Python
机器学习实战

决策树

决策树

国际权威学术组织ICDM在2006年12月评选出数据挖掘领域的十大经典算法

数据挖掘十大算法：一览表

IDMer

排名	挖掘主题	算法	得票数	发表时间	作者	讲解人
1	分类	C4.5	61	1993	Quinlan, J.R	*Hiroshi Motoda*
2	聚类	K-Means	60	1967	MacQueen, J.B	*Joydeep Ghosh*
3	统计学习	SVM	58	1995	Vapnik, V.N	*Qiang Yang*
4	关联分析	Apriori	52	1994	Rakesh Agrawal	*Christos Faloutsos*
5	统计学习	EM	48	2000	McLachlan, G	*Joydeep Ghosh*
6	链接挖掘	PageRank	46	1998	Brin, S.	*Christos Faloutsos*
7	集装与推进	AdaBoost	45	1997	Freund, Y.	*Zhi-Hua Zhou*
8	分类	kNN	45	1996	Hastie, T	*Vipin Kumar*
9	分类	Naïve Bayes	45	2001	Hand, D.J	*Qiang Yang*
10	分类	CART	34	1984	L.Breiman	*Dan Steinberg*

决策树

ID.3、C4.5、C5.0

CART（Classification and Regression Tree，分类回归树）

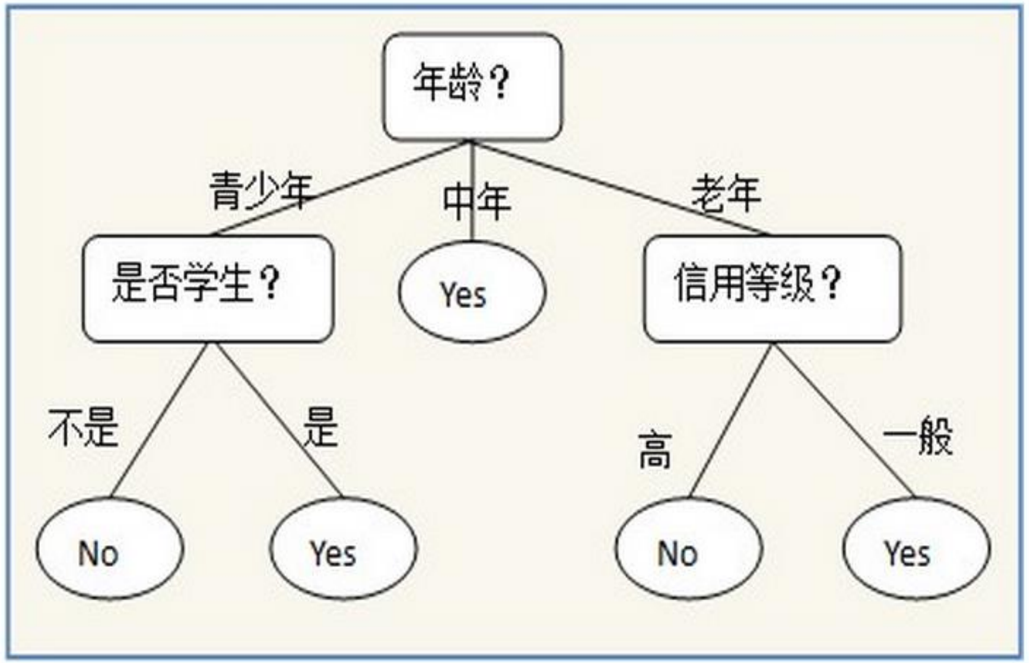
元模型

Bagging、Boosting、随机森林

决策树

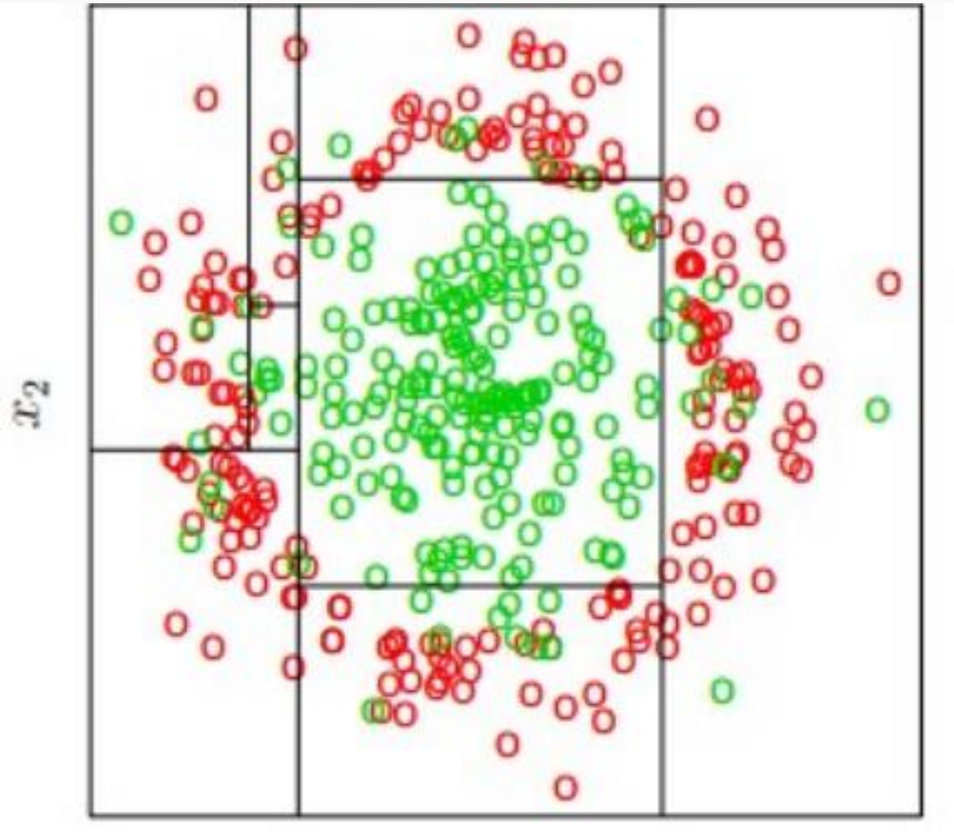
认识决策树

记录 ID	年龄	输入层次	学生	信用等级	是否购买电脑
1	青少年	高	否	一般	否
2	青少年	高	否	良好	否
3	中年	高	否	一般	是
4	老年	中	否	一般	是
5	老年	低	是	一般	是
6	老年	低	是	良好	否
7	中年	低	是	良好	是
8	青少年	中	否	一般	否
9	青少年	低	是	一般	是
10	老年	中	是	一般	是
11	青少年	中	是	良好	是
12	中年	中	否	良好	是
13	中年	高	是	一般	是
14	老年	中	否	良好	否



