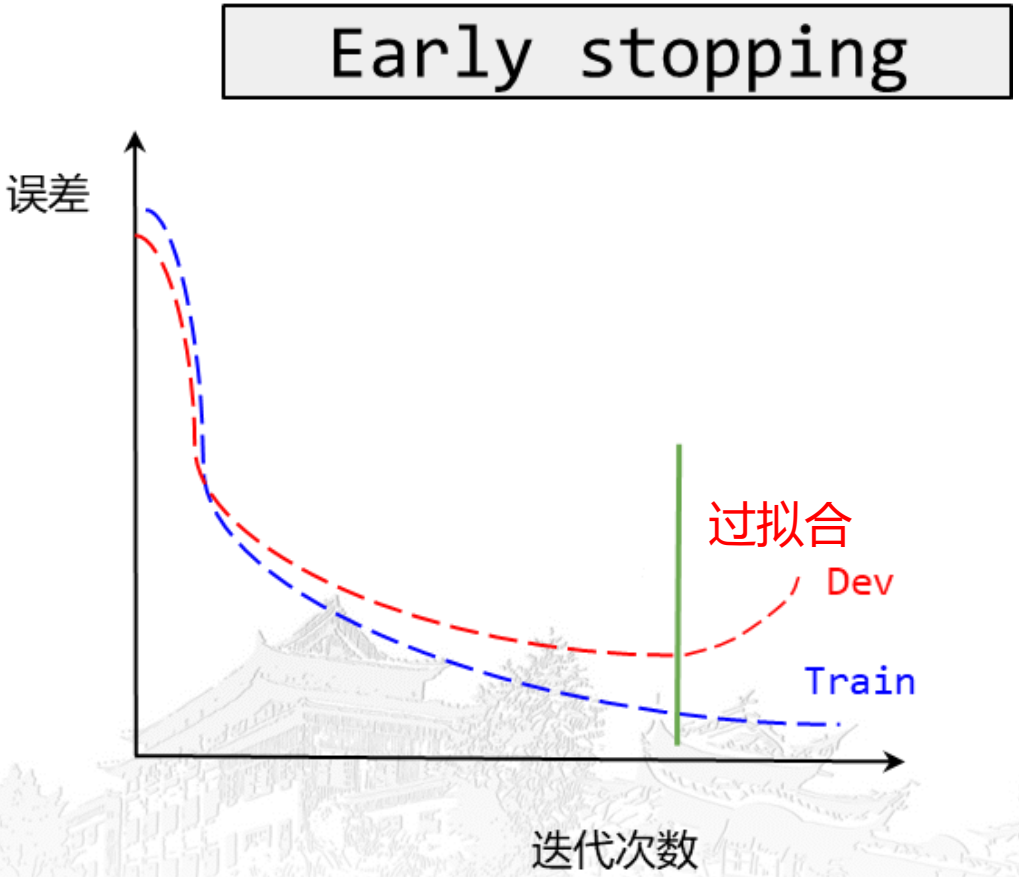
Dropout



Drop connect

正则化-训练过程

Early stopping代表提早停止训练模型

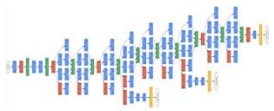


正则化-数据

- 数据收集困难
- 标注人力成本高

大数据

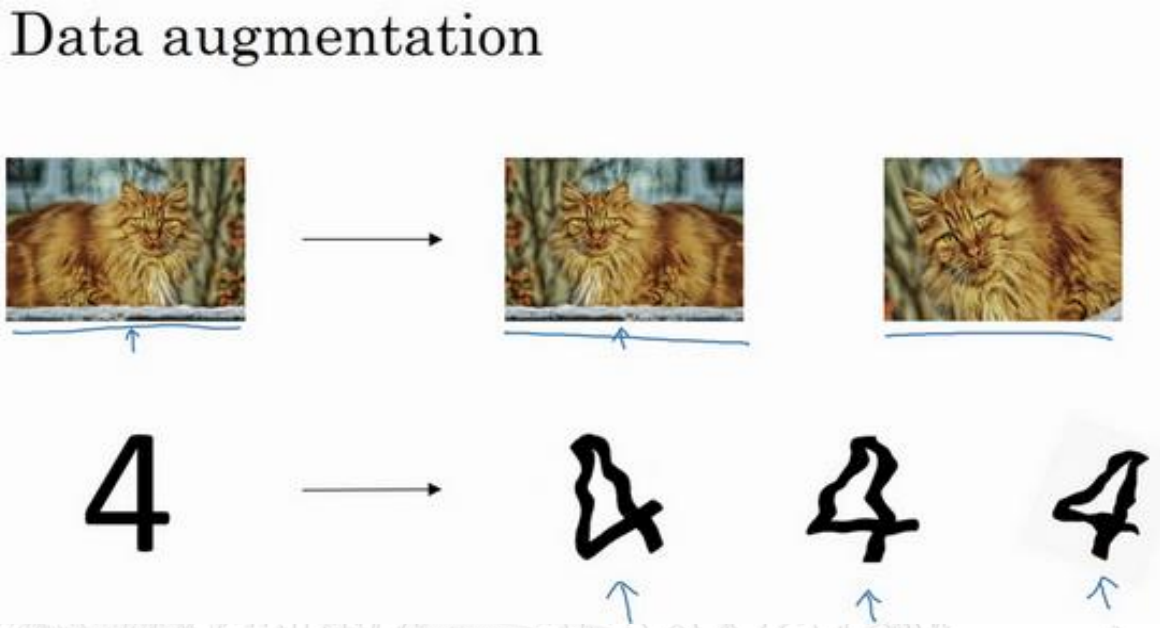
大模型



大算力

数据增广：人为增强数据的多样性

- 颜色空间：亮度、灰度、对比度等
- 几何空间：旋转、平移、缩放、弹性形变等



4.CART算法

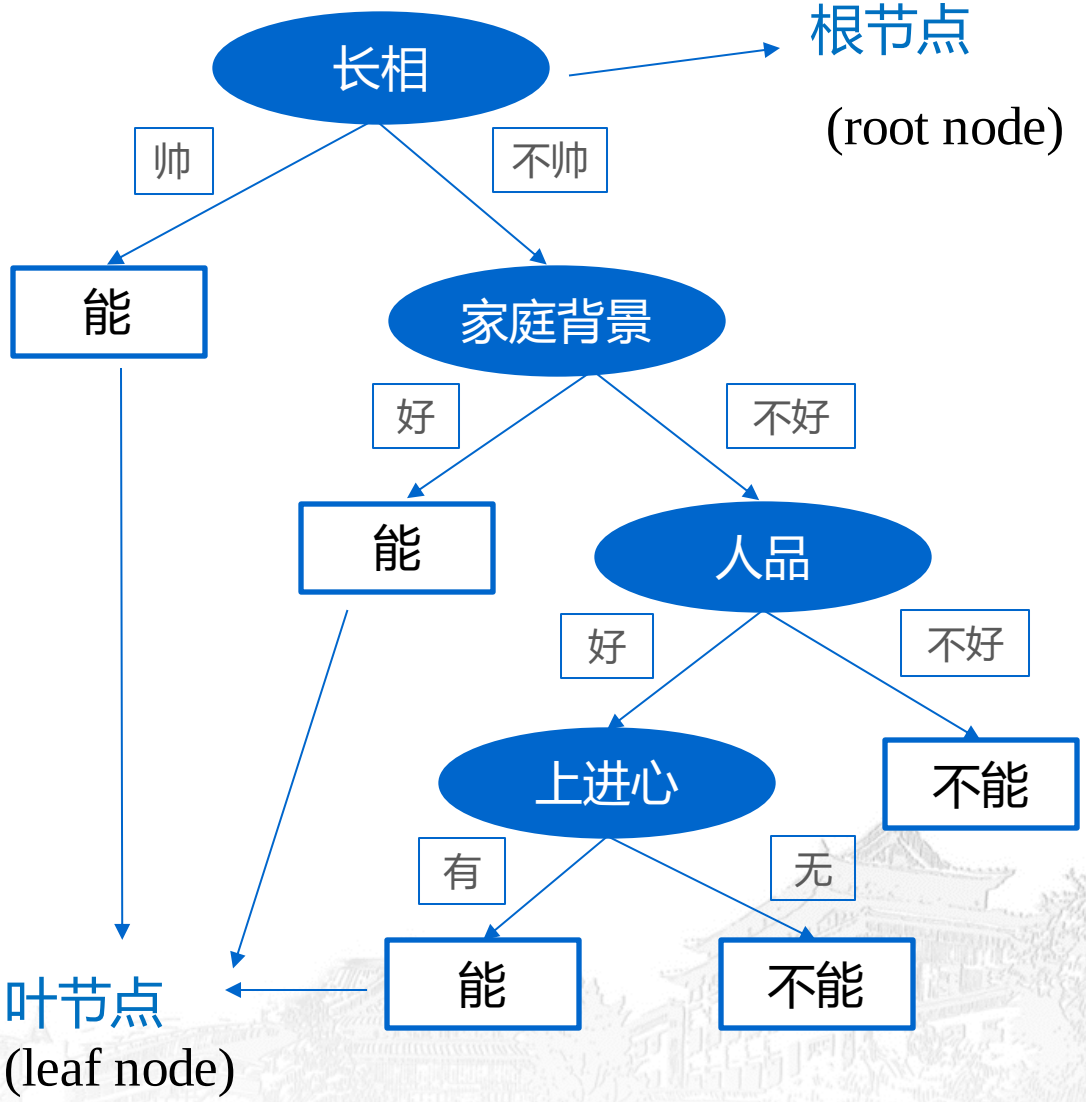
01 决策树原理

02 ID3算法

03 C4.5算法

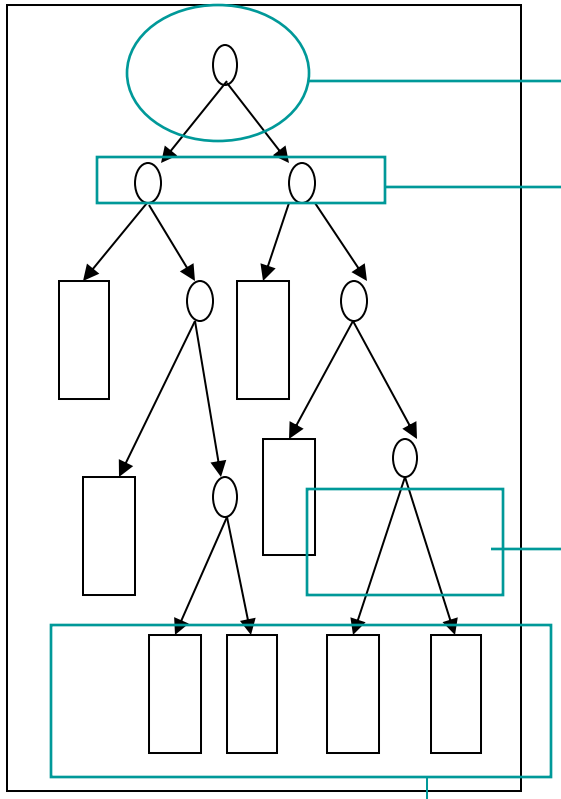
04 CART算法

1.决策树原理



- 决策树：从训练数据中学习得出一个树状结构的模型
- 决策树属于判别模型
- 决策树的决策过程是从根节点开始，测试待分类项对应的特征属性，并按照其值选择输出分支，直到叶节点，将叶节点存放的类别作为决策结果

1.决策树原理



根节点 (root node)

非叶节点 (non-leaf node)

(代表分类属性，如上进心)

分支 (branches)

(代表分类属性的取值，如帅/不帅)

