数据挖掘第五次作业

李晨昊 2017011466

2020 年 4 月 27 日

目录

1 决策树

下表给出了两组人的数据，每组数据分别包含 4 个和 5 个样本。

组次	id	身材	发色	年龄
第一组	1	矮	金色	老人
	2	高	红色	老人
	3	高	金色	老人
	4	矮	金色	成年
第二组	1	高	黑色	儿童
	2	矮	黑色	老人
	3	高	黑色	老人
	4	高	黑色	成年
	5	矮	金色	儿童

1. 使用任意一种决策树方法建立该数据集的二分类器，使它能正确区分这两组人，写出建立过程。

   我选择使用 ID3 算法。表格中的 id 对分类无任何影响，而且因为两组间有相同的 id，也不能用于区分元素，为了方便，我将第二组的编号 1-5 重新编号为 5-9，这样 id 就可以用来区分元素。

   - 剩余元素集合 $\{1,2,3,4,5,6,7,8,9\}$，剩余属性集合 $\{$身材, 发色, 年龄$\}$

   $$Info(D) = -(\frac{4}{9}\log_2\frac{4}{9} + \frac{5}{9}\log_2\frac{5}{9}) = 0.991$$
   $$Info_{身材} = -(\frac{4}{9}(\frac{2}{9}\log_2\frac{2}{9} + \frac{2}{9}\log_2\frac{2}{9}) + \frac{5}{9}(\frac{2}{9}\log_2\frac{2}{9} + \frac{3}{9}\log_2\frac{3}{9})) = 0.990$$
   $$Info_{发色} = -(\frac{4}{9}(\frac{3}{9}\log_2\frac{3}{9} + \frac{1}{9}\log_2\frac{1}{9}) + \frac{1}{9}(\frac{1}{9}\log_2\frac{1}{9} + \frac{0}{9}\log_2\frac{0}{9}) + \frac{4}{9}(\frac{0}{9}\log_2\frac{0}{9} + \frac{4}{9}\log_2\frac{4}{9})) = 0.662$$
   $$Info_{年龄} = -(\frac{5}{9}(\frac{3}{9}\log_2\frac{3}{9} + \frac{2}{9}\log_2\frac{2}{9}) + \frac{2}{9}(\frac{1}{9}\log_2\frac{1}{9} + \frac{1}{9}\log_2\frac{1}{9}) + \frac{2}{9}(\frac{0}{9}\log_2\frac{0}{9} + \frac{2}{9}\log_2\frac{2}{9})) = 0.825$$

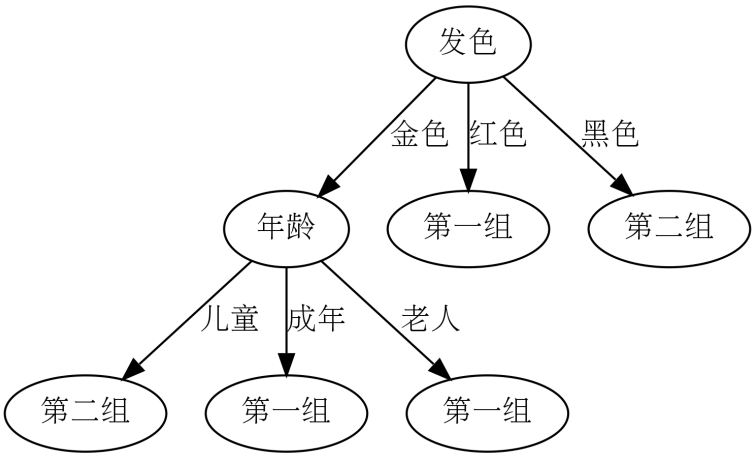
   选择发色，分出 $\{1,3,4,9\}$ 和 $\{2\}$ 和 $\{5,6,7,8\}$。