决策树

- 对样本数据不断分组过程
- 是个线性分类器



决策树

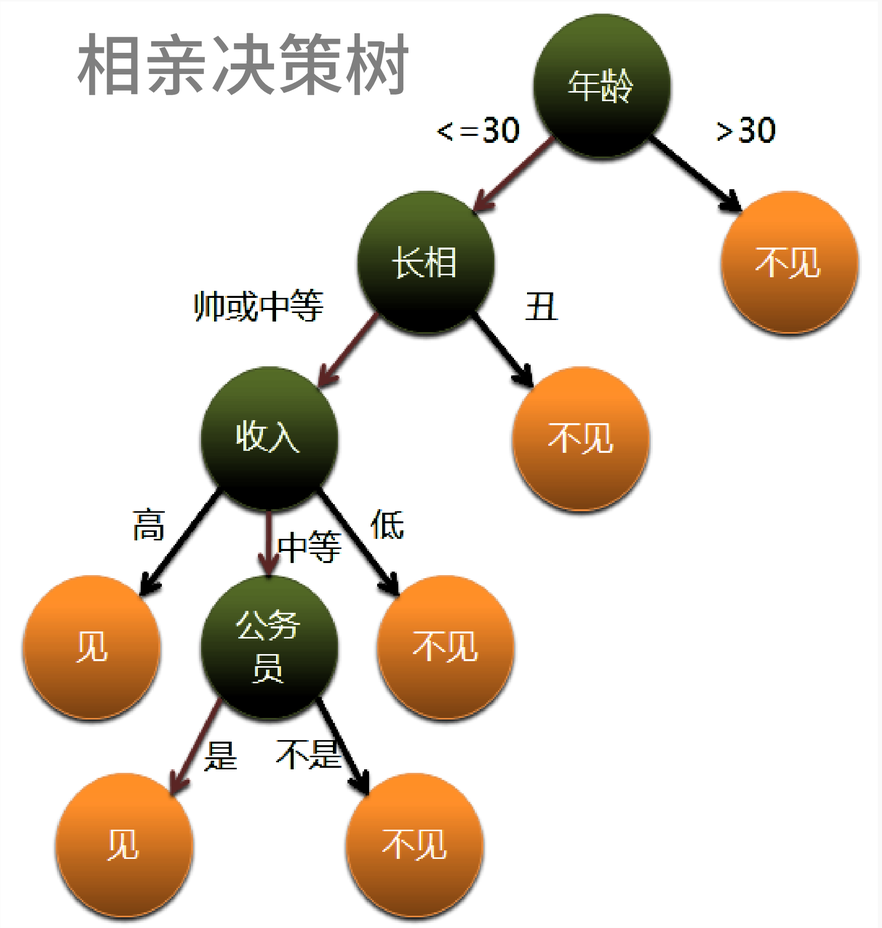
术语

根节点：一棵决策树只有一个根节点

叶节点：代表一个类别

中间节点：代表在一个属性上的测试

分支：代表一个测试输出



二叉树和多叉树

二叉树：每个节点最多有两个分支

多叉树：每个节点不止有两个分支

决策树

树的生长

- 采用分而治之的策略
- 选变量的顺序：如何从众多决策变量中选择一个当前最佳的决策变量；
- 最佳分离点在哪：如何从分组变量的众多取值中找到一个最佳的分割点；

树的修剪

避免过渡拟合：过于个性化、失去了一般化

（收入大于1万、年龄小于30岁、姓名是张三的）

决策树

ID3

- 信息增益
- 没有修剪

C4.5

- 信息增益率
- 悲观剪枝法

C5.0

- 信息增益率
- 自适应增强

CART（分类回归树）

- 基尼指数
- “代价复杂度”剪枝法

决策树

如何衡量混杂程度

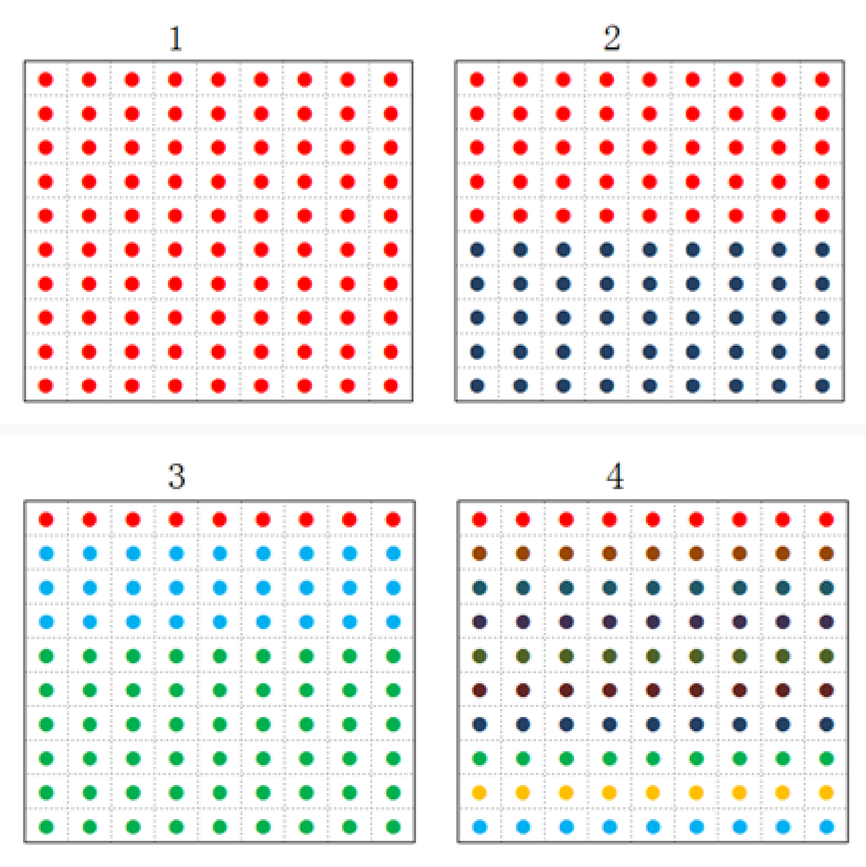
箱子1：100个红球

箱子2：50个红球、50个球

箱子3：10个红、30个篮、60个绿球

箱子4：10种颜色，每种颜色各10个

箱子5：有100种颜色的球



决策树

凭直觉，箱子1显然是最纯粹的。箱子5是最乱的，箱子3比箱子4要好一点，箱子2要再好一些。

如何把这种感觉量化呢，以数字化表示，进而计算机可以计算和处理。