叶节点 (leaf node)

(代表分类结果，如能/不能)

- 决策树算法是一种归纳分类算法，它通过对训练集的学习，挖掘出有用的规则
- 决策树归纳的基本算法是贪心算法，自顶向下来构建树
- 贪心算法：在每一步选择中都采取在当前状态下最优的选择
- 在决策树的生成过程中，属性选择的度量是关键

1.决策树原理

决策树的三种基本类型

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。建立决策树主要有以下三种算法： ID3(Iterative Dichotomiser)、C4.5、CART(Classification And Regression Tree)

算法	支持任务	树结构	特征选择
ID3	分类	多叉树	信息增益
C4.5	分类	多叉树	信息增益率
CART	分类 回归	二叉树	基尼指数 均方误差

4.CART算法

01 决策树原理

02 ID3算法

03 C4.5算法

04 CART算法

2. ID3算法

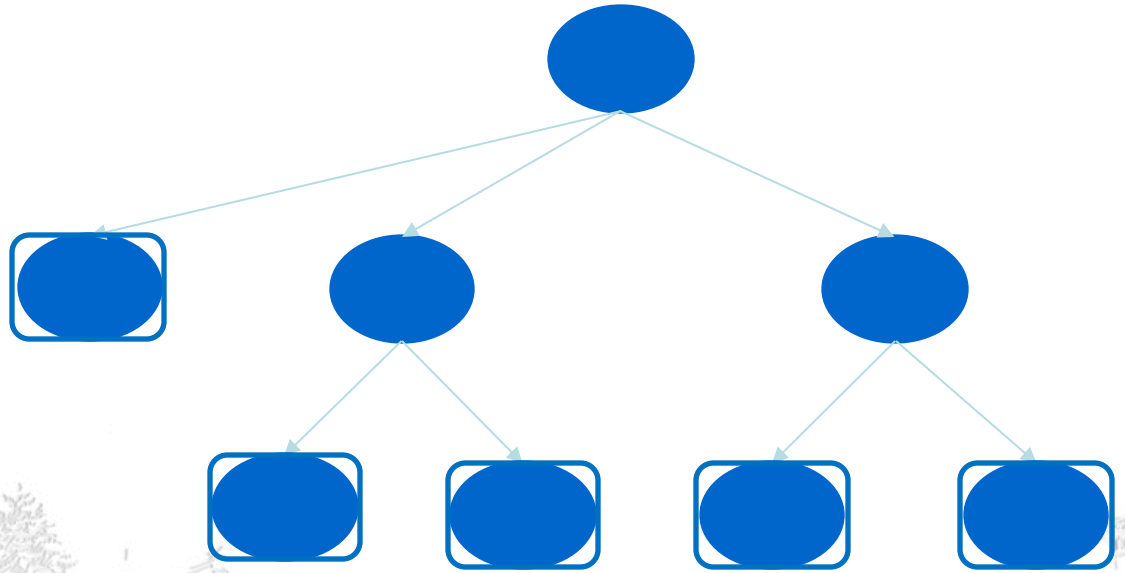
ID3 算法

- ID3算法最早是由罗斯.昆兰（J. Ross Quinlan）提出的一种决策树构建算法，算法的核心是信息熵
- ID3算法是以信息增益为衡量标准，实现对数据的归纳分类
- ID3算法计算每个属性的信息增益，并选取具有最高增益的属性作为给定的分类属性

