模式识别
(Pattern Recognition)

广东工业大学集成电路学院　邢延

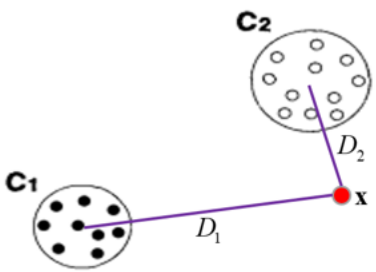
提醒

- 第9周做翻转课堂的同学请自带笔记本电脑！

回顾

● 已经讲过的分类器

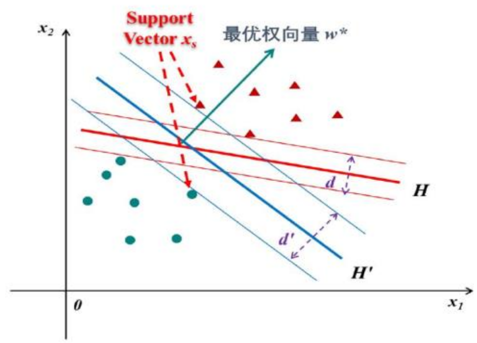
➤ 基于距离的分类器



$\because D_2 < D_1$

$\therefore \mathbf{x} \in C_2$

● 支持向量机(Support Vector Machine, SVM)



分类决策函数：

$$f(\boldsymbol{\phi}(\boldsymbol{x})) = sign(\sum_{j=1}^{J} \boldsymbol{\alpha}_j^* y_j(\phi(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{x}_j))),$$

其中$\boldsymbol{\alpha}_j^* > 0$

➤ 朴素Bayes分类器

$$p(C_i|\mathbf{x}) = \frac{p(\mathbf{x}|C_i)p(C_i)}{p(\mathbf{x})}, 1 \le i \le m$$

其中，$p(\mathbf{x}) = \sum_{i=1}^{m} p(\mathbf{x}|C_i)p(C_i)$，即$\mathbf{x}$的概率分布

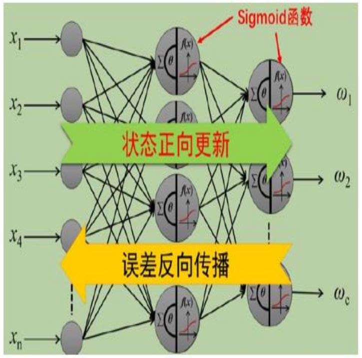
$$posterio = \frac{likelihood \times prior}{evidence}$$

后验概率主要由先验概率和似然函数的乘积所决定

给定训练数据集，evidence是个常数

➤ 反向传播（Back Propagation,BP）网络

- ♦ 典型的前馈型网络
- ♦ 相邻两层神经元间是全连接的
- ♦ 输入层的神经元个数为样本特征向量的维数
- ♦ 输出神经元的个数为类别数
- ♦ 激活函数采用Sigmoid函数



第六讲 决策树

（06 Decision Trees）

内容提要

- 决策树基本概念
- 决策树的经典算法
- 决策树的优、缺点

决策树基本概念

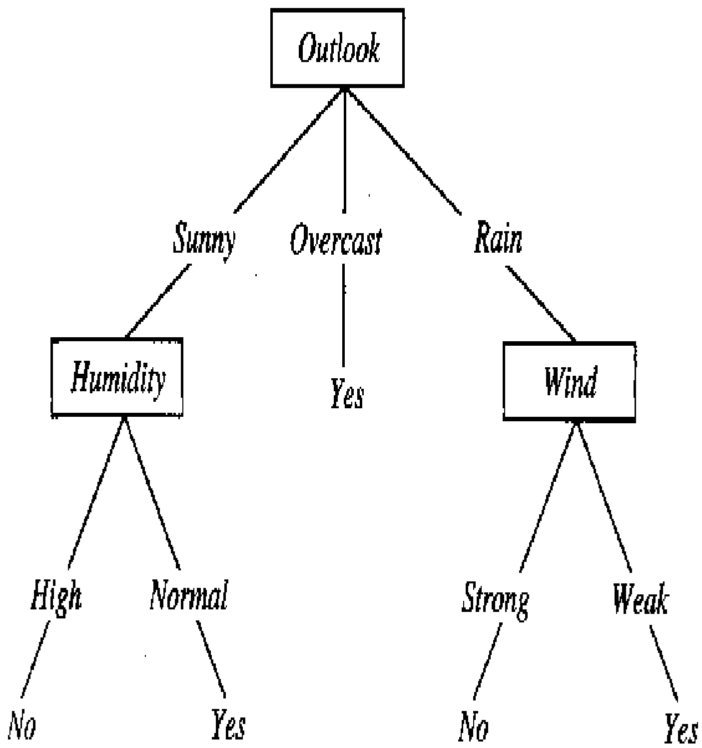
决策树如何运作:

