决策树分类方法

- 分类简介
- 决策树的用途
- 决策树的建立(例)
- 决策树的数据准备
- 小结

分类简介

谁加何种类型的油？

姓名	年龄	收入	种族	信誉	电话	地址	邮编	加何种油
张三	23	4000	亚裔	良	281-322-0328	2714 Ave. M	77388	Supreme
李四	34	2800	白人	优	713-239-7830	5606 Holly Cr	78766	Regular
王二	70	1900	西班牙	优	281-242-3222	2000 Bell Blvd.	70244	Plus
赵五	18	900	非洲	良	281-550-0544	100 Main Street	70244	Supreme
刘兰	34	2500	白人	优	713-239-7430	606 Holly Ct	78566	Regular
杨俊	27	8900	亚裔	优	281-355-7990	233 Rice Blvd.	70388	Plus
张毅	38	9500	亚裔	优	281-556-0544	399 Sugar Rd.	78244	Regular

分类简介

学校录取部门的困扰：新生录取以后会不会来报到？

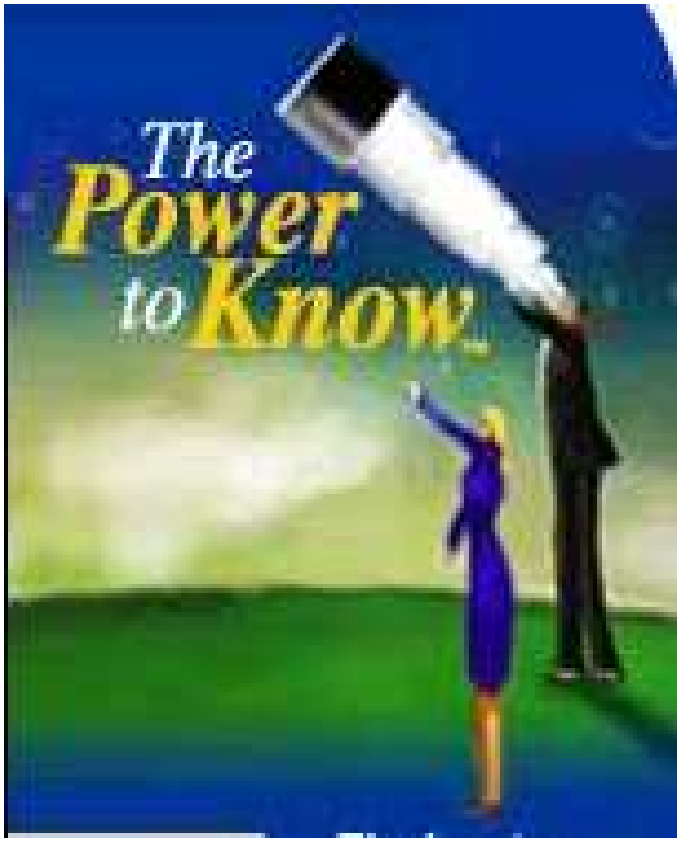
- 性别
- 年龄
- 种族
- 家庭人口
- 家庭收入
- 申请该校原因
- 家庭住址 ……

分类简介

你能判定他/她买计算机的可能性大不大吗？

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
	• • • • • •							

分类简介



1 我们拥有什么:

Huge amount of data (GTE:1TB/day)

2. 我们需要什么:

Information and knowledge

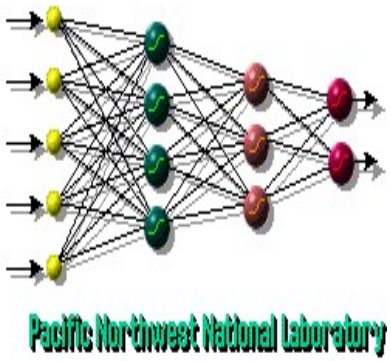
3. 我们应该怎么办:

Data Mining（Classfication…）

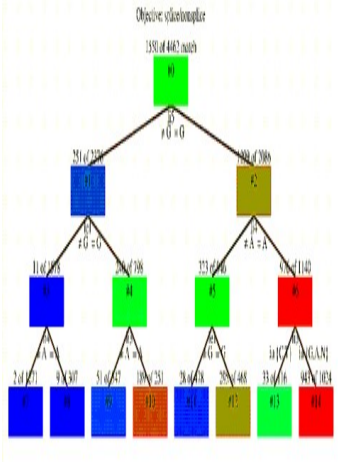
分类

- 分类的任务：通过分析训练集中数据表现出来的特性，找到一个能准确的描述这些数据的规则和模型
- 用于预测未来的数据趋势，辅助决策
- 分类知识发现已广泛有效的应用于科学试验，信贷审核，商业预测，医疗诊断等方面

分类技术



各八
显仙
神过
通海



- **Decision trees**
  决策树
- **Neural Networks**
  计算机神经网络
- **Bayes Classifier**
  贝叶斯分类
- **Association Rule**
- ……

内容概要

- 分类简介
- 决策树的用途
- 决策树的建立(例)
- 决策树的数据准备
- 小结

决策树的用途

- 决策树把数据归入可能对一个目标变量有不同效果的规则组。

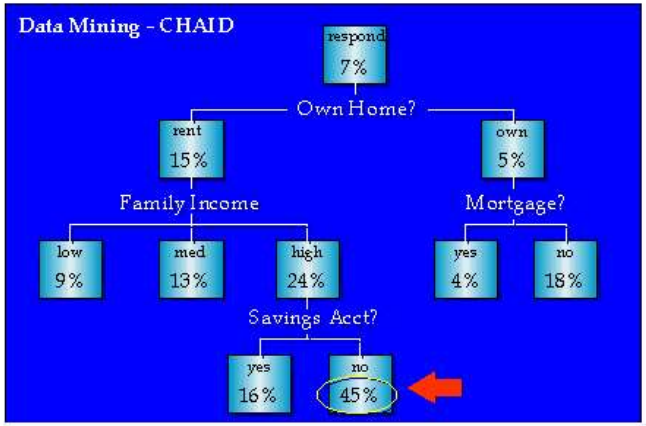
数据挖掘技术：决策树

决策树把数据归入可能对一个目标变量有不同效果的规则组。例如，我们希望发现可能会对直邮有反应的个人特点。这些特点可以解释为一组规则。

假设您是一个销售一种新的银行服务的直邮计划研究的负责人。为最大程度地获益，您希望确定基于前次促销活动的家庭细分最有可能响应相似的促销活动。通常这可以通过查找最能把响应前次促销的家庭和没有响应的家庭区分开的人口统计信息变量的组合来实现，种种技术称为“数据分段”或“分段建模”。

