1. **ID3算法 P130**

**例1**

S中有si个元组，对应类标签Ci （i={1,…,m}）

基于训练对象和已知类标号创建决策树，以信息增益为度量来为属性排序。

每个元组所需信息度量

根据属性A分裂后，区分所需信息量为

信息增益

|  |  |  |
| --- | --- | --- |
|  | 本科生 | 研究生 |
| 理科 | 42 | 84 |
| 工科 | 46 | 36 |
| 商科 | 42 | 0 |
| 汇总 | 130 | 120 |

1. 计算区分任一元组所需的期望信息量



1. 为每个属性，比如说major，计算熵

|  |  |  |  |
| --- | --- | --- | --- |
| For major=”Science”: | s11=84 | s21=42 | I(s11,s21)=0.9183 |
| For major=”Engineering”: | s12=36 | s22=46 | I(s12,s22)=0.9892 |
| For major=”Business”: | s13=0 | s23=42 | I(s13,s23)=0 |

1. 即已知major信息后，区分元组所需的信息量



1. 计算每个属性的信息增益



**例2：**

两个类标记：P（假设有p个元素） 和 N（假设有n个元素）

用来判断任一元素属于P还是N的信息量为：



根据属性A，集合S被划分为{*S1*, *S2* , …, *Sv*}

在每个Si中，属于类P的元素为Pi，属于N的元素为ni，用来区分的信息量（熵）为



在属性A上分支得到的信息增益为



**2. Apriori 算法 P152**

****

**3. FP树 P158**

**1．构建FP树的过程**

1. 扫描数据库，找到频繁1项集
2. 以频率递减的顺序对频繁项进行排序
3. 构建FP树

**2．挖掘FP树的过程**

（4）由每个长度为1的频繁模式（初始后缀模式）开始，构造它的**条件模式基**（一个“子数据库”，由FP树中与后缀模式一起出现的前缀路径集组成）

（5）然后构造它的**（条件）FP树**，并递归地对该树进行挖掘直到FP树为空或者只有单个频繁模式路径。模式增长通过后缀模式与条件FP树产生的频繁模式连接实现（abcdef频繁当且仅当abcde频繁，且f在包含abcde的事务中也是频繁的）。

1. **朴素贝叶斯（属性间独立）**



**贝叶斯定律：**

P(X) 对所有类都是常量。

要使P(C|X)最大，则P(X|C)·P(C)最大



**例子：**

|  |  |  |
| --- | --- | --- |
| **outlook** | |  |
| **P(sunny|p) = 2/9** | | **P(sunny|n) = 3/5** |
| **P(overcast|p) = 4/9** | | **P(overcast|n) = 0** |
| **P(rain|p) = 3/9** | | **P(rain|n) = 2/5** |
| **temperature** | |  |
| **P(hot|p) = 2/9** | | **P(hot|n) = 2/5** |
| **P(mild|p) = 4/9** | | **P(mild|n) = 2/5** |
| **P(cool|p) = 3/9** | | **P(cool|n) = 1/5** |
| **humidity** | |  |
| **P(high|p) = 3/9** | | **P(high|n) = 4/5** |
| **P(normal|p) = 6/9** | | **P(normal|n) = 2/5** |
| **windy** | |  |
| **P(true|p) = 3/9** | | **P(true|n) = 3/5** |
| **P(false|p) = 6/9** | | **P(false|n) = 2/5** |
| **P(p) = 9/14** |
| **P(n) = 5/14** |

**为sample X = <rain, hot, high, false> 分类**

* **P(X|p)·P(p) =   
  P(rain|p)·P(hot|p)·P(high|p)·P(false|p)·P(p) = 3/9·2/9·3/9·6/9·9/14 = 0.010582**
* **P(X|n)·P(n) =   
  P(rain|n)·P(hot|n)·P(high|n)·P(false|n)·P(n) = 2/5·2/5·4/5·2/5·5/14 = 0.018286**

**X被为N的概率更大，所以不适合打网球。**

**5. 贝叶斯信念网络**

****



**（1）计算HD的先验概率**



结论：不得心脏病的概率更大

**（2）BP=high的情况下，HD的概率**



结论：如果血压高，得心脏病概率更大

**（3）在BP=high,D=healthy,E=yes的情况下HD的概率（血压高，饮食健康，做运动的心脏病的概率）**



结论：饮食健康，做运动降低了心脏病可能性