老师随机写下0—100间的任意一个整数，请同学来猜，请问：

怎么猜效率最高?

决策树基本概念

决策树的表现形式：



决策树的基本组成部分：决策结点、分支和叶子。

决策树中最上面的结点称为根结点。是整个决策树的开始。

每个分支是一个新的决策结点，或者是树的叶子。

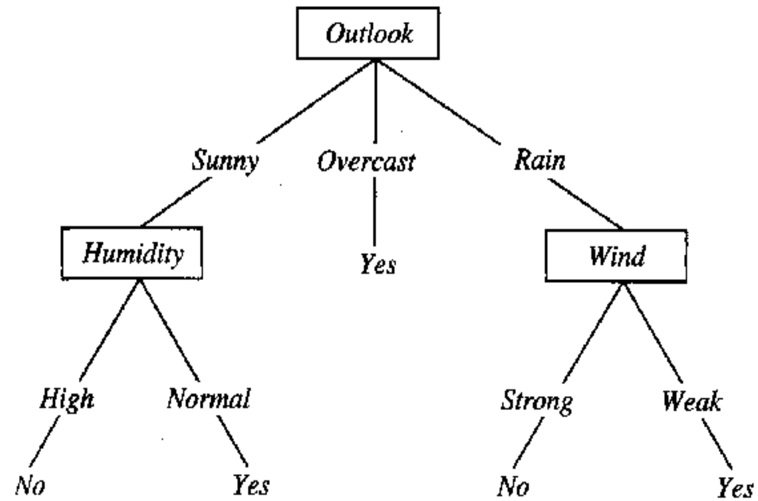
每个决策结点代表一个问题或者决策，通常对应待分类对象的属性。

每个叶结点代表一种可能的分类结果。

决策树基本概念

- **决策树所产生的规则**
  - **决策树代表实例属性值约束的合取的析取式。**
  - **从树根到树叶的每一条路径对应一组属性测试的合取**
  - **树本身对应这些合取的析取**

$$(Outlook = sunny \wedge Humidity = normal)$$
$$\vee (Outlook = overcast)$$
$$\vee (Outlook = rain \wedge Wind = weak)$$



决策树的经典算法

- **算法的发展历史**
  - **1979年, J.R. Quinlan 给出ID3算法，并在1983年和1986年对ID3 进行了总结和简化，使其成为决策树学习算法的典型。**
  - **1993年，Quinlan 进一步发展了ID3算法，改进成C4.5算法。**
  - **另一类决策树算法为CART（回归树），与C4.5不同的是，CART的决策树由二元逻辑问题生成，每个树节点只有两个分枝，分别包括学习实例的正例与反例。**
    - https://www.salford-systems.com/products/cart
    - 1984， Friedman & Breiman

决策树的经典算法

- **ID3算法**
  - **决策树学习方法中最具影响和最为典型的算法**
  - **使用“信息增益” 选择属性**
    - **信息增益：衡量属性对分类提供的信息的多少**

决策树的经典算法

- **ID3算法**

  - **信息增益的计算**

假设： S 是训练样本集合，$|S|$是训练样本数。样本划分为m个不同的类$C_1, C_2, \dots, C_m$，其样本数量分别为$|C_1|, |C_2|, \dots, |C_m|$，则$Gain(S, A)$是属性$A$在集合$S$上的信息增益:

