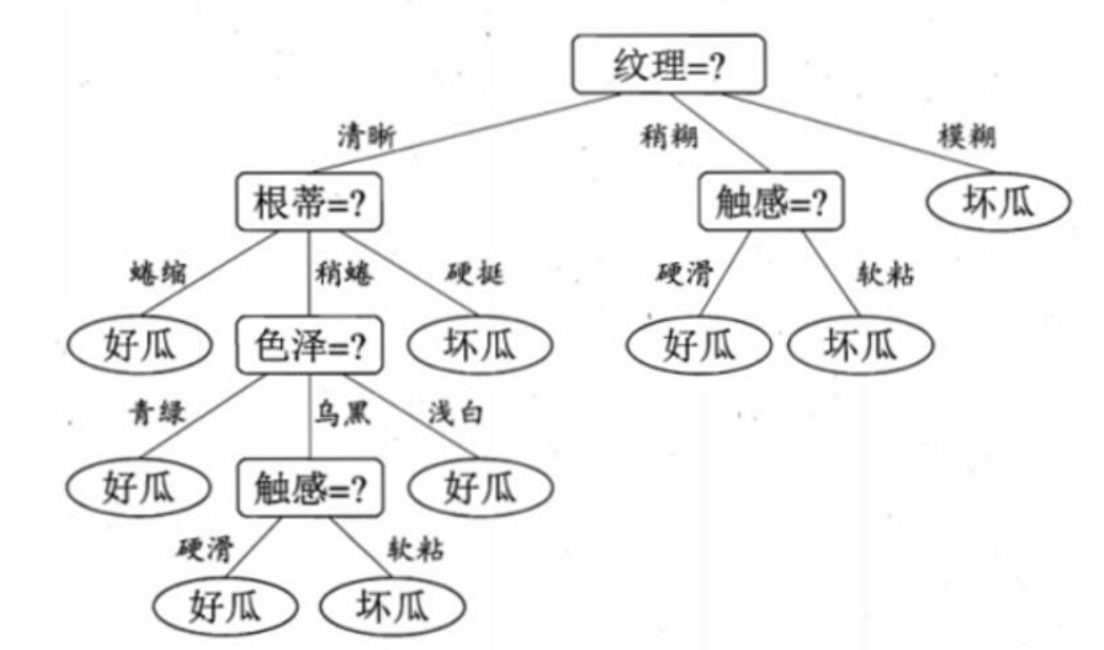
决策树

比较适合分析离散数据。 如果是连续数据要先转成离散数据再做分析。

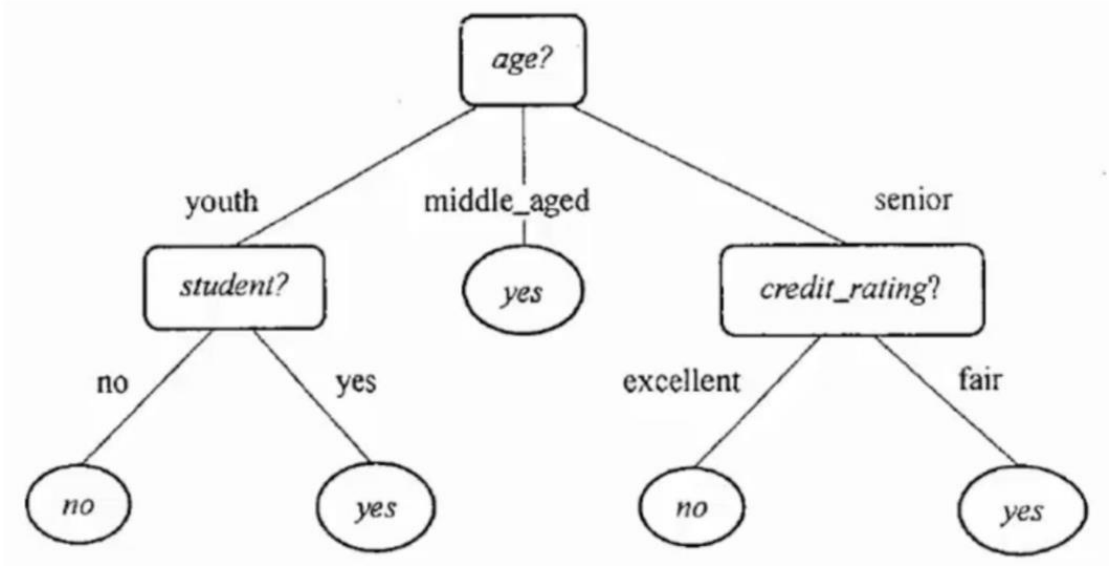


70 年代后期至 80 年代，Quinlan 开发了 ID3 算法。

Quinlan 改进了 ID3 算法，称为 C4.5 算法。

1984 年，多位统计学家提出了 CART 算法。

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no



熵(entropy)概念

1948 年，香农提出了“信息熵”的概念。

一条信息的信息量大小和它的不确定性有直接的关系， 要搞清楚一件非常非常不确定的事情，或者是我们一 无所知的事情，需要了解大量信息->信息量的度量就 等于不确定性的多少。