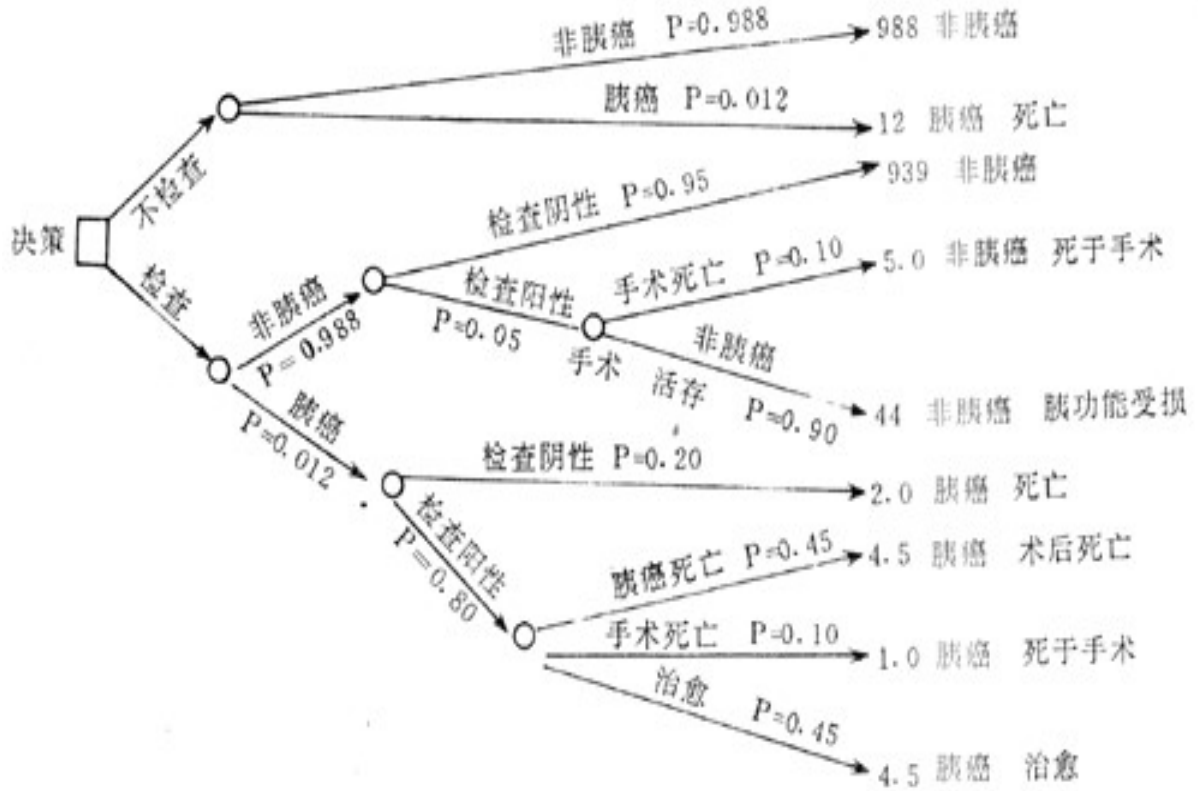
这一过程为您提供诸如谁会最好地响应新的促销等重要线索，并通过只邮寄给最有可能响应的人来最大程度地获得直邮效益，提高整体响应率，并极有希望同时增加销售。以下是在AnswerTree中用CHAID算法简化分段的过程：

下图中，可以看到所有收到直邮信件的人中有7%有响应。但是，如果分为有住房和无住房两组，则15%的租户有响应，而房主则只有5%。我们可以继续分组来发现最有可能响应的组群。这一组群可以表示为一个规则，如“如果收件人是租户，有较高的家庭收入，没有储蓄存款账户，那么他有45%的响应概率”。简单地说，有这些特点的组群中有45%可能会对直邮有响应。



决策树的用途

- 临床决策


决策
不检查
检查
非胰癌 P=0.988
988 非胰癌
胰癌 P=0.012
12 胰癌 死亡
检查阴性 P=0.95
939 非胰癌
非胰癌 P=0.988
检查阳性 P=0.05
手术死亡 P=0.10
5.0 非胰癌 死于手术
手术 活存
非胰癌 P=0.90
44 非胰癌 胰功能受损
胰癌 P=0.012
检查阴性 P=0.20
2.0 胰癌 死亡
检查阳性 P=0.80
胰癌死亡 P=0.45
4.5 胰癌 术后死亡
手术死亡 P=0.10
1.0 胰癌 死于手术
治愈 P=0.45
4.5 胰癌 治愈

