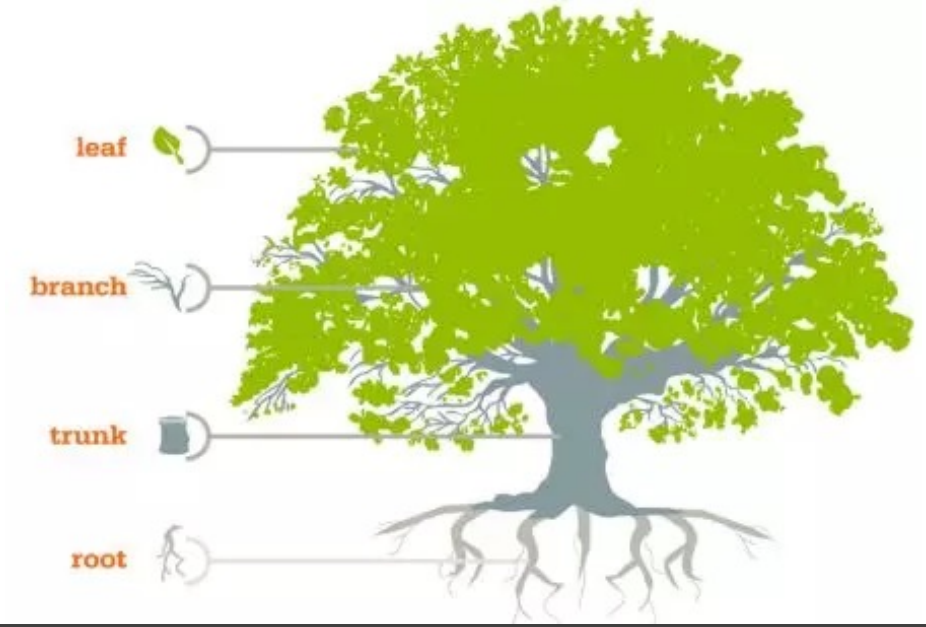
决策树

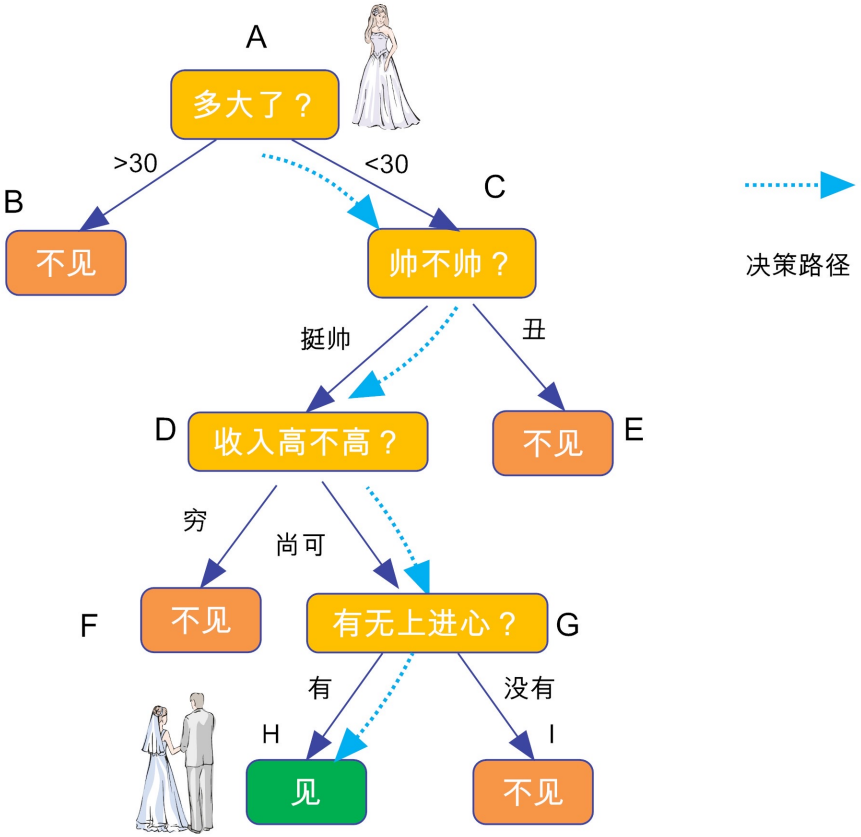
万永权



0. 问题引入

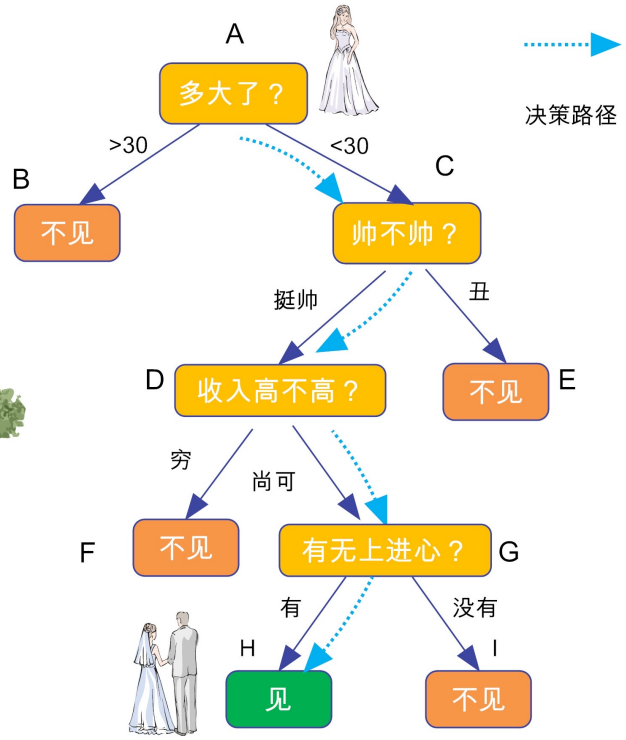
一段日常对话

妈妈操心翠花的终身大事，给介绍了一个相亲对象。
翠花随口一问：多大了？
妈妈：长你一岁，27。
翠花又问：帅不帅？
妈妈：挺帅的。
翠花：收入咋样？
妈妈：比上不足比下有余，中等。
翠花又问：有上进心吗？
妈妈：有，还写过几本上不错的书呢。
翠花：那好，去见见吧。



决策树看作有向图

- 决策树是一个分层结构，可以为每个节点赋予一个层次数。根节点的层次数为0， 子节点的层次数为父节点的层次数加1。树的深度的定义为所有节点的最大层次数。



- 根节点（root node）：没有入边，但有零条或多条出边。如图节点A。
- 内部节点（internal node）：恰好只有一条入边，但有一条或多条出边。如图节点C、D和G。
- 叶子节点（leaf node）：恰好只有一条入边，但没有任何出边。如图节点B、E、F、H和I。

决策树