2.ID3算法

ID3 算法核心思路

- 从根节点开始，计算所有属性的信息增益，选择信息增益最大的属性作为分类节点（如性别）
- 根据不同取值（如男/女）建立子节点
- 对子节点递归调用以上方法，直至属性为空



熵（Entropy）

- 熵（Entropy）：代表随机变量不确定性的度量。熵越大，随机变量的不确定性越大

设X是有限个取值的离散随机变量，其概率分布是$P(X = x_i) = p_i, i = 1,2, \dots, n$

随机变量X的熵定义为 $$H(X) = -\sum_{i=1}^{n} p_i \log p_i$$

熵依赖于X的分布，也可记成 $$H(p) = -\sum_{i=1}^{n} p_i \log p_i$$

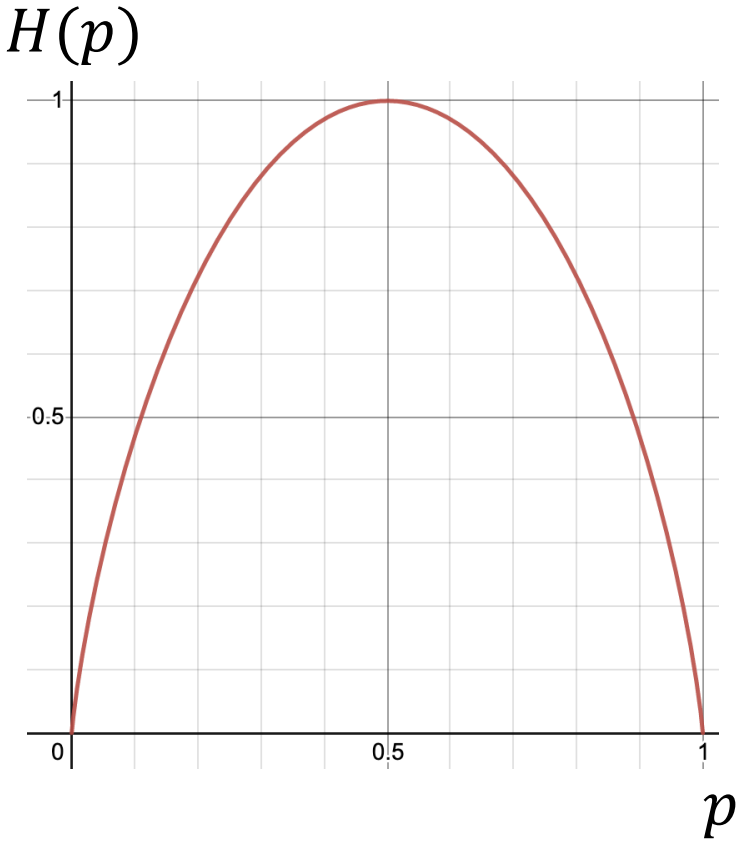
熵（Entropy）

- 设随机变量X只有1和0两种取值，即

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

$$H(p)=-p\log_2 p-(1-p)\log_2(1-p)$$

- 当$p=0$或$p=1$时，熵为0，随机变量没有不确定性
- 当$p=0.5$时，熵为1，随机变量不确定性最大



条件熵（Conditional entropy）

- 对于随机变量(X, Y)，其联合概率分布为

$$P(X = x_i, Y = y_i) = p_{ij}, i = 1,2, \dots, n, j = 1,2, \dots, m$$

- 条件熵$H(Y|X)$代表已知随机变量X的条件下随机变量Y的不确定性
- 条件熵的定义：X给定的条件下Y的条件概率分布的熵对X的数学期望

$$H(Y|X) = \sum_{i=1}^{n} p_i H(Y|X = x_i)$$

- 当熵和条件熵中的概率由数据估计得到时，所对应的熵与条件熵称为经验熵和经验条件熵
- 经验条件熵越小，代表已知X后Y的不确定性越小

信息增益（Information gain）

- 信息增益：得知特征X的信息使得类Y的信息不确定性减少的程度
- 特征A对训练数据集D的信息增益记为$g(D,A)$，其定义为经验熵$H(D)$与给定特征A条件下D的经验条件熵$H(D|A)$之差：

$$g(D,A) = H(D) - H(D|A)$$

- 对于分类任务而言，不同特征具有不同的信息增益。特征的信息增益越大，代表该特征具有更强的分类能力

经验熵

输入：年龄、工作、房子、信用

输出：是否同意贷款

经验熵 $$H(D) = -\sum_{k=1}^{K} \frac{|C_k|}{|D|} \log_2 \frac{|C_k|}{|D|}$$

D: 训练数据集, $|D|$: 训练样本数量

K: 类别数量, $|C_k|$: 类别C_k的样本数量 $$\sum_{k=1}^{K} |C_k| = |D|$$

右边数据中：

样本数量	同意	不同意	经验熵
15	9	6	0.971

$$H(D) = -\sum_{k=1}^{K} \frac{|C_k|}{|D|} \log_2 \frac{|C_k|}{|D|} = -\frac{9}{15} \log_2 \frac{9}{15} - \frac{6}{15} \log_2 \frac{6}{15} = 0.971$$

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

经验熵

按年龄划分

年龄	数量	同意	不同意	经验熵
青年	5	2	3	0.971
中年	5	3	2	0.971
老年	5	4	1	0.722

A_1	年龄
A_2	有工作
A_3	有房子
A_4	信用

- D: 训练数据集, $|D|$: 训练样本数量
- A: 某一特征（如年龄、有工作、有房子、信用）
- 假设A有n个不同的取值$\{a_1, a_2, \dots, a_n\}$，将D划分为n个子集 $D_1, D_2, \dots, D_n$
- $|D_i|$: D_i子集样本数目，$\sum_{i=1}^{n}|D_i| = |D|$
- D_{ik}: D_i子集中属于类C_k样本的集合，$|D_{ik}|$: D_{ik}包含的样本数目

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

经验熵

按年龄划分

年龄	数量	同意	不同意	经验熵
青年	5	2	3	0.971
中年	5	3	2	0.971
老年	5	4	1	0.722

A_1	年龄
A_2	有工作
A_3	有房子
A_4	信用

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

青年：$H(D_1) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971$

中年：$H(D_2) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.971$

老年：$H(D_3) = -\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5} = 0.722$

经验条件熵

经验条件熵 $H(D|A)=\sum_{i=1}^{n}\frac{|D_i|}{|D|}H(D_i)$

- A: 某一特征（如年龄、有工作、有房子、信用）
- 假设A有n个不同的取值$\{a_1,a_2,\dots,a_n\}$，将D划分为n个子集$D_1,D_2,\dots,D_n$

$$H(D|\text{年龄})=\sum_{i=1}^{n}\frac{|D_i|}{|D|}H(D_i)$$

$$=\frac{5}{15}\times 0.971+\frac{5}{15}\times 0.971+\frac{5}{15}\times 0.722$$

$$=0.888$$

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

信息增益

信息增益 $g(D,A)=H(D)-H(D|A)$

其中，$H(D|A)=\sum_{i=1}^{n}\frac{|D_i|}{|D|}H(D_i)=-\sum_{i=1}^{n}\frac{|D_i|}{|D|}\sum_{k=1}^{K}\frac{|D_{ik}|}{|D_i|}\log_2\frac{|D_{ik}|}{|D_i|}$