$$Gain(S, A) = Entropy(S) - Entropy(S, A) \qquad (1)$$

(1)式中：

$Entropy(S)$是训练样本的信息熵（即决策属性的信息熵），$Entropy(S, A)$ 是训练样本中属性A的信息熵，且有：

$$Entrop(S, A) = \sum_v \frac{|S_v|}{|S|} Entropy(S_v) \qquad (2)$$

(2)式中：

$\sum$是属性A的所有可能的值v，S_v是属性A 有v 值的S 子集，$|S_v|$是其样本数， $Entropy(S_v)$是其信息熵。

设$\left|C_j^v\right|$是S_v中属于j类样本的数量，则有信息熵的计算公式：

$$Entropy(S) = -\sum_{i=1}^{m} \frac{|C_i|}{|S|} \log_2 \frac{|C_i|}{|S|}$$

$$(3)$$

$$Entropy(S_v) = -\sum_{j=1}^{m} \frac{\left|C_j^v\right|}{|S_v|} \log_2 \frac{\left|C_j^v\right|}{|S_v|}$$

决策树的经典算法

- **ID3算法**
  - **信息增益的计算例子**

计数	年龄	收入	学生	信誉	归类：买计算机？
64	青	高	否	良	不买
64	青	高	否	优	不买
128	中	高	否	良	买
60	老	中	否	良	买
64	老	低	是	良	买
64	老	低	是	优	不买
64	中	低	是	优	买
128	青	中	否	良	不买
64	青	低	是	良	买
132	老	中	是	良	买
64	青	中	是	优	买
32	中	中	否	优	买
32	中	高	是	良	买
63	老	中	否	优	不买
1	老	中	否	优	买

决策树的经典算法

- **ID3算法**
  - 信息增益的计算例子

计数	年龄	收入	学生	信誉	归类：买计算机？
64	青	高	否	良	不买
64	青	高	否	优	不买
128	中	高	否	良	买
60	老	中	否	良	买
64	老	低	是	良	买
64	老	低	是	优	不买
64	中	低	是	优	买
128	青	中	否	良	不买
64	青	低	是	良	买
132	老	中	是	良	买
64	青	中	是	优	买
32	中	中	否	优	买
32	中	高	是	良	买
63	老	中	否	优	不买
1	老	中	否	优	买

第1步计算决策属性的熵，即$Entropy(S)$

$$Entropy(S) = -\sum_{i=1}^{m} \frac{|C_i|}{|S|} \log_2 \frac{|C_i|}{|S|} \qquad (3)$$

决策属性“买计算机？”，该属性分两类：买/不买

S_1(买)=641

S_2（不买）= 383

$S=S_1+S_2=1024$

$P_1=641/1024=0.6260$

$P_2=383/1024=0.3740$

由公式（3），得

$E(S_1,S_2)=E(641,383)$

$=-P_1Log_2P_1-P_2Log_2P_2$

$=-(P_1Log_2P_1+P_2Log_2P_2)$

$=0.9537$

决策树的经典算法

- **ID3算法**
  - **信息增益的计算例子**

计数	年龄	收入	学生	信誉	归类：买计算机？
64	青	高	否	良	不买
64	青	高	否	优	不买
128	中	高	否	良	买
60	老	中	否	良	买
64	老	低	是	良	买
64	老	低	是	优	不买
64	中	低	是	优	买
128	青	中	否	良	不买
64	青	低	是	良	买
132	老	中	是	良	买
64	青	中	是	优	买
32	中	中	否	优	买
32	中	高	是	良	买
63	老	中	否	优	不买
1	老	中	否	优	买

第2步计算条件属性的熵，即$Entropy(S,A)$

$$Entrop(S,A) = \sum_v \frac{|S_v|}{|S|} Entropy(S_v) \qquad (2)$$

条件属性共有4个。分别是年龄、收入、学生、信誉。
分别计算不同属性的信息增益。

决策树的经典算法

- **ID3算法**
  - **信息增益的计算例子**

计数	年龄	收入	学生	信誉	归类：买计算机？
64	青	高	否	良	不买
64	青	高	否	优	不买
128	中	高	否	良	买
60	老	中	否	良	买
64	老	低	是	良	买
64	老	低	是	优	不买
64	中	低	是	优	买
128	青	中	否	良	不买
64	青	低	是	良	买
132	老	中	是	良	买
64	青	中	是	优	买
32	中	中	否	优	买
32	中	高	是	良	买
63	老	中	否	优	不买
1	老	中	否	优	买

第2-1)步计算年龄的熵$Entropy(S_v)$和信息增益

$$Entropy(S_v) = -\sum_{j=1}^{m} \frac{|C_j^v|}{|S_v|} \log_2 \frac{|C_j^v|}{|S_v|} \tag{3}$$