- 决策树是一种树状结构，通过做出一系列决策（选择）来对数据进行划分，这类似于针对一系列问题进行选择。
- 决策树的决策过程就是从根节点开始，测试待分类项中对应的特征属性，并按照其值选择输出分支，直到叶子节点，将叶子节点的存放的类别作为决策结果。

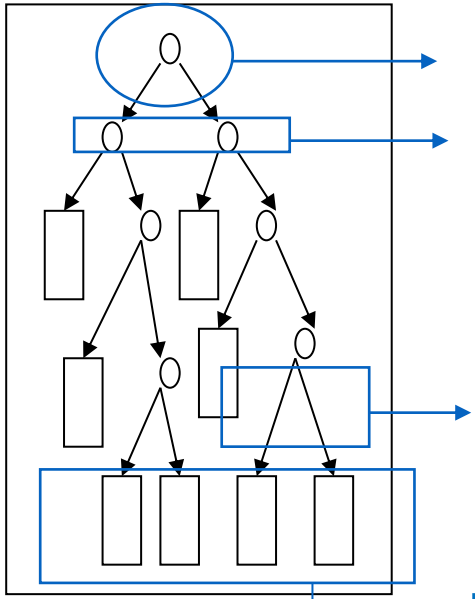
- 决策树：从训练数据中学习得出一个树状结构的模型。
- 决策树属于判别式模型。

决策树算法

- 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，用于对新数据进行预测。
- 决策树算法属于监督学习方法。

- 决策树归纳的基本算法是贪心算法，自顶向下来构建决策树。
- 贪心算法：在每一步选择中都采取在当前状态下最好/优的选择。
- 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。

决策树的结构



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)

决策树分类

1、训练阶段

从给定的训练数据集 TrainingSet，构造出一棵决策树

DecisionTree(TrainingSet)

