# 数据挖掘与机器学习

潘斌

panbin@nankai.edu.cn

范孙楼227

# 上节回顾

- 最近邻/k近邻分类器
- 决策树的构造

# 本节提要

- 决策树的节点划分
- 决策树的剪枝

# 实验四、使用决策树进行预测

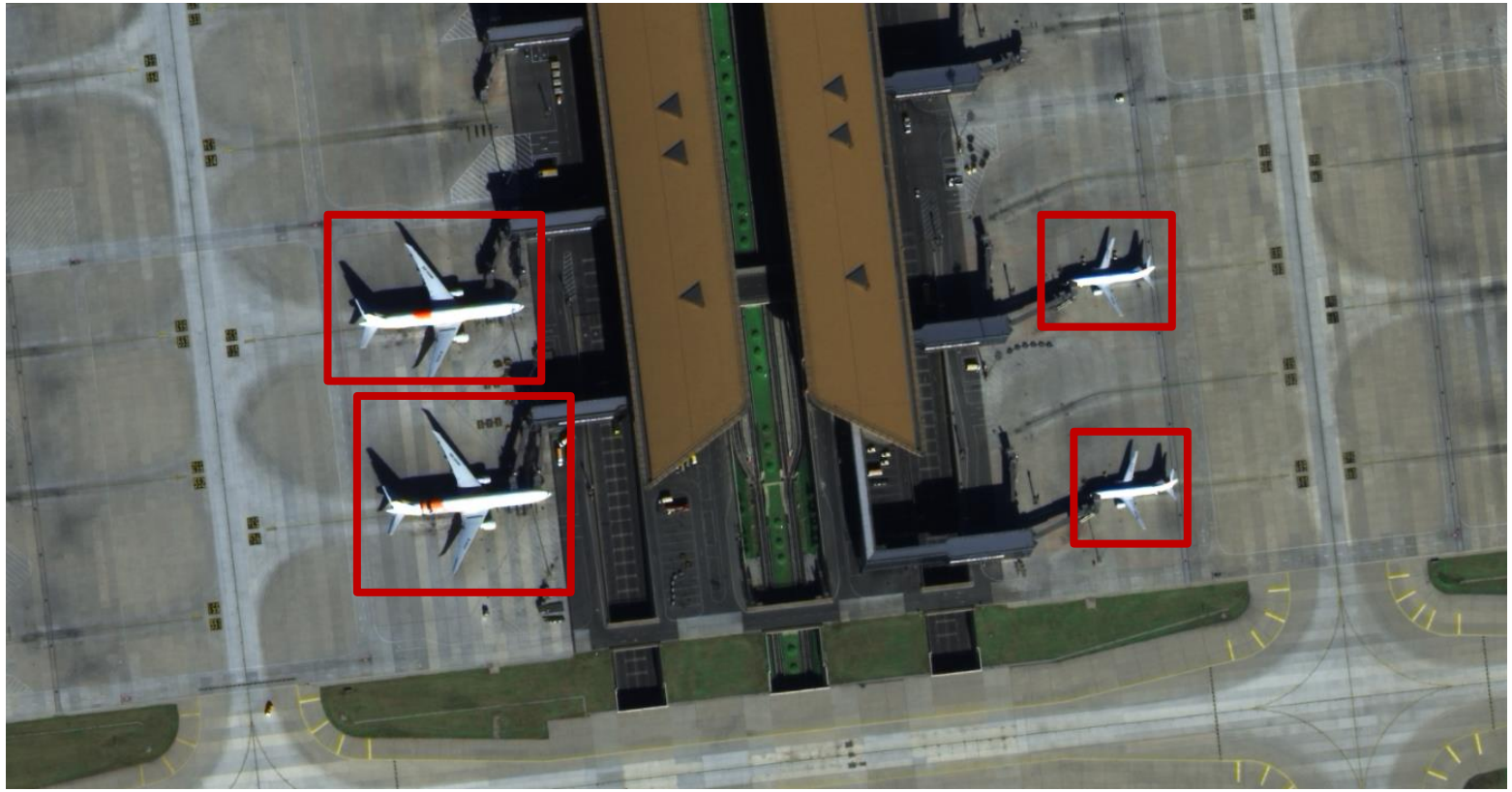
1. 给定一定时间内的犯罪数据集，使用决策树方法进行犯罪类型预测
2. 使用Python编程实现

# 大作业分组

- 自行分组，4-6人一组，1人为组长
- 5.4日前各组组长将本组名单及意向题目排序发给助教
- 助教提供一些示例样本，并及时公布汇报题目及顺序
- 可请助教协助组队

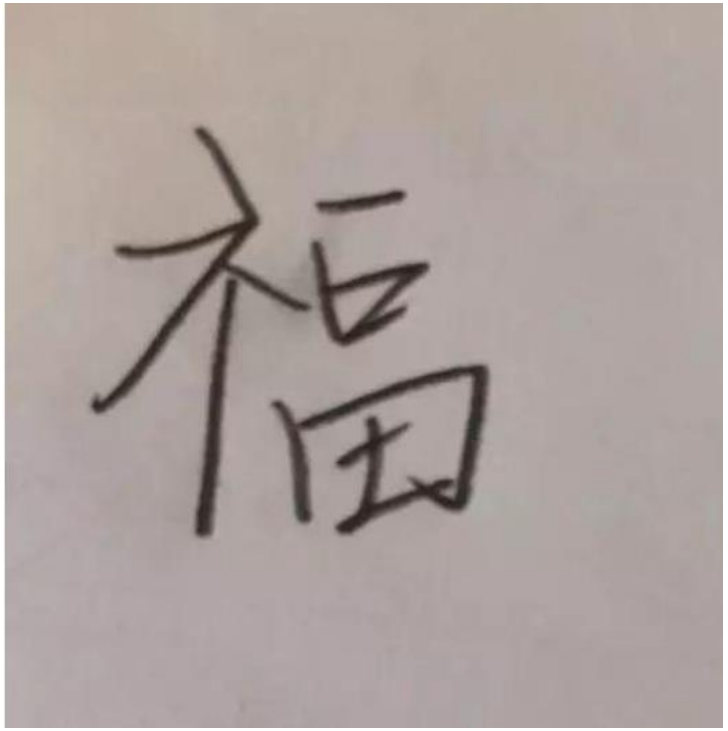
# 课程大作业（四选一）

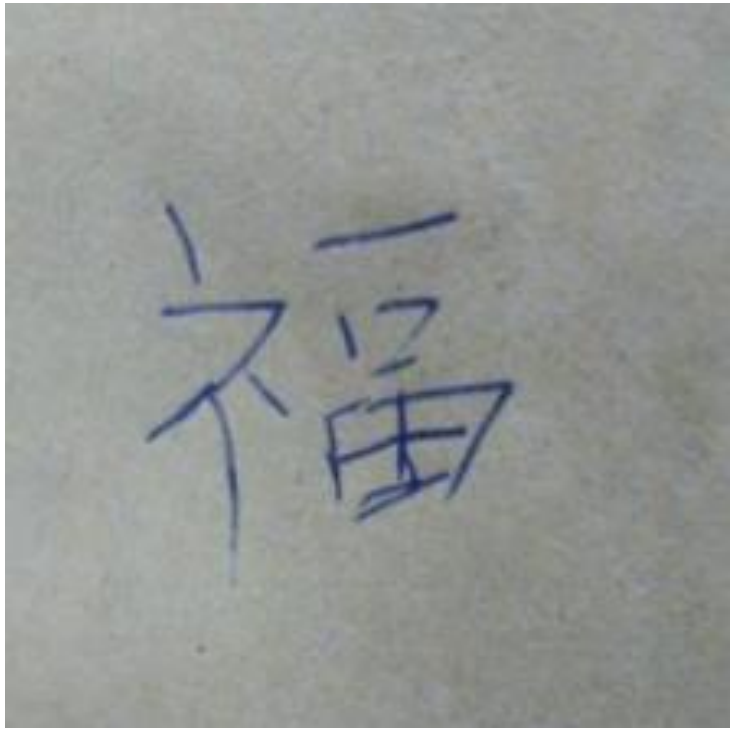
- 任务1：遥感图像飞机检测



# 课程大作业（四选一）

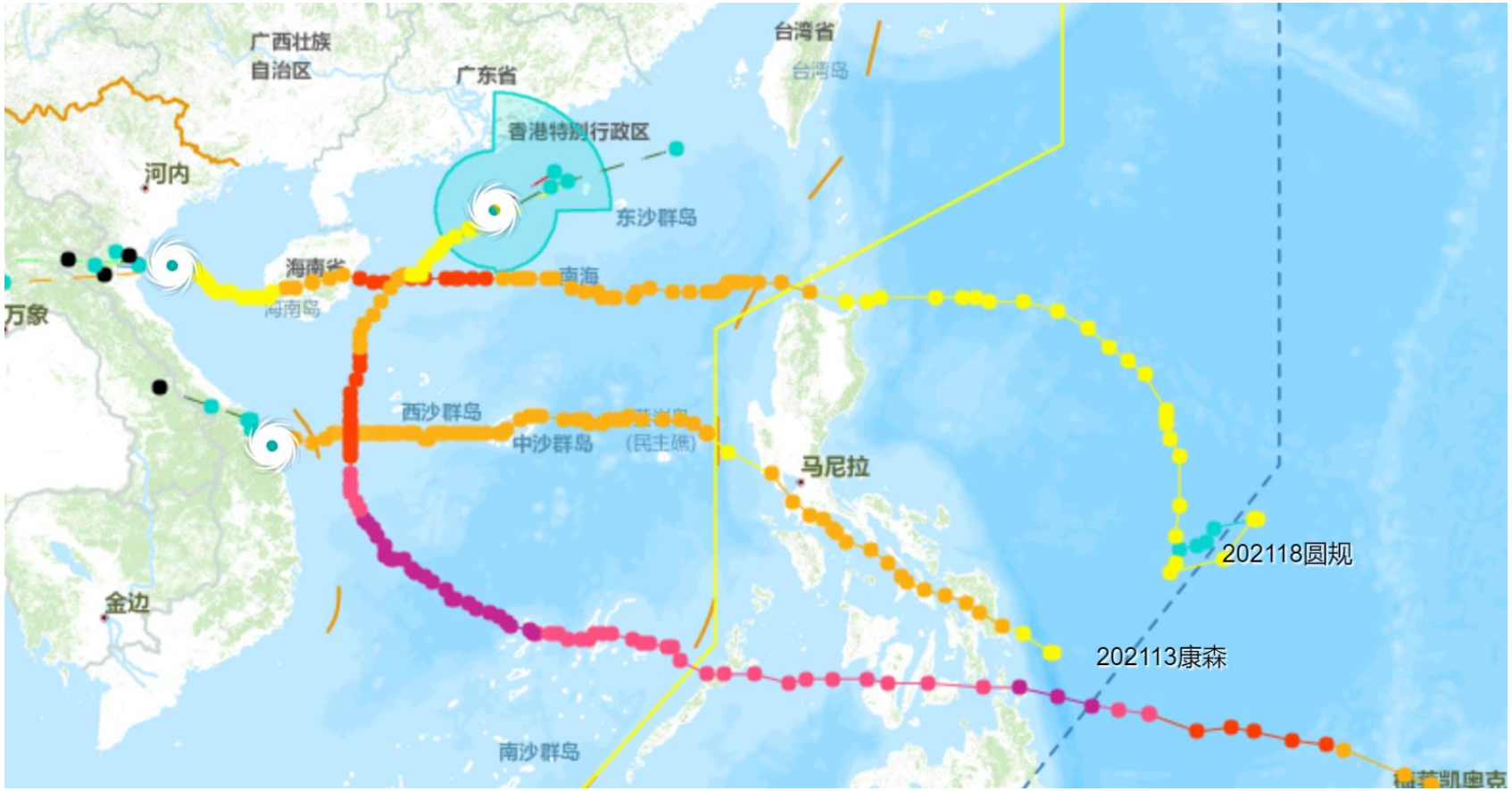
- 任务2：“福”字识别-解决类别不平衡问题





# 课程大作业（四选一）

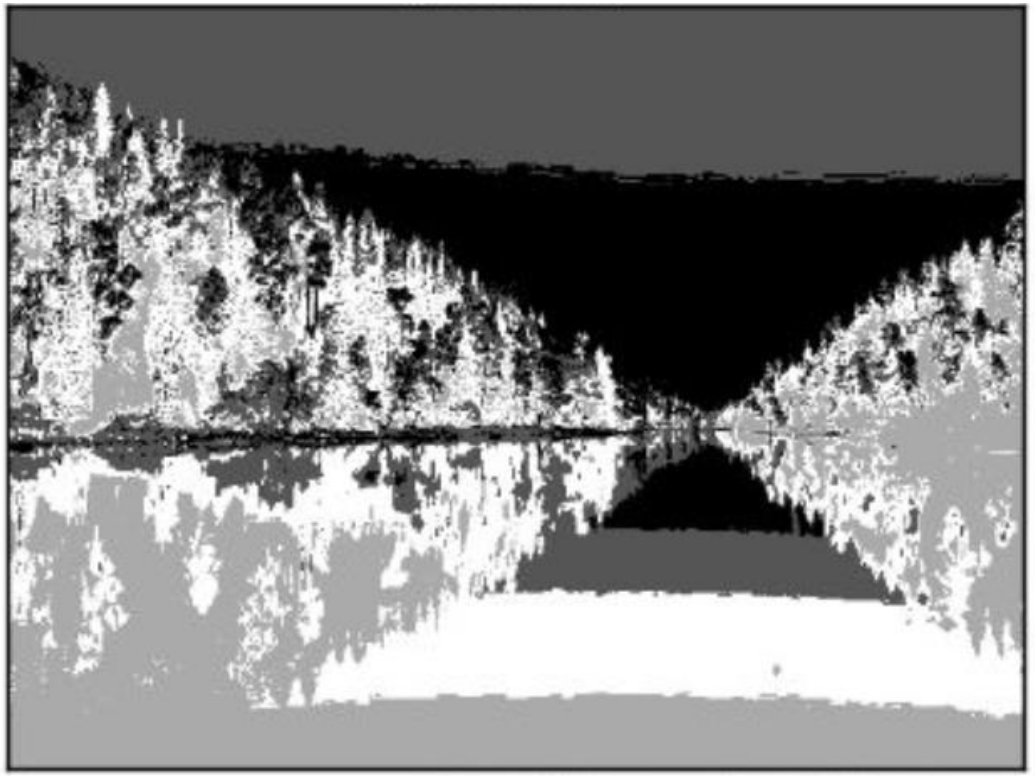
- 任务3：台风预报（https://tcdata.typhoon.org.cn/zjljsjj_zlhq.html）




```
66666 1815   21 0018 1815 0 6 LEEPI
2018081012 1 174 1448 1006     13
2018081018 1 179 1444 1004     15
2018081100 1 185 1441 1004     15
2018081106 1 192 1438 1004     15
2018081112 2 200 1435 1000     18
2018081118 2 208 1432  995     20
2018081200 2 217 1427  990     23
2018081206 2 227 1420  990     23
2018081212 3 235 1415  982     28
2018081218 3 245 1407  982     28
2018081300 3 253 1398  982     28
```

# 课程大作业（四选一）

- 任务4：图像区域分割提取-如何保持空间相关性?



