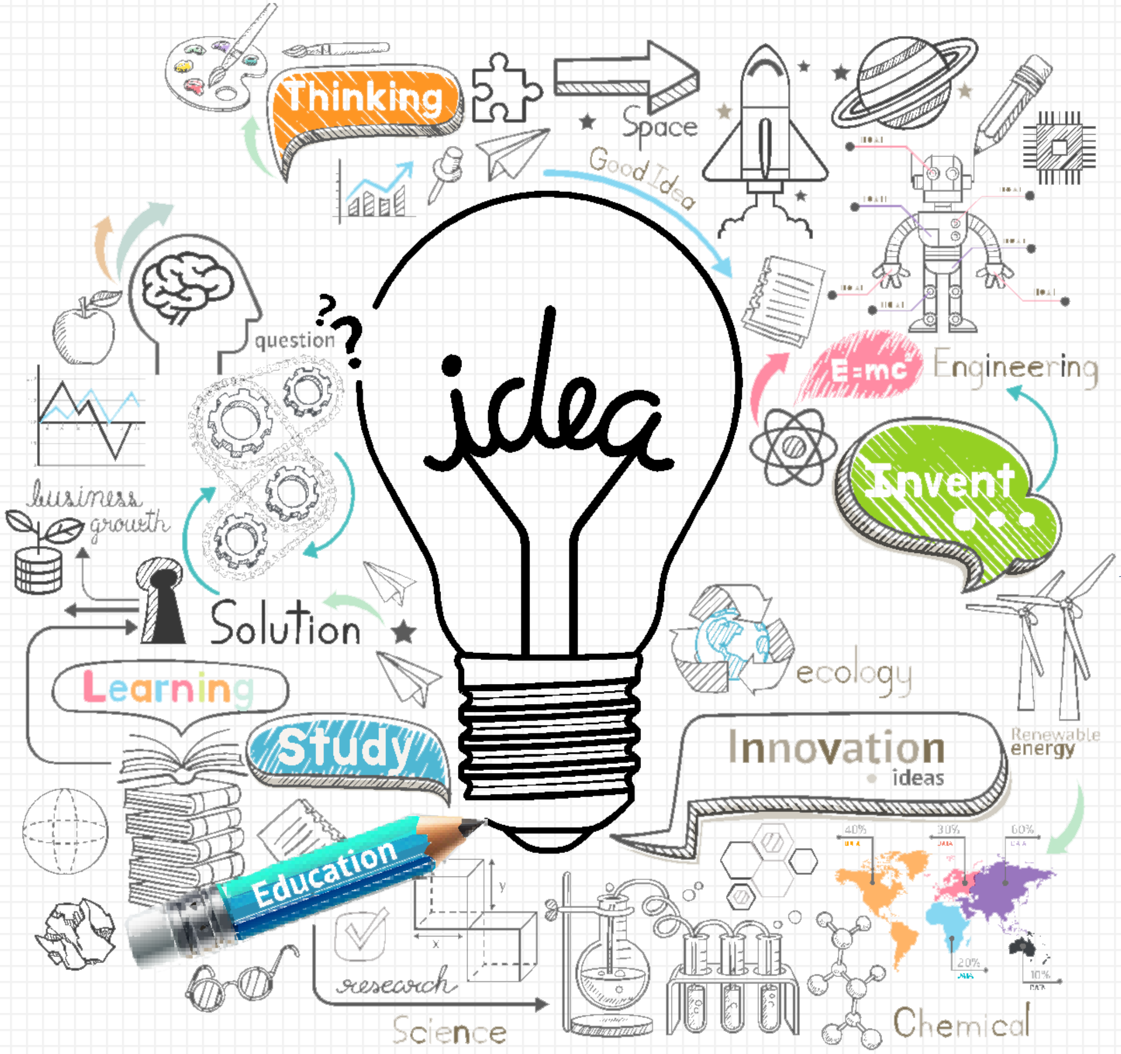
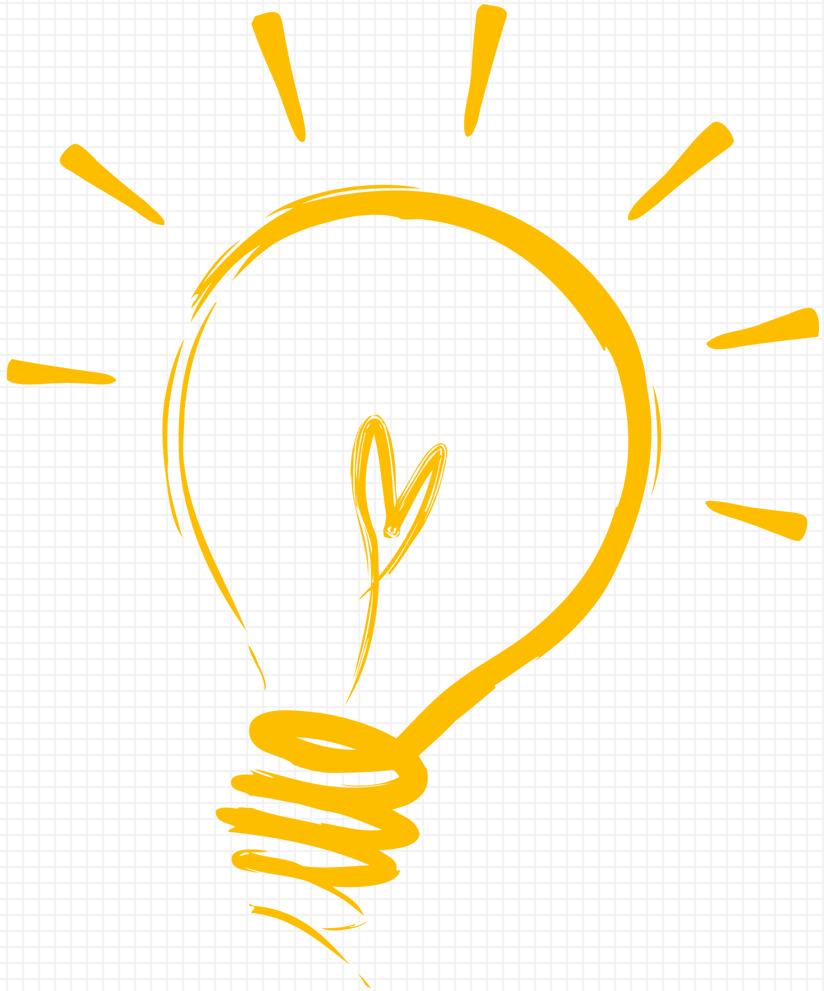