A=年龄，共分三个组：

v={青年、中年、老年}

v=青年，买与不买比例为128/256

S_1(买)=128

S_2（不买）= 256

$S=S_1+S_2=384$

$P_1=128/384$

$P_2=256/384$

由公式（3），得

$E(S_1,S_2)=E(128,256)$

$=-P_1Log_2P_1-P_2Log_2P_2$

$=-(P_1Log_2P_1+P_2Log_2P_2)$

$=0.9183$

决策树的经典算法

● ID3算法

第2-2)步计算年龄的熵$Entropy(S_v)$和信息增益

➢ 信息增益的计算例子

计数	年龄	收入	学生	信誉	归类：买计算机？
64	青	高	否	良	不买
64	青	高	否	优	不买
128	中	高	否	良	买
60	老	中	否	良	买
64	老	低	是	良	买
64	老	低	是	优	不买
64	中	低	是	优	买
128	青	中	否	良	不买
64	青	低	是	良	买
132	老	中	是	良	买
64	青	中	是	优	买
32	中	中	否	优	买
32	中	高	是	良	买
63	老	中	否	优	不买
1	老	中	否	优	买

$$Entropy(S_v) = -\sum_{j=1}^{m} \frac{|C_j^v|}{|S_v|} \log_2 \frac{|C_j^v|}{|S_v|} \quad (3)$$

A=年龄共分三个组：

v={青年、中年、老年}

V=中年,买与不买比例为256/0

S_1(买)=256

S_2（不买）= 0

$S=S_1+S_2=256$

$P_1=256/256$

$P_2=0/256$

$E(S_1,S_2)=E(256，0)$

$=-P_1Log_2P_1-P_2Log_2P_2$

$=-(P_1Log_2P_1+P_2Log_2P_2)$

$=0$

决策树的经典算法

- **ID3算法**

  - ➢ **信息增益的计算例子**

计数	年龄	收入	学生	信誉	归类：买计算机？
64	青	高	否	良	不买
64	青	高	否	优	不买
128	中	高	否	良	买
60	老	中	否	良	买
64	老	低	是	良	买
64	老	低	是	优	不买
64	中	低	是	优	买
128	青	中	否	良	不买
64	青	低	是	良	买
132	老	中	是	良	买
64	青	中	是	优	买
32	中	中	否	优	买
32	中	高	是	良	买
63	老	中	否	优	不买
1	老	中	否	优	买

第2-3)步计算年龄的熵$Entropy(S_v)$和信息增益

$$Entropy(S_v) = -\sum_{j=1}^{m} \frac{|C_j^v|}{|S_v|} \log_2 \frac{|C_j^v|}{|S_v|} \qquad (3)$$

A=年龄共分三个组：

v={青年、中年、老年}

V=老年,买与不买比例为257/127

S_1(买)=257

S_2（不买）=127

$S=S_1+S_2=384$

$P_1=257/384$

$P_2=127/384$

E(S1,S2)=E(257，127)

$=-P_1Log_2P_1-P_2Log_2P_2$

$=-(P_1Log_2P_1+P_2Log_2P_2)$

=0.9157

决策树的经典算法

- **ID3算法**

第2-4)步计算年龄的熵$Entropy(S_v)$和信息增益

➢ 信息增益的计算例子

计数	年龄	收入	学生	信誉	归类：买计算机？
64	青	高	否	良	不买
64	青	高	否	优	不买
128	中	高	否	良	买
60	老	中	否	良	买
64	老	低	是	良	买
64	老	低	是	优	不买
64	中	低	是	优	买
128	青	中	否	良	不买
64	青	低	是	良	买
132	老	中	是	良	买
64	青	中	是	优	买
32	中	中	否	优	买
32	中	高	是	良	买
63	老	中	否	优	不买
1	老	中	否	优	买

$$Entrop(S,A) = \sum_v \frac{|S_v|}{|S|} Entropy(S_v) \qquad (2)$$

A=年龄共分三个组：
v={青年、中年、老年}
所占比例
青年组 384/1024=0.375
中年组 256/1024=0.25
老年组 384/1024=0.375