- 选择最优分裂的度量通常是根据分裂后子女节点不纯性的程度，我们希望节点的不纯度逐渐降低
- 不纯的程度越低，类分布就越倾斜
- 度量不纯度的方法

  Entropy

  Gini Index

  Misclassification error

- 不纯度测量：熵

给定节点t:

$$Entropy(t) = -\sum_j p(j|t) \log_2 p(j|t)$$

p(j|t)是给定节点t中属于类j的记录所占的比例

某一节点t的熵值Entropy(t)越小，则该节点的纯度越高

- 不纯度测量：Gini Index

给定一个节点t的Gini Index为：

$$GINI(t) = 1 - \sum_{j} [p(j|t)]^2$$

p(j|t)是给定节点t中属于类j的记录所占的比例

反映了从节点t中随机抽取两个样本，其类别标记不一致的概率，Gini(t)越小，则节点t的纯度越高

- 不纯度测量：分类误差

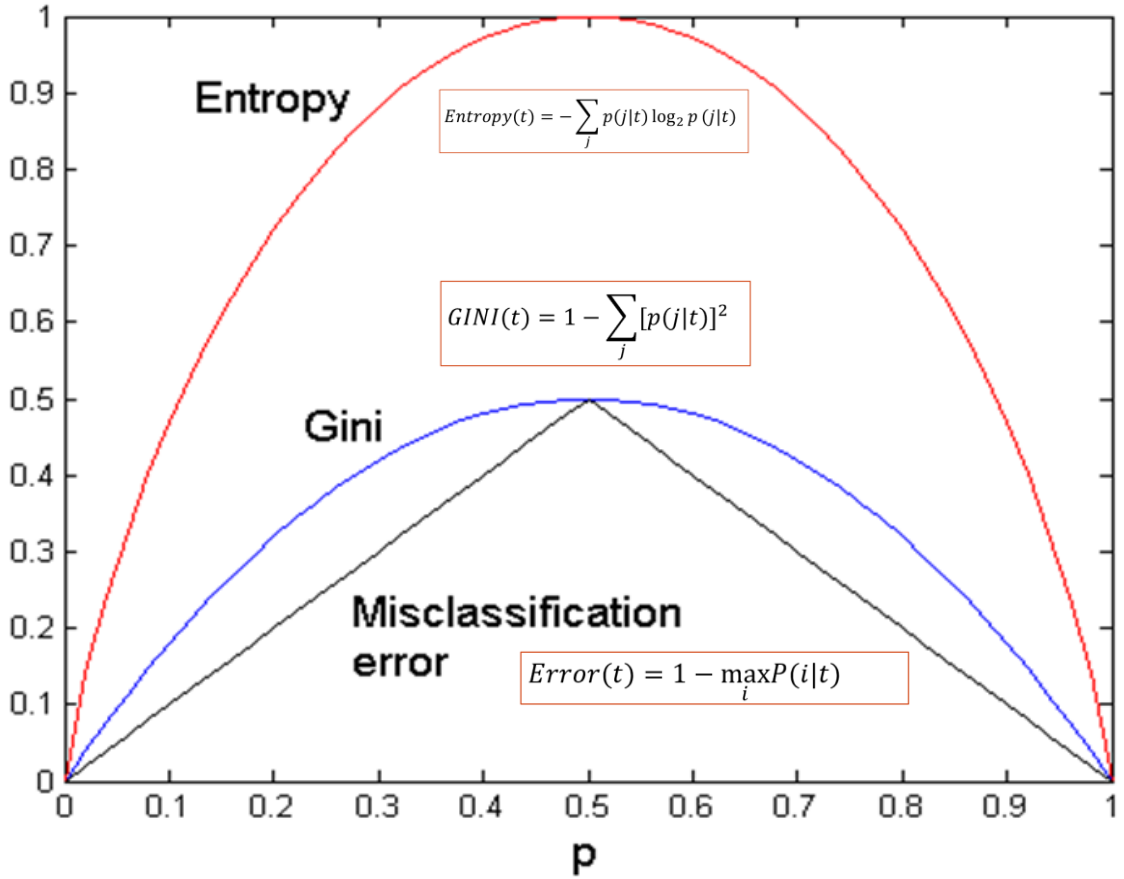
  给定一个节点t的分类误差为：

$$Error(t) = 1 - \max_i P(i|t)$$

p(i|t)是给定节点t中属于类i的记录所占的比例

用来表示某一节点的分类误差，如果所有记录都属于同一类别，该值为0，所以分类误差越小越好

# 二元分类问题不纯性度量之间的比较




基尼系数和香农熵对于p的改变更为敏感（表现在图上就是曲线的斜率更大，更为陡峭），因此，这两种度量方法旨在寻找更为"pure“的nodes

Misclassification error略微有些敏感度不足

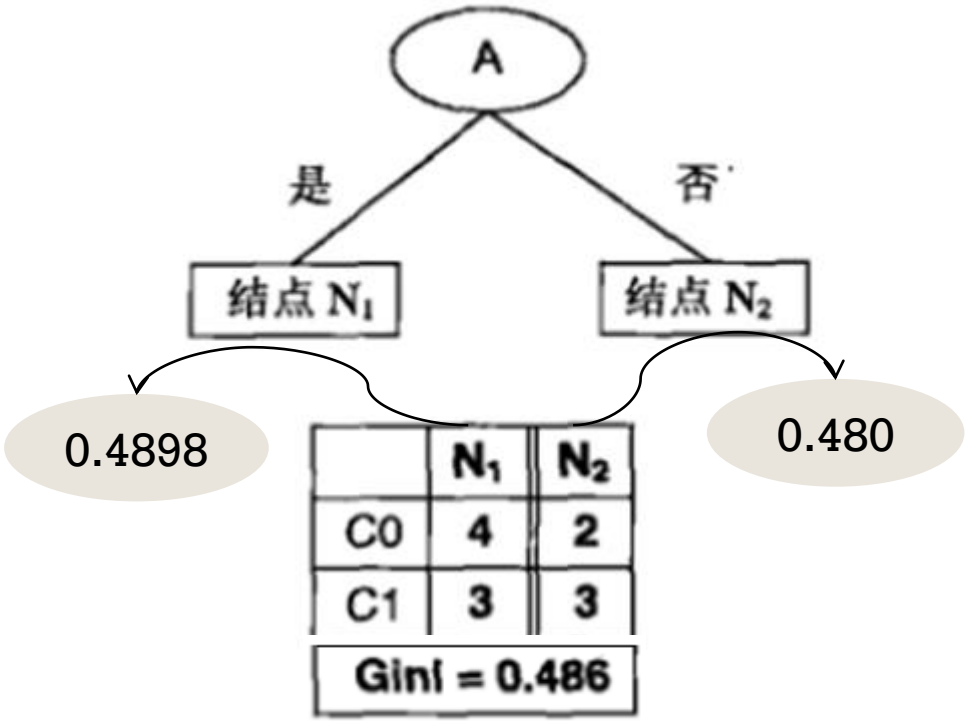
# 所有子女节点总不纯性的度量

- 用每个子女节点不纯度量的加权平均来表示

$$\sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j)$$

  - I(.)：给定节点的不纯性度量
  - N：父节点上的记录总数
  - k：属性值的个数
  - N(vj)：与子女节点vj相关联的记录个数

# 例（GINI INDEX）



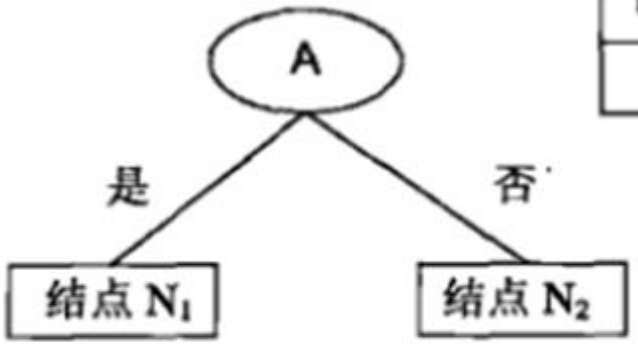

$$Gini(N_1) = 1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2 = \frac{24}{49}$$

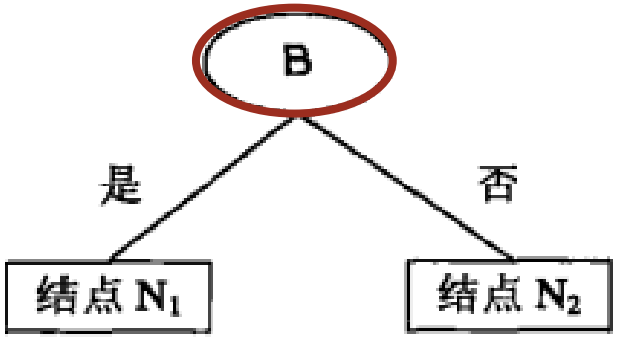
$$Gini(N_2) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = \frac{12}{25}$$

$$Gini = \frac{7}{12} \times \frac{24}{49} + \frac{5}{12} \times \frac{12}{25} = \frac{17}{35} \approx 0.486$$

# 节点的划分（GINI INDEX）

| | 父结点 |
|---|---|
| C0 | 6 |
| C1 | 6 |
| Gini = 0.500 | |




| | $N_1$ | $N_2$ |
|---|---|---|
| C0 | 4 | 2 |
| C1 | 3 | 3 |
| Gini = 0.486 | | |

| | $N_1$ | $N_2$ |
|---|---|---|
| C0 | 1 | 5 |
| C1 | 4 | 2 |
| Gini = 0.371 | | |

# 信息增益（INFORMATION GAIN）

$$GAIN_{split} = I(p) - \left( \sum_{i=1}^{k} \frac{n_i}{n} * I(i) \right)$$

- 划分前的样本不纯度-所有子女节点的总不纯度
- 信息增益既可以用熵也可以用GINI系数来计算

信息增益越大，说明用该特征来划分数据集的信息混乱程度越小。我们需要对样本的所有特征计算信息增益情况，选择信息增益大的特征来作为决策树的一个结点，或者可以说那些信息增益大的特征往往离根结点越近。

# 例：对所给数据集，根据信息增益准则选择最优特征。

| 贷款申请样本数据表 | | | | | |
|---|---|---|---|---|---|
| ID | 年龄 | 有工作 | 有房子 | 信贷情况 | 类别 |
| 1 | 青年 | 否 | 否 | 一般 | 否 |
| 2 | 青年 | 否 | 否 | 好 | 否 |
| 3 | 青年 | 是 | 否 | 好 | 是 |
| 4 | 青年 | 是 | 是 | 一般 | 是 |
| 5 | 青年 | 否 | 否 | 一般 | 否 |
| 6 | 中年 | 否 | 否 | 一般 | 否 |
| 7 | 中年 | 否 | 否 | 好 | 否 |
| 8 | 中年 | 是 | 是 | 好 | 是 |
| 9 | 中年 | 否 | 是 | 非常好 | 是 |
| 10 | 中年 | 否 | 是 | 非常好 | 是 |
| 11 | 老年 | 否 | 是 | 非常好 | 是 |
| 12 | 老年 | 否 | 是 | 好 | 是 |
| 13 | 老年 | 是 | 否 | 好 | 是 |
| 14 | 老年 | 是 | 否 | 非常好 | 是 |
| 15 | 老年 | 否 | 否 | 一般 | 否 |

| 贷款申请样本数据表 | | | | | |
|---|---|---|---|---|---|
| ID | 年龄 | 有工作 | 有房子 | 信贷情况 | 类别 |
| 1 | 青年 | 否 | 否 | 一般 | 否 |
| 2 | 青年 | 否 | 否 | 好 | 否 |
| 3 | 青年 | 是 | 否 | 好 | 是 |
| 4 | 青年 | 是 | 是 | 一般 | 是 |
| 5 | 青年 | 否 | 否 | 一般 | 否 |
| 6 | 中年 | 否 | 否 | 一般 | 否 |
| 7 | 中年 | 否 | 否 | 好 | 否 |
| 8 | 中年 | 是 | 是 | 好 | 是 |
| 9 | 中年 | 否 | 是 | 非常好 | 是 |
| 10 | 中年 | 否 | 是 | 非常好 | 是 |
| 11 | 老年 | 否 | 是 | 非常好 | 是 |
| 12 | 老年 | 否 | 是 | 好 | 是 |
| 13 | 老年 | 是 | 否 | 好 | 是 |
| 14 | 老年 | 是 | 否 | 非常好 | 是 |
| 15 | 老年 | 否 | 否 | 一般 | 否 |

在决策树开始学习时，根节点包含所有数据样例D，其中正例“是”占9/15，负例“否”占6/15，于是，计算出划分前的类别属性熵为：

$$\text{Entropy}(D)$$

$$= -(9/15)\log_2(9/15) - (6/15)\log_2(6/15)$$

$$= 0.971$$

| 贷款申请样本数据表 | | | | | |
|---|---|---|---|---|---|
| ID | 年龄 | 有工作 | 有房子 | 信贷情况 | 类别 |
| 1 | 青年 | 否 | 否 | 一般 | 否 |
| 2 | 青年 | 否 | 否 | 好 | 否 |
| 3 | 青年 | 是 | 否 | 好 | 是 |
| 4 | 青年 | 是 | 是 | 一般 | 是 |
| 5 | 青年 | 否 | 否 | 一般 | 否 |
| 6 | 中年 | 否 | 否 | 一般 | 否 |
| 7 | 中年 | 否 | 否 | 好 | 否 |
| 8 | 中年 | 是 | 是 | 好 | 是 |
| 9 | 中年 | 否 | 是 | 非常好 | 是 |
| 10 | 中年 | 否 | 是 | 非常好 | 是 |
| 11 | 老年 | 否 | 是 | 非常好 | 是 |
| 12 | 老年 | 否 | 是 | 好 | 是 |
| 13 | 老年 | 是 | 否 | 好 | 是 |
| 14 | 老年 | 是 | 否 | 非常好 | 是 |
| 15 | 老年 | 否 | 否 | 一般 | 否 |

# 接下来计算每个属性的信息增益：

## 1、“年龄”属性的信息增益

$G(D,A1)$

$=0.971-[5/15(-(2/5)\log_2(2/5)-(3/5)\log_2(3/5))+5/15(-(3/5)\log_2(3/5)-(2/5)\log_2(2/5))+5/15(-(4/5)\log_2(4/5)-(1/5)\log_2(1/5))]$

$=0.971-0.888$

$=\underline{0.083}$

| 贷款申请样本数据表 | | | | | |
|---|---|---|---|---|---|
| ID | 年龄 | 有工作 | 有房子 | 信贷情况 | 类别 |
| 1 | 青年 | 否 | 否 | 一般 | 否 |
| 2 | 青年 | 否 | 否 | 好 | 否 |
| 3 | 青年 | 是 | 否 | 好 | 是 |
| 4 | 青年 | 是 | 是 | 一般 | 是 |
| 5 | 青年 | 否 | 否 | 一般 | 否 |
| 6 | 中年 | 否 | 否 | 一般 | 否 |
| 7 | 中年 | 否 | 否 | 好 | 否 |
| 8 | 中年 | 是 | 是 | 好 | 是 |
| 9 | 中年 | 否 | 是 | 非常好 | 是 |
| 10 | 中年 | 否 | 是 | 非常好 | 是 |
| 11 | 老年 | 否 | 是 | 非常好 | 是 |
| 12 | 老年 | 否 | 是 | 好 | 是 |
| 13 | 老年 | 是 | 否 | 好 | 是 |
| 14 | 老年 | 是 | 否 | 非常好 | 是 |
| 15 | 老年 | 否 | 否 | 一般 | 否 |

## 2、“有工作”属性的信息增益

$G(D,A2)$

$=0.971-[5/15*0+ 10/15(-(4/10)\log_2(4/10)-(6/10)\log_2(6/10))]$

$=\underline{\boldsymbol{0.324}}$

| 贷款申请样本数据表 | | | | | |
|---|---|---|---|---|---|
| ID | 年龄 | 有工作 | 有房子 | 信贷情况 | 类别 |
| 1 | 青年 | 否 | 否 | 一般 | 否 |
| 2 | 青年 | 否 | 否 | 好 | 否 |
| 3 | 青年 | 是 | 否 | 好 | 是 |
| 4 | 青年 | 是 | 是 | 一般 | 是 |
| 5 | 青年 | 否 | 否 | 一般 | 否 |
| 6 | 中年 | 否 | 否 | 一般 | 否 |
| 7 | 中年 | 否 | 否 | 好 | 否 |
| 8 | 中年 | 是 | 是 | 好 | 是 |
| 9 | 中年 | 否 | 是 | 非常好 | 是 |
| 10 | 中年 | 否 | 是 | 非常好 | 是 |
| 11 | 老年 | 否 | 是 | 非常好 | 是 |
| 12 | 老年 | 否 | 是 | 好 | 是 |
| 13 | 老年 | 是 | 否 | 好 | 是 |
| 14 | 老年 | 是 | 否 | 非常好 | 是 |
| 15 | 老年 | 否 | 否 | 一般 | 否 |