数据挖掘算法

ID3分类算法



CONTENTS

di yi zhang jie

第一章节

算法背景

1.1 前置知识

监督学习（英语：Supervised learning），又叫有监督学习，监督式学习，是机器学习的一种方法，可以由训练资料中学到或建立一个模式（函数 / learning model），并依此模式推测新的实例[1]。训练资料是由输入物件（通常是向量）和预期输出所组成。

常见的有监督学习:分类,回归分析等.

无监督学习（英语：unsupervised learning）是机器学习的一种方法，没有给定事先标记过的训练示例，自动对输入的资料进行分类或分群。

最常见的无监督学习:聚类.

简单来看,有监督学习就是"对数据进行预先学习",然后再去解决任务.无监督学习就是没有预先的学习行为,直接解决任务.

1.2 分类与聚类的区别与联系

分类：按照种类、等级或性质分别归类。

聚类：将物理或抽象对象的集合分成由类似的对象组成的多个类的过程。

分类简单来说，就是根据文本的特征或属性，划分到已有的类别中。这些**类别是已知的**，**通过对已知分类的数据进行训练和学习**，找到这些不同类的特征，再对未分类的数据进行分类。聚类的理解更简单，就是你压根不知道数据会分为几类，通过聚类分析将数据或者说用户聚合成几个群体，那就是聚类了。**聚类不需要对数据进行训练和学习**[2]。