则年龄的信息熵$Entropy(S,A)$：
E（年龄）=0.375*0.9183+
0.25*0+
0.375*0.9157
=0.6877

- **ID3算法**

第2-5)步计算年龄的熵$Entropy(S_v)$和信息增益

➢ 信息增益的计算例子

计数	年龄	收入	学生	信誉	归类：买计算机？
64	青	高	否	良	不买
64	青	高	否	优	不买
128	中	高	否	良	买
60	老	中	否	良	买
64	老	低	是	良	买
64	老	低	是	优	不买
64	中	低	是	优	买
128	青	中	否	良	不买
64	青	低	是	良	买
132	老	中	是	良	买
64	青	中	是	优	买
32	中	中	否	优	买
32	中	高	是	良	买
63	老	中	否	优	不买
1	老	中	否	优	买

$$Gain(S,A) = Entropy(S) - Entropy(S,A) \qquad (1)$$

年龄的信息增益：

G（年龄） =0.9537-0.6877

=0.2660　（1）

决策树的经典算法

- **ID3算法**

  - **信息增益的计算例子**

计数	年龄	收入	学生	信誉	归类：买计算机？
64	青	高	否	良	不买
64	青	高	否	优	不买
128	中	高	否	良	买
60	老	中	否	良	买
64	老	低	是	良	买
64	老	低	是	优	不买
64	中	低	是	优	买
128	青	中	否	良	不买
64	青	低	是	良	买
132	老	中	是	良	买
64	青	中	是	优	买
32	中	中	否	优	买
32	中	高	是	良	买
63	老	中	否	优	不买
1	老	中	否	优	买

同理，可计算出其它三个属性的信息增益：

A=收入，共分三个组：
={高、中、低}
E（收入）=0.9361
G（收入） =0.9537-0.9361
=0.0176 (2)

A=学生，共分二个组：
v={学生、非学生}
E（学生）=0.7811
G（学生） =0.9537-0.7811
=0.1726 （3）

A=信誉，分二个组：
v={良好，优秀}
E（信誉）= 0.9048
G（信誉） =0.9537-0.9048
=0.0453 （4）

决策树的经典算法

- **ID3算法**

第3步 选择节点，应该选取信息增益最大的那个属性

➢ **信息增益的计算例子**

计数	年龄	收入	学生	信誉	归类：买计算机？
64	青	高	否	良	不买
64	青	高	否	优	不买
128	中	高	否	良	买
60	老	中	否	良	买
64	老	低	是	良	买
64	老	低	是	优	不买
64	中	低	是	优	买
128	青	中	否	良	不买
64	青	低	是	良	买
132	老	中	是	良	买
64	青	中	是	优	买
32	中	中	否	优	买
32	中	高	是	良	买
63	老	中	否	优	不买
1	老	中	否	优	买