决策树的用途

- 根据起因分类设备故障

DELL™ SERVICE

故障诊断

决策树的用途

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

假定公司收集了左表数据，那么对于任意给定的客人（测试样例），你能帮助公司将这位客人归类吗？

即：你能预测这位客人是属于“买”计算机的那一类，还是属于“不买”计算机的那一类？

又：你需要多少有关这位客人的信息才能回答这个问题？

决策树可以帮助你解决*好*这个问题

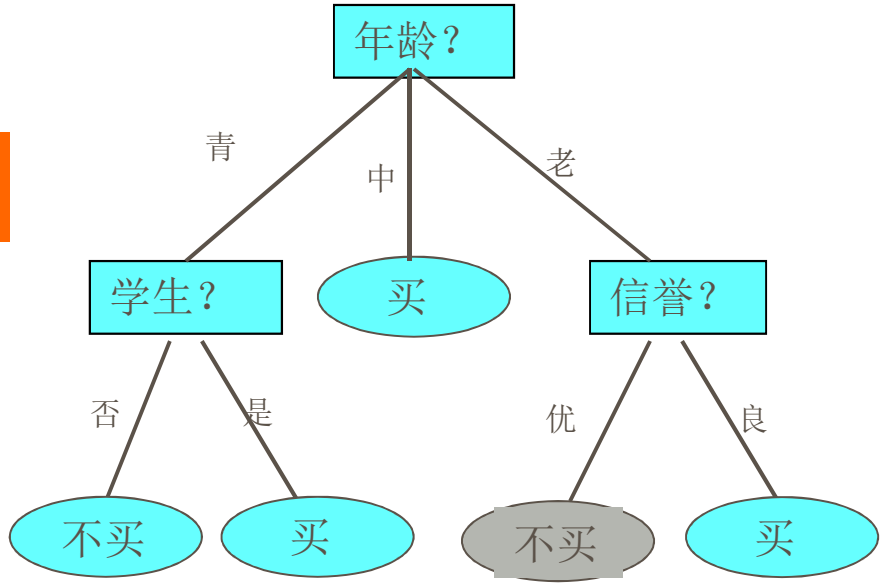
决策树

- 内部节点上选用一个属性进行分割（测试）
- 分枝表示测试输出
- 叶子节点表示类

谁在买计算机？

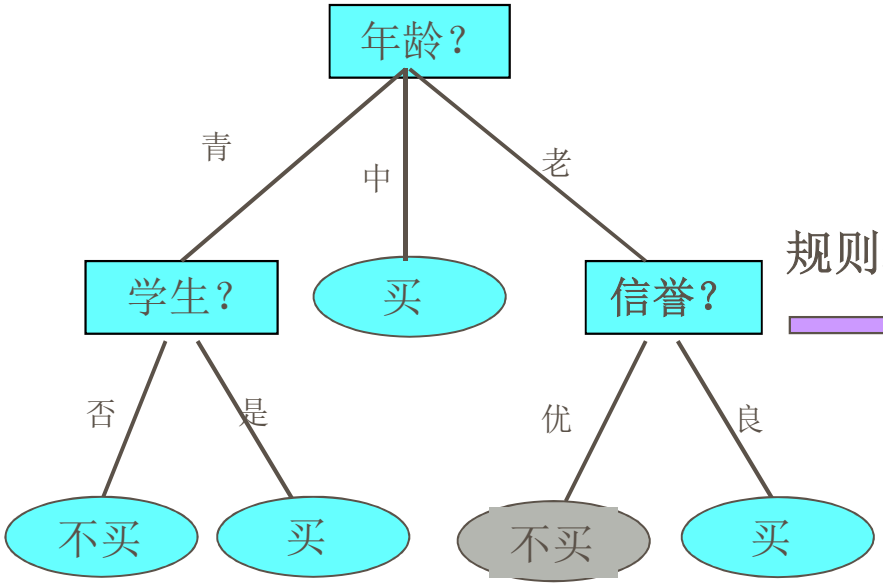
类似情况 → 学习

他/她会买计算机吗？



决策树学习

- 决策树算法对数据处理过程中，将数据按树状结构分成若干分枝形成决策树，从根到树叶的每条路径创建一个规则。



规则表示

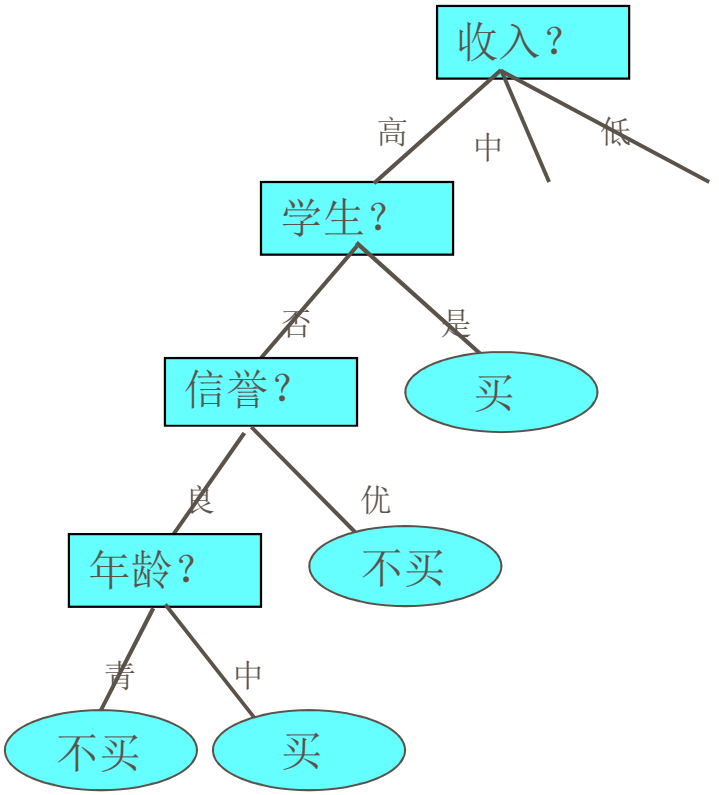
If (年龄＝中)**or** (年龄＝老 **and** 信誉＝良) **or**(年龄＝青 **and** 学生＝是) **then** 买计算机

If (年龄＝老 **and** 信誉＝优) **or** (年龄＝青 **and** 学生＝否)
then 不买计算机

反例

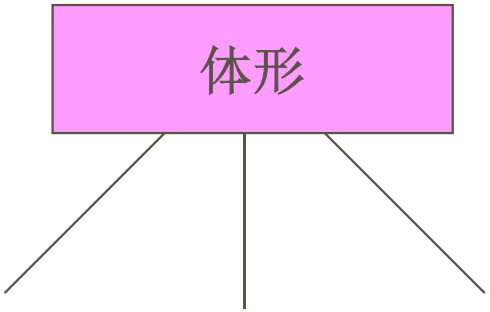
计数	年龄	收入	学生	信誉	归类：买计算机？
64	青	高	否	良	不买
64	青	高	否	优	不买
128	中	高	否	良	买
60	老	中	否	良	买
64	老	低	是	良	买
64	老	低	是	优	不买
64	中	低	是	优	买
128	青	中	否	良	不买
64	青	低	是	良	买
132	老	中	是	良	买
64	青	中	是	优	买
32	中	中	否	优	买
32	中	高	是	良	买
63	老	中	否	优	不买
1	老	中	否	优	买