分类算法有很多种,比如朴素贝叶斯，Logistic回归，决策树，支持向量机等等。

由于**决策树算法**容易理解与解释，结果可以直观展示，所以有大量的应用。

经典的决策树算法莫过于**ID3算法**。

di yi zhang jie

第二章节

算法原理

2.1 案例引导

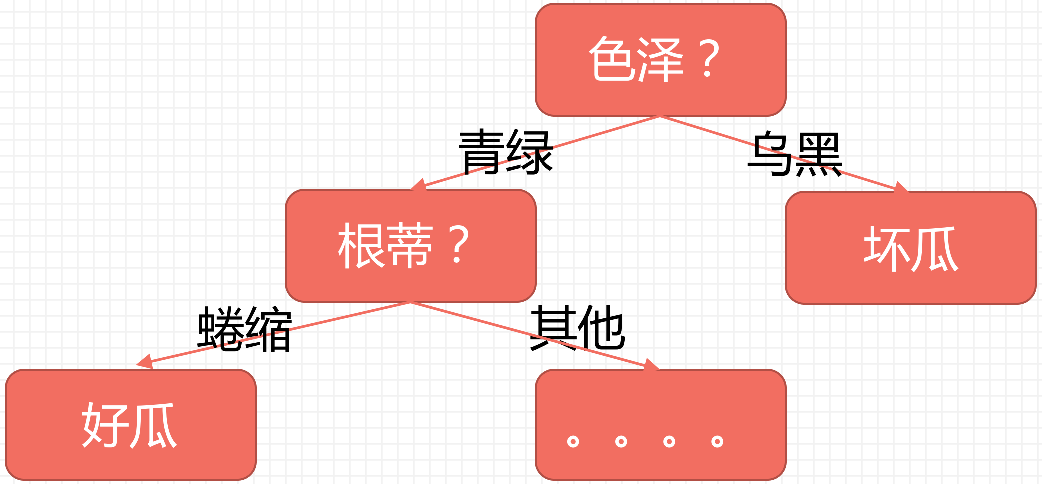
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	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

左表为某品种西瓜不同属性记录以及对应瓜果质量,共17条记录。

请根据已有数据设计算法，能够自动判断小明买的西瓜{青绿，稍蜷，沉闷，稍糊，凹陷，硬滑}是否为好瓜。

2.1 案例引导

我们的想法是，能不能构造一颗决策树去帮助我们做决定或者分类？比如下边这样的：



最后根据构造好的决策树去判断新的记录的类别，以达到分类（预测）的作用。

而决策树的构造，即首先根据哪个因素判断，然后再根据哪个因素判断，基于什么样的原理去选择呢？这就是我们要讲的ID3算法。

在接触ID3算法之前，我们要先了解一些重要的知识。

2.2 必要知识[2,3]

随机现象：现实生活中，一个动作或一件事情，在一定条件下，所得的结果不能预先完全确定，而只能确定是多种可能结果中的一种，称这种现象为随机现象。比如明天的天气，或者是抛硬币的结果等，这种现象在生活中非常常见，就不做过多解释了。

随机试验：试验结果呈现出不确定性的试验，且满足以下三个条件

（1）试验可在相同条件下重复进行；（2）试验的可能结果不止一个，且所有可能结果可事先预知；（3）每次试验的结果只有一个，但不能事先预知。

有了随机现象和随机试验的概念，我们就可以再引出三个概念，那就是样本空间，样本点和随机事件了。

2.2 必要知识[2,3]

样本空间：随机试验的所有可能结果组成的集合称为样本空间。对于抛掷硬币试验，样本空间 = { 正面，反面 } 。

样本点：样本空间中的元素称为样本点， “正面”就是上述样本空间的一个样本点。

随机事件：样本空间的子集。在每次试验中，当且仅当该子集中的任意一个元素发生时，称该随机事件发生。

例如在掷一次骰子的随机试验中，设随机事件A= “偶数点朝上” 。

那么当我们在掷出{2,4,6}这个集合中的任意一个点的时候，我们都可以说：随机事件A发生了。

2.2 必要知识[2,3]