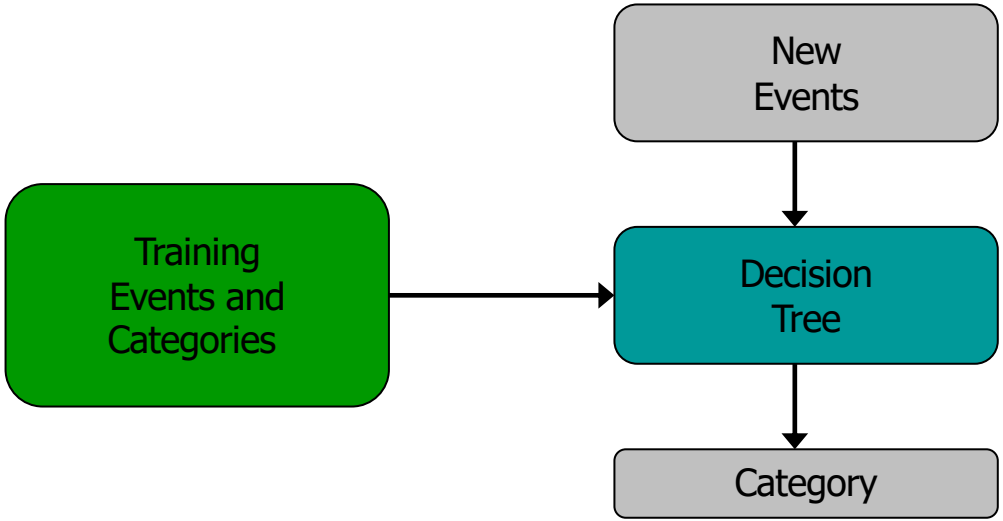
2、分类阶段

从根节点开始，按照决策树的分类属性逐层往下划分，直到叶节点，获得分类结果。

$$\hat{y} = \text{DecisionTree}(x)$$

决策树

- 使用决策树来预测新事件（样本）的分类
- 使用训练数据构建决策树

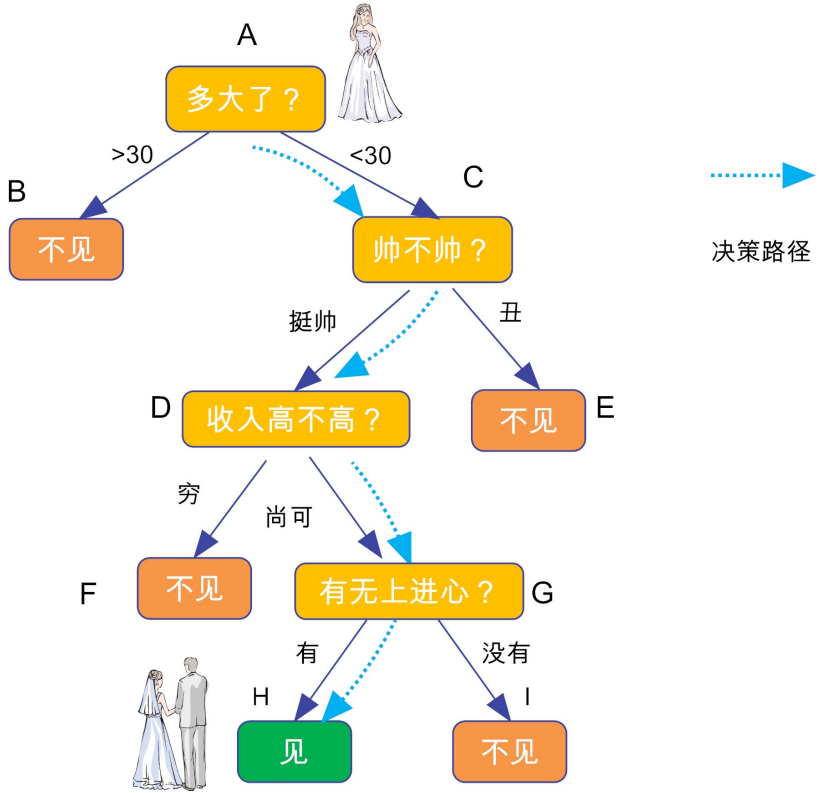


决策树模型

- 典型的决策树有ID3、C4.5、CART (Classification And Regression Tree， 分类树与回归树）。
- 它们的区别在于树的结构与构造算法不同。
- CART既支持分类问题，也可以用于回归问题。
- 决策树是一种判别模型，天然支持多种分类问题。

决策树与if-then规则

- 由决策树的根结点到叶结点的每一条路径构建一条规则;
- 路径上内部结点的特征对应着规则的条件，而叶结点的类对应着规则的结论。
- If-then规则集合的一重要性质：互斥并且完备



决策树的应用领域

例子:

- 设备或医疗诊断
- 信用风险分析
- 日历调度偏好

案例分析

- 为了更好地理解决策树算法的原理与学习如何用决策树算法解决实际问题，我们先来举一个例子。
- 一天，老师问了个问题：只根据头发和声音怎么判断一位同学的性别？
- 为了解决这个问题，同学们马上简单地统计了7位同学的相关特征，男女特征统计表如表2.2所示。