一棵很糟糕的决策树



决策树生成过程

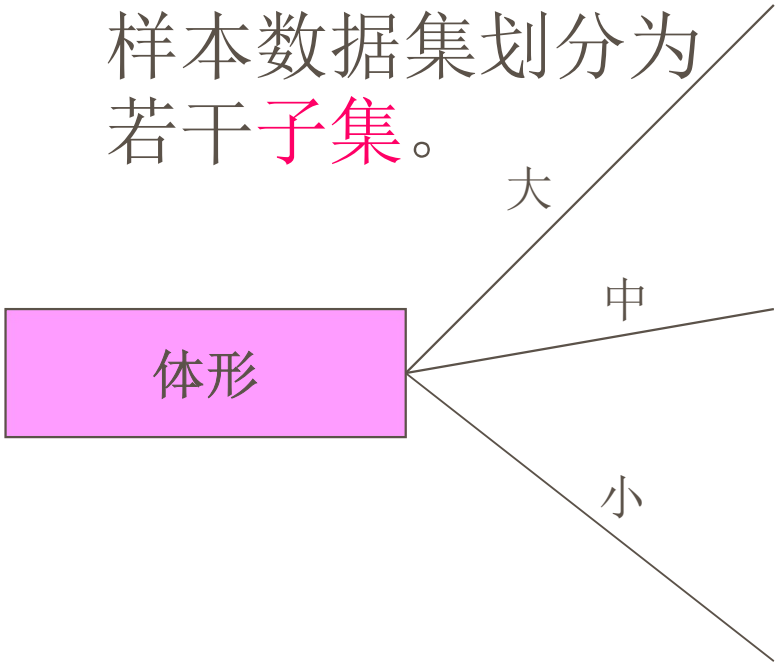
(1)在条件属性集中选择最有分类标识能力的属性作为决策树当前节点。



实例序号	颜色	体形	毛型	类别
1	黑	大	卷毛	危险
2	棕	大	光滑	危险
3	棕	中	卷毛	不危险
4	黑	小	卷毛	不危险
5	棕	中	光滑	危险
6	黑	大	光滑	危险
7	棕	小	卷毛	危险
8	棕	小	光滑	不危险
9	棕	大	卷毛	危险
10	黑	中	卷毛	不危险
11	黑	中	光滑	不危险
12	黑	小	光滑	不危险

决策树生成过程

(2) 根据当前决策属性取值不同，将训练样本数据集划分为若干子集。



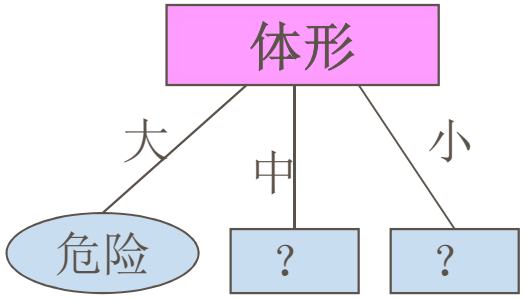
实例序号	颜色	体形	毛型	类别
1	黑	大	卷毛	危险
2	棕	大	光滑	危险
6	黑	大	光滑	危险
9	棕	大	卷毛	危险

实例序号	颜色	体形	毛型	类别
3	棕	中	卷毛	不危险
5	棕	中	光滑	危险
10	黑	中	卷毛	不危险
11	黑	中	光滑	不危险

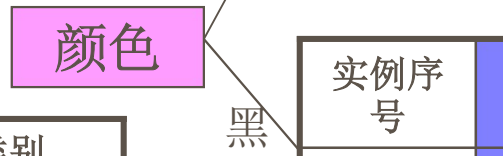
实例序号	颜色	体形	毛型	类别
4	黑	小	卷毛	不危险
7	棕	小	卷毛	危险
8	棕	小	光滑	不危险
12	黑	小	光滑	不危险