随机变量：是定义在样本空间上的映射。通常是将样本空间映射到数字空间，这样做的目的是方便引入高等数学的方法来研究随机现象。

例如，在抛掷硬币试验中，将正面与1对应，反面与0对应，那么样本空间＝{ 正面，反面 } 与随机变量X＝{ 1，0 } 之间建立起了一一对应的关系。

需要指出的是，对于随机事件A，P（A）表示随机事件发生的概率；对于随机变量X，我们一般这么写，P（X＝x）表示随机变量取值为x的概率。其中x是一个确定的值。

2.2 必要知识[4,5,6,7]

克劳德·艾尔伍德·香农（英语：Claude Elwood Shannon，1916年4月30日－2001年2月26日），美国数学家、电子工程师和密码学家，被誉为信息论的创始人。香农是密歇根大学学士，麻省理工学院博士。

1948年，香农发表了划时代的论文——**通信的数学原理**，奠定了现代信息论的基础。不仅如此，香农还被认为是数字计算机理论和数字电路设计理论的创始人。1937年，21岁的香农是麻省理工学院的硕士研究生，他在其硕士论文中提出，将布尔代数应用于电子领域，能够构建并解决任何逻辑和数值关系，被誉为有史以来最具水平的硕士论文之一。二战期间，香农为军事领域的密码分析——密码破译和保密通信——做出了很大贡献。

2.2 必要知识[4,5,6,7]

1928年R.V.L.哈特莱首先提出信息定量化的初步设想，但对信息量作深入而系统研究，还是从香农的工作开始的。

（自）信息量:是概率空间中的单一事件或离散随机变量的值相关的信息的量度。

比如吃瓜群众经常说某某明星的某条微博信息量很大，这里“信息量” 指的就是这个定义。

那么到底如何计算信息量呢？由定义，当信息被拥有它的实体传递给接收它的实体时，仅当接收实体不知道信息的先验知识时信息才得到传递。如果接收实体事先知道了消息的内容，这条消息所传递的信息量就是0。只有当接收实体对消息对先验知识少于100%时，消息才真正传递信息。

因此，一个随机产生的事件ω_n所包含的自信息数量，只与事件发生的几率相关。事件发生的几率越低，在事件真的发生时，接收到的信息中，包含的自信息越大。

2.2 必要知识[4,5,6,7]

既然信息量和随机事件发生的概率有关，不妨将ω_n的信息量写作

$$I(\omega_n) = f[p(\omega_n)]$$

由之前的分析，我们知道如果$p(\omega_n) = 1$，$I(\omega_n) = 0$，且当$p(\omega_n) < 1$，$I(\omega_n) > 0$

此外，根据定义，自信息的量度是非负的而且是可加的。如果事件C是两个独立事件A和B的交集，那么宣告C发生的信息量就等于分别宣告A和事件B的信息量的和：

$$I(C) = I(A \cap B) = I(A) + I(B)$$

因为A和B是独立事件，所以

$$P(C) = P(A \cap B) = P(A) \cdot P(B)$$

应用函数$f(x)$，$I(C) = I(A) + I(B) \Leftrightarrow f[p(C)] = f[p(A)] + f[p(B)] = f[P(A) \cdot P(B)]$

所以，函数$f(x)$有如下性质

$$f[x] + f[y] = f[x \cdot y]$$

2.2 必要知识[4,5,6,7]

那么问题来了，哪个函数满足$f[x \cdot y] = f[x] + f[y]$的性质呢？

对数函数正好有这个性质，不同的底的对数函数之间的区别只差一个常数，也就是说：

$$f(x) = K\log(x)$$

由于事件的概率（即这里的x）总是在0和1之间，而信息量必须是非负的，所以有K<0。

综上所述，定义随机事件的自信息$I(\omega_n)$：

$$I(\omega_n) = -\log[P(\omega_n)] = \log(\frac{1}{P(\omega_n)})$$

在上面的定义中，没有指定的对数的基底：如果以 2 为底，单位是bit。当使用以 e 为底的对数时，单位将是 nat。对于基底为 10 的对数，单位是 hart。**决策树算法使用以2为底，后续不再提及。**