# 3、“有房子”属性的信息增益

$G(D,A3)$

$=0.971-[6/15*0+ 9/15(-(3/9)\log_2(3/9)-(6/9)\log_2(6/9))]$

$= \underline{\boldsymbol{0.420}}$

| 贷款申请样本数据表 | | | | | |
|---|---|---|---|---|---|
| ID | 年龄 | 有工作 | 有房子 | 信贷情况 | 类别 |
| 1 | 青年 | 否 | 否 | 一般 | 否 |
| 2 | 青年 | 否 | 否 | 好 | 否 |
| 3 | 青年 | 是 | 否 | 好 | 是 |
| 4 | 青年 | 是 | 是 | 一般 | 是 |
| 5 | 青年 | 否 | 否 | 一般 | 否 |
| 6 | 中年 | 否 | 否 | 一般 | 否 |
| 7 | 中年 | 否 | 否 | 好 | 否 |
| 8 | 中年 | 是 | 是 | 好 | 是 |
| 9 | 中年 | 否 | 是 | 非常好 | 是 |
| 10 | 中年 | 否 | 是 | 非常好 | 是 |
| 11 | 老年 | 否 | 是 | 非常好 | 是 |
| 12 | 老年 | 否 | 是 | 好 | 是 |
| 13 | 老年 | 是 | 否 | 好 | 是 |
| 14 | 老年 | 是 | 否 | 非常好 | 是 |
| 15 | 老年 | 否 | 否 | 一般 | 否 |

# 4、“信贷情况”属性的信息增益

G(D,A3)

=0.971-[5/15(-1/5log2(1/5)-4/5log2(4/5))+6/15(-4/6log2(4/6)-2/6log(2/6))+4/15*0]

=***0.363***

| 属性 | 信息增益 |
| --- | --- |
| 年龄 | 0.083 |
| 有工作 | 0.324 |
| 有房子 | 0.420 |
| 信贷情况 | 0.363 |

选择特征“有房子”作为最优特征。

然后呢?

# 决策树的三种常用算法:

- ID3算法
- C4.5算法
- CART算法

# ID3（ITERATIVE DICHOTOMISER 3）算法

- ID3是Quinlan提出的，机器学习中一种广为人知的算法，主要用于学习布尔函数。它的提出开创了决策树算法的先河，而且是国际上最早最有影响的决策树方法。
- 自顶向下构造决策树，**用信息增益作为属性选择标准。**
- ID3的过程
  1. 分类能力最好的属性被选作树的根节点
  2. 根节点的每个可能值产生一个分支
  3. 训练样例排列到适当的分支
  4. 重复上面的过程直到所有训练样本使用完毕

# 应用示例

- Target：是否交往？

| 年龄 | 学历 | 房产 | 年薪 | 交往 |
|---|---|---|---|---|
| 小 | 研究生 | 有 | 高 | 是 |
| 小 | 研究生 | 有 | 低 | 是 |
| 大 | 研究生 | 有 | 高 | 否 |
| 中 | 本科 | 有 | 高 | 否 |
| 中 | 本科以下 | 无 | 高 | 否 |
| 中 | 本科以下 | 无 | 低 | 是 |
| 大 | 本科以下 | 无 | 低 | 否 |
| 小 | 本科 | 有 | 高 | 是 |
| 小 | 本科以下 | 无 | 高 | 否 |
| 中 | 本科 | 无 | 高 | 否 |
| 小 | 本科 | 无 | 低 | 否 |
| 大 | 本科 | 有 | 低 | 否 |
| 大 | 研究生 | 无 | 高 | 否 |
| 中 | 本科 | 有 | 低 | 是 |

- 类别属性信息熵的计算
  - 由于未划分前，训练数据集中共有14个实例，其中有9个实例属于“否”类（不交往），5个实例属于“是”类（交往），因此划分前类别属性的熵为：

$$\text{Entropy} = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.940$$

- 非类别属性信息熵的计算
  - 若选择“年龄”为测试属性，则

$$
\begin{aligned}
E(\text{年龄}) = & \frac{5}{14}\left(-\frac{2}{5}\log_2\frac{2}{5}-\frac{3}{5}\log_2\frac{3}{5}\right) \\
& +\frac{4}{14}\left(-\frac{4}{4}\log_2\frac{4}{4}-\frac{0}{4}\log_2\frac{0}{4}\right) \\
& +\frac{5}{14}\left(-\frac{3}{5}\log_2\frac{3}{5}-\frac{2}{5}\log_2\frac{2}{5}\right) \\
= & \ 0.694
\end{aligned}
$$

| 年龄 | 交往 |
|---|---|
| 小 | 是 |
| 小 | 是 |
| 大 | 否 |
| 中 | 否 |
| 中 | 否 |
| 中 | 是 |
| 大 | 否 |
| 美 | 是 |
| 美 | 否 |
| 中 | 否 |
| 美 | 否 |
| 大 | 否 |
| 大 | 否 |
| 中 | 是 |

- 因此，这种划分的信息增益是：

$$Gain(\text{年龄}) = \text{Entropy} - E(\text{年龄}) = 0.940 - 0.694 = 0.246$$

- 同理，可以求出其它三个属性的信息增益

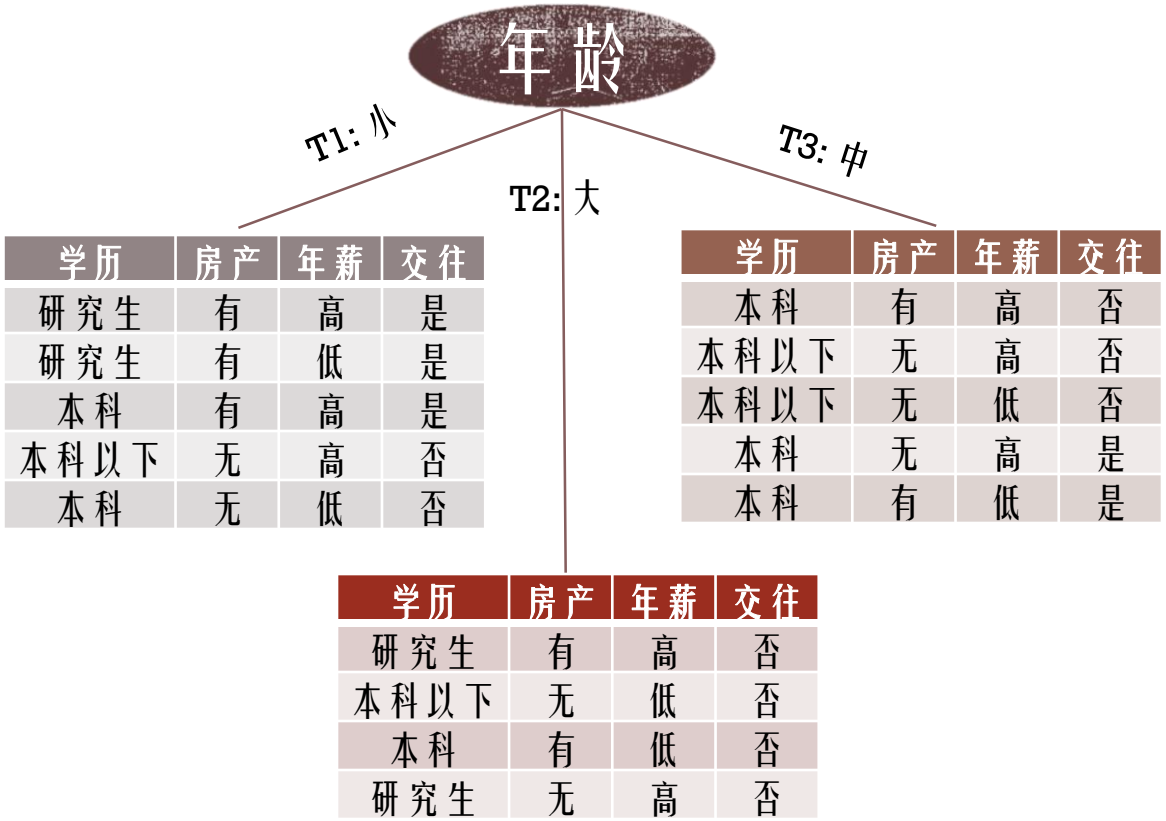
$Gain(\text{学历}) = 0.029$

$Gain(\text{房产}) = 0.151$

$Gain(\text{年薪}) = 0.048$

- 由上可知，属性值“年龄”在各个属性中具有最高的增益，被选作测试属性（根节点）。
  - 选择测试属性后，将训练实例集分为三个子集，生成三个子节点，对每个子节点递归采用上述过程进行分类直至每个节点都成为叶节点。

- 属性值“年龄”在各个属性中具有最高的增益，被选作分裂属性，并对于每个属性值生长一个分支：




| 学历 | 房产 | 年薪 | 交往 |
|---|---|---|---|
| 研究生 | 有 | 高 | 是 |
| 研究生 | 有 | 低 | 是 |
| 本科 | 有 | 高 | 是 |
| 本科以下 | 无 | 高 | 否 |
| 本科 | 无 | 低 | 否 |

| 学历 | 房产 | 年薪 | 交往 |
|---|---|---|---|
| 本科 | 有 | 高 | 否 |
| 本科以下 | 无 | 高 | 否 |
| 本科以下 | 无 | 低 | 否 |
| 本科 | 无 | 高 | 是 |
| 本科 | 有 | 低 | 是 |

| 学历 | 房产 | 年薪 | 交往 |
|---|---|---|---|
| 研究生 | 有 | 高 | 否 |
| 本科以下 | 无 | 低 | 否 |
| 本科 | 有 | 低 | 否 |
| 研究生 | 无 | 高 | 否 |

| 学历 | 房产 | 年薪 | 交往 |
| --- | --- | --- | --- |
| 研究生 | 有 | 高 | 是 |
| 研究生 | 有 | 低 | 是 |
| 本科 | 有 | 高 | 是 |
| 本科以下 | 无 | 高 | 否 |
| 本科 | 无 | 低 | 否 |

- 分析图中的“T1:小”分支，计算其子属性的信息增益值来决定子分裂属性形成子分支之一。
  - 针对“T1:小”中的子训练数据集分支，有两个类别，该分支中有3个实例属于“是”类，有2个实例属于“否”类，因此该分支的信息熵为：

$$\text{Entropy(T1)} = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971$$

| 学历 | 房产 | 年薪 | 交往 |
| --- | --- | --- | --- |
| 研究生 | 有 | 高 | 是 |
| 研究生 | 有 | 低 | 是 |
| 本科 | 有 | 高 | 是 |
| 本科以下 | 无 | 高 | 否 |
| 本科 | 无 | 低 | 否 |

- 若以“T1:小”分支中的属性“学历”为测试属性，则测试属性“学历”的信息量为：

$$E_{T_1}(\text{学历}) = \frac{2}{5}\left(-\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2}\right) + \frac{2}{5}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{1}{5}\left(-\frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1}\right) = 0.400$$

- 其信息增益为：

$$\text{Gain}_{T_1}(\text{学历}) = 0.971 - 0.400 = 0.571$$

| 学历 | 房产 | 年薪 | 交往 |
| --- | --- | --- | --- |
| 研究生 | 有 | 高 | 是 |
| 研究生 | 有 | 低 | 是 |
| 本科 | 有 | 高 | 是 |
| 本科以下 | 无 | 高 | 否 |
| 本科 | 无 | 低 | 否 |

- 若以“T1:小”分支中的属性“房产”为测试属性，则测试属性“房产”的信息量为：

$$E_{T_1}(\text{房产}) = \frac{3}{5}\left(-\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3}\right) + \frac{2}{5}\left(-\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2}\right)$$
$$= 0.00$$

- 其信息增益为：

$$Gain_{T_1}(\text{房产}) = 0.971 - 0.00 = 0.971$$

| 学历 | 房产 | 年薪 | 交往 |
| --- | --- | --- | --- |
| 研究生 | 有 | 高 | 是 |
| 研究生 | 有 | 低 | 是 |
| 本科 | 有 | 高 | 是 |
| 本科以下 | 无 | 高 | 否 |
| 本科 | 无 | 低 | 否 |

- 若以“T1:小”分支中的属性“年薪”为测试属性，则测试属性“年薪”的信息量为：

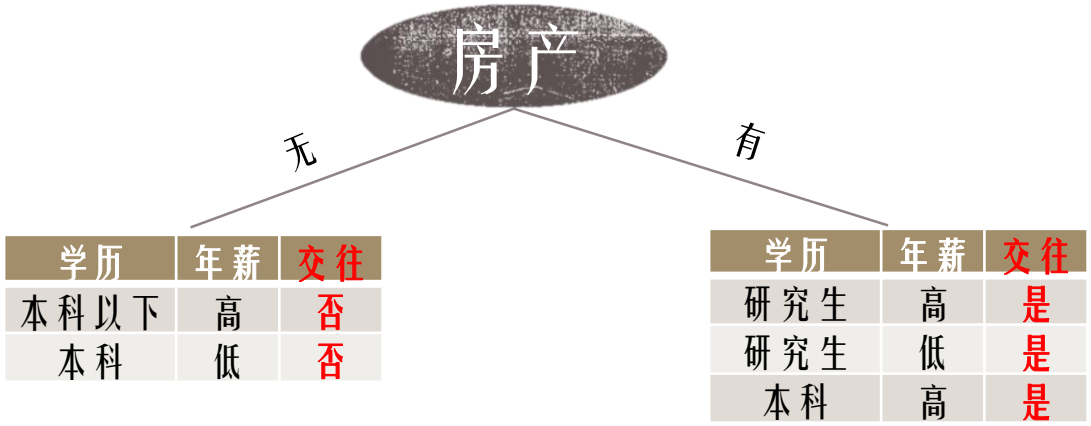
$$E_{T_1}(\text{年薪}) = \frac{3}{5}\left(-\frac{2}{3}\log_2\frac{2}{3}-\frac{1}{3}\log_2\frac{1}{3}\right)+\frac{2}{5}\left(-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2}\right)$$
$$= 0.468$$

