我们统计了14天的气象数据(指标包括outlook，temperature，humidity，windy)，并已知这些天气是否打球(play)。如果给出新一天的气象指标数据:sunny,cool,high,TRUE，判断一下会不会去打球。

table 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **outlook** | **temperature** | **humidity** | **windy** | **play** |
| sunny | hot | high | FALSE | no |
| sunny | hot | high | TRUE | no |
| overcast | hot | high | FALSE | yes |
| rainy | mild | high | FALSE | yes |
| rainy | cool | normal | FALSE | yes |
| rainy | cool | normal | TRUE | no |
| overcast | cool | normal | TRUE | yes |
| sunny | mild | high | FALSE | no |
| sunny | cool | normal | FALSE | yes |
| rainy | mild | normal | FALSE | yes |
| sunny | mild | normal | TRUE | yes |
| overcast | mild | high | TRUE | yes |
| overcast | hot | normal | FALSE | yes |
| rainy | mild | high | TRUE | no |

这个问题当然可以用[朴素贝叶斯法](http://www.cnblogs.com/zhangchaoyang/articles/2586402.html)求解，分别计算在给定天气条件下打球和不打球的概率，选概率大者作为推测结果。

现在我们使用ID3归纳决策树的方法来求解该问题。

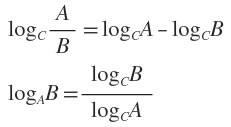
**预备知识：信息熵**

熵是无序性（或不确定性）的度量指标。假如事件A的全概率划分是（A1,A2,...,An），每部分发生的概率是(p1,p2,...,pn)，那信息熵定义为：

http://pic002.cnblogs.com/images/2012/103496/2012073015414651.png

通常以2为底数，所以信息熵的单位是bit。

补充两个对数公式：



**ID3算法**

构造树的基本想法是随着树深度的增加，节点的熵迅速地降低。熵降低的速度越快越好，这样我们有望得到一棵高度最矮的决策树。

在没有给定任何天气信息时，根据历史数据，我们只知道新的一天打球的概率是9/14，不打的概率是5/14。此时的熵为：

http://pic002.cnblogs.com/images/2012/103496/2012073015584424.png

属性有4个：outlook，temperature，humidity，windy。我们首先要决定哪个属性作树的根节点。

对每项指标分别统计：在不同的取值下打球和不打球的次数。

table 2

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **outlook** | | | **temperature** | | | **humidity** | | | **windy** | | | **play** | |
|  | yes | no |  | yes | no |  | yes | no |  | yes | no | yes | no |
| sunny | 2 | 3 | hot | 2 | 2 | high | 3 | 4 | FALSE | 6 | 2 | 9 | 5 |
| overcast | 4 | 0 | mild | 4 | 2 | normal | 6 | 1 | TRUR | 3 | 3 |  |  |
| rainy | 3 | 2 | cool | 3 | 1 |  |  |  |  |  |  |  |  |

下面我们计算当已知变量outlook的值时，信息熵为多少。

outlook=sunny时，2/5的概率打球，3/5的概率不打球。entropy=0.971

outlook=overcast时，entropy=0

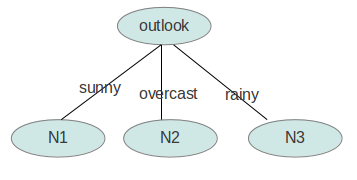
outlook=rainy时，entropy=0.971

而根据历史统计数据，outlook取值为sunny、overcast、rainy的概率分别是5/14、4/14、5/14，所以当已知变量outlook的值时，信息熵为：5/14 × 0.971 + 4/14 × 0 + 5/14 × 0.971 = 0.693

这样的话系统熵就从0.940下降到了0.693，信息增溢gain(outlook)为0.940-0.693=0.247

同样可以计算出gain(temperature)=0.029，gain(humidity)=0.152，gain(windy)=0.048。

gain(outlook)最大（即outlook在第一步使系统的信息熵下降得最快），所以决策树的根节点就取outlook。



接下来要确定N1取temperature、humidity还是windy?在已知outlook=sunny的情况，根据历史数据，我们作出类似table 2的一张表，分别计算gain(temperature)、gain(humidity)和gain(windy)，选最大者为N1。

依此类推，构造决策树。当系统的信息熵降为0时，就没有必要再往下构造决策树了，此时叶子节点都是纯的--这是理想情况。最坏的情况下，决策树的高度为属性（决策变量）的个数，叶子节点不纯（这意味着我们要以一定的概率来作出决策）。