n是特征A的取值个数，K为类别数目

1. 年龄A_1对D的信息增益

$g(D,A_1)=H(D)-H(D|A_1)$

$=0.971-0.888=0.083$

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

信息增益

信息增益 $g(D,A)=H(D)-H(D|A)$

有工作	数量	同意	不同意
是(D_1)	5	5	0
否(D_2)	10	4	6

2. 有工作A_2对D的信息增益

$$g(D,A_2)=H(D)-H(D|A_2)$$

$$=0.971-\left[\frac{5}{15}H(D_1)+\frac{10}{15}H(D_2)\right]$$

$$=0.971-\left[\frac{5}{15}\times\left[-\frac{5}{5}\log_2\frac{5}{5}-\frac{0}{5}\log_2\frac{0}{5}\right]+\frac{10}{15}\left[-\frac{4}{10}\log_2\frac{4}{10}-\frac{6}{10}\log_2\frac{6}{10}\right]\right]$$

$$=0.324$$

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

信息增益

信息增益 $g(D,A)=H(D)-H(D|A)$

3. 有房子A_3对D的信息增益

有房子	数量	同意	不同意
是(D_1)	6	6	0
否(D_2)	9	3	6

$$g(D,A_3)=H(D)-H(D|A_3)$$

$$H(D)=-\sum_{k=1}^{K}\frac{|C_k|}{|D|}\log_2\frac{|C_k|}{|D|}=0.971$$

$$H(D|A)=\sum_{i=1}^{n}\frac{|D_i|}{|D|}H(D_i)=-\sum_{i=1}^{n}\frac{|D_i|}{|D|}\sum_{k=1}^{K}\frac{|D_{ik}|}{|D_i|}\log_2\frac{|D_{ik}|}{|D_i|}$$

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

信息增益

信息增益 $g(D,A)=H(D)-H(D|A)$

3. 有房子A_3对D的信息增益

有房子	数量	同意	不同意
是(D_1)	6	6	0
否(D_2)	9	3	6

$$g(D,A_3)=H(D)-H(D|A_3)$$

$$=0.971-\left[\frac{6}{15}H(D_1)+\frac{9}{15}H(D_2)\right]$$

$$=0.971-\left[\frac{6}{15}\times\left[-\frac{6}{6}\log_2\frac{6}{6}-\frac{0}{6}\log_2\frac{0}{6}\right]+\frac{9}{15}\left[-\frac{3}{9}\log_2\frac{3}{9}-\frac{6}{9}\log_2\frac{6}{9}\right]\right]$$

$$=0.420$$

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

信息增益

信息增益 $g(D,A)=H(D)-H(D|A)$

4. 信用A_4对D的信息增益

信用	数量	同意	不同意
一般(D_1)	5	1	4
好(D_2)	6	4	2
非常好(D_3)	4	4	0

$$g(D,A_4)=H(D)-H(D|A_4)$$

$$H(D)=-\sum_{k=1}^{K}\frac{|C_k|}{|D|}\log_2\frac{|C_k|}{|D|}=0.971$$

$$H(D|A)=\sum_{i=1}^{n}\frac{|D_i|}{|D|}H(D_i)=-\sum_{i=1}^{n}\frac{|D_i|}{|D|}\sum_{k=1}^{K}\frac{|D_{ik}|}{|D_i|}\log_2\frac{|D_{ik}|}{|D_i|}$$

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

信息增益

信息增益 $g(D,A)=H(D)-H(D|A)$

信用	数量	同意	不同意
一般(D_1)	5	1	4
好(D_2)	6	4	2
非常好(D_3)	4	4	0

4. 信用A_4对D的信息增益

$$g(D,A_4)=H(D)-H(D|A_4)$$

$$=0.971-\left[\frac{5}{15}H(D_1)+\frac{6}{15}H(D_2)+\frac{4}{15}H(D_3)\right]$$

$$=0.971-\left[\frac{5}{15}\times\left[-\frac{1}{5}\log_2\frac{1}{5}-\frac{4}{5}\log_2\frac{4}{5}\right]+\frac{6}{15}\left[-\frac{4}{6}\log_2\frac{4}{6}-\frac{2}{6}\log_2\frac{2}{6}\right]+0\right]$$

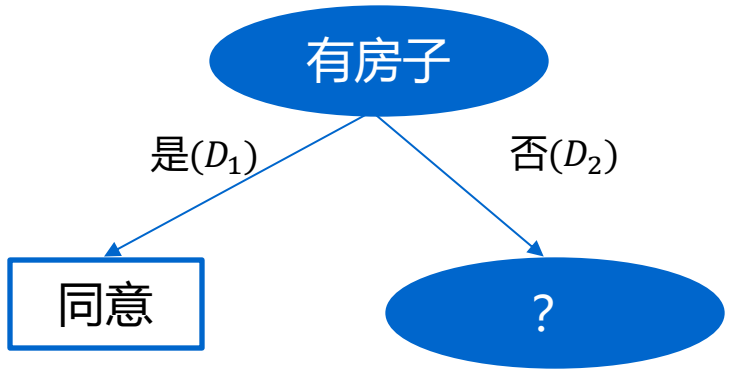
$$=0.363$$

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

信息增益

特征	信息增益
年龄	0.083
有工作	0.324
有房子	0.420
信用	0.363

- 特征有房子信息增益最大，选择该特征作为分类节点



- 有房子的样本均同意，因此为叶结点
- 针对D_2，计算年龄、有工作、信用的信息增益

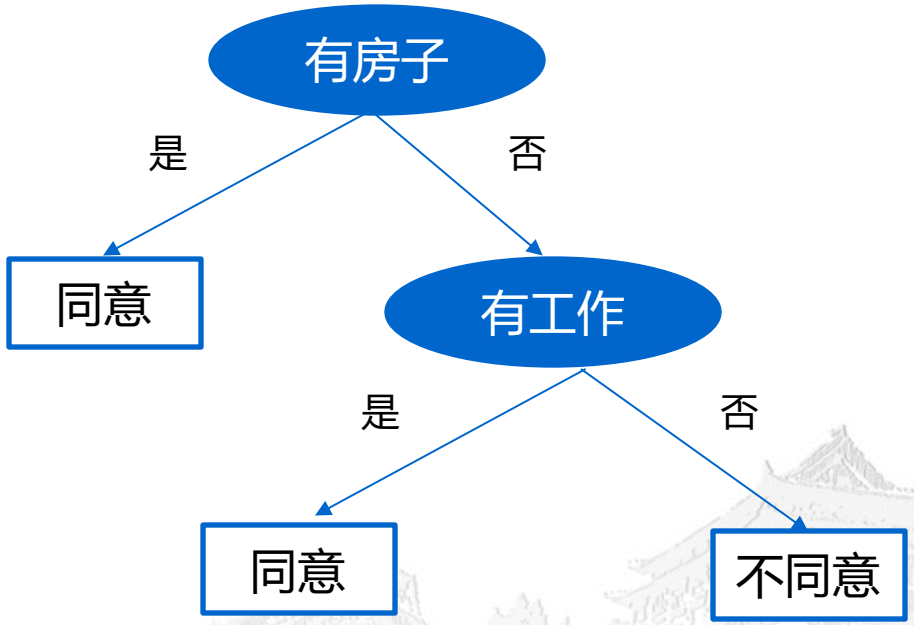
	年龄	有工作	有房子	信用	类别
1	青年	否	否	一般	不同意
2	青年	否	否	好	不同意
3	青年	是	否	好	同意
4	青年	是	是	一般	同意
5	青年	否	否	一般	不同意
6	中年	否	否	一般	不同意
7	中年	否	否	好	不同意
8	中年	是	是	好	同意
9	中年	否	是	非常好	同意
10	中年	否	是	非常好	同意
11	老年	否	是	非常好	同意
12	老年	否	是	好	同意
13	老年	是	否	好	同意
14	老年	是	否	非常好	同意
15	老年	否	否	一般	不同意