显然的，该式符合人们的直观概念，某事件发生的概率越小，那么当该事件发生时所带来的信息量就越大。

2.2 必要知识[4,5,6,7]

在1948年，克劳德·艾尔伍德·香农将热力学的熵，引入到信息论，因此它又被称为香农熵(Shannon entropy)。热力学的熵和信息论中的熵，虽然表达式不同，但是描述的本质相同。

香农把随机变量X的熵值 H定义如下：

$$\boldsymbol{H(X) = E[I(X)] = \sum_{i}^{n}[p(x_i) \cdot I(x_i)] = \sum_{i}^{n}[p(x_i) \cdot \log(\frac{1}{p(x_i)})]}$$

约定$0 \cdot \log 0 = 0$

解释：对于随机变量X，以一定的概率$p(x_i)$取值为x_i，那么当我们计算随机变量X的自信息量时，由于我们不知道X的具体取值，所以只能考虑所有X取到每一个x_i的情况，而对于每一个x_i的自信息量我们是可以计算的：$\log(\frac{1}{p(x_i)})$，所以$H(X) = \sum_i^n[p(x_i) \cdot \log(\frac{1}{p(x_i)})]$，称$H(X)$为随机变量X的熵。

2.2 必要知识[4,5,6,7]

说了一大堆，我们到底要讲什么？我们要看的是信息论中的熵所反映的实际意义，换句话说，这个熵，它能解释什么，它描述的是什么。我们来看熵值的式子：

$$H(X)=E[I(X)]=\sum_{i}^{n}[p(x_i)\cdot I(x_i)]=\sum_{i}^{n}[p(x_i)\cdot \log(\frac{1}{p(x_i)})]$$

前边说道，信息量取决于概率，随机事件发生的概率越小，那么当该事件发生时，所蕴含的信息量越大。同时，随机事件发生的概率小⇔这件事不易控制，不易预测，即这件事的**不确定度高。**

再换句话说，某事件的信息量可以描述该事件**不确定的程度**，而熵是随机变量信息量的期望，所以我们说：

熵可以描述随机变量的不确定程度。

熵可以描述随机变量的不确定程度。

熵可以描述随机变量的不确定程度。

2.2 必要知识[4,5,6,7]

条件熵：条件熵描述了在已知第二个随机变量 X 的值的前提下，随机变量 Y 的信息熵还有多少。基于随机变量X条件下的Y的熵，表示为$H(Y|X)$。

如果用$H(Y|X=x)$表示为变量Y在变量X取特定值x条件下的熵，那么$H(Y|X)$就是X在取遍所有的x后取平均的结果。即

$$H(Y|X)=\sum_{x\in X}[p(x)\cdot H(Y|X=x)]$$

$$=\sum_{x\in X}\left\{p(x)\cdot\sum_{i}^{n}[p(y|x)\cdot\log(\frac{1}{p(y|x)})]\right\}$$

条件熵可以描述在某个随机变量确定的情况下，另一个随机变量的不确定程度。

2.2 必要知识

信息增益：

$$Gain(Y,X) = H(Y) - H(Y|X)$$

对式子的解释：在随机变量X确定的条件下，随机变量Y的熵值较没有任何条件确定时减少了多少。

本次课程的目标是我们要完成以下的任务

不确定瓜的好坏⟹确定瓜的好坏

这是一个从模糊到清晰，是一个**不确定度越来越小的过程。**

在决策树构造的过程中，最重要的步骤就是决策树节点属性的选择，我们只需要一步步的，依次找出哪个属性确定后，我们的研究目标的熵会相对下降的最多，即该属性对目标属性的信息增益最大，我们就先把这个属性确定下来，这样我们的目标就会逐渐清晰了，这就是ID3算法的核心思想。