表 2.2　男女特征统计表

头　发	声　音	性　别
长	粗	男
短	粗	男
短	粗	男
长	细	女
短	细	女
短	粗	女
长	粗	女
长	粗	女

案例分析

- 机智的同学A想了想，先根据头发判断性别，若判断不出，再根据声音判断性别，于是画了一幅图，如图2.16所示

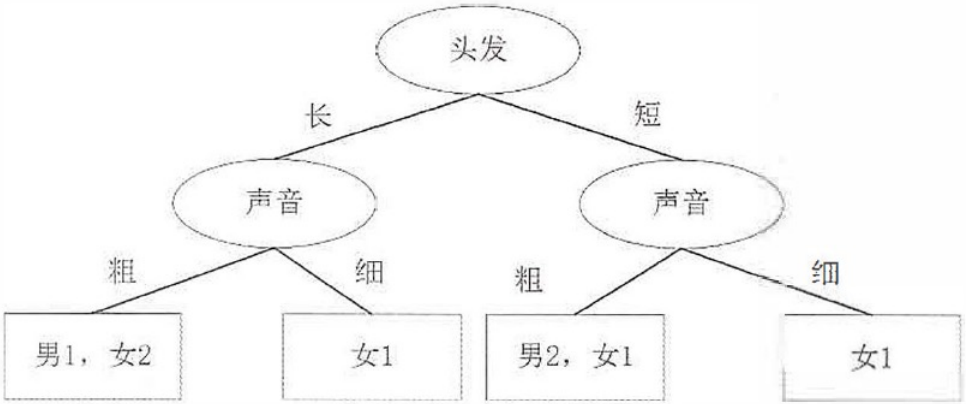


图 2.16　决策树 A

- 于是，一个简单、直观的决策树就这么得出了。头发长、声音粗就是男生；头发长、声音细就是女生：头发短、声音粗是男生；头发短、声音细是女生。

案例分析

- 这时同学B提出，想先根据声音判断，再根据头发来判断，于是同学B大手一挥，也画了个决策树，如图2.17所示。

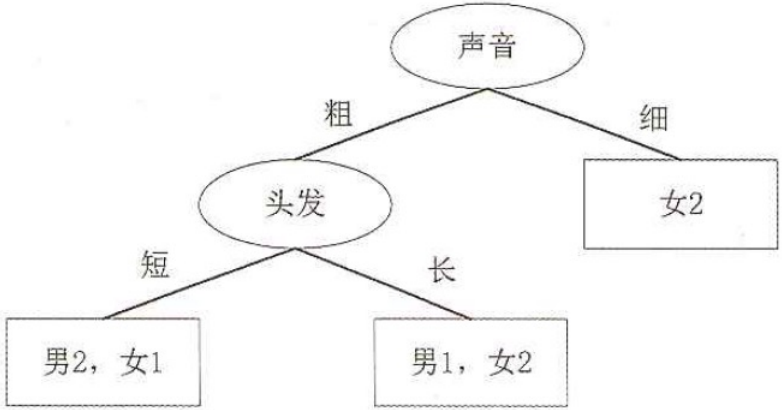


图 2.17　决策树 B

- 同学B的决策树：首先判断声音，声音细，就是女生：声音粗、头发长的是男生：声音粗、头发长的是女生。

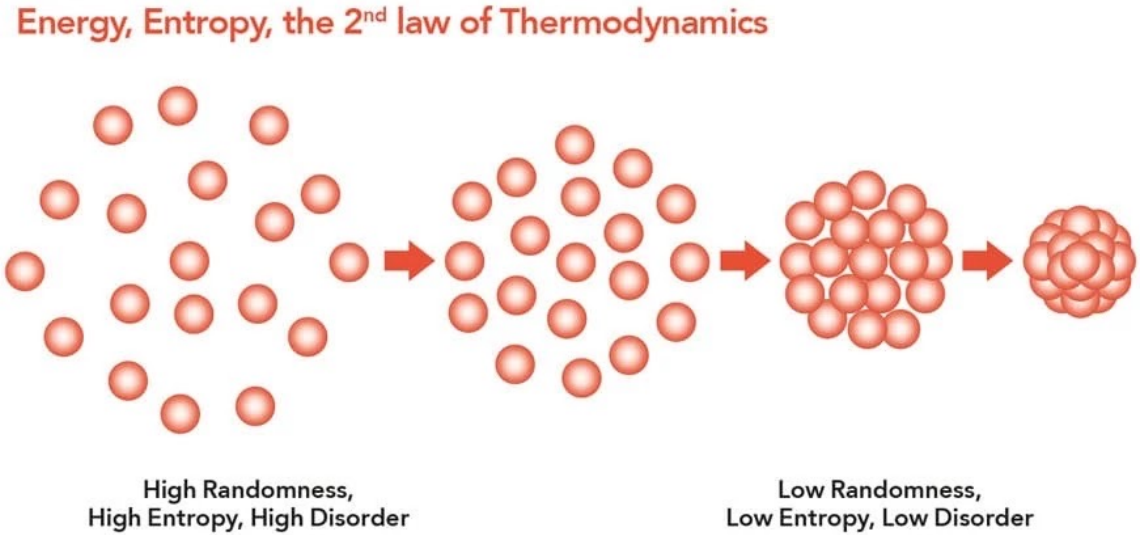
如何构建决策树

- 那么问题来了：同学A和同学B的决策树，谁的更好些？计算机做决策树的时候，面对多个特征，该选哪个特征作为最优分类特征？

ID3 算法

- ID3 算法最早是由罗斯昆（J. Ross Quinlan）于1975年提出的一种决策树构建算法，算法的核心是“信息熵”。

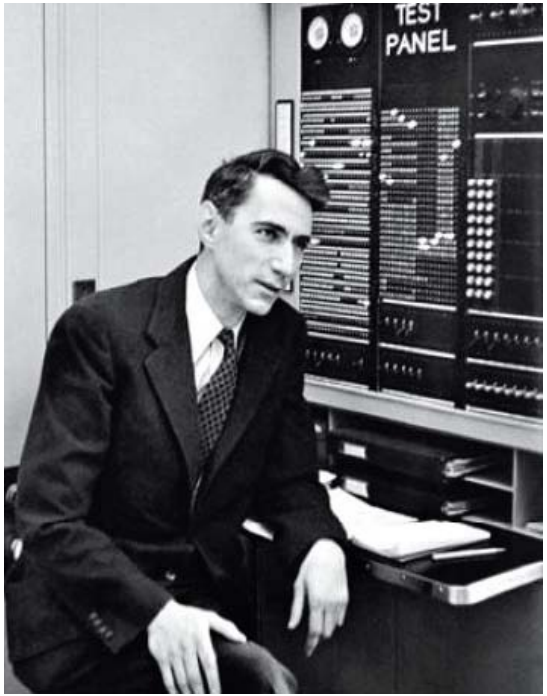
- ID3 算法是以信息论为基础，以信息增益为衡量标准，从而实现对数据的归纳分类。
- ID3 算法计算每个属性的信息增益，并选取具有最高增益的属性作为给定的测试属性。


Energy, Entropy, the 2nd law of Thermodynamics
High Randomness,
High Entropy, High Disorder
Low Randomness,
Low Entropy, Low Disorder