G(年龄)=0.9537-0.6877
=0.2660 （1）

G(收入)=0.9537-0.9361
=0.0176 （2）

G(学生）=0.9537-0.7811
=0.1726 （3）

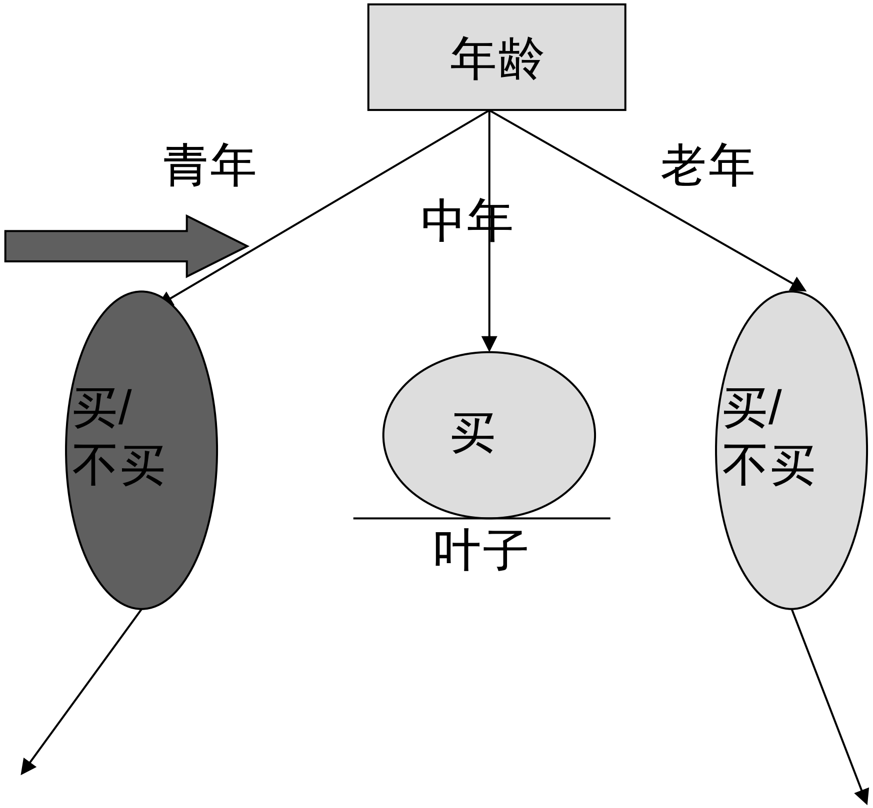
G(信誉)=0.9537-0.9048
=0.0453 （4）

决策树的经典算法

- **ID3算法**
  - **信息增益的计算例子**

计数	年龄	收入	学生	信誉	归类：买计算机？
64	青	高	否	良	不买
64	青	高	否	优	不买
128	青	中	否	良	不买
64	青	低	是	良	买
64	青	中	是	优	买

第4步：根节点及其分支，针对每个不是叶节点的分支，**根据划入该分支的数据**，继续重复上述计算



决策树的经典算法

- **ID3算法**
  - **信息增益的计算例子**

计数	年龄	收入	学生	信誉	归类：买计算机？
64	青	高	否	良	不买
64	青	高	否	优	不买
128	青	中	否	良	不买
64	青	低	是	良	买
64	青	中	是	优	买

第4-1步：

$$Entropy(S) = -\sum_{i=1}^{m} \frac{|C_i|}{|S|} \log_2 \frac{|C_i|}{|S|} \tag{3}$$

买与不买比例为128/256

S_1(买)=128
S_2（不买）= 256
S=S1+S2=384

P_1=128/384
P_2=256/384

$E(S_1,S_2)=E(128,256)$
$=-P_1Log_2P_1-P_2Log_2P_2$
$=-(P_1Log_2P_1+P_2Log_2P_2)$
$=0.9183$

决策树的经典算法

- **ID3算法**
  - **信息增益的计算例子**

计数	年龄	收入	学生	信誉	归类：买计算机？
64	青	高	否	良	不买
64	青	高	否	优	不买
128	青	中	否	良	不买
64	青	低	是	良	买
64	青	中	是	优	买