信息增益

$$g(D_2, A_1) = H(D_2) - H(D_2|A_1) = 0.918 - 0.667 = 0.251$$

$$g(D_2, A_2) = H(D_2) - H(D_2|A_2) = 0.918 - 0 = 0.918$$

$$g(D_2, A_4) = H(D_2) - H(D_2|A_4) = 0.918 - 0.444 = 0.474$$



- 有工作的样本均同意，因此为叶结点
- 没有工作的样本均不同意，因此为叶结点

	年龄 (A_1)	有工作 (A_2)	有房子 (A_3)	信用 (A_4)	类别
1	青年	否	否	一般	不同意
2	青年	否	否	好	不同意
3	青年	是	否	好	同意
4	青年	否	否	一般	不同意
5	中年	否	否	一般	不同意
6	中年	否	否	好	不同意
7	老年	是	否	好	同意
8	老年	是	否	非常好	同意
9	老年	否	否	一般	不同意

- 只用了两个特征完成了决策树的建立

2.ID3算法

ID3 算法

输入：训练数据集D，属性集合A，信息增益阈值ε

输出：决策树T

（1）若D中所有实例属于同一类C_k，则T为单节点树，并将C_k作为该节点的类标记，返回T

（2）若A为空集，则T为单节点树，并将D中实例最多的类C_k作为该节点的类标记，返回T

（3）若A不为空集，则计算A中各属性对D的信息增益，选择信息增益最大的属性A_g

（4）如果A_g的信息增益小于阈值ε，则T为单节点树，将D中实例最多的类C_k作为该节点类标记，返回T

（5）否则对A_g的每一可能值a_i，依$A_g = a_i$将D划分为若干非空子集D_i，将D_i中实例数最大的类作为标记，构建子节点，由节点及其子节点构成树T，并返回T

（6）对第i个子节点，以D_i为训练集，以$A - \{A_g\}$为属性集合，递归地调用（1）-（5）步，得到子树T_i，返回T_i

4.CART算法

信息增益

特征	信息增益
年龄	0.083
有工作	0.324
有房子	0.420
信用	0.363
编号	0.971

$$g(D,\text{编号}) = H(D) - H(D|\text{编号})$$

$$= 0.971 - \left[\frac{1}{15}H(D_1) + \frac{1}{15}H(D_2) + \cdots + \frac{1}{15}H(D_{15})\right]$$

$$= 0.971$$

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

3.C4.5算法

C4.5 算法

C4.5 算法对 ID3 算法的改进

- 用信息增益比来选择属性。ID3选择属性用的是子树的信息增益，而C4.5用的是信息增益比
- 在决策树构造过程中进行剪枝

信息增益比

信息增益比 $g_R(D,A)=\dfrac{g(D,A)}{H_A(D)}$ 训练数据集D关于特征A的熵

其中，$H_A(D)=-\sum_{i=1}^{n}\dfrac{|D_i|}{|D|}log_2\dfrac{|D_i|}{|D|}$，$n$是特征$A$的取值个数

$$g(D,A=\text{年龄})=H(D)-H(D|A=\text{年龄})=0.971-0.888=0.083$$

$$g_R(D,A=\text{年龄})=\frac{g(D,A=\text{年龄})}{H_A(D)}=\frac{0.083}{-\sum_{i=1}^{n}\frac{|D_i|}{|D|}\log_2\frac{|D_i|}{|D|}}$$

$$=\frac{0.083}{-\frac{5}{15}\log_2\frac{5}{15}-\frac{5}{15}\log_2\frac{5}{15}-\frac{5}{15}\log_2\frac{5}{15}}=0.052$$

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

信息增益 $g(D,A)=H(D)-H(D|A)$

C4.5的剪枝

决策树的过拟合

- 为了尽可能正确分类训练样本，节点的划分过程会不断重复直到不能再分，这样就可能对训练样本学习的“太好”了，把训练样本的一些特点当做所有数据都具有的一般性质，从而导致过拟合
- 剪枝使用类别代替原本基于属性的判断过程，让树更简单
- 通过剪枝处理去掉一些分支来降低过拟合的风险
- 剪枝的基本策略有“预剪枝”（prepruning）和“后剪枝”（postpruning）

C4.5的剪枝

预剪枝（prepruning）

预剪枝不仅可以降低过拟合的风险而且还可以减少训练时间，但另一方面它是基于“贪心”策略，仅考虑当前最优，会带来欠拟合风险

- 划分验证集来评估决策树的分类准确率
- 比较节点展开前后验证集的准确率，根据准确率的高低决定是否展开该节点

训练集

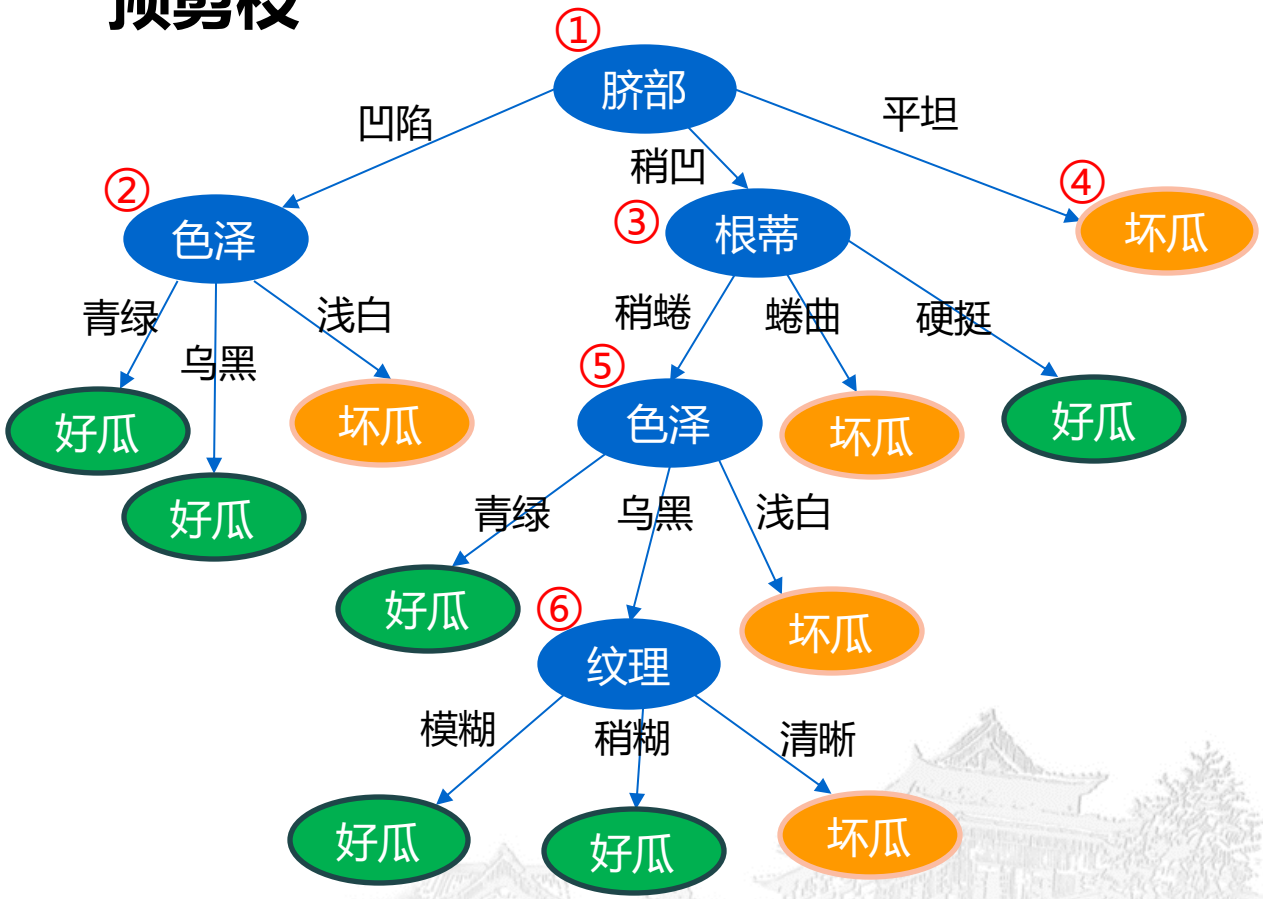
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