1. 熵的概念

信息熵

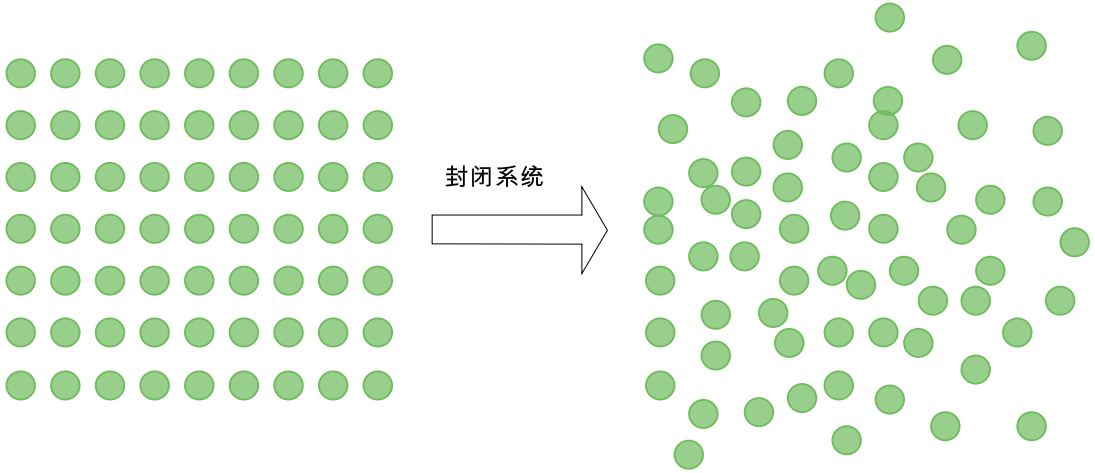
- 1948年，香农提出了“信息熵”的概念，解决了对信息的量化度量问题。
- 信息熵定义为离散随机事件的出现概率。
- 变量的不确定性越大，熵也就越大，把它搞清楚所需要的信息量也就越大。
- 熵的概念源自热物理学。



克劳德·艾尔伍德·香农，美国数学家、信息论的创始人

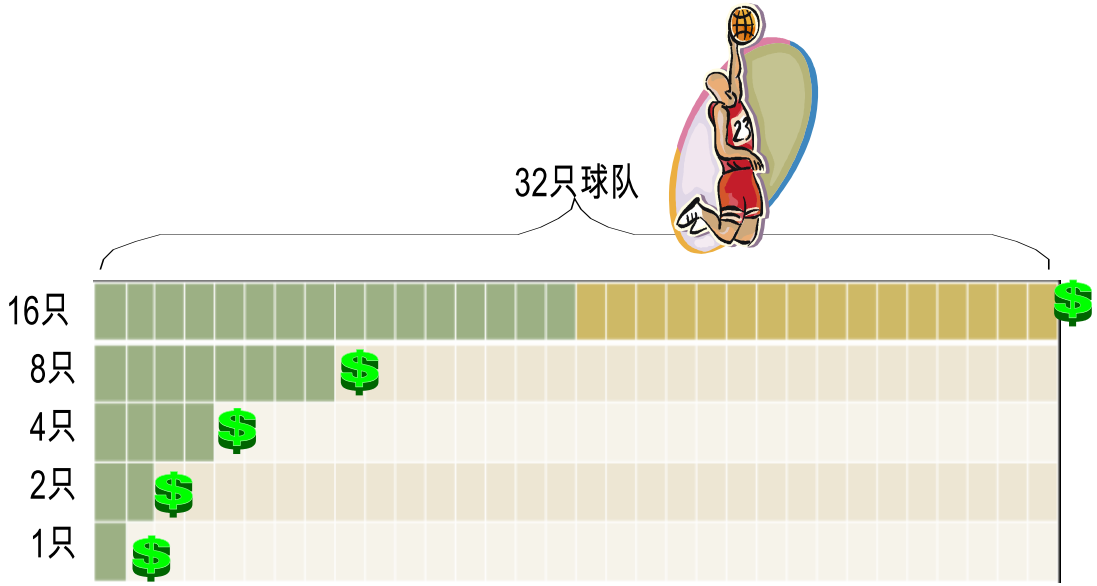
决策树的智慧（熵）

- 所以，信息熵也可以说是系统有序化程度的一个度量。



（a）高序、低熵　　（b）低序、高熵

人工智能算法的本质就在于，在利用更多信息，消除不确定性（即保证熵减）



冠军球队是1-16号球队吗？是或不是 （消除一半不确定性，花费1块钱） 。

冠军球队是1-8号球队吗？是或不是 （同上）

冠军球队是1-4号球队吗？是或不是 （同上）

冠军球队是1-2号球队吗？是或不是 （同上）

冠军球队是1号球队吗？是或不是 （同上）

因此，知道谁是冠军这条信息，值5块钱

熵减：消除不确定性

选择特征

- 划分数据集的大原则：将无序的数据变得更加有序。
- 如果能测量数据的复杂度，对比按不同特征分类后的数据复杂度，若按某一特征分类后复杂度降低得更多，那么这个特征就是最优分类特征。

- Claude Shannon定义了熵（Entropy）和信息增益（Information Gain），用熵来表示信息的复杂度，熵越大，则信息越复杂。
- 设X是一个取有限个值的离散随机变量，其概率分布为

$$P(X = x_i) = p_i, i = 1,2, \dots, n$$

则随机变量X的熵定义为

$$H(X) = -\sum_{i=1}^{n} p_i log p_i$$

熵的概念

- 熵：在信息论中，是接收的每条消息中包含的信息的平均量；（比较不可能发生的事情，当它发生了，会提供更多的信息）
- 熵：表示随机变量不确定性的度量。熵越大，随机变量的不确定性就越大。（举例：抛硬币）