信息熵公式:

$$H = -\sum_{x} P(x)\log_2 P(x)$$

假如有一个普通骰子 A，仍出 1-6 的概率都是 1/6

有一个骰子 B，扔出 6 的概率是 50%，扔出 1-5 的概率都是 10%

有一个骰子 C，扔出 6 的概率是 100%。

骰子 A：- 1 6× log2 1/6× 6 ≈ 2.585

骰子 B：- 1 10× log2 1/10× 5 −1 2× log2 1/2≈ 2.161

骰子 C：- 1 × log2/1 = 0

决策树会选择最大化信息增益来对结点进行划分。 信息增益计算:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

$$Gain(A) = Info(D) - Info_A(D)$$

信息增益(Information Gain)：Gain(A) = Info(D) - Infor_A(D)

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

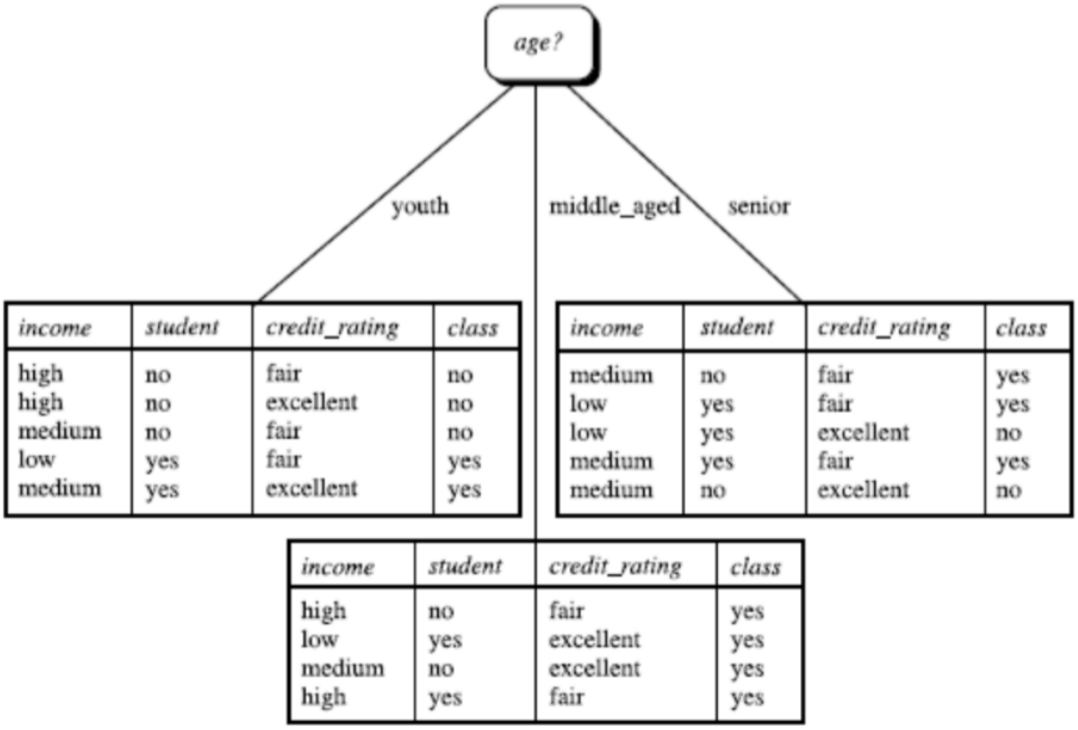
$$Info(D) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.940 \text{ bits.}$$

$$\begin{aligned} Info_{age}(D) &= \frac{5}{14} \times \left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) \\ &+ \frac{4}{14} \times \left(-\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}\right) \\ &+ \frac{5}{14} \times \left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right) \\ &= 0.694 \text{ bits.} \end{aligned}$$

$$Gain(age) = Info(D) - Info_{age}(D) = 0.940 - 0.694 = 0.246 \text{ bits.}$$

类似：
Gain(income) = 0.029,
Gain(student) = 0.151,
Gain(credit_rating)=0.048

选择根节点-ID3 算法



income	student	credit_rating	class
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes

income	student	credit_rating	class
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

income	student	credit_rating	class
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

连续变量处理

C4.5 算法

信息增益的方法倾向于首先选择因子数较多的变量 信息增益的改进：增益率

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)$$

$$GrianRate(A) = \frac{Grain(A)}{SplitInfo_A(D)}$$