2.3 算法流程

输入 ： 训练集（学习样本、训练样本）

Step1：对当前样本集合，计算所有属性的信息增益；

Step2：选择信息增益最大的属性作为测试属性，把测试属性取值相同的样本划为同一个子样本集；

Step3：若子样本集的类别属性只含有单个属性，则分支为叶子节点，判断其属性值并标上相应的符号，然后返回调用处；否则对子样本集递归调用本算法。

输出 ：输出一颗决策树

注意：算法结束的条件一般不是把所有的属性划分的特别详细。这是为了防止算法的过拟合，即对训练集拟合的“太好了”，结果在对新的未知样本进行分类的时候准确率反而下降了。

这里的分裂终止条件，剪枝等比较简单，自己网络上看一下就明白，这里不再赘述。本教程重在讲解ID3算法的思想。

2.3 算法流程

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
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
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
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
10	青绿	硬挺	清脆	清晰	平坦	软粘	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
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

“好瓜”，即Y的信息熵：

$$H(Y) = P(Y=\text{“是”})\cdot\log\left(\frac{1}{P(Y=\text{“是”})}\right) + P(Y=\text{“否”})\cdot\log\left(\frac{1}{P(Y=\text{“否”})}\right)$$

$$= \frac{8}{17}\cdot\log\left(\frac{17}{8}\right) + \frac{9}{17}\cdot\log\left(\frac{17}{9}\right) = 0.9975$$

令随机变量X为“色泽”，则X的取值为{青绿、乌黑、浅白}

概率分别为：$\frac{6}{17}\ \frac{6}{17}\ \frac{5}{17}$，而

$$H(Y|X=\text{“青绿”}) = \frac{3}{6}\cdot\log\left(\frac{6}{3}\right) + \frac{3}{6}\cdot\log\left(\frac{6}{3}\right) = 1$$

$$H(Y|X=\text{“乌黑”}) = \frac{4}{6}\cdot\log\left(\frac{6}{4}\right) + \frac{2}{6}\cdot\log\left(\frac{6}{2}\right) = 0.9183$$

$$H(Y|X=\text{“浅白”}) = \frac{1}{6}\cdot\log\left(\frac{5}{1}\right) + \frac{4}{5}\cdot\log\left(\frac{5}{4}\right) = 0.7219$$

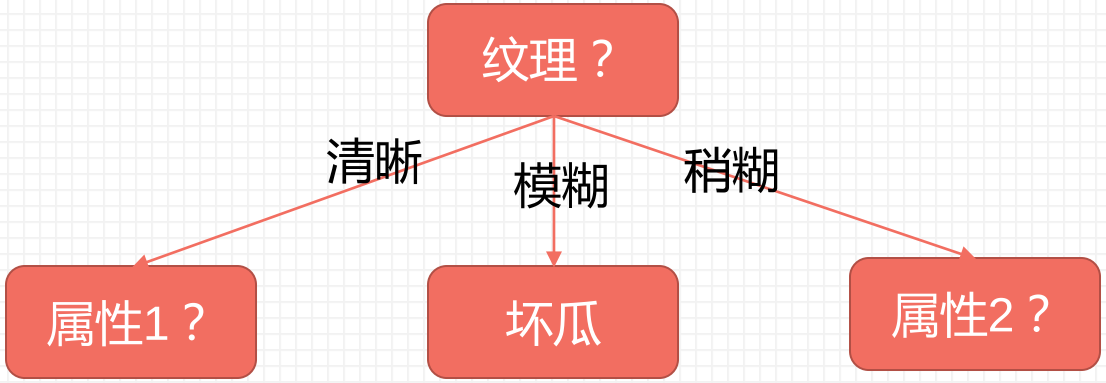
则$H(Y|X) = \frac{6}{17}\cdot 1 + \frac{6}{17}\cdot 0.9183 + \frac{5}{17}\cdot 0.7219 = 0.8894$

$Gain(Y,X) = H(Y) - H(Y|X) = 0.9975 - 0.8894 = 0.1081$

2.3 算法流程

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
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
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
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
10	青绿	硬挺	清脆	清晰	平坦	软粘	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
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

即属性“色泽”的信息增益为0.1081，同理计算其他属性的信息增益分为别：0.143，0.141，0.381，0.289，0.006，故决策树根节点选择信息增益为0.381对应的属性，即“纹理”，于是决策树如下：