熵的计算

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_b P(x_i)$$

投掷硬币一次，正反面的概率是0.5。

$$H(\text{硬币}) = -\left(0.5 \cdot \log_2 0.5 + 0.5 \cdot \log_2 0.5\right)$$

计算过程

由于 $\log_2 0.5 = -1$，所以：

$$H(\text{硬币}) = -\left(0.5 \times -1 + 0.5 \times -1\right) = -\left(-0.5 - 0.5\right) = 1$$

正面概率0.2，反面概率0.8。

$$H(\text{硬币}) = -\left(0.2 \cdot \log_2 0.2 + 0.8 \cdot \log_2 0.8\right)$$

计算过程

首先，计算 $\log_2 0.2$ 和 $\log_2 0.8$ 的值：

$$\log_2 0.2 \approx -2.32$$

$$\log_2 0.8 \approx -0.32$$

代入后得到：

$$H(\text{硬币}) = -\left(0.2 \cdot (-2.32) + 0.8 \cdot (-0.32)\right)$$

$$= -(-0.464 - 0.256)$$

$$= 0.72$$

熵变小了，不确定性变小了。

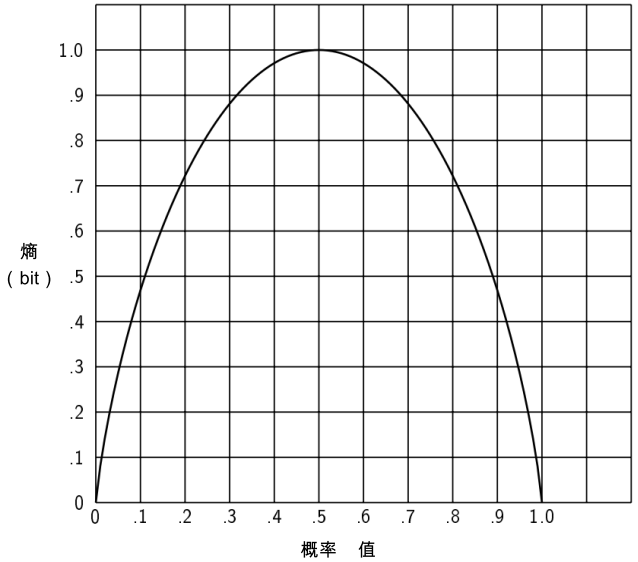
熵

- 当随机变量只有两个值，例如1,0时，即X的分布为

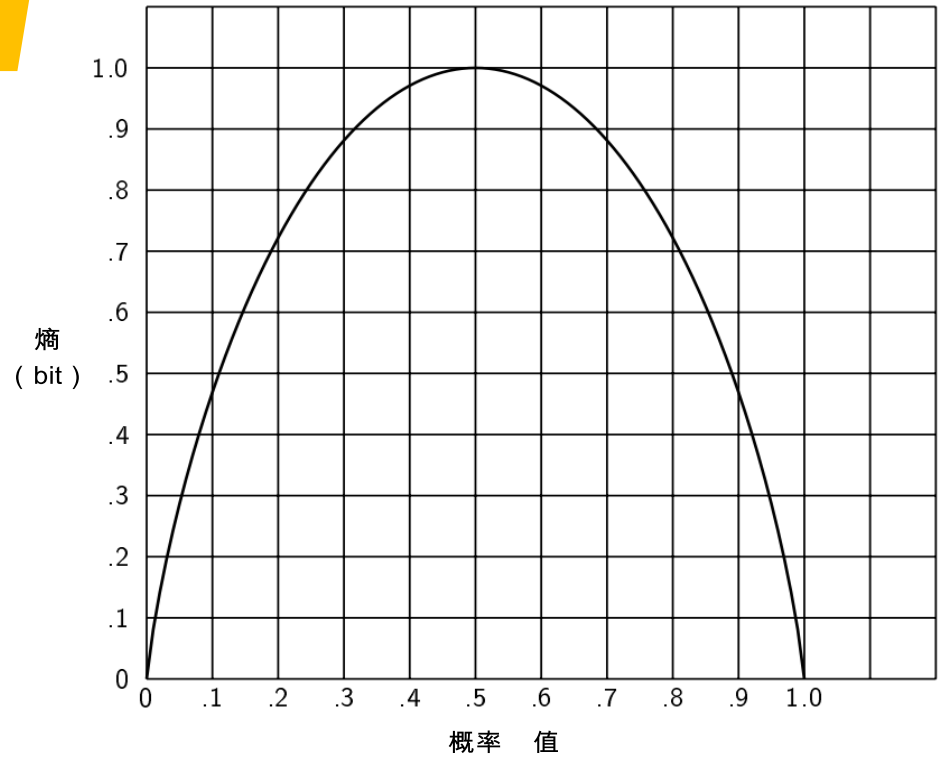
P(X=1)=p , P(X=0)=1-p , 0<=p<=1.

