Machine Learning 2019

袁欣

2019 年 3 月 29 日

1 决策树

1.1 西瓜数据

从 csv 文件中读取西瓜数据，并进行展示。

- 数据展示：

编号	色泽	根蒂	敲声	纹理	脐部	触感	label
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	1
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	1
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	1
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	1
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	1
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	1
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	1
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	1
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0
10	青绿	硬挺	清脆	清晰	平坦	软粘	0
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0

1.2 决策树

- 信息增益

```
## [1] "||Boot {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17} Ent = 0.998"
## [1] "..1)纹理=模糊 {11,12,16} Bad"
## [1] "..1)纹理=清晰 {1,2,3,4,5,6,8,10,15} "
## [1] ".....2)根蒂=蜷缩 {1,2,3,4,5} Good"
## [1] ".....2)根蒂=硬挺 {10} Bad"
## [1] "..1)纹理=稍糊 {7,9,13,14,17} "
## [1] ".....2)触感=软粘 {7} Good"
## [1] ".....2)触感=硬滑 {9,13,14,17} Bad"
## [1] ".....2)根蒂=稍蜷 {6,8,15} "
## [1] ".....3)色泽=青绿 {6} Good"
## [1] ".....3)色泽=乌黑 {8,15} "
## [1] ".....4)触感=软粘 {15} Bad"
## [1] ".....4)触感=硬滑 {8} Good"
```

- 增益率

```
## [1] "||Boot {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17} Ent = 0.998"
## [1] "..1)纹理=模糊 {11,12,16} Bad"
## [1] "..1)纹理=清晰 {1,2,3,4,5,6,8,10,15} "
## [1] ".....2)触感=硬滑 {1,2,3,4,5,8} Good"
## [1] "..1)纹理=稍糊 {7,9,13,14,17} "
## [1] ".....2)触感=软粘 {7} Good"
## [1] ".....2)触感=硬滑 {9,13,14,17} Bad"
## [1] ".....2)触感=软粘 {6,10,15} "
## [1] ".....3)色泽=乌黑 {15} Bad"
## [1] ".....3)色泽=青绿 {6,10} "
## [1] ".....4)根蒂=稍蜷 {6} Good"
## [1] ".....4)根蒂=硬挺 {10} Bad"
```

- 基尼系数

```
## [1] "||Boot {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17} Ent = 0.998"
## [1] "..1)纹理=模糊 {11,12,16} Bad"
## [1] "..1)纹理=清晰 {1,2,3,4,5,6,8,10,15} "
## [1] ".....2)根蒂=蜷缩 {1,2,3,4,5} Good"
## [1] ".....2)根蒂=硬挺 {10} Bad"
## [1] "..1)纹理=稍糊 {7,9,13,14,17} "
## [1] ".....2)触感=软粘 {7} Good"
## [1] ".....2)触感=硬滑 {9,13,14,17} Bad"
## [1] ".....2)根蒂=稍蜷 {6,8,15} "
## [1] ".....3)色泽=青绿 {6} Good"
## [1] ".....3)色泽=乌黑 {8,15} "
```

```
## [1] "..........4)触感=软粘 {15} Bad"
## [1] "..........4)触感=硬滑 {8} Good"
```

通过对比发现，ID3 与 CART 算法得到了相同得决策树。C4.5 算法得到的决策树在第二层中选择了较少类别的属性（触感）。

1.2.1 剪枝处理

- 预剪枝

```
## [1] "||Boot {1,2,3,6,7,10,14,15,16,17} Ent = 1"
## [1] "..1)脐部=平坦 {10,16} Bad"
## [1] "..1)脐部=凹陷 {1,2,3,14} "
## [1] ".....2)色泽=浅白 {14} Bad"
## [1] ".....2)色泽=青绿 {1} Good"
## [1] ".....2)色泽=乌黑 {2,3} Good"
## [1] "..1)脐部=稍凹 {6,7,15,17} "
## [1] ".....2)根蒂=蜷缩 {17} Bad"
## [1] ".....2)根蒂=稍蜷 {6,7,15} "
## [1] ".......3)色泽=青绿 {6} Good"
## [1] ".......3)色泽=乌黑 {7,15} "
## [1] "..........4)纹理=清晰 {15} Bad"
## [1] "..........4)纹理=稍糊 {7} Good"
```

剪枝后的树为：

```
## [1] "||Boot {1,2,3,6,7,10,14,15,16,17} Ent = 1"
## [1] "..1)脐部=平坦 {10,16} Bad"
## [1] "..1)脐部=凹陷 {1,2,3,14} "
## [1] "Class:Good"
## [1] "No increase in accuracy, terminate classification"
## [1] "..1)脐部=稍凹 {6,7,15,17} "
## [1] "Class:Bad"
## [1] "No increase in accuracy, terminate classification"
```

- 后剪枝

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	label
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	1
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	1

```
## [1] "||Boot {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17} Ent = 0.998"
```

```
## [1] "..1)纹理=模糊 {11,12,16} Bad"
## [1] "..1)纹理=清晰 {1,2,3,4,5,6,8,10,15} "
## [1] ".....2)密度=<= 0.3815 {10,15} Bad"
## [1] ".....2)密度=> 0.3815 {1,2,3,4,5,6,8} Good"
## [1] "..1)纹理=稍糊 {7,9,13,14,17} "
## [1] ".....2)触感=软粘 {7} Good"
## [1] ".....2)触感=硬滑 {9,13,14,17} Bad"
```