4.4连续值与缺失值

4.4.1连续值处理

C4.5算法：

1. 二分法，每次将数据分隔为2部分
2. 找到使得信息增益Gain(D,a)最大的阈值
3. 确定是否要分叉

4.4.3缺失值处理

C4.5例子

表 4.4 西瓜数据集 2.0α

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

思路： 将缺失值去除后的样本$\bar{D}$计算Ent($\bar{D}$)，取出后剩余14个样本，对不同属性计算Ent($\hat{D}^i$)，可得信息增益，找到信息增益最大的那个属性作为根进行划分。

做法：

在学习开始时, 根结点包含样本集 D 中全部 17 个样例, 各样例的权值均为 1. 以属性“色泽”为例, 该属性上无缺失值的样例子集 $\tilde{D}$ 包含编号为 $\{2,3,4,6,7,8,9,10,11,12,14,15,16,17\}$ 的 14 个样例. 显然, $\tilde{D}$ 的信息熵为

$$\begin{aligned}\operatorname{Ent}(\tilde{D}) &= -\sum_{k=1}^{2} \tilde{p}_k \log_2 \tilde{p}_k \\ &= -\left(\frac{6}{14}\log_2\frac{6}{14} + \frac{8}{14}\log_2\frac{8}{14}\right) = 0.985 .\end{aligned}$$

令 $\tilde{D}^1$, $\tilde{D}^2$ 与 $\tilde{D}^3$ 分别表示在属性“色泽”上取值为“青绿”“乌黑”以及“浅白”的样本子集, 有

$$\operatorname{Ent}(\tilde{D}^1) = -\left(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}\right) = 1.000 ,$$

$$\operatorname{Ent}(\tilde{D}^2) = -\left(\frac{4}{6}\log_2\frac{4}{6} + \frac{2}{6}\log_2\frac{2}{6}\right) = 0.918 ,$$

$$\operatorname{Ent}(\tilde{D}^3) = -\left(\frac{0}{4}\log_2\frac{0}{4} + \frac{4}{4}\log_2\frac{4}{4}\right) = 0.000 ,$$

因此, 样本子集 $\tilde{D}$ 上属性“色泽”的信息增益为

$$\begin{aligned}\operatorname{Gain}(\tilde{D}, \text{色泽}) &= \operatorname{Ent}(\tilde{D}) - \sum_{v=1}^{3} \tilde{r}_v \operatorname{Ent}(\tilde{D}^v) \\ &= 0.985 - \left(\frac{4}{14}\times 1.000 + \frac{6}{14}\times 0.918 + \frac{4}{14}\times 0.000\right) \\ &= 0.306 .\end{aligned}$$

于是, 样本集 D 上属性“色泽”的信息增益为

$$\operatorname{Gain}(D, \text{色泽}) = \rho \times \operatorname{Gain}(\tilde{D}, \text{色泽}) = \frac{14}{17} \times 0.306 = 0.252 .$$

类似地可计算出所有属性在 D 上的信息增益:

$$\operatorname{Gain}(D, \text{色泽}) = 0.252; \quad \operatorname{Gain}(D, \text{根蒂}) = 0.171;$$

$$\operatorname{Gain}(D, \text{敲声}) = 0.145; \quad \operatorname{Gain}(D, \text{纹理}) = 0.424;$$

$$\operatorname{Gain}(D, \text{脐部}) = 0.289; \quad \operatorname{Gain}(D, \text{触感}) = 0.006.$$

（疑问：ρ为什么是$\frac{14}{17}$而不是$\frac{17}{14}$）

其他方法

离散值

- 众数填充
- 相关性最高的列填充

连续值

- 中位数
- 相关性最高的列做线性回归进行估计