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

C4.5的剪枝

预剪枝



基于训练集生成未剪枝的决策树

剪枝策略

在节点划分前来确定是否继续增长，及早停止增长

C4.5的剪枝

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

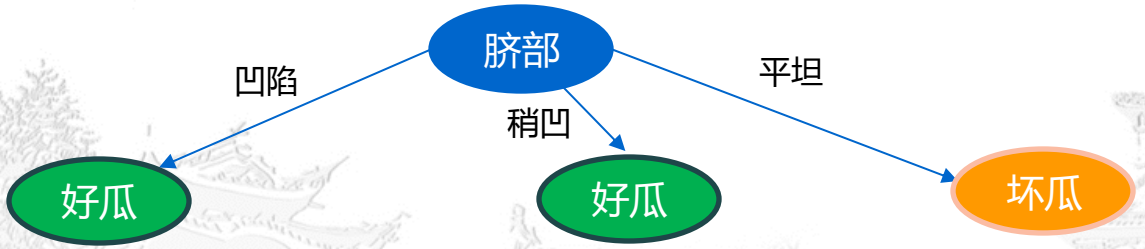
验证集

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

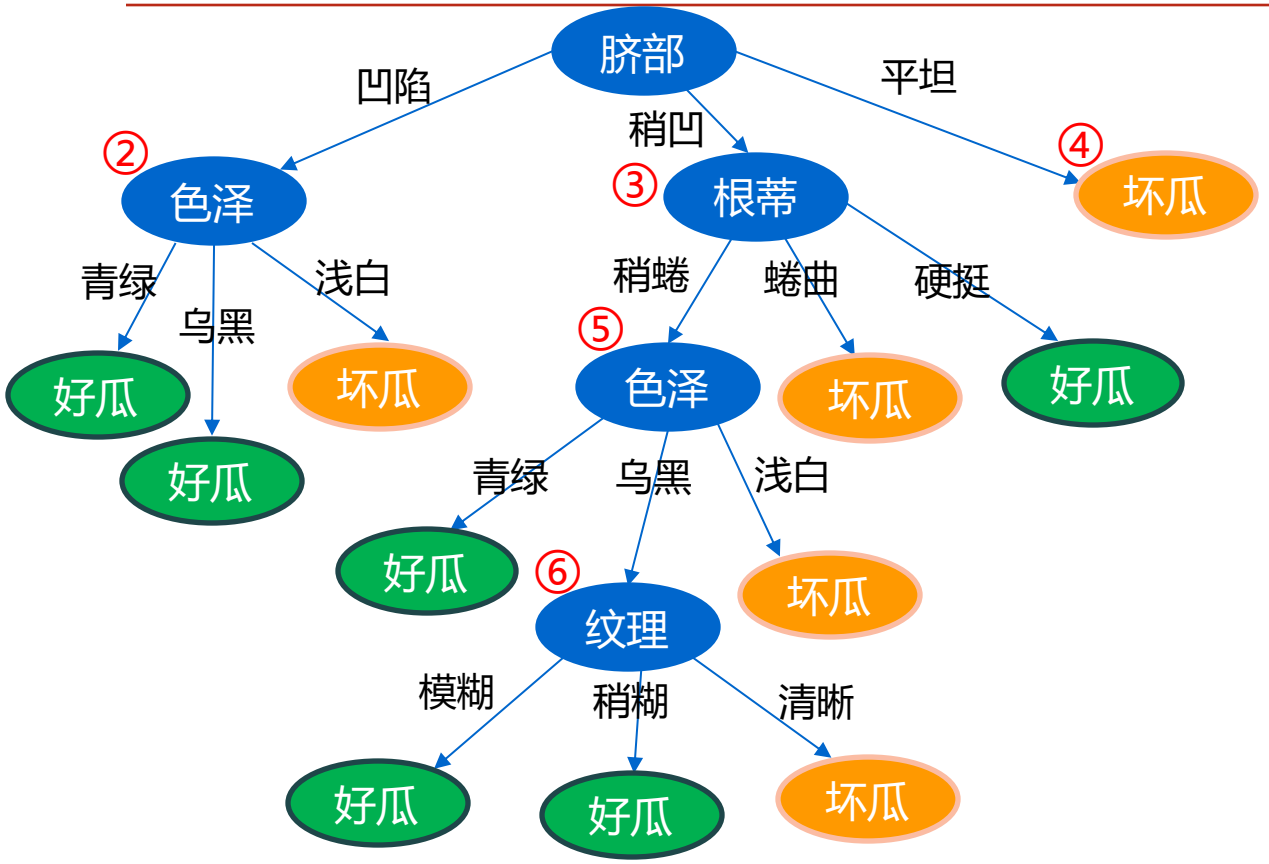
划分前：

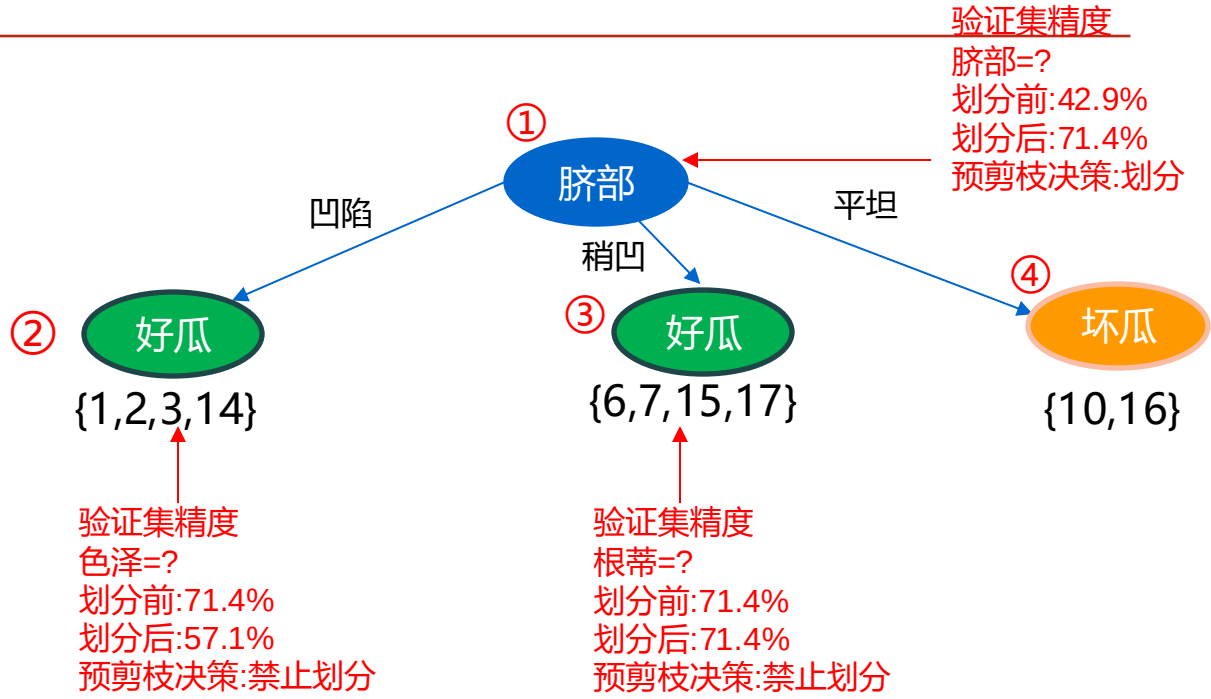