- 其信息增益为：

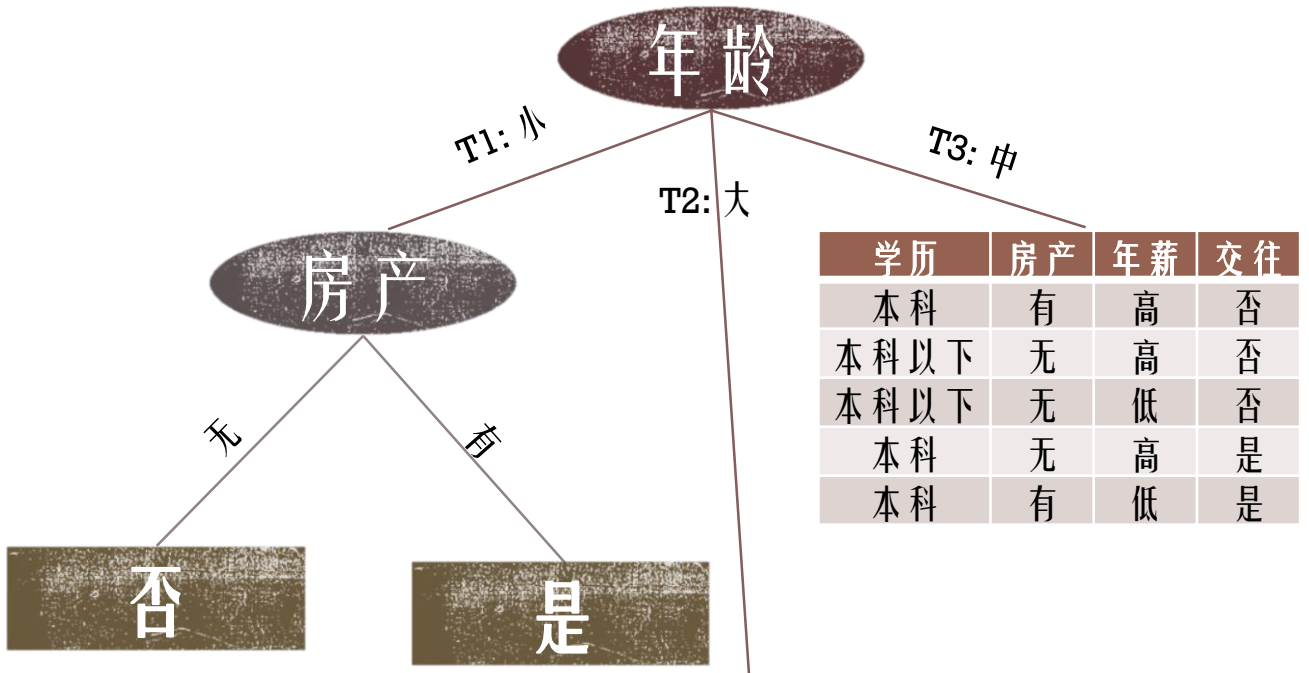
$$Gain_{T_1}(\text{年薪}) = 0.971 - 0.468 = 0.493$$

| 学历 | 房产 | 年薪 | 交往 |
| --- | --- | --- | --- |
| 研究生 | 有 | 高 | 是 |
| 研究生 | 有 | 低 | 是 |
| 本科 | 有 | 高 | 是 |
| 本科以下 | 无 | 高 | 否 |
| 本科 | 无 | 低 | 否 |

- 三个属性中“房产”属性的信息增益最高，因而将其作为属性测试条件



- 这时生成的决策树为：



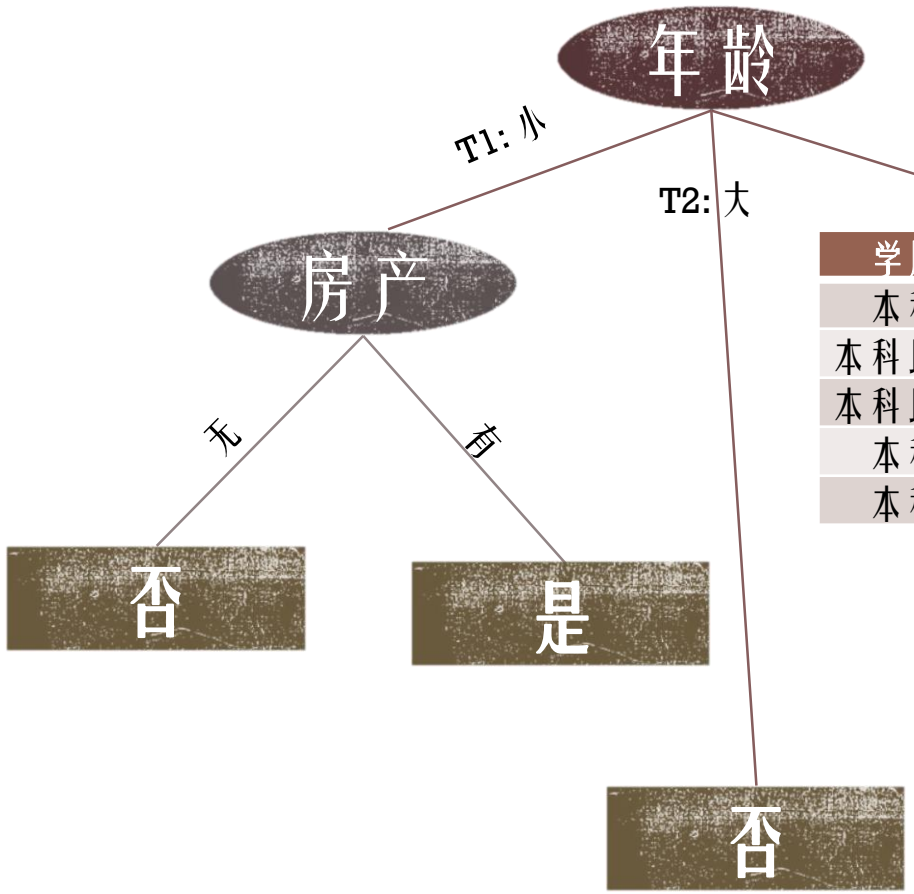

| 学历 | 房产 | 年薪 | 交往 |
|---|---|---|---|
| 本科 | 有 | 高 | 否 |
| 本科以下 | 无 | 高 | 否 |
| 本科以下 | 无 | 低 | 否 |
| 本科 | 无 | 高 | 是 |
| 本科 | 有 | 低 | 是 |

| 学历 | 房产 | 年薪 | 交往 |
|---|---|---|---|
| 研究生 | 有 | 高 | 否 |
| 本科以下 | 无 | 低 | 否 |
| 本科 | 有 | 低 | 否 |
| 研究生 | 无 | 高 | 否 |

| 学历 | 房产 | 年薪 | 交往 |
| --- | --- | --- | --- |
| 研究生 | 有 | 高 | 否 |
| 本科以下 | 无 | 低 | 否 |
| 本科 | 有 | 低 | 否 |
| 研究生 | 无 | 高 | 否 |

- 分析图中的“T2:大”分支
  - 该分支中的实例都属于“否”类，因而直接作为叶节点，类标签为“否”

- 这时生成的决策树为：




| 学历 | 房产 | 年薪 | 交往 |
|---|---|---|---|
| 本科 | 有 | 高 | 否 |
| 本科以下 | 无 | 高 | 否 |
| 本科以下 | 无 | 低 | 否 |
| 本科 | 无 | 高 | 是 |
| 本科 | 有 | 低 | 是 |

| 学历 | 房产 | 年薪 | 交往 |
| --- | --- | --- | --- |
| 本科 | 有 | 高 | 否 |
| 本科以下 | 无 | 高 | 否 |
| 本科以下 | 无 | 低 | 否 |
| 本科 | 无 | 高 | 是 |
| 本科 | 有 | 低 | 是 |

- 分析“T3:中”分支，计算其子属性的信息增益值来确定子分裂属性形成子分支之三。
  - 针对“T3:中”中的子训练数据集分支，有两个类别，该分支中有3个实例属于“是”类，有2个实例属于“否”类，因此针对分支的信息熵为：

$$\text{Entropy(T3)} = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971$$

| 学历 | 房产 | 年薪 | 交往 |
| --- | --- | --- | --- |
| 本科 | 有 | 高 | 否 |
| 本科以下 | 无 | 高 | 否 |
| 本科以下 | 无 | 低 | 否 |
| 本科 | 无 | 高 | 是 |
| 本科 | 有 | 低 | 是 |

- 若以“T3:中”分支中的属性“学历”为测试属性，则测试属性“学历”的信息量为：

$$E_{T_3}(\text{学历}) = \frac{3}{5}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) + \frac{2}{5}\left(-\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2}\right)$$
$$= 0.285$$

- 其信息增益值为：

$$Gain_{T_3}(\text{学历}) = 0.971 - 0.285 = 0.696$$

| 学历 | 房产 | 年薪 | 交往 |
| --- | --- | --- | --- |
| 本科 | 有 | 高 | 否 |
| 本科以下 | 无 | 高 | 否 |
| 本科以下 | 无 | 低 | 否 |
| 本科 | 无 | 高 | 是 |
| 本科 | 有 | 低 | 是 |

- 若以“T3:中”分支中的属性“房产”为测试属性，则测试属性“房产”的信息量为：

$$E_{T_3}(\text{房产}) = \frac{3}{5}\left(-\frac{2}{3}\log_2\frac{2}{3}-\frac{1}{3}\log_2\frac{1}{3}\right)+\frac{2}{5}\left(-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2}\right)$$
$$= 0.951$$

- 其信息增益值为：

$$Gain_{T_3}(\text{房产}) = 0.971-0.950=0.020$$

| 学历 | 房产 | 年薪 | 交往 |
| --- | --- | --- | --- |
| 本科 | 有 | 高 | 否 |
| 本科以下 | 无 | 高 | 否 |
| 本科以下 | 无 | 低 | 否 |
| 本科 | 无 | 高 | 是 |
| 本科 | 有 | 低 | 是 |

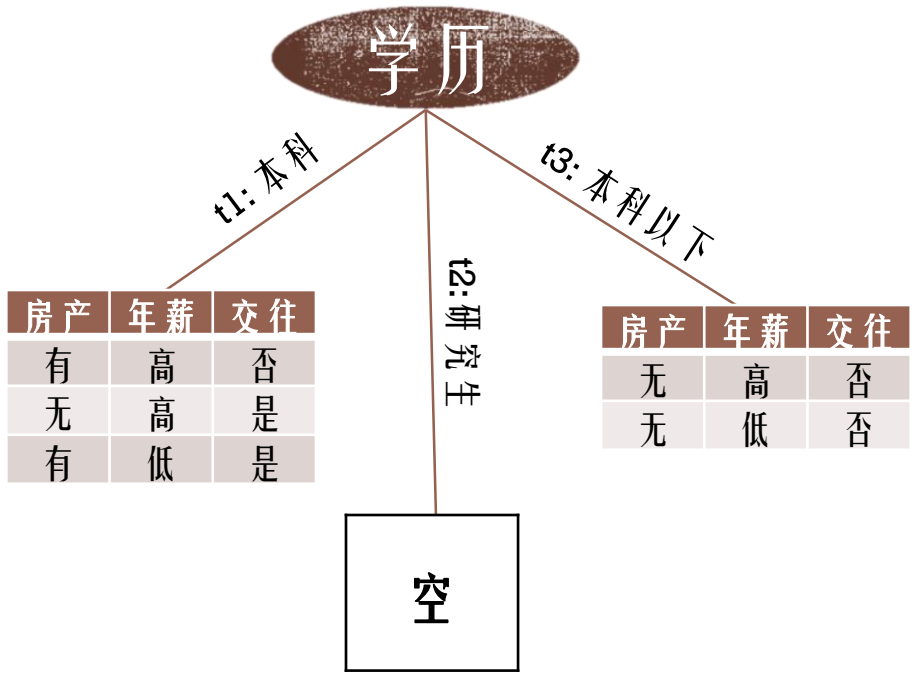
- 若以“T3:中”分支中的属性“年薪”为测试属性，则测试属性“年薪”的信息量为：

$$E_{T_3}(\text{年薪}) = \frac{3}{5}\left(-\frac{2}{3}\log_2\frac{2}{3}-\frac{1}{3}\log_2\frac{1}{3}\right)+\frac{2}{5}\left(-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2}\right)$$
$$= 0.951$$

- 其信息增益值为：

$$Gain_{T_3}(\text{年薪}) = 0.971-0.951 = 0.020$$

- 根据上述步骤所求得的信息增益值，属性“学历”的信息增益值最大，故选为属性测试条件

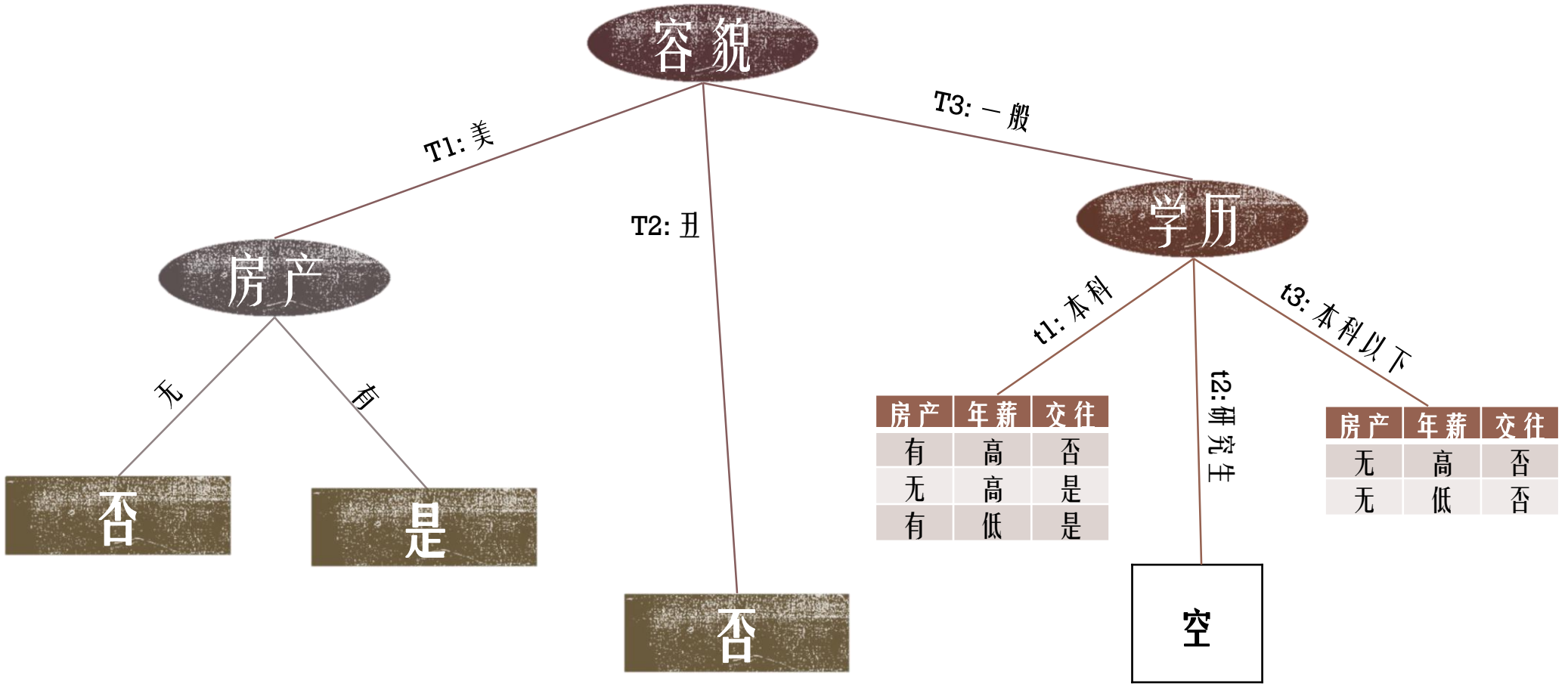



| 房产 | 年薪 | 交往 |
|---|---|---|
| 有 | 高 | 否 |
| 无 | 高 | 是 |
| 有 | 低 | 是 |

| 房产 | 年薪 | 交往 |
|---|---|---|
| 无 | 高 | 否 |
| 无 | 低 | 否 |

- 这时生成的决策树为：

何时停止分裂？
- 所有记录属于同一类
- 各属性值所占比例相同
- 没有相匹配的记录



| 房产 | 年薪 | 交往 |
| --- | --- | --- |
| 有 | 高 | 否 |
| 无 | 高 | 是 |
| 有 | 低 | 是 |

- 对“t1：本科”分支，若假设“房产”为测试属性，则分支t1总的信息熵和测试属性“房产”信息量分别为：

$$\text{Entropy(t1)} = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.918$$

$$E_{t1}(\text{房产}) = \frac{1}{3}\left(-\frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1}\right) + \frac{2}{3}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) = 0.666$$