决策树生成过程

(3) 针对上一步得到每一个子集，重复上述过程，直到子集中所有元组都属于同一类，不能再进一步划分为止。

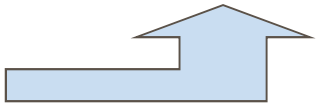


实例序号	颜色	体形	毛型	类别
3	棕	中	卷毛	不危险
5	棕	中	光滑	危险
10	黑	中	卷毛	不危险
11	黑	中	光滑	不危险

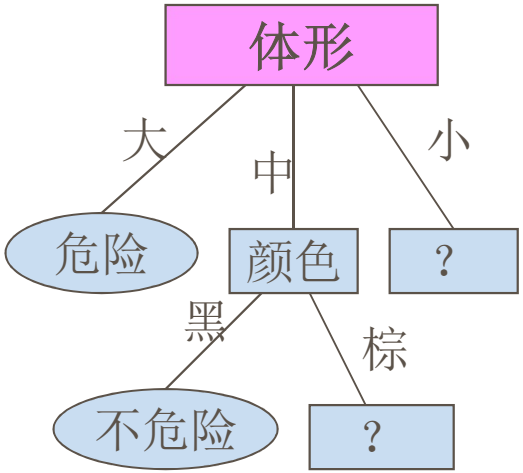


实例序号	颜色	体形	毛型	类别
3	棕	中	卷毛	不危险
5	棕	中	光滑	危险

实例序号	颜色	体形	毛型	类别
10	黑	中	卷毛	不危险
11	黑	中	光滑	不危险



决策树生成过程



实例序号	颜色	体形	毛型	类别
4	黑	小	卷毛	不危险
7	棕	小	卷毛	危险
8	棕	小	光滑	不危险
12	黑	小	光滑	不危险

颜色

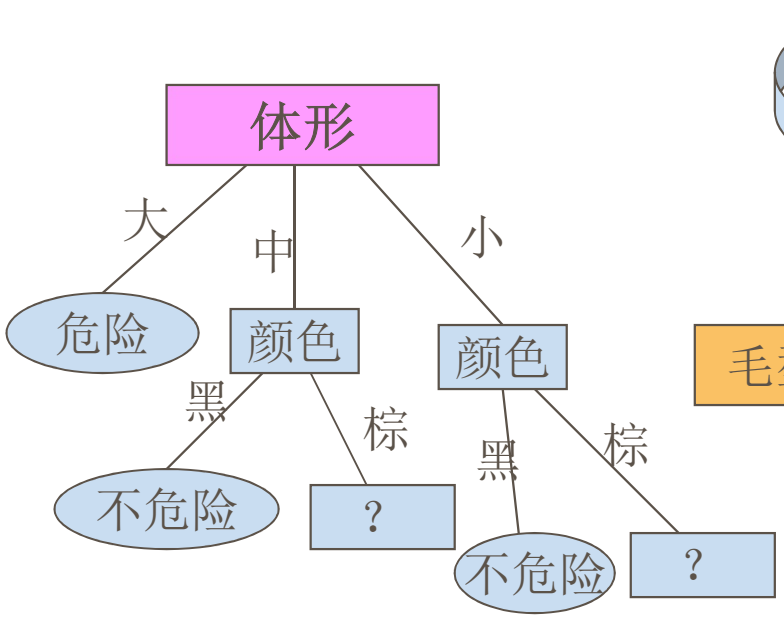
黑

实例序号	颜色	体形	毛型	类别
4	黑	小	卷毛	不危险
12	黑	小	光滑	不危险

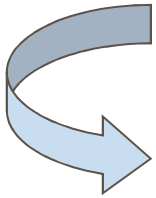
棕

实例序号	颜色	体形	毛型	类别
7	棕	小	卷毛	危险
8	棕	小	光滑	不危险

决策树生成过程



实例序号	颜色	体形	毛型	类别
7	棕	小	卷毛	危险
8	棕	小	光滑	不危险



实例序号	颜色	体形	毛型	类别
3	棕	中	卷毛	不危险
5	棕	中	光滑	危险

毛型

卷毛

实例序号	颜色	体形	毛型	类别
3	棕	中	卷毛	不危险

光滑

实例序号	颜色	体形	毛型	类别
5	棕	中	光滑	危险

毛型

卷毛

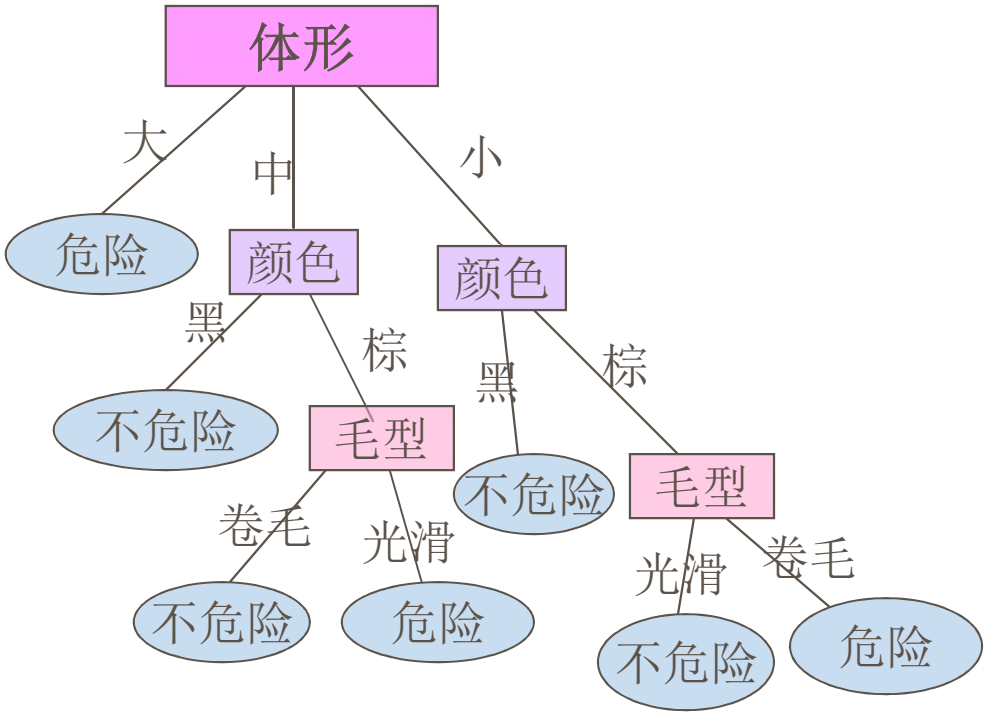
实例序号	颜色	体形	毛型	类别
3	棕	小	卷毛	危险

光滑

实例序号	颜色	体形	毛型	类别
5	棕	小	光滑	不危险

决策树生成过程

- 最终生成的决策树



内容概要

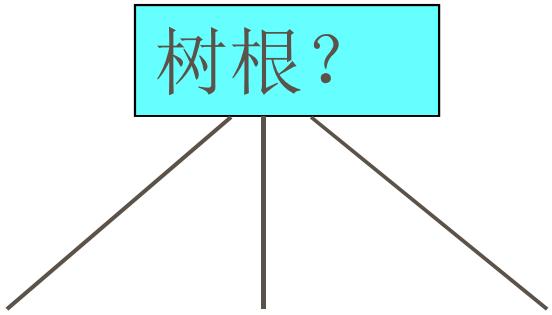
- 分类简介
- 决策树的用途
- 决策树的建立(例)
- 决策树的数据准备
- 小结

决策树的建立

-- 决策树建立的关键

实例序号	颜色	体形	毛型	类别
1	黑	大	卷毛	危险
2	棕	大	光滑	危险
3	棕	中	卷毛	不危险
4	黑	小	卷毛	不危险
5	棕	中	光滑	危险
6	黑	大	光滑	危险
7	棕	小	卷毛	危险
8	棕	小	光滑	不危险
9	棕	大	卷毛	危险
10	黑	中	卷毛	不危险
11	黑	中	光滑	不危险
12	黑	小	光滑	不危险

建立一个好的决策树的关键是决定树根和子树根的属性



决策树分类算法—ID3算法

- 基本思想：

  按一定准则选择一个条件属性作为根节点，根据其属性取值将整个例子空间划分为几个子空间，然后递归使用这一准则继续划分，直到所有底层子空间只含有一类例子，决策树构造结束。

ID3学习算法

- 1 熵 度量样例的纯度 (度量标准)

  熵定义：设S是n个数据样本的集合，将样本划分为c个不同的类，每个类含样本数 n_i，则S划分为c个类的熵为

$$E(S) = -\sum_{i=1}^{c} \frac{n_i}{n} \log_2 (\frac{n_i}{n}) = -\sum_{i=1}^{c} p_i \log_2 p_i$$

ID3学习算法

- 分为两类，“危险”的类有6个，“不危险”的类有6个，则划分为两类的信息熵为：

$$E(S) = -\frac{6}{12}\log_2(\frac{6}{12}) - \frac{6}{12}\log_2(\frac{6}{12}) = \frac{1}{2} + \frac{1}{2} = 1$$

类别
危险
危险
不危险
不危险
危险
危险
危险
不危险
危险
不危险
不危险
不危险

•2 信息增益（Information Gain）衡量属性区分训练样例的能力:一个属性的信息增益就是由于使用这个属性分割样例而导致的熵的降低

•属性A相对样例集合S的信息增益定义：

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

ID3学习算法

- 根据“体形”取值可分为3个子树，每类划分为2类，每个子树进行划分的信息熵为：

$$E(S_1) = -\frac{0}{4}\log_2(\frac{0}{4}) - \frac{4}{4}\log_2(\frac{4}{4}) = 0 + 0 = 0$$

$$E(S_2) = -\frac{1}{4}\log_2(\frac{1}{4}) - \frac{3}{4}\log_2(\frac{3}{4}) = 0.5 + 0.0637 = 0.5637$$

$$E(S_3) = -\frac{1}{4}\log_2(\frac{1}{4}) - \frac{3}{4}\log_2(\frac{3}{4}) = 0.5 + 0.0637 = 0.5637$$

$$E(X,S) = \frac{4}{12}E(S_1) + \frac{4}{12}E(S_2) + \frac{4}{12}E(S_3)$$

$$= \frac{4}{12}\times 0 + \frac{4}{12}\times 0.5637 + \frac{4}{12}\times 0.5637 = 0.3758$$