对于这个数字，我们希望，集合越纯粹，数字越小。不妨想象成就一个集合的“纯洁度”给“差评”，越不纯洁，差评分数越高，越纯洁，分数越低，直到0差评。

度量信息混杂程度的指标

- 熵

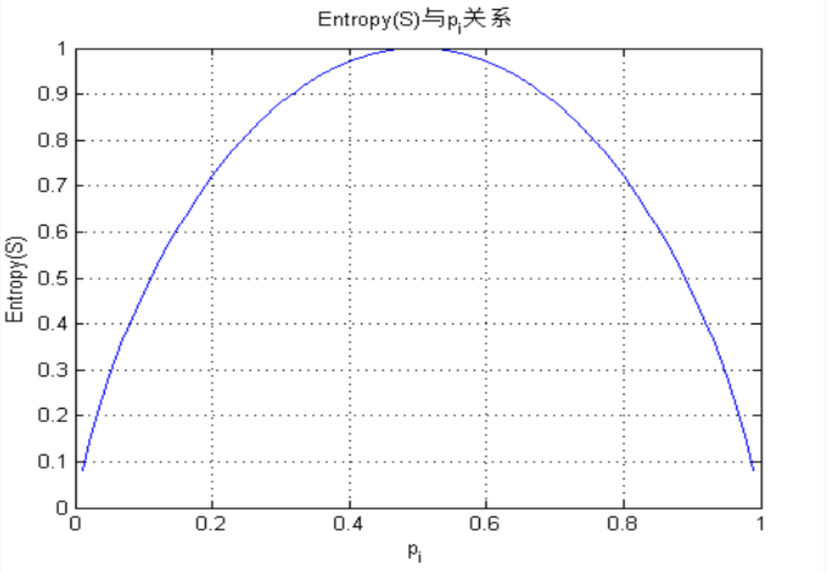
$$H(X) = -\sum_{x \in X} p(x) \ln p(x)$$

- 基尼系数

$$Gini(p) = \sum_{k=1}^{K} p_k (1 - p_k) = 1 - \sum_{k=1}^{K} p_k^2$$

熵的理解

- 事件X的信息量的期望
- 事件的X的结果发生的概率越小，信息量越大
- 定义一个结果的信息量：h(x)=-log x
- 所有结果的信息量，或者事件X的信息量的期望



熵是随机变量不确定性的量化指标，越不确定，熵越大

决策树

条件熵

在知道了条件Y之后，X事件仍然具有的不确定性

$$H(X|Y) = H(X,Y) - H(Y)$$

$$= \sum_x p(x) H(Y \mid X = x)$$

信息增益（互信息）

- 在知道了事件Y之后，X事件不确定性减少程度
- Gain = 熵 – 条件熵

选择信息增益最大的属性作为决策属性（ID3）

决策树

- 是否购买电脑的熵

$$H(D) = -\frac{5}{14}\log_2\frac{5}{14} - \frac{9}{14}\log_2\frac{9}{14} = 0.94$$