- 则其信息增益为：

$$Gain_{t1}(\text{房产}) = 0.918 - 0.666 = 0.252$$

| 房产 | 年薪 | 交往 |
| --- | --- | --- |
| 有 | 高 | 否 |
| 无 | 高 | 是 |
| 有 | 低 | 是 |

- 若假设“年薪”为测试属性，则测试属性“年薪”的信息量为：

$$E_{t1}(\text{年薪}) = \frac{1}{3}\left(-\frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1}\right) + \frac{2}{3}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) = 0.666$$

- 其信息增益为：

$$Gain_{t1}(\text{年薪}) = 0.918 - 0.666 = 0.252$$

“房产”和“年薪”二者信息增益相同！

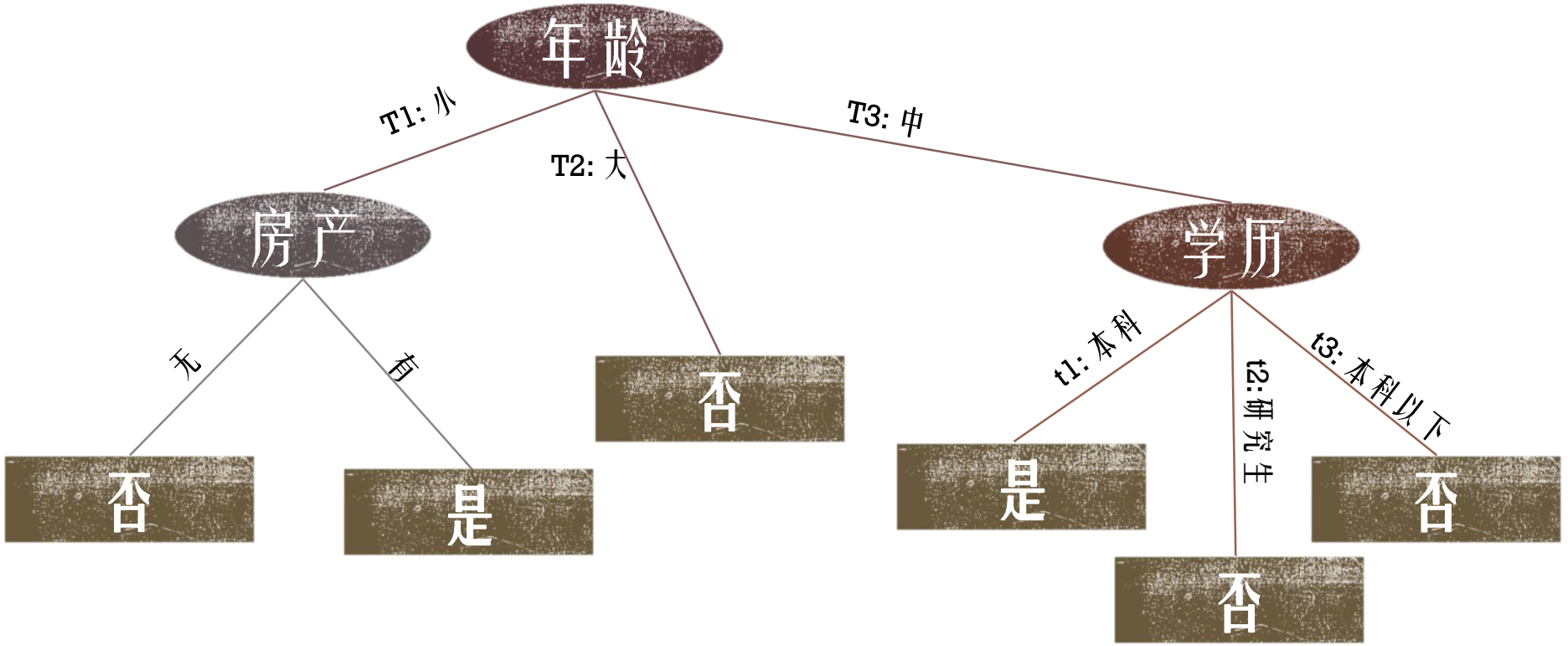
| 房产 | 年薪 | 交往 |
| --- | --- | --- |
| 有 | 高 | 否 |
| 无 | 高 | 是 |
| 有 | 低 | 是 |

何时停止分裂?

- 所有记录属于同一类
- 各属性值所占比例相同
- 没有相匹配的记录

- 在t1分支的记录中，“是”的比例比“否”的大，因而该节点为叶节点，类标签为“是”

- 最终得到的决策树图如图所示：



- ID3算法的优势
  - 算法理论清晰
  - 方法简单
  - 学习能力较强
- ID3算法存在的主要不足
  - **过度拟合问题**
  - **处理连续属性值问题**（最初定义ID3限制为取离散值的属性，但对于连续值属性，我们可以通过动态的定义新的离散值属性来实现。例如，对于连续值的属性A，可动态的创建一个新的布尔属性F，如果A<c，那么F为真，否则为假）
  - **处理缺少属性值问题**（可供使用的数据可能缺少某些属性的值）
  - **属性选择的度量标准问题**（信息增益法会倾向于选择取值比较多的特征）

# 决策树的三种常用算法:

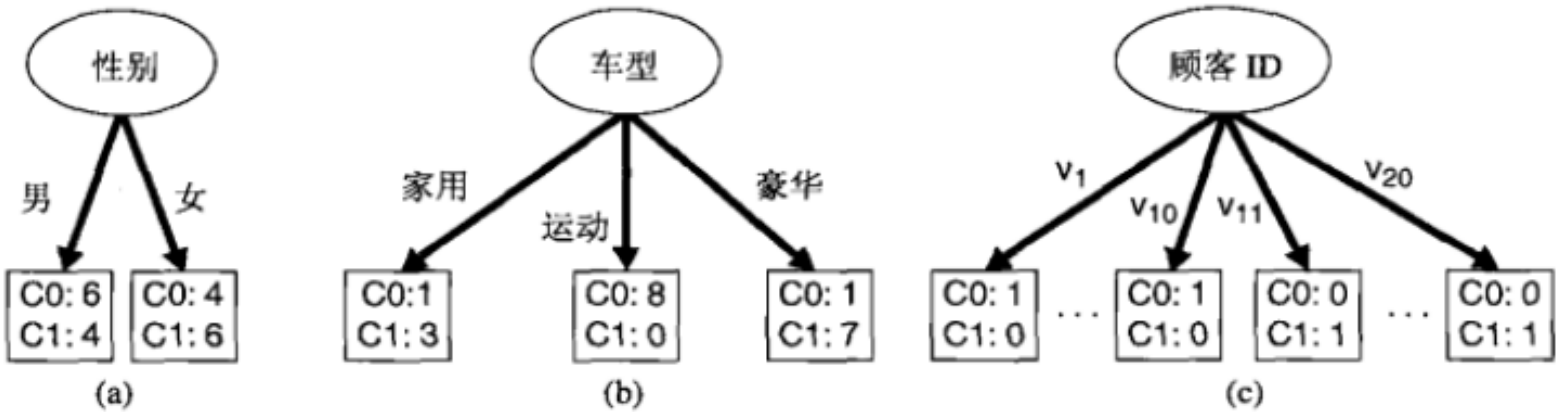
- ID3算法
- C4.5算法
- CART算法

# C4.5算法

- 针对ID3算法的不足有很多改进算法，其中Quinlan在1993年开发出的C4.5算法流行最广。
- C4.5改进的主要体现
  - 新的属性选择度量标准
  - 连续属性值的学习问题
  - 不完整数据（含缺失值）的处理
  - 避免树无节制增长，避免过度拟合数据

# 1、属性选择其他度量标准

- 信息增益度量存在一个内在偏置，偏向具有较多值的属性
  - 熵和Gini Index等不纯性度量倾向有利于具有大不同值的属性




  - (c)看来产生更纯的划分
  - 形成一棵深度为一级但是却非常宽的树
  - 完美地分割了训练数据

- 避免方法：加入惩罚项
  - 分裂信息

$$Entropy(t) = -\sum_j p(j|t) \log_2 p(j|t)$$

$$\sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j)$$

$$SplitINFO = -\sum_{i=1}^{k} \frac{n_i}{n} \log_2 \frac{n_i}{n}$$

Parent Node p is split into k partitions

$n_i$ is the number of records in partition i

  - 惩罚分裂属性自身的不纯度

- （增益比率）Gain Ratio:

$$GainRATIO_{split} = \frac{GAIN_{Split}}{SplitINFO}$$

  - 通过惩罚项SplitINFO调整各属性的增益比率，对于高不纯度的属性惩罚较大

# 2、连续属性值的处理

- 对于“年薪”属性，若它是连续型变量，应将该属性化为离散型属性
- 测试条件“年薪<=v”

如何确定阈值v?

# 排序法

把区间 $[a^i, a^{i+1})$ 的中位点 $\frac{a^i+a^{i+1}}{2}$ 作为候选划分点.

| 年薪 | 125K | 100K | 70K | 120K | 95K | 60K | 220K | 85K | 75K | 90K |
|---|---|---|---|---|---|---|---|---|---|---|
| 类 | No | No | No | No | Yes | No | No | Yes | No | Yes |

| 类 | No | No | No | Yes | Yes | Yes | No | No | No | No |
|---|---|---|---|---|---|---|---|---|---|---|
| 排序后值 | 年薪 | | | | | | | | | |
| | 60 | 70 | 75 | 85 | 90 | 95 | 100 | 120 | 125 | 220 |

| 类 | No | No | No | Yes | Yes | Yes | No | No | No | No |
|---|---|---|---|---|---|---|---|---|---|---|
| 排序后值 | 年薪 | | | | | | | | | |
| | No | No | No | Yes | Yes | Yes | No | No | No | No |
| 排序后值 | 年薪 | | | | | | | | | |
| | < 60 | 70 | 75 | 85 | 90 | 95 | 100 | 120 | 125 | 220 |
| Yes | ( 55 | 65 | 72 | 80 | 87 | 92 | 97 | 110 | 122 | 172 | 230 |
| No | ( <= > | <= > | <= > | <= > | <= > | <= > | <= > | <= > | <= > | <= > | <= > |
| Gini | 0.420 | 0.400 | 0.375 | 0.343 | 0.417 | 0.400 | 0.300 | 0.343 | 0.375 | 0.400 | 0.420 |

# 例

| 属性1 | 属性2 | 属性3 | 类 |
|---|---|---|---|
| A | 70 | 真 | 1 |
| A | 90 | 真 | 2 |
| A | 85 | 假 | 2 |
| A | 95 | 假 | 2 |
| A | 70 | 假 | 1 |
| B | 90 | 真 | 1 |
| B | 78 | 假 | 1 |
| B | 65 | 真 | 1 |
| B | 75 | 假 | 1 |
| C | 80 | 真 | 2 |
| C | 70 | 真 | 2 |
| C | 80 | 假 | 1 |
| C | 80 | 假 | 1 |
| C | 96 | 假 | 1 |

| 属性1 | 属性2 | 属性3 | 类 |
|---|---|---|---|
| A | 70 | 真 | 1 |
| A | 90 | 真 | 2 |
| A | 85 | 假 | 2 |
| A | 95 | 假 | 2 |
| A | 70 | 假 | 1 |
| B | 90 | 真 | 1 |
| B | 78 | 假 | 1 |
| B | 65 | 真 | 1 |
| B | 75 | 假 | 1 |
| C | 80 | 真 | 2 |
| C | 70 | 真 | 2 |
| C | 80 | 假 | 1 |
| C | 80 | 假 | 1 |
| C | 96 | 假 | 1 |

- 9个样本属于类1，5个属于类2，因此划分前的熵为

Entropy

$=-9/14\log_2(9/14)-5/14\log_2(5/14)$

$=0.940$

| 属性1 | 属性2 | 属性3 | 类 |
|---|---|---|---|
| A | 70 | 真 | 1 |
| A | 90 | 真 | 2 |
| A | 85 | 假 | 2 |
| A | 95 | 假 | 2 |
| A | 70 | 假 | 1 |
| B | 90 | 真 | 1 |
| B | 78 | 假 | 1 |
| B | 65 | 真 | 1 |
| B | 75 | 假 | 1 |
| C | 80 | 真 | 2 |
| C | 70 | 真 | 2 |
| C | 80 | 假 | 1 |
| C | 80 | 假 | 1 |
| C | 96 | 假 | 1 |

- 按属性1划分可得子集的熵的加权和：

Entropy (A1)

$=5/14(-2/5\,log_2(2/5)-3/5\,log_2(3/5))+4/14(-4/4\,log_2(4/4)-0/4\,log_2(0/4))+5/14(-3/5\,log_2(3/5)-2/5\,log_2(2/5))$

$=0.694$

相应的信息增益：Gain(A1)=0.94−0.694=0.246

| 属性1 | 属性2 | 属性3 | 类 |
|---|---|---|---|
| A | 70 | 真 | 1 |
| A | 90 | 真 | 2 |
| A | 85 | 假 | 2 |
| A | 95 | 假 | 2 |
| A | 70 | 假 | 1 |
| B | 90 | 真 | 1 |
| B | 78 | 假 | 1 |
| B | 65 | 真 | 1 |
| B | 75 | 假 | 1 |
| C | 80 | 真 | 2 |
| C | 70 | 真 | 2 |
| C | 80 | 假 | 1 |
| C | 80 | 假 | 1 |
| C | 96 | 假 | 1 |

- 按属性3划分可得子集的熵的加权和：

$$\text{Entropy (A3)} = 6/14(-3/6\log_2(3/6)-3/6\log_2(3/6)) + 8/14(-6/8\log_2(6/8)-2/8\log_2(2/8)) = 0.892$$

相应的信息增益：Gain(A3)=0.94−0.892=0.048

| 属性1 | 属性2 | 属性3 | 类 |
|---|---|---|---|
| A | 70 | 真 | 1 |
| A | 90 | 真 | 2 |
| A | 85 | 假 | 2 |
| A | 95 | 假 | 2 |
| A | 70 | 假 | 1 |
| B | 90 | 真 | 1 |
| B | 78 | 假 | 1 |
| B | 65 | 真 | 1 |
| B | 75 | 假 | 1 |
| C | 80 | 真 | 2 |
| C | 70 | 真 | 2 |
| C | 80 | 假 | 1 |
| C | 80 | 假 | 1 |
| C | 96 | 假 | 1 |

- 由于属性2是数值型的连续数据，不能简单按上面方式计算。属性2的值的集合是:

  {65,70,75,78,80,85,90,95,96}

  可能的阈值Z的集合是:

  {67,72,77,79,82,87,92,95}。

  从其中选择最优的阈值，可计算得到最优的阈值为Z=82。

| 属性1 | 属性2 | 属性3 | 类 |
|---|---|---|---|
| A | 70 | 真 | 1 |
| A | 90 | 真 | 2 |
| A | 85 | 假 | 2 |
| A | 95 | 假 | 2 |
| A | 70 | 假 | 1 |
| B | 90 | 真 | 1 |
| B | 78 | 假 | 1 |
| B | 65 | 真 | 1 |
| B | 75 | 假 | 1 |
| C | 80 | 真 | 2 |
| C | 70 | 真 | 2 |
| C | 80 | 假 | 1 |
| C | 80 | 假 | 1 |
| C | 96 | 假 | 1 |

- 对应属性2的测试条件(属性2≤82和属性2>82)的信息增益计算：

$$\text{Entropy (A2)}= 9/14(-7/9\log_2(7/9)-2/9\log_2(2/9))$$
$$+5/14(-2/5\log_2(2/5)-3/5\log_2(3/5))$$
$$=0.837$$

