

决策树

- 例

ID	年龄	有工作	有自己的房子	信贷情况	类别
1	青年	否	否	一般	否
2	青年	否	否	好	否
3	青年	是	否	好	是
4	青年	是	是	一般	是
5	青年	否	否	一般	否
6	中年	否	否	一般	否
7	中年	否	否	好	否
8	中年	是	是	好	是
9	中年	否	是	非常好	是
10	中年	否	是	非常好	是
11	老年	否	是	非常好	是
12	老年	否	是	好	是
13	老年	是	否	好	是
14	老年	是	否	非常好	是
15	老年	否	否	一般	否

$$\begin{aligned} H(D) &= - \sum_{k=1}^K \frac{|D_k|}{|D|} \log_2 \frac{|D_k|}{|D|} \\ &= -\frac{6}{15} \log_2 \frac{6}{15} - \frac{9}{15} \log_2 \frac{9}{15} \\ &= 0.971 \end{aligned}$$

决策树

- 例： $H(D) = 0.971$

ID	年龄	有工作	有自己的房子	信贷情况	类别
1	青年	否	否	一般	否
2	青年	否	否	好	否
3	青年	是	否	好	是
4	青年	是	是	一般	是
5	青年	否	否	一般	否
6	中年	否	否	一般	否
7	中年	否	否	好	否
8	中年	是	是	好	是
9	中年	否	是	非常好	是
10	中年	否	是	非常好	是
11	老年	否	是	非常好	是
12	老年	否	是	好	是
13	老年	是	否	好	是
14	老年	是	否	非常好	是
15	老年	否	否	一般	否

$$\begin{aligned}
 H(D|A) &= \sum_{i=1}^n \frac{|D_i|}{|D|} H(D_i) \\
 &= - \sum_{i=1}^n \frac{|D_i|}{|D|} \sum_{k=1}^K \frac{|D_{ik}|}{|D_i|} \log_2 \frac{|D_{ik}|}{|D_i|}
 \end{aligned}$$

以年龄为例：

$$\begin{aligned}
 &= -\frac{5}{15} \left(\frac{3}{5} \log_2 \frac{3}{5} + \frac{2}{5} \log_2 \frac{2}{5} \right) \\
 &\quad - \frac{5}{15} \left(\frac{2}{5} \log_2 \frac{2}{5} + \frac{3}{5} \log_2 \frac{3}{5} \right) \\
 &\quad - \frac{5}{15} \left(\frac{1}{5} \log_2 \frac{1}{5} + \frac{4}{5} \log_2 \frac{4}{5} \right) \\
 &= \mathbf{0.888}
 \end{aligned}$$

决策树

- 例： $H(D) = 0.971$, $H(D|A = \text{年龄}) = 0.888$

ID	年龄	有工作	有自己的房子	信贷情况	类别
1	青年	否	否	一般	否
2	青年	否	否	好	否
3	青年	是	否	好	是
4	青年	是	是	一般	是
5	青年	否	否	一般	否
6	中年	否	否	一般	否
7	中年	否	否	好	否
8	中年	是	是	好	是
9	中年	否	是	非常好	是
10	中年	否	是	非常好	是
11	老年	否	是	非常好	是
12	老年	否	是	好	是
13	老年	是	否	好	是
14	老年	是	否	非常好	是
15	老年	否	否	一般	否

以有工作为特征：

$$\begin{aligned} &= -\frac{5}{15} \left(\frac{5}{5} \log_2 \frac{5}{5} \right) \\ &\quad - \frac{10}{15} \left(\frac{4}{10} \log_2 \frac{4}{10} + \frac{6}{10} \log_2 \frac{6}{10} \right) \\ &= 0.647 \end{aligned}$$

决策树

• 例 : $H(D) = 0.971$ $H(D|A = \text{年龄}) = 0.888$ $H(D|A = \text{有工作}) = 0.647$

ID	年龄	有工作	有自己的房子	信贷情况	类别
1	青年	否	否	一般	否
2	青年	否	否	好	否
3	青年	是	否	好	是
4	青年	是	是	一般	是
5	青年	否	否	一般	否
6	中年	否	否	一般	否
7	中年	否	否	好	否
8	中年	是	是	好	是
9	中年	否	是	非常好	是
10	中年	否	是	非常好	是
11	老年	否	是	非常好	是
12	老年	否	是	好	是
13	老年	是	否	好	是
14	老年	是	否	非常好	是
15	老年	否	否	一般	否

$$H(D|A = \text{有房子}) = 0.551 \quad \checkmark$$

$$H(D|A = \text{信贷情况}) = 0.608$$