- 按照年龄这一列分裂

$$H_{age}(D_{youth}) = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.971$$

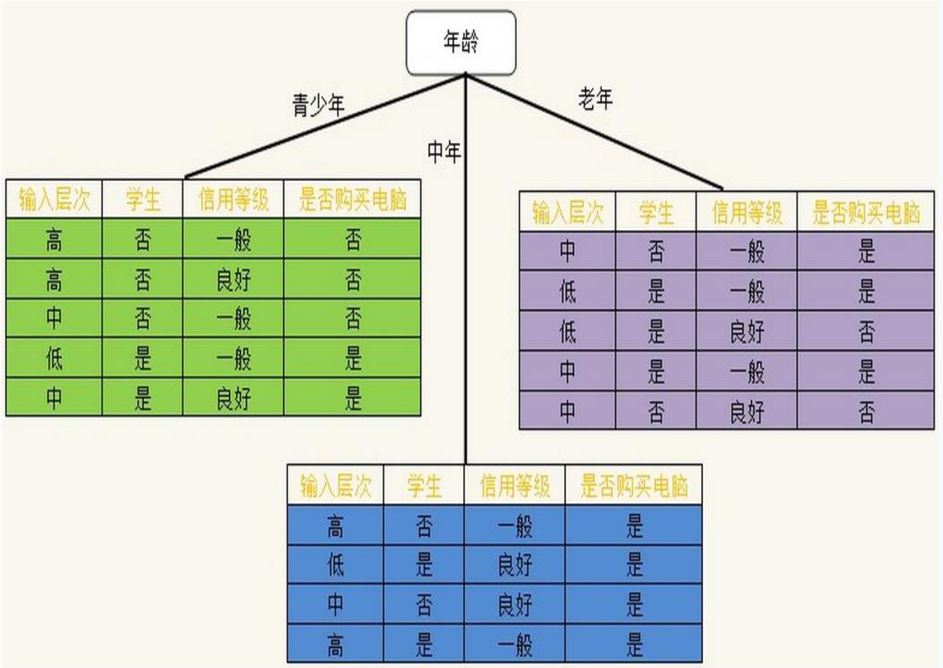
$$H_{age}(D_{youth}) = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.971$$

- 分裂后的条件熵

$$H_{age}(D) = \frac{5}{14}*0.971 + \frac{4}{14}*0 + \frac{5}{14}*0.971 = 0.694$$

- 用年龄作为决策属性后的信息增益

$$Gain(age) = H(D) - H_{age}(D) = 0.246$$



输入层次	学生	信用等级	是否购买电脑
高	否	一般	否
高	否	良好	否
中	否	一般	否
低	是	一般	是
中	是	良好	是

输入层次	学生	信用等级	是否购买电脑
中	否	一般	是
低	是	一般	是
低	是	良好	否
中	是	一般	是
中	否	良好	否

输入层次	学生	信用等级	是否购买电脑
高	否	一般	是
低	是	良好	是
中	否	良好	是
高	是	一般	是

决策树

分别计算四个属性各自的信息增益

- Gain(age) = 0.246
- Gain(income)=0.029
- Gain(stduent)=0.151
- Gain(credit_rating)=0.048

于是把年龄作为根结点的测试属性，根据青少年、中年、老年分为三个分支

信息增益率

- 信息增益倾向于选择更混杂的属性
- 信息增益的改进：增益率
- C4.5

$$g_r(D,A) = g(D,A) / H(A)$$

信息增益率的例子

Gain(income)=0.029

H(income) = $-\frac{4}{14}\times\log_2\frac{4}{14}-\frac{6}{14}\times\log_2\frac{6}{14}-\frac{4}{14}\times\log_2\frac{4}{14}=1.557$

Gr(income) = 0.029 / 1.557 = 0.019

实战