相应的信息增益: Gain(A2)=0.94−0.837=0.103

| 属性 | 信息增益 |
| --- | --- |
| 属性1 | 0.246 |
| 属性2 | 0.103 |
| 属性3 | 0.048 |

- 属性1的信息增益最高，选择该属性作为根节点。

  每个属性值具有一个分支，产生3个分支。

| 属性1 | 属性2 | 属性3 | 类 |
|---|---|---|---|
| A | 70 | 真 | 1 |
| A | 90 | 真 | 2 |
| A | 85 | 假 | 2 |
| A | 95 | 假 | 2 |
| A | 70 | 假 | 1 |
| B | 90 | 真 | 1 |
| B | 78 | 假 | 1 |
| B | 65 | 真 | 1 |
| B | 75 | 假 | 1 |
| C | 80 | 真 | 2 |
| C | 70 | 真 | 2 |
| C | 80 | 假 | 1 |
| C | 80 | 假 | 1 |
| C | 96 | 假 | 1 |

$$SplitINFO = -\sum_{i=1}^{k} \frac{n_i}{n} \log_2 \frac{n_i}{n}$$

- 用信息增益比率
  - 各属性的分裂信息与信息增益比率为：
    - SplitINFO(A1)=-5/14log$_2$(5/14)-4/14log$_2$(4/14)-5/14log$_2$(5/14)=1.577bit

      GainRATIO(A1)=0.156
    - SplitINFO(A2)=-5/14log$_2$(5/14)- 9/14log$_2$(9/14)=0.940bit

      GainRATIO(A2)=0.110
    - SplitINFO(A3)=-6/14log$_2$(6/14)- 8/14log$_2$(8/14)=0.985bit

      GainRATIO(A3)=0.049
  - 仍然是属性1的增益比率最大

# 3、缺失值问题

$$S=\left\{\begin{array}{l}(\cdots,x,\cdots)+\\(\cdots,x,\cdots)-\\(\cdots,x,\cdots)+\\(\cdots,y,\cdots)-\\(\cdots,y,\cdots)+\\(\cdots,y,\cdots)+\\(\cdots,y,\cdots)+\\(\cdots,z,\cdots)+\\(\cdots,z,\cdots)+\\(\cdots,z,\cdots)-\\(\cdots,?,\cdots)-\\(\cdots,?,\cdots)+\end{array}\right\}$$

- 在C4.5算法中，有未知值的样本是按照已知值的相对频率随机分布的
  - 例如，给定一个布尔属性A，如果节点n包含6个已知A=1和4个A=0的实例，那么A(x)=1的概率是0.6，而A(x)=0的概率是0.4。于是，实例x的60%被分配到A=1的分支，40%被分配到另一个分支。
- 用系数F合理地修正增益参数
  - F = **数据集中一个给出的属性值具有已知值的样本的数量**/**数据集中样本数量总和**。

- 新的增益标准有以下形式：

$$Gain(A)=F\cdot(Entropy(p)-Entropy_p(A))$$

- 把具有未知值的样本看作分区的一个附加组修改SplitINFO(A)。如果属性A有n个输出，SplitINFO(A)按照属性把数据集分区成n+1个子集计算。

| 属性1 | 属性2 | 属性3 | 类 |
|---|---|---|---|
| A | 70 | 真 | 1 |
| A | 90 | 真 | 2 |
| A | 85 | 假 | 2 |
| A | 95 | 假 | 2 |
| A | 70 | 假 | 1 |
| ? | 90 | 真 | 1 |
| B | 78 | 假 | 1 |
| B | 65 | 真 | 1 |
| B | 75 | 假 | 1 |
| C | 80 | 真 | 2 |
| C | 70 | 真 | 2 |
| C | 80 | 假 | 1 |
| C | 80 | 假 | 1 |
| C | 96 | 假 | 1 |

- 例：该例有14个样本，属性1有一个丢失值，用“?”表示。只有13个样本数据完整。

| 属性1 | 属性2 | 属性3 | 类 |
|---|---|---|---|
| A | 70 | 真 | 1 |
| A | 90 | 真 | 2 |
| A | 85 | 假 | 2 |
| A | 95 | 假 | 2 |
| A | 70 | 假 | 1 |
| ? | 90 | 真 | 1 |
| B | 78 | 假 | 1 |
| B | 65 | 真 | 1 |
| B | 75 | 假 | 1 |
| C | 80 | 真 | 2 |
| C | 70 | 真 | 2 |
| C | 80 | 假 | 1 |
| C | 80 | 假 | 1 |
| C | 96 | 假 | 1 |

划分前的熵为：

Entropy

$=-8/13\log_2(8/13) -5/13\log_2(5/13)$

$=0.961$

| 属性1 | 属性2 | 属性3 | 类 |
|---|---|---|---|
| A | 70 | 真 | 1 |
| A | 90 | 真 | 2 |
| A | 85 | 假 | 2 |
| A | 95 | 假 | 2 |
| A | 70 | 假 | 1 |
| ? | 90 | 真 | 1 |
| B | 78 | 假 | 1 |
| B | 65 | 真 | 1 |
| B | 75 | 假 | 1 |
| C | 80 | 真 | 2 |
| C | 70 | 真 | 2 |
| C | 80 | 假 | 1 |
| C | 80 | 假 | 1 |
| C | 96 | 假 | 1 |

Entropy(A1)

$=5/13(-2/5\log_2(2/5)-3/5\log_2(3/5))$

$+3/13(-3/3\log_2(3/3)-0/3\log_2(0/3))$

$+5/13(-3/5\log_2(3/5)-2/5\log_2(2/5))$

$=0.747$

- 用F(F=13/14)修正信息增益：
  - Gain(A1)=13/14(0.961-0.747)=0.199

| 属性1 | 属性2 | 属性3 | 类 |
|---|---|---|---|
| A | 70 | 真 | 1 |
| A | 90 | 真 | 2 |
| A | 85 | 假 | 2 |
| A | 95 | 假 | 2 |
| A | 70 | 假 | 1 |
| ? | 90 | 真 | 1 |
| B | 78 | 假 | 1 |
| B | 65 | 真 | 1 |
| B | 75 | 假 | 1 |
| C | 80 | 真 | 2 |
| C | 70 | 真 | 2 |
| C | 80 | 假 | 1 |
| C | 80 | 假 | 1 |
| C | 96 | 假 | 1 |

- SplitINFO(A1)=

$-5/14\log_2(5/14)-3/14\log_2(3/14)$

$-5/14\log_2(5/14)-1/14\log_2(1/14))$

$=1.876$

- 和没有缺失值时的分裂信息相比较有何变化?

- 若属性1作为当前节点的测试属性，则将分为三个子分支，而有缺失值的样本将以一定的概率分别分配给每个子分支。

$T_1$：(属性 1=A)

| 属性 2 | 属性 3 | 类 | W |
|---|---|---|---|
| 70 | 真 | 类 1 | 1 |
| 90 | 真 | 类 2 | 1 |
| 85 | 假 | 类 2 | 1 |
| 95 | 假 | 类 2 | 1 |
| 70 | 假 | 类 1 | 1 |
| 90 | 真 | 类 1 | 5/13 |

$T_2$：(属性 1=B)

| 属性 2 | 属性 3 | 类 | W |
|---|---|---|---|
| 90 | 真 | 类 1 | 3/13 |
| 78 | 假 | 类 1 | 1 |
| 65 | 真 | 类 1 | 1 |
| 75 | 假 | 类 1 | 1 |

$T_3$：(属性 1=C)

| 属性 2 | 属性 3 | 类 | W |
|---|---|---|---|
| 80 | 真 | 类 2 | 1 |
| 70 | 真 | 类 2 | 1 |
| 80 | 假 | 类 1 | 1 |
| 80 | 假 | 类 1 | 1 |
| 96 | 假 | 类 1 | 1 |
| 90 | 真 | 类 1 | 5/13 |

$$Entropy(t) = -\sum_j p(j|t)\log_2 p\,(j|t)$$


$T_1$：(属性 1=A)

| 属性 2 | 属性 3 | 类 | W |
|---|---|---|---|
| 70 | 真 | 类 1 | 1 |
| 90 | 真 | 类 2 | 1 |
| 85 | 假 | 类 2 | 1 |
| 95 | 假 | 类 2 | 1 |
| 70 | 假 | 类 1 | 1 |
| 90 | 真 | 类 1 | 5/13 |

$T_2$：(属性 1=B)

| 属性 2 | 属性 3 | 类 | W |
|---|---|---|---|
| 90 | 真 | 类 1 | 3/13 |
| 78 | 假 | 类 1 | 1 |
| 65 | 真 | 类 1 | 1 |
| 75 | 假 | 类 1 | 1 |

(属性 1=C)

| 属性 2 | 属性 3 | 类 | W |
|---|---|---|---|
| 80 | 真 | 类 2 | 1 |
| 70 | 真 | 类 2 | 1 |
| 80 | 假 | 类 1 | 1 |
| 80 | 假 | 类 1 | 1 |
| 96 | 假 | 类 1 | 1 |
| 90 | 真 | 类 1 | 5/13 |

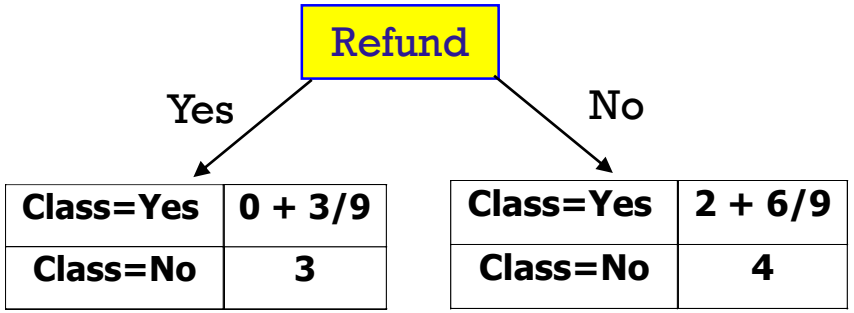
# 举例：

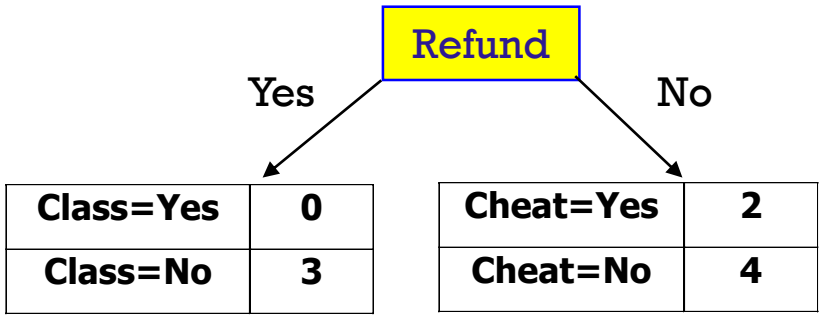
| Tid | Refund | Marital Status | Taxable Income | Class |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | ? | Married | 90K | Yes |

Probability that Refund=Yes is 3/9

Probability that Refund=No is 6/9

Assign record to the left child with weight = 3/9 and to the right child with weight = 6/9



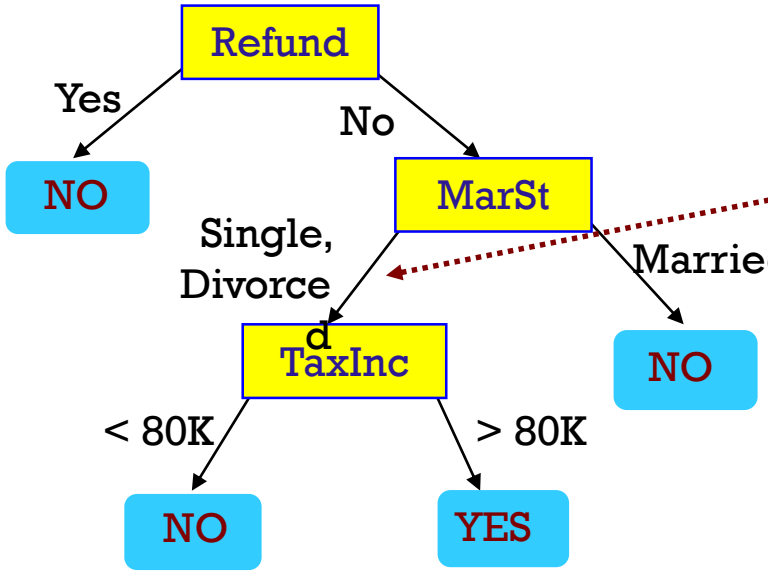



# 决策树构造完成后，如果测试样本的属性值不完整，该如何确定该样本的类别?

New record:

| *Tid* | Refund | Marital Status | Taxable Income | Class |
|---|---|---|---|---|
| 11 | No | ? | 85K | ? |

| | Married | Single | Divorced | Total |
|---|---|---|---|---|
| Class=No | 3 | 1 | 0 | 4 |
| Class=Yes | 6/9 | 1 | 1 | 2.67 |
| Total | 3.67 | 2 | 1 | 6.67 |




Probability that Marital Status = Married is 3.67/6.67

Probability that Marital Status ={Single,Divorced} is 3/6.67

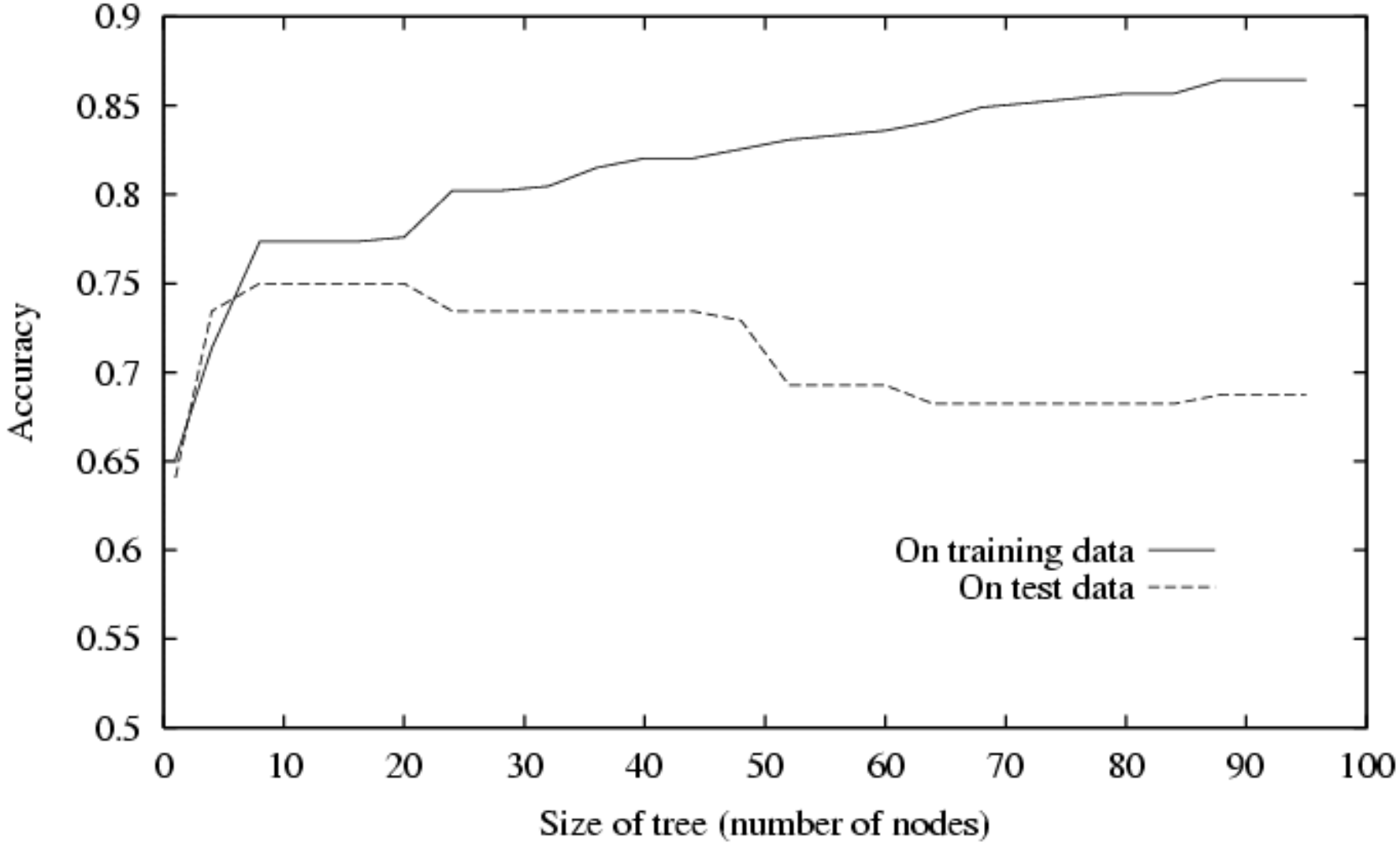
| *Tid* | Refund | Marital Status | Taxable Income | Class |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | ? | Married | 90K | Yes |