实例序号	颜色	体形	毛型	类别
1	黑	大	卷毛	危险
2	棕	大	光滑	危险
6	黑	大	光滑	危险
9	棕	大	卷毛	危险

实例序号	颜色	体形	毛型	类别
3	棕	中	卷毛	不危险
5	棕	中	光滑	危险
10	黑	中	卷毛	不危险
11	黑	中	光滑	不危险

实例序号	颜色	体形	毛型	类别
4	黑	小	卷毛	不危险
7	棕	小	卷毛	危险
8	棕	小	光滑	不危险
12	黑	小	光滑	不危险

ID3学习算法

按属性”体形“取值划分的信息增益为：

$$Gain(X,S) = E(S) - E(X,S) = 1 - 0.3758 = 0.6242$$

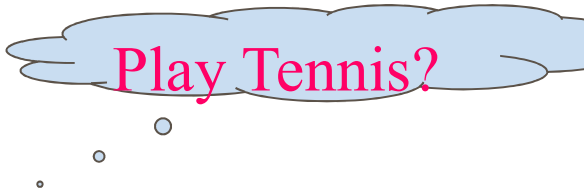
“颜色” “毛型”划分..

选取信息增益值最大的属性作为最佳属性（树根），进行分类

ID3算法

1. 决定分类属性
2. 对目前的数据表，建立一个节点N。
3. 如果数据表中的数据都属于同一类，N就是树叶，在树叶上标上所属的那一类。
4. 如果数据表中没有其他属性可以考虑，N也是树叶，按照少数服从多数的原则在树叶上标上所属类别。
5. 否则，根据平均信息期望值E或Gain值选出一个最佳属性作为节点N的测试属性A。
6. 节点属性选定以后，对于该属性的每一个值ai：
   - 从N生成一个A=ai的分支， 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，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。

ID3算法举例



Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

ID3算法举例

- 对样本分类的信息熵为：

$$E(S) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) = 0.94$$

- 以属性“outlook”为例计算信息增益

  属性“outlook”有3个取值，分别为Sunny，Overcast，Rain

Outlook	Play
Sunny	No
Sunny	No
Overcast	Yes
Rain	Yes
Rain	Yes
Rain	Yes
Overcast	No
Sunny	Yes
Sunny	No
Rain	Yes
Sunny	Yes
Overcast	Yes
Overcast	Yes
Rain	Yes
	No

ID3算法举例

Outlook	Play
Sunny	No
Sunny	No
Sunny	No
Sunny	Yes
Sunny	Yes

$$E(S_{sunny}) = -\frac{2}{5}\log_2(\frac{2}{5}) - \frac{3}{5}\log_2(\frac{3}{5}) = 0.971$$

Outlook	Play
Overcast	Yes
Overcast	Yes
Overcast	Yes
Overcast	Yes

$$E(S_{Overcast}) = -\frac{4}{4}\log_2(\frac{4}{4}) - \frac{0}{4}\log_2(\frac{0}{4}) = 0$$

$$E(S_{Rain}) = -\frac{3}{5}\log_2(\frac{3}{5}) - \frac{2}{5}\log_2(\frac{2}{5}) = 0.971$$

Outlook	Play
Rain	Yes
Rain	Yes
Rain	No
Rain	Yes
Rain	No

$$E(S, Outlook) = \frac{5}{14}E(S_{Sunny}) + \frac{4}{14}E(S_{Overcast}) + \frac{5}{14}E(S_{Rain}) = 0.694$$

ID3算法举例

- 属性”**Outlook**“的信息增益：

$$Gain(S,Outlook)=E(S)-E(S,Outlook)=0.94-0.694=0.246$$

- 同理通过计算，得**Humidity**，**Temperature**，**Wind**属性的信息增益：

$$Gain(S,Humidity)=0.151$$

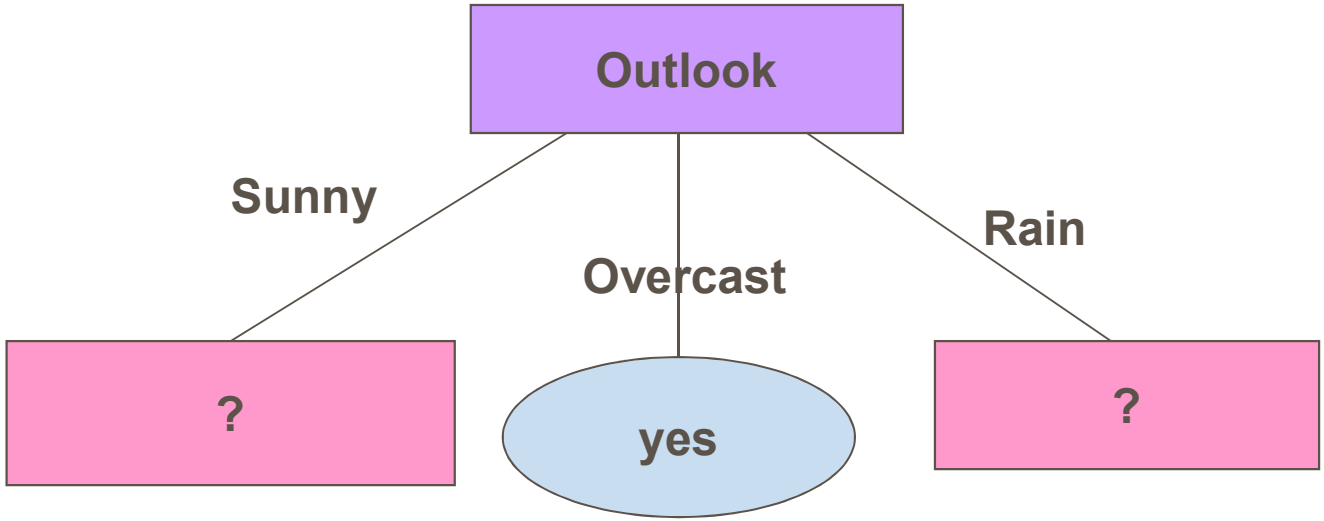
$$Gain(S,Temperature)=0.029$$

$$Gain(S,Wind)=0.048$$

通过比较，选择信息增益最大的属性”**Outlook**”作为根节点。

ID3算法举例

- 初步生成的决策树：



ID3算法举例

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes

Day	Outlook	Temperature	Humidity	Wind	Play
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D10	Rain	Mild	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