则熵$H(p) = -plog_2p - (1-p)log_2(1-p)$

- 熵H(p)随概率p变化的曲线如右图：



- 可知，当p=0或p=1时，H(p)=0，随机变量完全没有不确定性。



两状态下的熵与概率p之间的关系

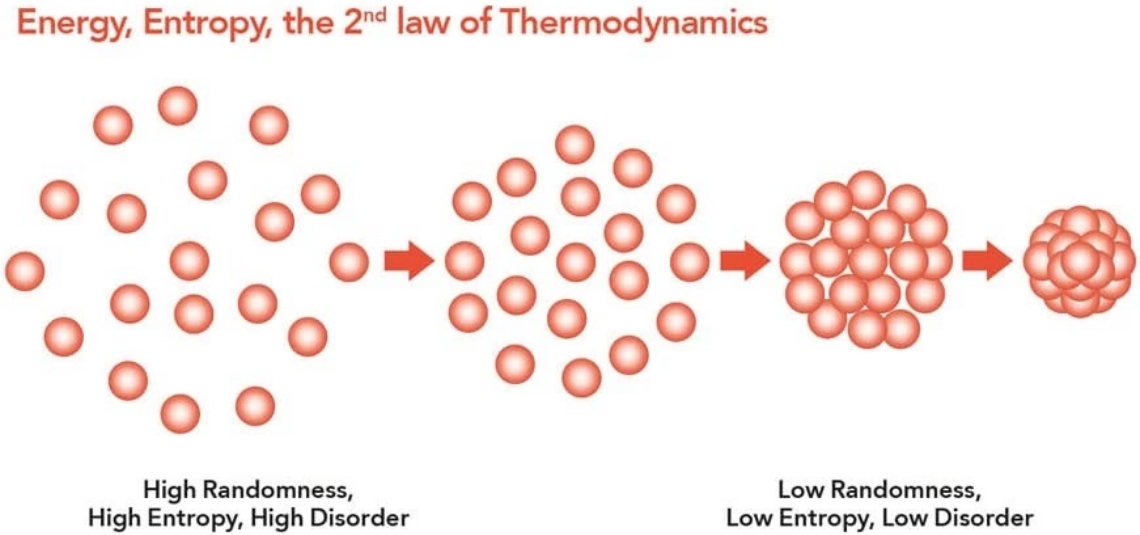
结论：永远不要听信那些正确率总是50%的专家的建议，因为那相当于什么都没说，他们没有提供任何减少“信息熵”的信息量

条件熵

- 设有随机变量(X,Y),其联合概率分布为

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

- 条件熵H(Y|X)表示在已知随机变量X的条件下随机变量Y的不确定性。
- 定义：$H(Y|X) = \sum_{i=1}^{n} p_i H(Y|X = x_i), \quad p_i = P(X = x_i), i = 1,2,\dots,n.$
- 当熵和条件熵中的概率由数据估计（特别是极大似然估计）得到时，所对应的熵分别称为经验熵和经验条件熵。



2. 如何用熵来划分决策树

信息增益

教材P142开始

- 信息增益表示得知特征X的信息而使得类Y的信息不确定性减少的程度。

- 定义（信息增益）特征A对训练数据集D的信息增益g(D,A),定义为集合D的经验熵H(D)与特征A给定条件下D的经验条件熵H(D|A)之差，即
$g(D,A) = H(D) - H(D|A)$

- 根据信息增益准则的特征选择方法是：对训练数据集（或子集）D，计算其每个特征的信息增益，并比较它们的大小，选择信息增益最大的特征。

信息增益的具体公式

- 设训练数据集为D，|D|表示其样本容量，即样本个数。
- 设有K个类C_k，k=1,2,⋯,K. $|C_k|$为属于类C_k的样本个数$\sum_{k=1}^{K}|C_k| = |D|$.
- 设特征A有n个不同的取值 $\{a_1,a_2,\cdots, a_n\}$，根据特征A的取值将D划分为n个子集$D_1, D_2,\cdots, D_n$，$|D_t|$为D_i的样本个数，$\sum_{i=1}^{n}|D_i| = |D|$.
- 记子集D_i 中属于类C_k 的样本的集合为D_{ik}，$|D_{ik}|$为D_{ik}的样本个数。

信息增益算法

- 输入：训练数据集D和特征A;
- 输出：特征A对训练数据集D的信息增益g(D,A).
- (1) 计算数据集D的经验熵H(D)

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

(2)计算特征A对数据集D的经验条件熵H(D|A)

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

(3)计算信息增益 g(D,A)=H(D)-H(D|A)

Information Gain = Entropy before splitting - Entropy after splitting

案例分析

- 信息增益表示两个信息熵的差值，首先计算分类前的熵，如共有8位同学，其中，男生有3位，女生有5位。

$$\text{熵（总）}=-\frac{3}{8}\log_2\frac{3}{8}-\frac{5}{8}\log_2\frac{5}{8}=0.9544$$

- 然后分别计算对同学A和同学B分类后的信息熵。
- 同学A按头发分类，分类后的结果：长头发中有1男3女；短头发中有2男2女。

$$\text{熵（同学 A 长发）}=-1/4\log2(1/4)-3/4\log2(3/4)=0.8113$$
$$\text{熵（同学 A 短发）}=-2/4\log2(2/4)-2/4\log2(2/4)=1$$
$$\text{熵（同学 A）}=（4/8）0.8113+4/81=0.9057$$

- 信息增益(同学A） = 熵(总） -熵 （同学A)=0.9544-.09507=0.0487

案例分析

- 同理，同学B按声音特征来分类，分类后的结果：声音粗中有3男3女；声音细中有0男2女。

$$\text{熵（同学 B 声音粗）}=-3/6\log2(3/6)-3/6\log2(3/6)=1$$
$$\text{熵（同学 B 声音粗）}=-2/2\log2(2/2)=0$$
$$\text{熵（同学 B）}=(6/8)1+2/8\times0=0.75$$

- 信息增益（同学B）= 熵（总）-熵（同学B)=0.9544-0.75=0.2044
- 将同学B按声音特征分类，信息增益更大，区分样本的能力更强，更具有代表性。
- 以上就是决策树ID3算法的核心思想。