# 4、决策树的剪枝

- 过拟合（Overfitting，过学习）
  - 定义：给定一个假设空间H，一个假设h∈H，如果存在其他的假设h'∈H，使得在训练样例上h的错误率比h'小，但在整个实例分布上h'的错误率比h小，那么就说假设h过度拟合训练数据。

$$error_{train}(h) < error_{train}(h')$$

$$error_{\mathcal{D}}(h) > error_{\mathcal{D}}(h')$$


0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Accuracy
0
10
20
30
40
50
60
70
80
90
100
Size of tree (number of nodes)
On training data
On test data

# 导致过度拟合的原因

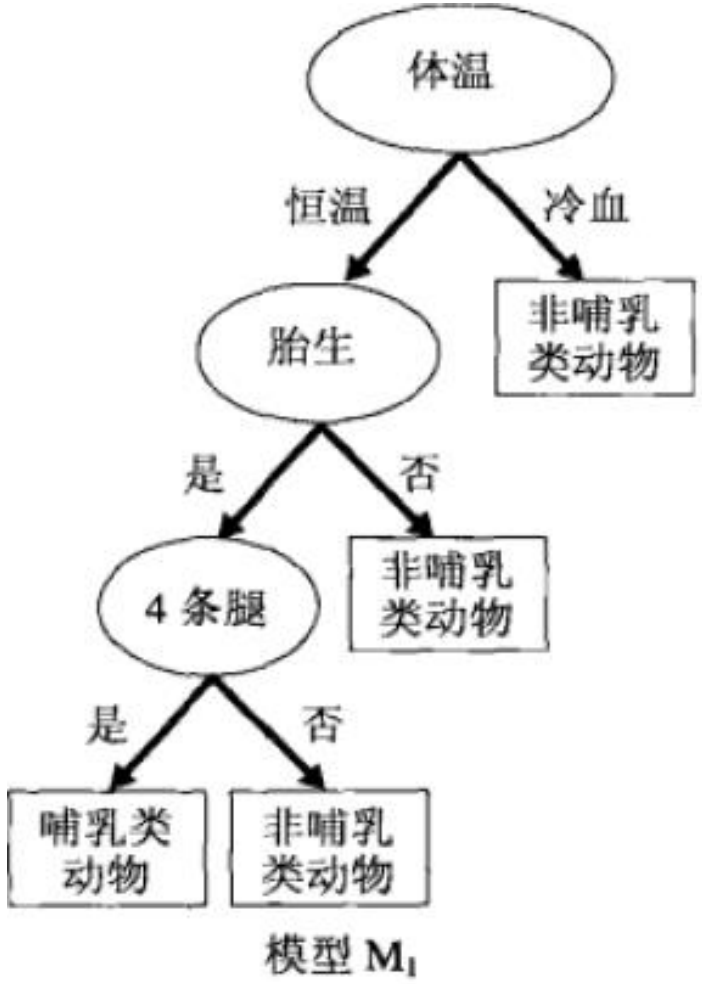
- 训练样例含有随机错误或噪声
- 当少量的样例被关联到叶节点时，很可能出现巧合的规律性。

# 实例1：噪声导致的过度拟合

哺乳类动物分类的训练数据集样本。打星号的类标号代表错误标记的记录

| 名称 | 体温 | 胎生 | 4条腿 | 冬眠 | 类标号 |
|---|---|---|---|---|---|
| 豪猪 | 恒温 | 是 | 是 | 是 | 是 |
| 猫 | 恒温 | 是 | 是 | 否 | 是 |
| 蝙蝠 | 恒温 | 是 | 否 | 是 | 否* |
| 鲸 | 恒温 | 是 | 否 | 否 | 否* |
| 蝾螈 | 冷血 | 否 | 是 | 是 | 否 |
| 科莫多巨蜥 | 冷血 | 否 | 是 | 否 | 否 |
| 蟒蛇 | 冷血 | 否 | 否 | 是 | 否 |
| 鲑鱼 | 冷血 | 否 | 否 | 否 | 否 |
| 鹰 | 恒温 | 否 | 否 | 否 | 否 |
| 虹鳉 | 冷血 | 是 | 否 | 否 | 否 |

- 决策树M1
- 完全拟合训练数据



模型 $M_1$

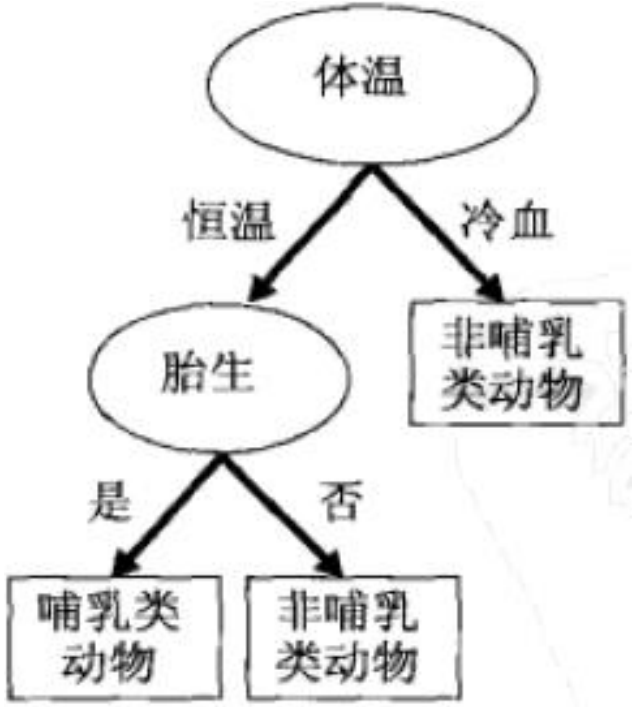
哺乳类动物分类的训练数据集样本。打星号的类标号代表错误标记的记录

| 名称 | 体温 | 胎生 | 4 条腿 | 冬眠 | 类标号 |
|---|---|---|---|---|---|
| 豪猪 | 恒温 | 是 | 是 | 是 | 是 |
| 猫 | 恒温 | 是 | 是 | 否 | 是 |
| 蝙蝠 | 恒温 | 是 | 否 | 是 | 否* |
| 鲸 | 恒温 | 是 | 否 | 否 | 否* |
| 蝾螈 | 冷血 | 否 | 是 | 是 | 否 |
| 科莫多巨蜥 | 冷血 | 否 | 是 | 否 | 否 |
| 蟒蛇 | 冷血 | 否 | 否 | 是 | 否 |
| 鲑鱼 | 冷血 | 否 | 否 | 否 | 否 |
| 鹰 | 恒温 | 否 | 否 | 否 | 否 |
| 虹鳉 | 冷血 | 是 | 否 | 否 | 否 |

- 决策树M2
- 训练误差为20%

哺乳类动物分类的训练数据集样本。打星号的类标号代表错误标记的记录

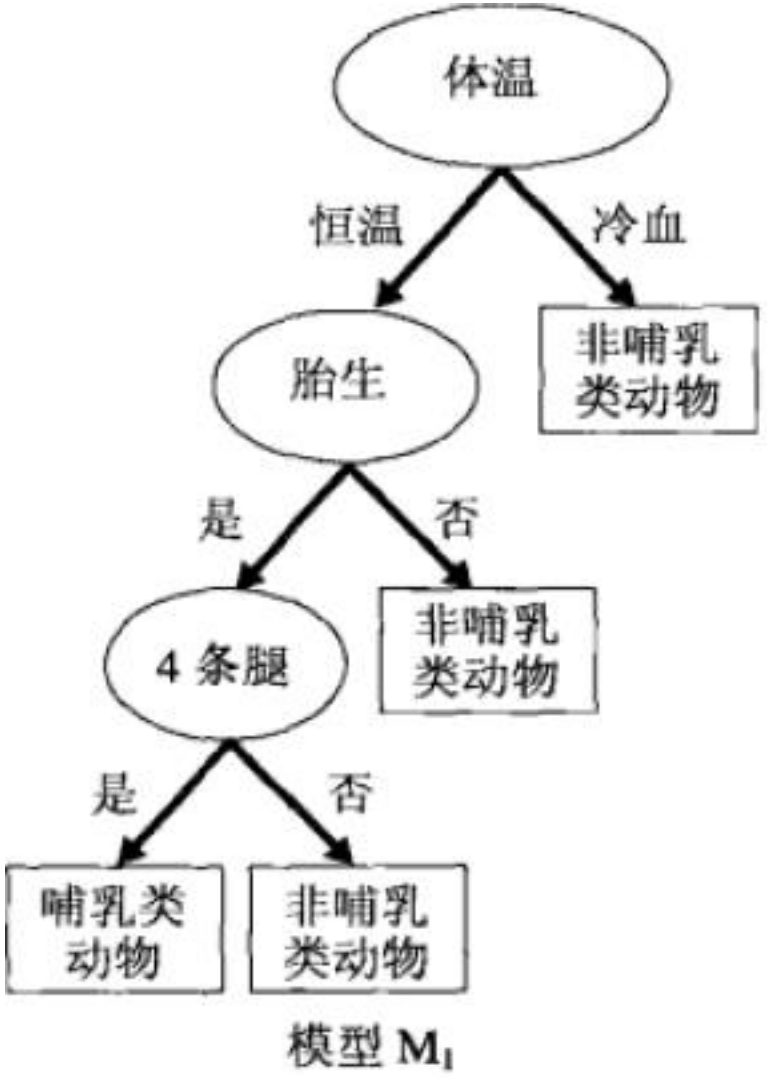
| 名称 | 体温 | 胎生 | 4条腿 | 冬眠 | 类标号 |
|---|---|---|---|---|---|
| 豪猪 | 恒温 | 是 | 是 | 是 | 是 |
| 猫 | 恒温 | 是 | 是 | 否 | 是 |
| 蝙蝠 | 恒温 | 是 | 否 | 是 | 否* |
| 鲸 | 恒温 | 是 | 否 | 否 | 否* |
| 蝾螈 | 冷血 | 否 | 是 | 是 | 否 |
| 科莫多巨蜥 | 冷血 | 否 | 是 | 否 | 否 |
| 蟒蛇 | 冷血 | 否 | 否 | 是 | 否 |
| 鲑鱼 | 冷血 | 否 | 否 | 否 | 否 |
| 鹰 | 恒温 | 否 | 否 | 否 | 否 |
| 虹鳉 | 冷血 | 是 | 否 | 否 | 否 |




模型 $M_2$

哺乳类动物分类的检验数据集样本

| 名称 | 体温 | 胎生 | 4 条腿 | 冬眠 | 类标号 |
|---|---|---|---|---|---|
| 人 | 恒温 | 是 | 否 | 否 | 是 |
| 鸽子 | 恒温 | 否 | 否 | 否 | 否 |
| 象 | 恒温 | 是 | 是 | 否 | 是 |
| 豹纹鲨 | 冷血 | 是 | 否 | 否 | 否 |
| 海龟 | 冷血 | 否 | 是 | 否 | 否 |
| 企鹅 | 冷血 | 否 | 否 | 否 | 否 |
| 鳗 | 冷血 | 否 | 否 | 否 | 否 |
| 海豚 | 恒温 | 是 | 否 | 否 | 是 |
| 针鼹 | 恒温 | 否 | 是 | 是 | 是 |
| 希拉毒蜥 | 冷血 | 否 | 是 | 是 | 否 |