- 剩余元素集合 $\{1,3,4,9\}$，剩余属性集合 $\{$身材, 年龄$\}$

  $$Info(D) = -(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}) = 0.811$$
  $$Info_{\text{身材}} = -(\frac{3}{4}(\frac{2}{4}\log_2\frac{2}{4} + \frac{1}{4}\log_2\frac{1}{4}) + \frac{1}{4}(\frac{1}{4}\log_2\frac{1}{4} + \frac{0}{4}\log_2\frac{0}{4})) = 0.875$$
  $$Info_{\text{年龄}} = -(\frac{2}{4}(\frac{2}{4}\log_2\frac{2}{4} + \frac{0}{4}\log_2\frac{0}{4}) + \frac{1}{4}(\frac{1}{4}\log_2\frac{1}{4} + \frac{0}{4}\log_2\frac{0}{4}) + \frac{1}{4}(\frac{0}{4}\log_2\frac{0}{4} + \frac{1}{4}\log_2\frac{1}{4})) = 0.500$$

  选择年龄，分出 $\{1,3\}$ 和 $\{4\}$ 和 $\{9\}$。
- 剩余元素集合 $\{1,3\}$，剩余属性集合 $\{$身材$\}$ 。由于元素种类唯一，不再分类。
- 剩余元素集合 $\{4\}$，剩余属性集合 $\{$身材$\}$ 。由于元素种类唯一，不再分类。
- 剩余元素集合 $\{9\}$，剩余属性集合 $\{$身材$\}$ 。由于元素种类唯一，不再分类。
- 剩余元素集合 $\{2\}$，剩余属性集合 $\{$身材, 年龄$\}$ 。由于元素种类唯一，不再分类。
- 剩余元素集合 $\{5,6,7,8\}$，剩余属性集合 $\{$身材, 年龄$\}$ 。由于元素种类唯一，不再分类。

综上，决策树表示为：



2. 用所建分类器说明给定样例 (矮，金色，成年) 是属于第几组。

   第一组。

2 朴素贝叶斯

对于上表中的数据：

1. 用朴素贝叶斯建立二分类器，写出建立过程（不用考虑平滑）。

$$
\begin{aligned}
P(\text{高}|\text{第一组}) &= 0.5 \\
P(\text{矮}|\text{第一组}) &= 0.5 \\
P(\text{金色}|\text{第一组}) &= 0.75 \\
P(\text{红色}|\text{第一组}) &= 0.25 \\
P(\text{黑色}|\text{第一组}) &= 0 \\
P(\text{儿童}|\text{第一组}) &= 0 \\
P(\text{成年}|\text{第一组}) &= 0.25 \\
P(\text{老人}|\text{第一组}) &= 0.75 \\
P(\text{高}|\text{第二组}) &= 0.6 \\
P(\text{矮}|\text{第二组}) &= 0.4 \\
P(\text{金色}|\text{第二组}) &= 0.2 \\
P(\text{红色}|\text{第二组}) &= 0 \\
P(\text{黑色}|\text{第二组}) &= 0.8 \\
P(\text{儿童}|\text{第二组}) &= 0.4 \\
P(\text{成年}|\text{第二组}) &= 0.2 \\
P(\text{老人}|\text{第二组}) &= 0.4
\end{aligned}
$$

2. 用所建分类器对给定样例 (矮，金色，成年) 分类。

$$
\begin{aligned}
&P((\text{矮，金色，成年})|\text{第一组})P(\text{第一组}) \\
&= P(\text{矮}|\text{第一组})P(\text{金色}|\text{第一组})P(\text{成年}|\text{第一组})P(\text{第一组}) \\
&= 0.5 * 0.75 * 0.25 * \frac{4}{9} \\
&= 0.042
\end{aligned}
$$

$$
\begin{aligned}
&P((\text{矮，金色，成年})|\text{第二组})P(\text{第二组}) \\
&= P(\text{矮}|\text{第二组})P(\text{金色}|\text{第二组})P(\text{成年}|\text{第二组})P(\text{第二组}) \\
&= 0.4 * 0.2 * 0.2 * \frac{5}{9} \\
&= 0.009
\end{aligned}
$$

故是第一组。