- 划分前：脐部为根节点，选择训练样本数目最多的类别（当数目相等时，可任选一类，此处优先考虑好瓜），将叶结点标记为“好瓜”，验证集准确率：$\frac{3}{7} = 42.9\%$
- 划分后：$\frac{5}{7} = 71.4\%$

划分后：



C4.5的剪枝





预剪枝的决策树

- 预剪枝使决策树很多分支没有展开，减少了决策树的训练时间开销
- 有些分支的当前划分虽不能提升泛化性能，但在此基础上的后续划分有可能显著提升性能，基于贪心策略可能带来欠拟合风险

C4.5的剪枝

后剪枝

- 在已经生成的决策树上进行剪枝，从而得到简化版的剪枝决策树
- 后剪枝决策树通常比预剪枝决策树保留了更多的分支。一般情况下，后剪枝的欠拟合风险更小，泛化性能往往优于预剪枝决策树

训练集

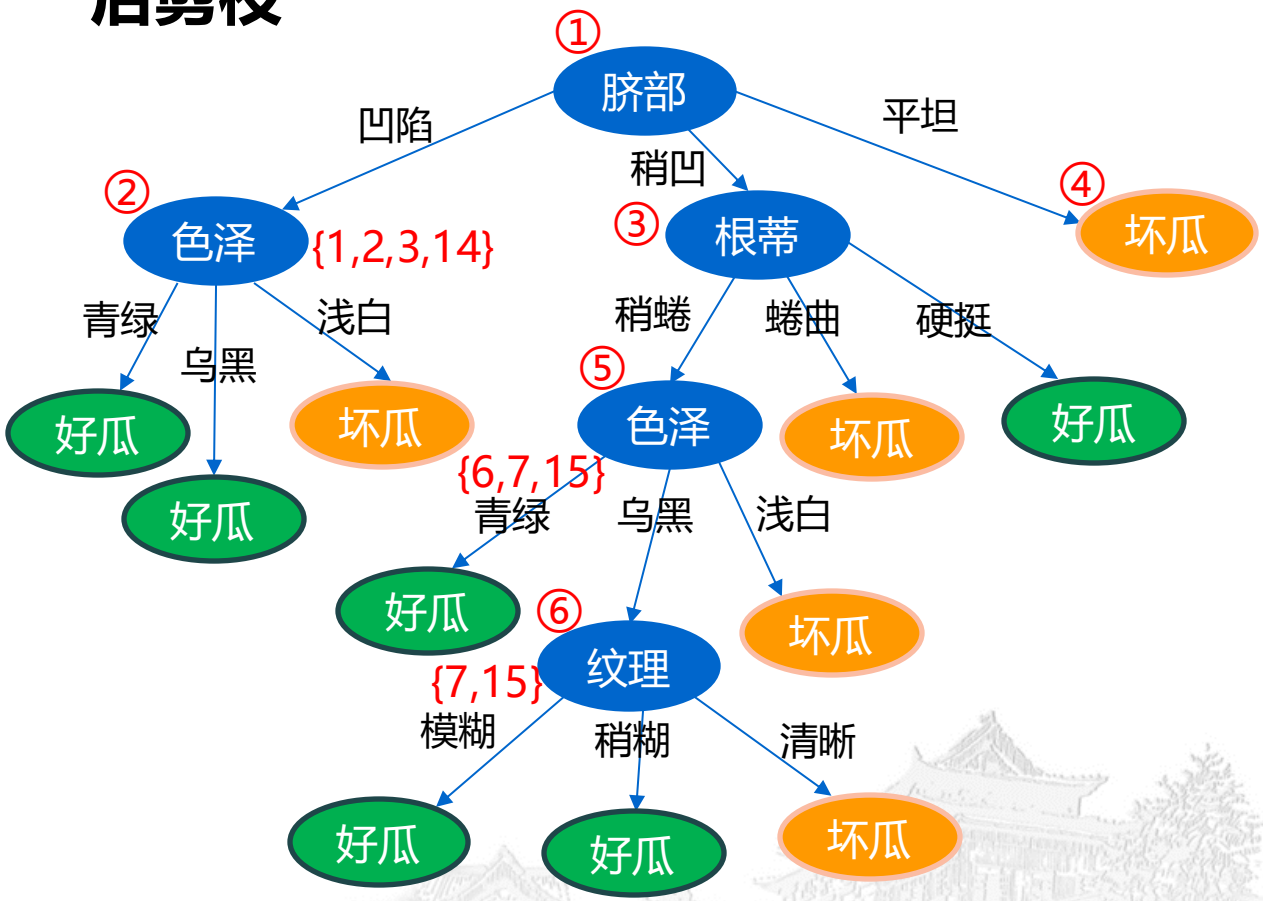
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