ID3算法举例

- 以outlook=“sunny”对应的节点为例继续划分。

对样本划分的信息熵：

$$E(S) = -\frac{3}{5}\log_2(\frac{3}{5}) - \frac{2}{5}\log_2(\frac{2}{5}) = 0.971$$

Temperature	Humidity	Wind	Play
Hot	High	Weak	No
Hot	High	Strong	No
Mild	High	Weak	No
Cool	Normal	Weak	Yes
Mild	Normal	Strong	Yes

ID3算法举例

- 以属性”temperature”为例计算信息增益，有3个属性值hot，mild，cool。

$$E(S_{Hot}) = -\frac{2}{2}\log_2(\frac{2}{2}) - \frac{0}{2}\log_2(\frac{0}{2}) = 0$$

$$E(S_{Mild}) = -\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2}) = 1$$

$$E(S_{Cool}) = -\frac{1}{1}\log_2(\frac{1}{1}) - \frac{0}{1}\log_2(\frac{0}{1}) = 0$$

$$E(S, Temperature) = \frac{2}{5}E(S_{Hot}) + \frac{2}{5}E(S_{Mild}) + \frac{1}{5}E(S_{Cool})$$

$$= \frac{2}{5}\times 0 + \frac{2}{5}\times 1 + \frac{1}{5}\times 0 = 0.4$$

Temperature	Play
Hot	No
Hot	No

Temperature	Play
Mild	No
Mild	Yes

Temperature	Play
Cool	Yes

ID3算法举例

- 属性“temperature”的信息增益

$$Gain(S,Temperature)=E(S)-E(S,Temperature)=0.971-0.4=0.571$$

同理通过计算，得**Humidity**，**Temperature**，**Wind**属性的信息增益：

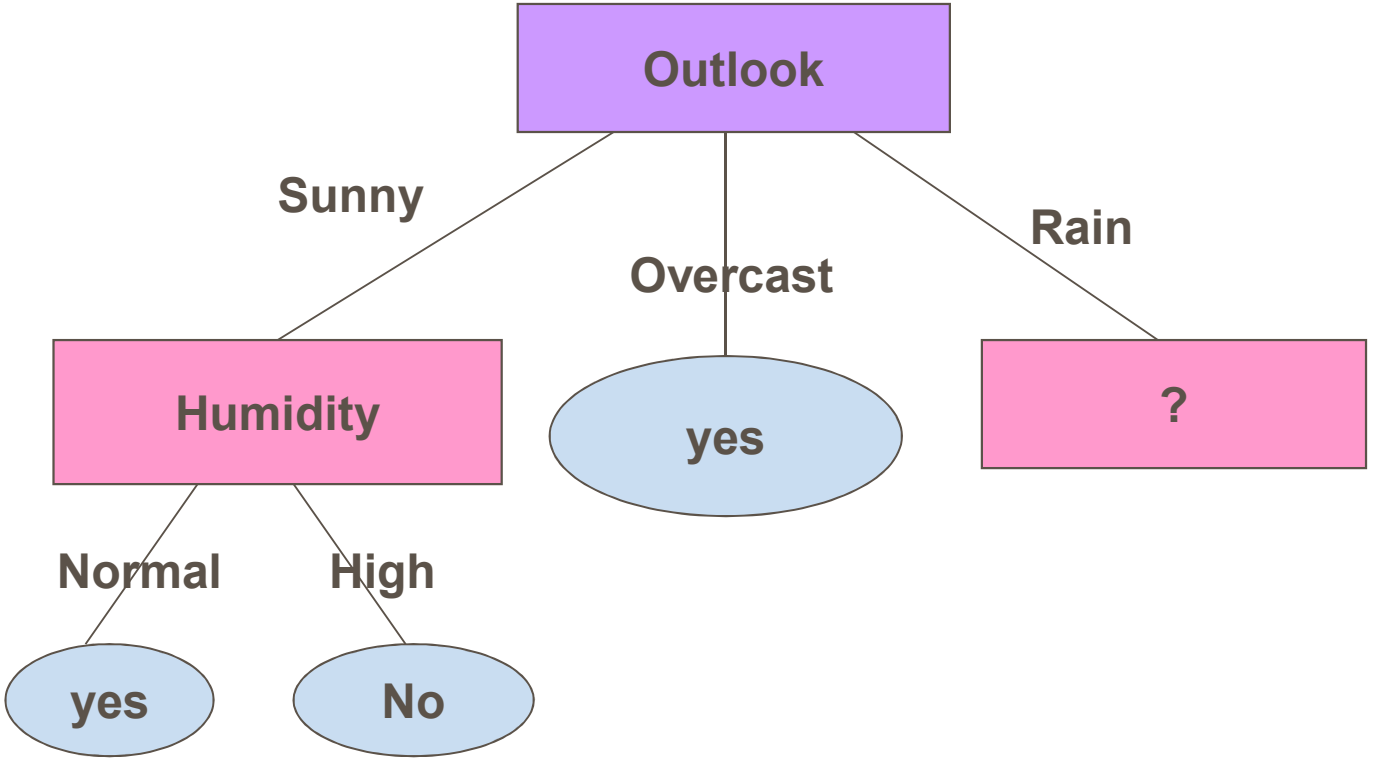
$$Gain(S,Humidity)=0.971$$

$$Gain(S,Wind)=0.02$$

通过比较，选择信息增益最大的属性”**Humidity**”作为当前节点。

ID3算法举例

- 进一步生成的决策树：



Humidity	Play
High	No
High	No
High	No
Normal	Yes
Normal	Yes

ID3算法举例

- 以“outlook=‘Rain’”对应的节点为例继续划分。
- 对样本划分的信息熵：

$$E(S) = -\frac{3}{5}\log_2(\frac{3}{5}) - \frac{2}{5}\log_2(\frac{2}{5}) = 0.971$$

Temperature	Humidity	Wind	Play
Mild	High	Weak	Yes
Cool	Normal	Weak	Yes
Cool	Normal	Strong	No
Mild	Normal	Weak	Yes
Mild	High	Strong	No

