

[illegible][illegible]

连续与缺失值

连续值处理 二分法，使得信息增益最大的阈值

缺失值处理

对于集合D，按照不同的属性进行分类，这些属性的属性值均属于连续值

例：

根据属性值对西瓜数据集分类

- 1) 西瓜数据集大小和实际西瓜甜度成正比关系
- 2) 通过计算信息增益，查找一次为最佳的属性值。

继续以剪切处理的西瓜数据集D为例，即数据集为：

编号	A (色泽)	B (纹理)
1-5	青绿	青
6-7
8-11	青白	白

其中，1-8是正例，9-17是反例。

缺失值处理

- 1) 以色泽属性为例，缺失了1号，5号，13号在色泽上的数据，即无缺失值的样例子集 D' 中只有：

6	青	青
7	青	青
8	青	青
9	青	青
10	青	青
11	青	青
12	青	青
14	青	青
15	青	青
16	青	青
17	青	青
- 2) 分别表示色泽属性上取值青绿、青白和白的样本子集 $En(D)$ 、 $En(D')$ 、 $En(D'')$
- 3) 因此，样本子集 D' 上色泽的信息增益为 $G_{gain}(D, 色泽) = En(D) - \frac{1}{|D|} \cdot [En(D') + En(D'')]$
- 4) 于是，样本集D上色泽属性的信息增益为 $G_{gain}(D, 色泽) = \frac{1}{|D|} \cdot [G_{gain}(D, 色泽) + G_{gain}(D', 色泽)]$

编号	A-F	概率
1-8	...	是
9-17	...	否

多变量决策树

对于正常基于连续值生成的决策树，它们不同的属性最后往往呈线性的关系，因为每次的取值都是一次二分类。

多变量决策树

$-5 \times \text{密度} - 0.99 \times \text{含糖率} \leq -0.343$
 是 否
 否 否
 $-0.365 \times \text{密度} + 0.345 \times \text{含糖率} \leq -0.158$
 是 否
 否 否

关于密度和含糖率生成的决策树

多变量决策树的分界点不再是对某个属性，而是对属性的线性组合进行一种测试。

多变量决策树对应的分类边界

```

graph TD
    A["含糖率 < 0.125?"] -- 是 --> B((环纹))
    A -- 否 --> C["密度 < 0.28?"]
    C -- 是 --> D((环纹))
    C -- 否 --> E["含糖率 < 0.20?"]
    E -- 是 --> F["密度 < 0.16?"]
    E -- 否 --> G((散纹))
    F -- 是 --> H((环纹))
    F -- 否 --> I((散纹))
  
```

关于干物质和含糖率生成的决策树