那么接下来“清晰”这条支线下的属性1怎么确定？同理，我们仍然采用之前的方法继续在纹理“清晰”的样本中计算各属性的信息增益。“稍糊”支线同理，这里不再赘述。只以“清晰”这条支线举例。

2.3 算法流程

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
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
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
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
10	青绿	硬挺	清脆	清晰	平坦	软粘	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
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

那么既然是“清晰”支线的样本，所以我们先把“纹理清晰”的样本都挑出来，作为我们新的总的样本，此时样本个数为9。

此时“好瓜”，即Y的新的信息熵：

$$H(Y) = P(Y=\text{“是”})\cdot\log\left(\frac{1}{P(Y=\text{“是”})}\right) + P(Y=\text{“否”})\cdot\log\left(\frac{1}{P(Y=\text{“否”})}\right)$$

$$= \frac{7}{9}\cdot\log\left(\frac{9}{7}\right) + \frac{2}{9}\cdot\log\left(\frac{9}{2}\right) = 0.7642$$

现在新的样本可分类属性为：{色泽，根蒂，敲声，脐部，触感}和之前算法原理一致，计算新样本下的各属性信息增益随后进行划分即可，直至不可划分，不再赘述。

di san zhang jie

第三章节

算法分析

3.1 算法分析

优点：

1）原理比较简单，实现也是很容易。

2）健壮性好，不受噪声影响。

3）算法的可解释度比较强。

缺点：

1）ID3算法对于连续值、缺失值没有进行考虑。

2）**ID3算法最大的缺点是，它倾向于选择包含特征更多的属性。**

解释：从概念来讲,信息增益反映的给定一个条件以后不确定性减少的程度(由于某个属性而使得数据集的分类不确定性减少的程度), **必然是分得越细的数据集确定性更高**,也就是条件熵越小,信息增益越大

di si zhang jie

第四章节

算法拓展

4.1 算法拓展

针对经典ID3算法的缺点，很多前辈提出或者改进出了更加高效的分类算法，如在ID3算法基础上的C4.5算法。以及CART决策树，贝叶斯分类，随机森林，SVM等等。

以下是几篇这方面的博客文章，大家感兴趣可以看看。

[1] https://zhuanlan.zhihu.com/p/85731206

[2] https://www.cnblogs.com/wj-1314/p/9628303.html

[3] https://blog.csdn.net/tyh70537/article/details/76768802

[4] https://www.zybuluo.com/77qingliu/note/1137445

di wu zhang jie

第五章节

案例实操

5.1 案例实操

这一部分在下次视频中讲解，我们使用Python中经典机器学习第三方库scikit-learn，来操作一下。

需要做的准备是：

首先要有Python环境，最好可以安装上Pycharm。

然后安装一下必要的库（以后也需要的）：如numpy，pandas，xlrd，scikit-learn库等

如果你现在没有Python环境以及pycharm，或者不清楚第三方库如何安装，以及pip文件如何设置，Pycharm的一些设置等，可以看一下我之前的基于Python实现网络爬虫的视频的第一课：

https://www.bilibili.com/video/BV1WV411U7LQ

同时如果有一些其他问题，可以联系我

QQ ： 1366420642，Q群：1019030249

欢迎大佬萌新加入

参　考　资　料

【1】https://zh.wikipedia.org/wiki/监督学习

【2】https://flashgene.com/archives/129177.html

【3】https://blog.csdn.net/weixin_41353276/article/details/78877194

【4】https://zh.wikipedia.org/wiki/克劳德·艾尔伍德·香农

【5】https://baike.baidu.com/item/%E8%87%AA%E4%BF%A1%E6%81%AF%E9%87%8F

【6】https://zh.wikipedia.org/wiki/%E7%86%B5_(%E4%BF%A1%E6%81%AF%E8%AE%BA)

【7】https://zh.wikipedia.org/wiki/%E6%9D%A1%E4%BB%B6%E7%86%B5

THANKS

谢谢观看