第4-2步：

依据相同的计算步骤，选择A=收入作为节点
V={高、中、低}

E(0,128)=0
比例:
128/384=0.3333

E(64,128)=0.9183
比例: 192/384=0.5

E(64,0)=0
比例:
64/384=0.1667

信息熵：

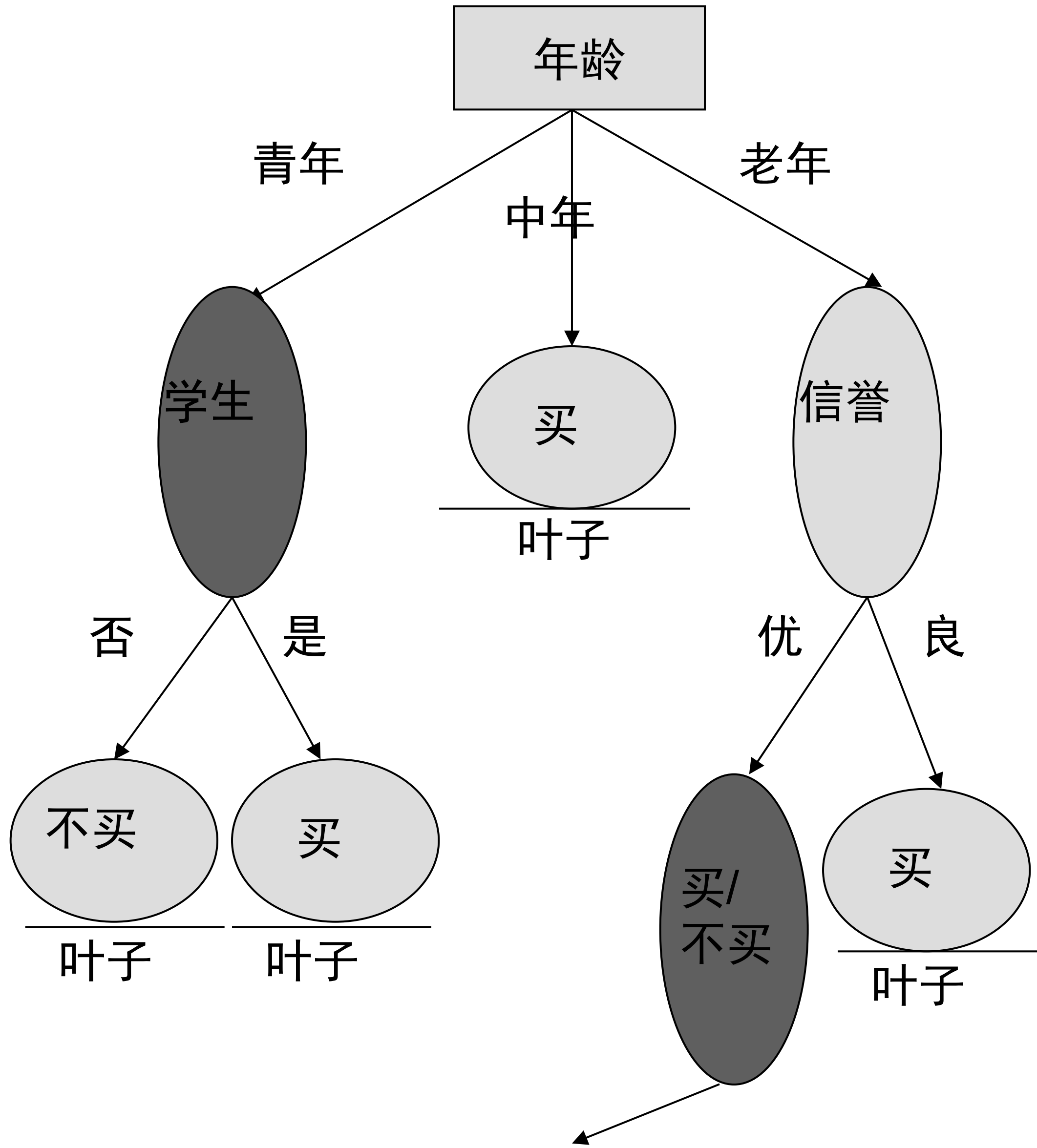
E(收入） = 0.3333 * 0 + 0.5 * 0.9183 + 0.1667 * 0 = 0.4592

Gain(收入) = E(128, 256) - E(收入)=0.9183 – 0.4592 = 0.4591

决策树的经典算法

- **ID3算法**
  - **信息增益的计算例子**

第5)步 分支1# 的计算，选择“学生”属性，继续树的生长.



决策树的经典算法

- **ID3算法**
  - **算法步骤**

1）决定分类属性；

2） 对目前的数据表，建立一个节点N

3） 如果数据库中的数据都属于同一个类，N就是树叶，在树叶上标出所属的类

4） 如果数据表中没有其他属性可以考虑，则N也是树叶，按照少数服从多数的原则在树叶上标出所属类别

5） 否则，根据信息增益GAIN值选出一个最佳属性作为节点N的测试属性

6） 节点属性选定后，对于该属性中的每个值：

- **6a)从N生成一个分支，并将数据表中与该分支有关的数据收集形成分支节点的数据表，**
- **6b)在表中删除节点属性那一栏，如果分支数据表非空，则运用以上算法从该节点建立子树。**

决策树的经典算法

- **ID3算法**
  - **数据准备**

原始表

姓名	年龄	收入	学生	信誉	电话	地址	邮编	买计算机
张三	23	4000	是	良	281-322-0328	2714 Ave. M	77388	买
李四	34	2800	否	优	713-239-7830	5606 Holly Cr	78766	买
王二	70	1900	否	优	281-242-3222	2000 Bell Blvd.	70244	不买
赵五	18	900	是	良	281-550-0544	100 Main Street	70244	买
刘兰	34	2500	否	优	713-239-7430	606 Holly Ct	78566	买
杨俊	27	8900	否	优	281-355-7990	233 Rice Blvd.	70388	不买
张毅	38	9500	否	优	281-556-0544	399 Sugar Rd.	78244	买
。。。	。。							
。。。								