ID3算法举例

- 以属性“temperature”为例计算信息增益，有2个属性值mild，cool。

$$E(S_{Mild}) = -\frac{2}{3}\log_2(\frac{2}{3}) - \frac{1}{3}\log_2(\frac{1}{3}) = 0.918$$

$$E(S_{Cool}) = -\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2}) = 1$$

$$E(S, Temperature) = \frac{3}{5}E(S_{Mild}) + \frac{2}{5}E(S_{Cool})$$

$$= \frac{3}{5} \times 0.918 + \frac{2}{5} \times 1 = 0.951$$

Temperature	Play
Mild	Yes
Mild	Yes
Mild	No

Temperature	Play
Cool	Yes
Cool	No

ID3算法举例

- 属性“temperature”的信息增益

$$Gain(S,Temperature)=E(S)-E(S,Temperature)=0.971-0.951=0.02$$

同理通过计算，得**Humidity**，**Wind**属性的信息增益：

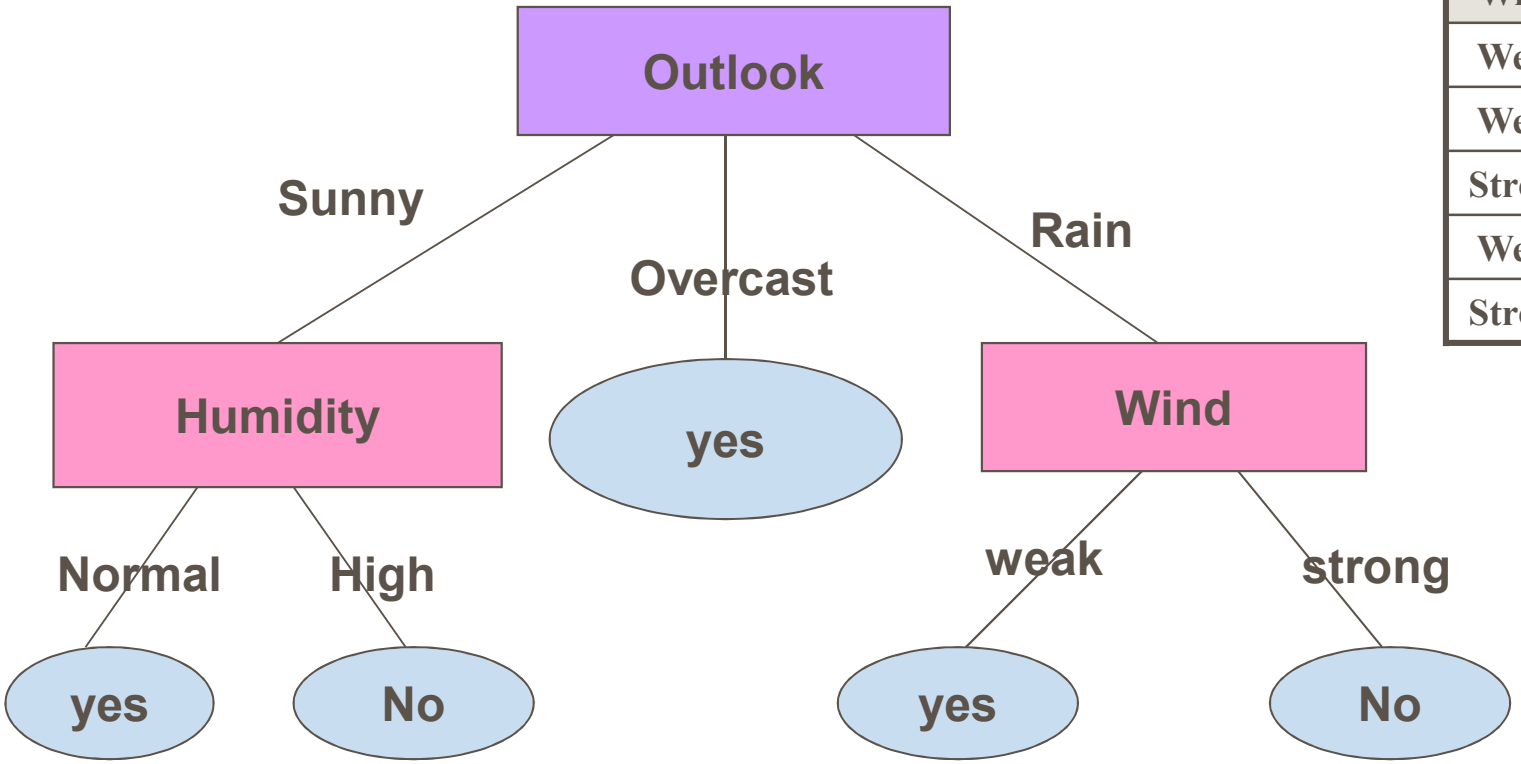
$$Gain(S,Humidity)=0.02$$

$$Gain(S,Wind)=0.971$$

通过比较，选择信息增益最大的属性”**Wind**”作为当前节点。

ID3算法举例

- 最终生成的决策树



Wind	Play
Weak	Yes
Weak	Yes
Strong	No
Weak	Yes
Strong	No

ID3算法举例

提取规则（或知识）：

- 通过对样本的学习，可以得到如下知识：

If (outlook=sunny And Humidity=Normal) Or (outlook=Overcast) Or (outlook=Rain And Wind=weak) Then play=yes

If (outlook=sunny And Humidity=high) Or (outlook=Rain And Wind=strong) Then play=No

其他决策树建立方法

- CART算法

  CART (classification and regression tree)即分类和回归树算法，它是仅有的一种通用的树生长方法。

- C4.5算法

  C4.5算法是ID3的后继和改进，也是最流行的分类树算法

内容概要

- 分类简介
- 决策树的用途
- 决策树的建立(例)
- 决策树的数据准备
- 小结

决策树的数据准备

- 数据清理Data cleaning

  删除/减少噪声（noise)，补填空缺值(missing values)

- 相关性分析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

- 数据变换Data transformation

  数据标准化（data normalization）

  数据归纳（generalize data to higher-level concepts using concept hierarchies）

  控制每个属性的可能值不超过七种（最好不超过五种）

Conclusion (杂谈)

- 这个世界上每天产生的新数据速度惊人
- Data Mining就是要从数据中挖出知识—超出人的脑力所及的知识
- 每天都有公司破产，每年都有公司崛起，公司间的竞争之一将是Data Mining的能力的竞争
- 我们每个人都自觉不自觉地为产生数据而奉献，而这些数据又成为控制我们的工具

当信息控制人类的时候，

人类就进入了信息时代

Thanks a lot!