决策树基本知识总结

//参考http://blog.csdn.net/suipingsp/article/details/41927247 李航第五章

**决策树**模型呈树状结构，在分类问题中，表示基于特征对实力进行分类的过程。它可以认为是if-then规则的集合，也可以认为是定义在特征空间与类空间上的条件概率分布。

**常用**决策树**类型**有ID3,C4.5,CART

决策树的**学习过程**：特征选择，决策树生成，剪纸

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 类型 | 选择属性 | 连续分布 | 剪枝 | 规模 | 效率 |
| ID3 | 信息增益 | 不支持 | 无 | 大 | 低 |
| C4.5 | 信息增益比 | 支持 | 预剪枝 | 大 | 低 |
| CART | Gini指数  平方误差 | 支持 | 后剪枝 | 小 | 高 |

**http://img.blog.csdn.net/20141214191529906信息熵**：若待分类的事物可能划分在N类中，分别是x1，x2，……，xn，每一种取到的概率分别是P1，P2，……，Pn，那么X的熵就定义为：

信息熵代表了数据包含的信息量，代表了数据集的类别的不确定程度，代表了一个集合标号需要的信息量。

http://img.blog.csdn.net/20141214191555257随机变量(X,Y)联合概率分布为:P(X=xi,Y=yi)=pij,i=1,2,⋯,n;j=1,2,⋯,m则**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，其定义为X在给定条件下Y的条件概率分布的熵对X的数学期望:

http://img.blog.csdn.net/20141214191618937**信息增益**表示得知特征X的信息后，而使得Y的不确定性减少的程度。

定义为:

**http://images.cnitblog.com/i/633472/201408/121923196868742.gif信息增益比**：

http://images.cnitblog.com/blog/431880/201308/20122208-8f5c244da4e54fa9832cb260dacff56c.jpghttp://images2015.cnblogs.com/blog/816079/201511/816079-20151101183710560-1461821857.png也叫做拆分信息

http://images.cnitblog.com/blog/431880/201308/20123020-badbe42a81b947078f093dc8cd878993.jpg**基尼指数**： 　 m表示数据集D中类别C的个数，Pi表示D中任意一个记录属于Ci的概率，计算时Pi=(D中属于Ci类的集合的记录个数/|D|)。

以属性R分裂后的基尼系数为：

D1为D的一个非空真子集，D2为D1在D的补集，即D1+D2=D，对于属性R来说，有多个真子集，即GiniR(D)有多个值，但我们选取最小的那么值作为R的基尼指数。

http://images.cnitblog.com/blog/431880/201308/20123358-2c50747c3fc644448a6dc8de69b5abc8.jpg