ID3算法的学习过程:

首先以整个例子集作为决策树的根节点S，并计算S关于每个属性的期望熵（即条件熵）；

然后选择能使S的期望熵为最小的一个属性对根节点进行分裂，得到根节点的一层子节点;

接着再用同样的方法对这些子节点进行分裂，直至所有叶节点的熵值都下降为0为止。

这时，就可得到一棵与训练例子集对应的熵为0的决策树，即ID3算法学习过程所得到的最终决策树。该树中每一条从根节点到叶节点的路径，都代表了一个分类过程，即决策过程。

A	B	C	D	E	F
日期ID	天气	温度	湿度	有风	类别
1	晴天	高温	高	否	否
2	晴天	高温	高	是	否
3	阴天	高温	高	否	是
4	下雨	中温	高	否	是
5	下雨	低温	正常	否	是
6	下雨	低温	正常	是	否
7	阴天	低温	正常	是	是
8	晴天	中温	高	否	否
9	晴天	低温	正常	否	是
10	下雨	中温	正常	否	是
11	晴天	中温	正常	是	是
12	阴天	中温	高	是	是
13	阴天	高温	正常	否	是
14	下雨	中温	高	是	否

H3 1 f_x =COUNTIF($F$2:$F$15,"否")/COUNTA($F$2:$F$15)

	A	B	C	D	E	F	G	H	I
1	日期ID	天气	温度	湿度	有风	类别		未分割前类分割比例	
2	1	晴天	高温	高	否	否		C1	C2
3	2	晴天	高温	高	是	否		0.357	0.643
4	3	阴天	高温	高	否	是		分割前的熵:	
5	4	下雨	中温	高	否	是		0.940	
6	5	下雨	低温	正常	否	是			
7	6	下雨	低温	正常	是	否			
8	7	阴天	低温	正常	是	是			
9	8	晴天	中温	高	否	否			
10	9	晴天	低温	正常	否	是			
11	10	下雨	中温	正常	否	是			
12	11	晴天	中温	正常	是	是			
13	12	阴天	中温	高	是	是			
14	13	阴天	高温	正常	否	是			
15	14	下雨	中温	高	是	否			

$$H(D) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) = 0.940$$

分割前“熵”的计算

计算“天晴”情况下的子树之熵

0.643	0.357
分割前的熵:	
0.940	
按“天气”分割后	

晴天		阴天		下雨天	
C1	C2	C1	C2	C1	C2
0.4	0.6	1	0	0.6	0.4
晴天下的子树之熵		阴天下的子树之熵		下雨天下的子树之熵	
0.971		0		0.971	

三棵子树的加权比例			
晴天	阴天	下雨天	
0.357	0.286	0.357	

分割之后的熵

0.694

信息增益（即熵减）

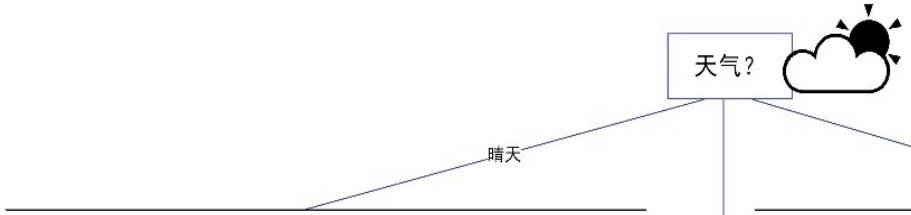
0.247

$$gain(D,温度)=0.029$$

$$gain(D,湿度)=0.152$$

$$gain(D,有风)=0.048$$

ID3决策树核心思想：如果哪个特征分割后带来的信息增益越大，就选择这个特征作为决策树的划分



日期 ID	天气	温度	湿度	有风	类别
1	晴天	高温	高	否	否
2	晴天	高温	高	是	否
8	晴天	中温	高	否	否
9	晴天	低温	正常	否	是
11	晴天	中温	正常	是	是

日期 ID	天气	温度	湿度	有风	类别
4	下雨	中温	高	否	是
5	下雨	低温	正常	否	是
6	下雨	低温	正常	是	否
10	下雨	中温	正常	否	是
14	下雨	中温	高	是	否

日期 ID	天气	温度	湿度	有风	类别
3	阴天	高温	高	否	是
7	阴天	低温	正常	是	是
12	阴天	中温	高	是	是
13	阴天	高温	正常	否	是

“天气”特征带来的信息增益最大，因此作为决策树分支

信息增益比

教材p148

- 以信息增益作为划分训练数据集的特征，存在偏向于选择取值较多的特征的问题。

- 定义（信息增益比）特征A对训练数据集D的信息增益比$g_R(D,A)$定义为其信息增益$g(D,A)$与训练数据集D关于特征A的值的熵$H_A(D)$之比，即 $g_R(D,A)=\frac{g(D,A)}{H_A(D)}$

C4.5的生成算法

- C4.5算法是由Ross Quinlan开发的用于产生决策树的算法。该算法是对Ross Quinlan之前开发的ID3算法的一个扩展。
- C4.5算法与ID3算法相似，C4.5算法对ID3算法进行了改进.C4.5在生成的过程中，用信息增益比来选择特征。
- **C4.5 算法最大的特点是克服了 ID3 对特征数目的偏好这一缺点**，引入信息增益率来作为分类标准。