决策树的经典算法

- **ID3算法**
  - **数据准备**

整理后的数据表

计数	年龄	收入	学生	信誉	归类：买计算机?
64	青	高	否	良	不买
64	青	高	否	优	不买
128	中	高	否	良	买
60	老	中	否	良	买
64	老	低	是	良	买
64	老	低	是	优	不买
64	中	低	是	优	买
128	青	中	否	良	不买
64	青	低	是	良	买
。。。					

- Data cleaning
  删除/减少noise，
  补填missing values
- Data transformation
  数据标准化（data normalization）
  数据归纳（generalize data to higher-level concepts using concept hierarchies）

例如：年龄归纳为老、中、青三类

控制每个属性的可能值不超过七种（最好不超过五种）

- Relevance analysis
  对于与问题无关的属性：删
  对于属性的可能值大于七种又不能归纳的属性：删

决策树的经典算法

- **ID3算法**
  - **算法的缺点**
    - 属性必须是离散值
    - 不能有缺失值
    - 容易过拟合
  - **思考题**
    - ID3生成模型最极端的情况是怎样的?

决策树的经典算法

- **C4.5算法**
  - **在ID3算法的基础上，进行了以下改进：**
    - **用信息增益率替代信息增益**
      - 信息增益的缺陷：取值越多的属性，其信息增益越大
      - 例如：日期
    - **在树的构造过程中进行剪枝**
      - 防止过拟合
    - **能够对连续属性进行离散化处理**
      - 自动进行数据预处理
    - **能够对缺失数据进行处理**
      - 自动进行数据预处理

决策树的经典算法

- **C4.5算法**

  - **信息增益率的计算**

假设： S 是训练样本集合，$|S|$是训练样本数。样本划分为m个不同的类$C_1, C_2, \ldots, C_m$，其样本数量分别为$|C_1|, |C_2|, \ldots, |C_m|$，则$GainRatio(S, A)$是属性$A$在集合$S$上的信息增益率：

$$GainRatio(S, A) = \frac{Gain(S,A)}{SplitInformation(S,A)} \qquad (4)$$

(4)式中：

$Gain(S, A)$是属性A在集合S上的信息增益，计算见公式(1)(2)(3)

SplitInformation(S, A)是分裂信息，用于衡量属性分裂数据的广度和均匀性，且有：

$$SplitInformation(S, A) = \sum_v \frac{|S_v|}{|S|} \log_2 \frac{|S_v|}{|S|} \qquad (5)$$

(5)式中：

$\sum$是属性A的所有可能的值v，S_v是属性A 有v 值的S 子集，$|S_v|$是其样本数。

决策树的经典算法

- **C4.5算法**
  - **算法步骤**
  - **过程一：生成树**
    - **用信息增益率作为度量选取测试属性**
    - **按照ID3递归算法生成树**
    - **允许过拟合**
    - **算法停止条件：**
      - **如果某一节点的分支所覆盖的样本都属于同一类的时候,那么递归就可以终止,该分支就会产生一个叶子节点；**
      - **如果某一分支覆盖的样本的个数如果小于一个阈值，那么也可产生叶子节点，该叶子节点所覆盖的样本哪个类占大多数，那么该叶子节点的类别就是那个占大多数的类。**

- **C4.5算法**
  - **算法步骤**
  - **过程二：剪枝**
    - **错误率降低剪枝(Reduced-Error Pruning ,REP)**
      - 数据量足够时，使用与训练样例截然不同的验证样本，来评估通过后修剪方法从树上修剪节点的效用
    - **悲观剪枝(Pessimistic Error Pruning ,PEP)**
      - 数据量有限时，使用所有可用的数据进行训练，但进行统计测试来估计修剪一个特定的节点是否有可能改善在训练集合外的实例上的性能

决策树的经典算法

- **C4.5算法**
  - **优点**
    - 原理较简单直观
    - 模型具有可解释性
  - **缺点**
    - 计算效率较低，特别是针对含有连续属性值的训练样本时表现的尤为突出；
    - 在选择分裂属性时没有考虑属性间的相关性，有可能影响到属性选择的正确性；
    - 非全局最优解

模式识别算法编程实例演示

- **实例6：决策树**
  - **参见Jupyter Notebook文档目录：lesson06**