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

C4.5的剪枝

后剪枝

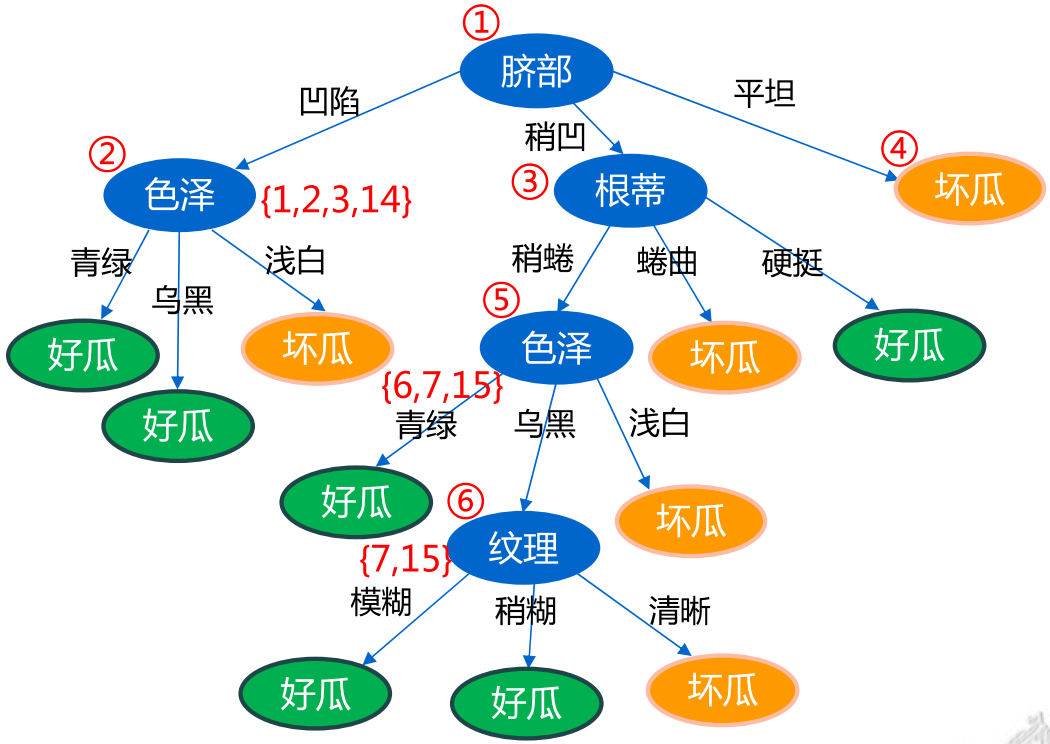


未剪枝的决策树

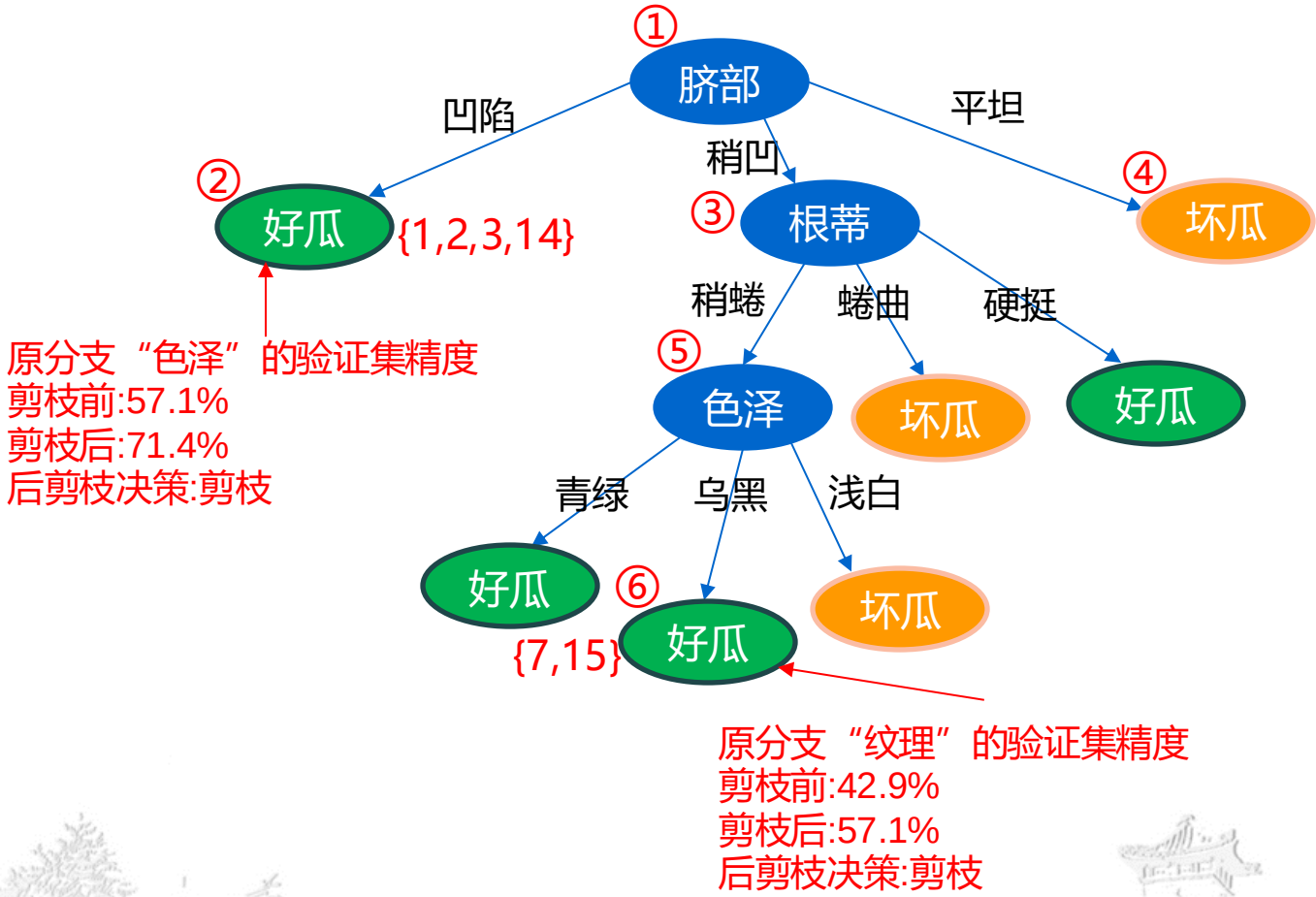
剪枝方法

- 从底往上针对每个非叶节点，评估用叶节点代替该节点前后，验证集准确率是否有提升
- 如有提升，则剪枝。否则，不剪枝

C4.5的剪枝



未剪枝的决策树



后剪枝的决策树

4.CART算法

01 决策树原理

02 ID3算法

03 C4.5算法

04 CART算法

4.CART算法

CART

- Classification And Regression Tree (CART) 是二叉树，属于决策树的一种
- 用基尼指数来选择属性（分类），或用均方误差来选择属性（回归）
- 顾名思义，CART算法既可以用于创建分类树，也可以用于创建回归树，两者在构建的过程中稍有差异
- 如果目标变量是离散的，称为分类树
- 如果目标变量是连续的，称为回归树

CART算法-分类

基尼指数

- 分类问题中假设有K个类别，样本属于第k的概率为p_k，则概率分布的基尼指数定义如下

$$Gini(p) = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2$$

对于二分类：$Gini(p) = 2p(1-p)$

- 给定样本集合D，其基尼指数定义为

$$Gini(D) = 1 - \sum_{k=1}^{K} \left(\frac{|C_k|}{|D|}\right)^2$$

其中C_k指D中第k类样本子集，$|C_k|$指子集大小

- 基尼指数反映了从数据集中随机抽取两个样本，其类别不一致的概率。基尼指数越小，数据集的不确定性越小，与熵类似

CART算法-分类

基尼指数

- 数据集D根据属性A是否取可能值a被划分为D_1和D_2两部分，则在特征A的条件下，集合D的基尼指数定义为

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

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

$$Gini(p)=\sum_{k=1}^{K}p_k(1-p_k)=1-\sum_{k=1}^{K}p_k^2$$

对于二分类：$Gini(p)=2p(1-p)$

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

作业

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

针对左侧数据，利用信息增益比生成决策树

谢 谢!