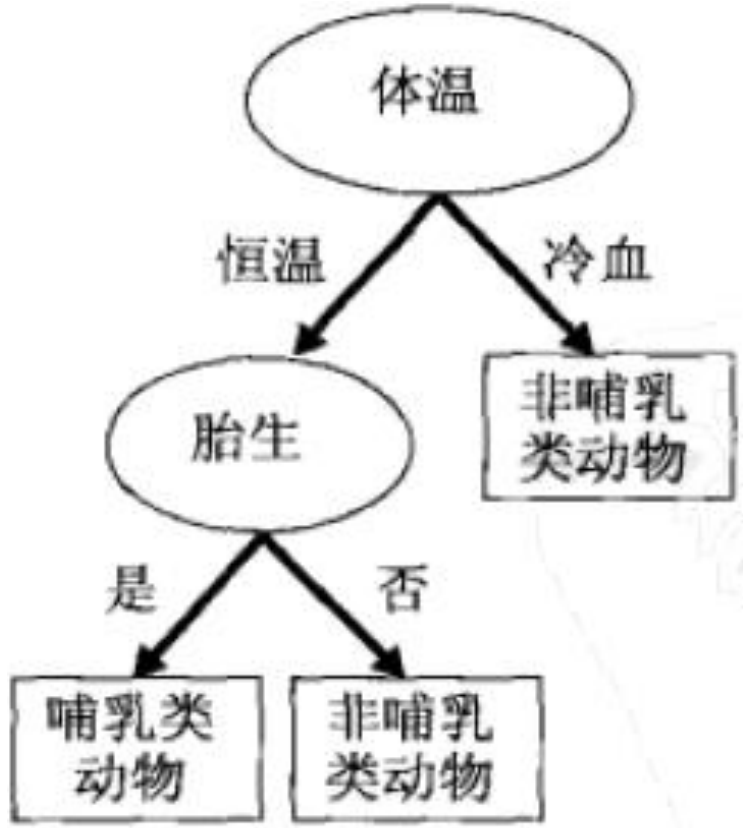


模型 $M_1$

- 利用模型M1
- 人、海豚、针鼹被错误分类，检验误差为30%

哺乳类动物分类的检验数据集样本

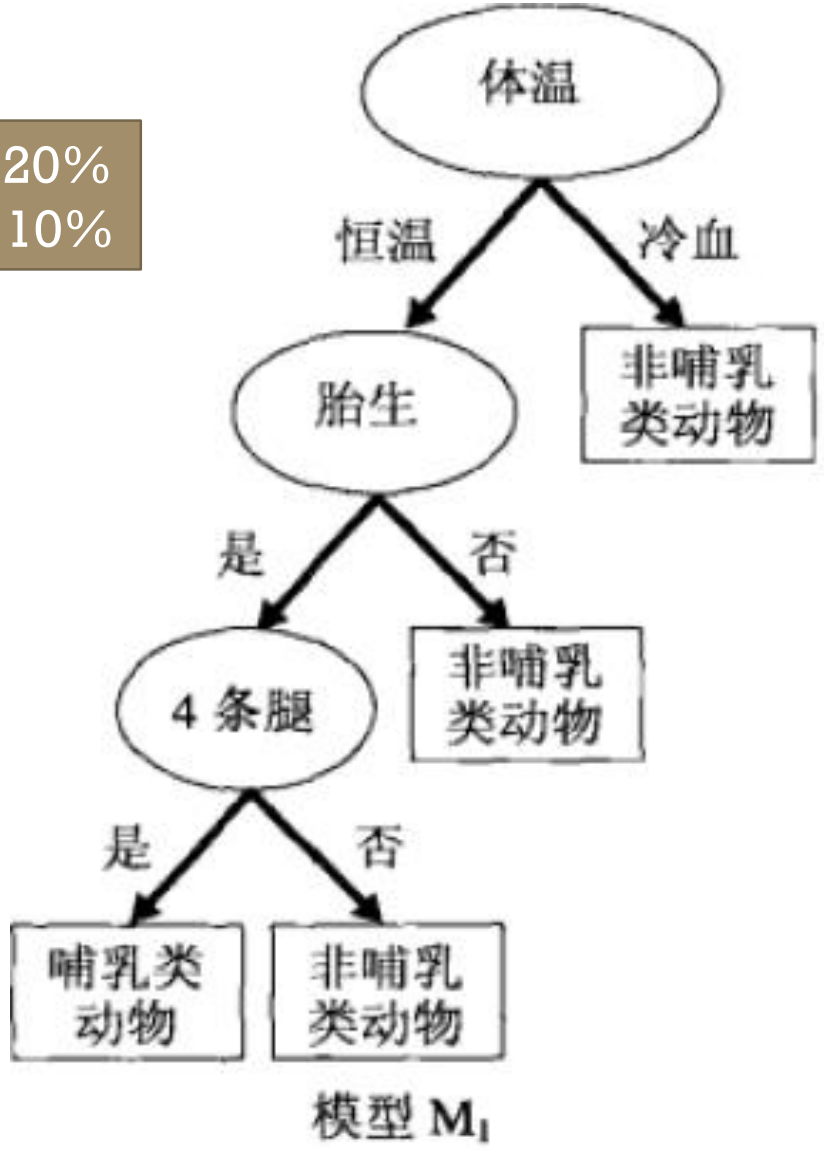
| 名称 | 体温 | 胎生 | 4 条腿 | 冬眠 | 类标号 |
|---|---|---|---|---|---|
| 人 | 恒温 | 是 | 否 | 否 | 是 |
| 鸽子 | 恒温 | 否 | 否 | 否 | 否 |
| 象 | 恒温 | 是 | 是 | 否 | 是 |
| 豹纹鲨 | 冷血 | 是 | 否 | 否 | 否 |
| 海龟 | 冷血 | 否 | 是 | 否 | 否 |
| 企鹅 | 冷血 | 否 | 否 | 否 | 否 |
| 鳗 | 冷血 | 否 | 否 | 否 | 否 |
| 海豚 | 恒温 | 是 | 否 | 否 | 是 |
| 针鼹 | 恒温 | 否 | 是 | 是 | 是 |
| 希拉毒蜥 | 冷血 | 否 | 是 | 是 | 否 |



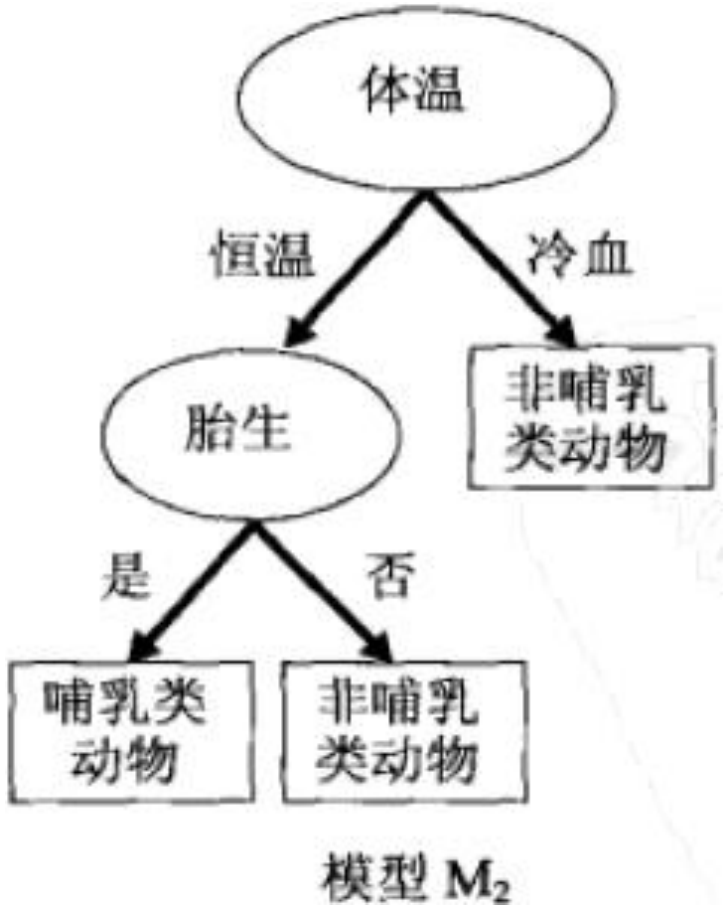
模型 $M_2$

- 利用模型M2
- 针鼹被错误分类，检验误差为10%

训练误差，M1 0%，M2 20%
检验误差，M1 30%，M2 10%



模型 $M_1$



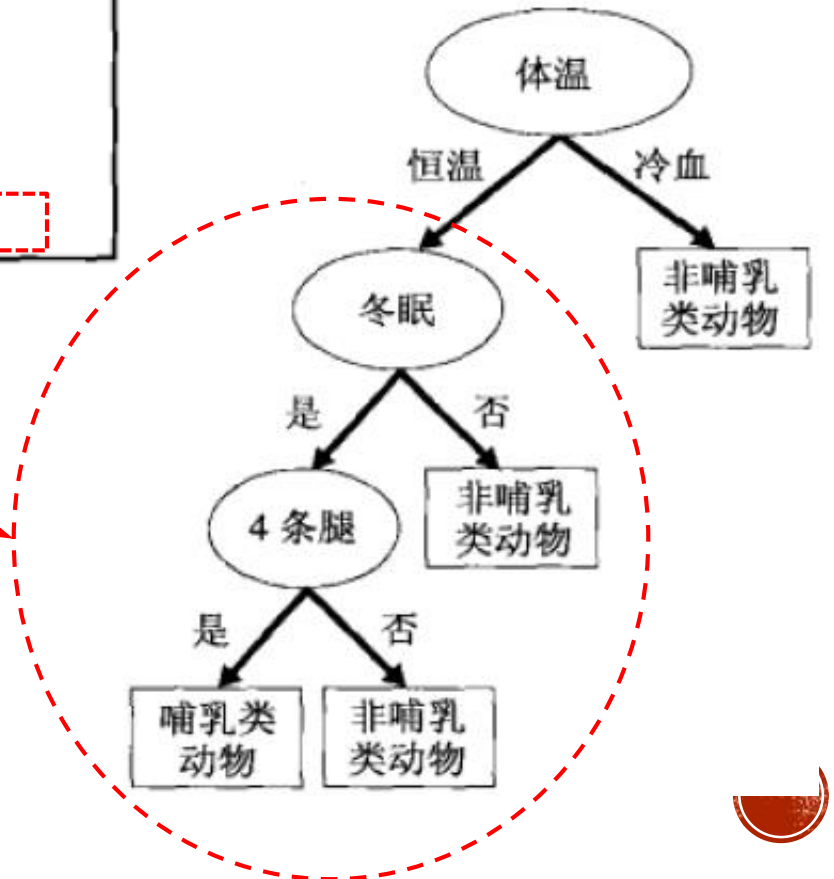
模型 $M_2$

模型M1过拟合

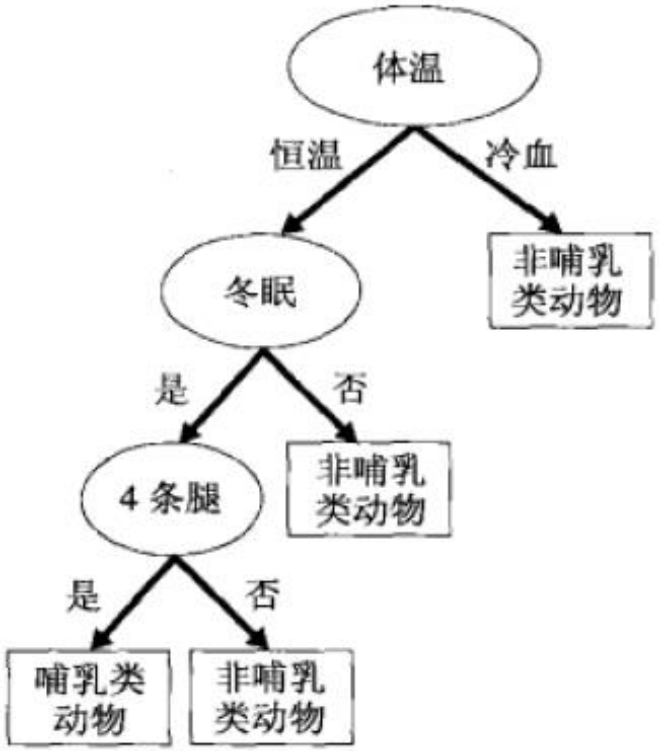
- 实例2：缺乏代表性样本导致的过度拟合

哺乳动物分类的训练集样本

| 名称 | 体温 | 胎生 | 4 条腿 | 冬眠 | 类标号 |
|---|---|---|---|---|---|
| 蝾螈 | 冷血 | 否 | 是 | 是 | 否 |
| 虹鳉 | 冷血 | 是 | 否 | 否 | 否 |
| 鹰 | 恒温 | 否 | 否 | 否 | 否 |
| 弱夜鹰 | 恒温 | 否 | 否 | 是 | 否 |
| 鸭嘴兽 | 恒温 | 否 | 是 | 是 | 是 |




- 完全拟合数据




哺乳类动物分类的检验数据集样本

| 名称 | 体温 | 胎生 | 4条腿 | 冬眠 | 类标号 |
|---|---|---|---|---|---|
| 人 | 恒温 | 是 | 否 | 否 | 是 |
| 鸽子 | 恒温 | 否 | 否 | 否 | 否 |
| 象 | 恒温 | 是 | 是 | 否 | 是 |
| 豹纹鲨 | 冷血 | 是 | 否 | 否 | 否 |
| 海龟 | 冷血 | 否 | 是 | 否 | 否 |
| 企鹅 | 冷血 | 否 | 否 | 否 | 否 |
| 鳗 | 冷血 | 否 | 否 | 否 | 否 |
| 海豚 | 恒温 | 是 | 否 | 否 | 是 |
| 针鼹 | 恒温 | 否 | 是 | 是 | 是 |
| 希拉毒蜥 | 冷血 | 否 | 是 | 是 | 否 |

- 人、象、海豚都被错误分类
- 这是因为决策树把恒温但不冬眠的脊椎动物划分为非哺乳动物，决策树的决策是由于只有一个训练记录（鹰）具有这些特性。

# 决策树中避免过拟合——剪枝

- **预剪枝（pre-prune），及早停止增长树法，在完美分类训练数据之前停止增长树**，这种方法实际中的效果并不好。

- **后剪枝（post-prune），即允许树过度拟合数据，然后对这棵树后修剪**。对拥有同样父节点的一组节点进行检查，判断是否可以合并一个节点。后剪枝是目前最普遍的做法。

# 错误率降低修剪（REDUCED-ERROR PRUNING）

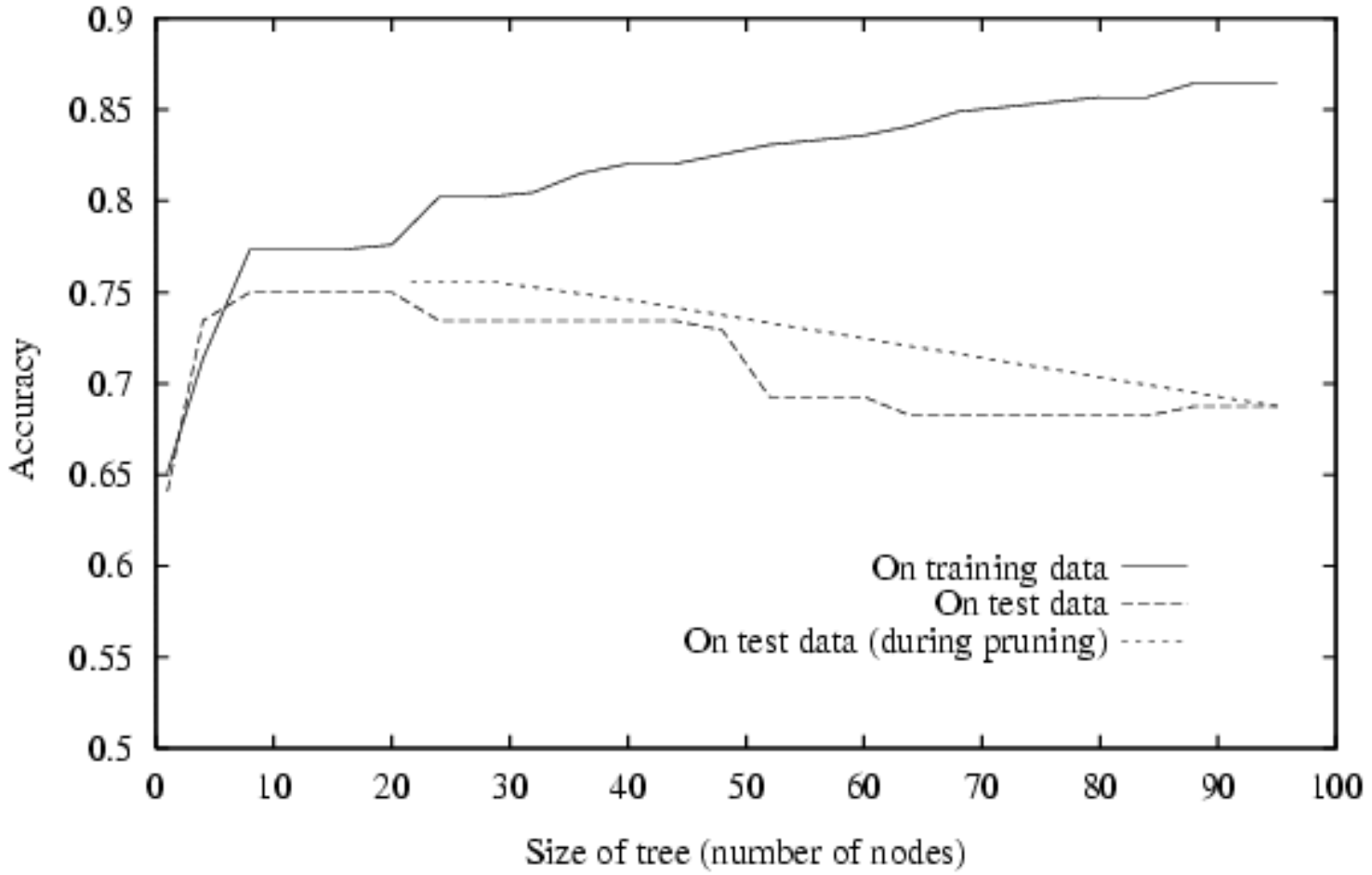
- 将树上的每一个节点作为修剪的候选对象
- 数据分成3个子集
  - 训练样例：形成决策树
  - 验证样例：修剪决策树
  - 测试样例：提供在未来的未见实例上的精度的无偏估计

# 错误率降低修剪（REDUCED-ERROR PRUNING）

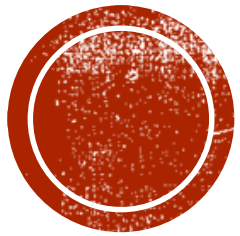
修剪步骤

- 删除以此节点为根的子树，使它成为叶节点
- 把和该节点相关联的训练样例的最常见分类赋给它
- 反复修剪节点，每次总是选取那些删除后可以最大提高决策树在验证集合上的精度的节点
- 继续修剪，直到进一步的修剪是有害的为止

# 决策树学习中错误率降低修剪的效果




这个方法的主要缺点是当数据有限时，从中保留一部分用作验证集合进一步减少了训练可以使用的样例



# THE END!