火龙果·整理
uml.org.cn

聚类分析的含义

- 聚类（Clustering）用于发现在数据库中未知的对象类
- 聚类方法对象类划分的依据是“物以类聚”，即考察个体或数据对象间的相似性
- 在聚类之前，对象类划分的数量与类型均是未知的

火龙果·整理
uml.org.cn

欧几里德距离

- 欧几里德距离（Euclidean distance） 类似空间距离的计算。

$$d(i,j)=\sqrt{|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+\cdots+|x_{im}-x_{jm}|^2}$$

$$i\in\{1,2,\cdots,n\};\ j\in\{1,2,\cdots,n\}$$

火龙果 · 整理
uml.org.cn

分割聚类方法概述

- 分割聚类方法是一种基于原型（Prototype）的聚类方法。
- 其本质是首先从数据集中随机地选择几个对象作为聚类的原型，然后将其它对象分别分配到由原型所代表的最相似、也就是距离最近的类中。
- 分割聚类方法通过迭代控制策略对原型不断地进行调整，从而使得整个聚类得到优化。

火龙果·整理
uml.org.cn

k-means算法的思路

1.首先随机地选择k个对象代表k个类，每一个对象作为一个类的原型，根据距离原型最近的原则将其它对象分配到各个类中。

火龙果·整理
uml.org.cn

k-means算法的思路

2.以每一个类所有对象的平均值（mean）作为该类新的原型，迭代进行对象的再分配，直到没有变化为止，从而得到最终的个类。

火龙果·整理
uml.org.cn

k-means算法步骤

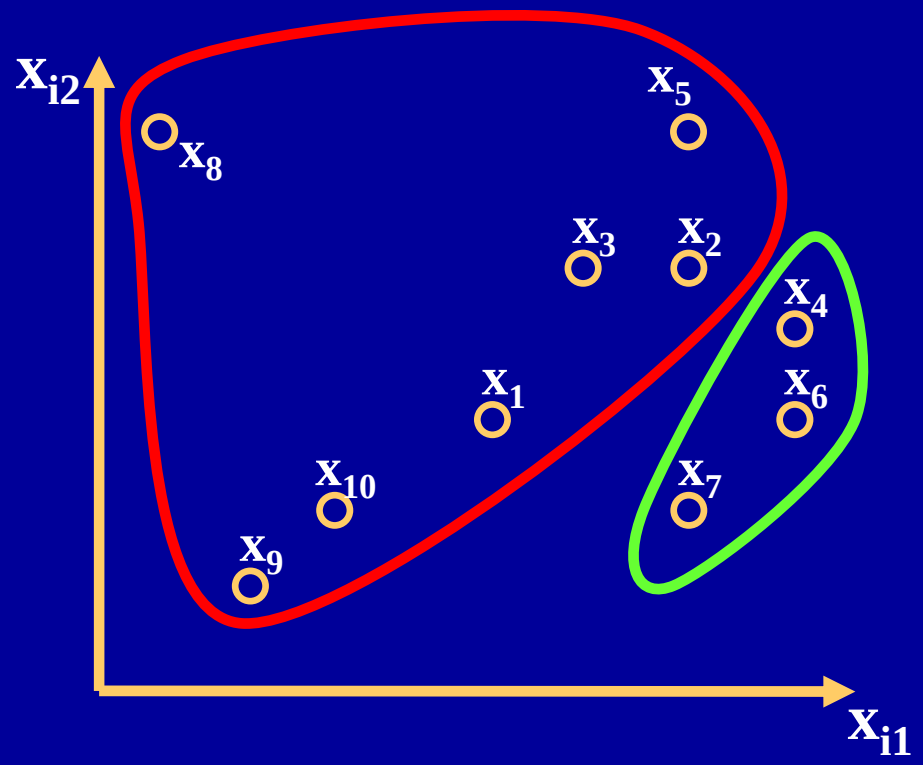
1.首先随机地选择k个对象，每一个对象作为一个类的“中心”，分别代表将分成的k个类。

2.根据距离“中心”最近的原则，寻找与各对象最为相似的类，将其它对象分配到各个相应的类中。

火龙果·整理
uml.org.cn

k-means算法例

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_{i1}	5	7	6	8	7	8	7	1	2	3
x_{i2}	4	6	6	5	8	4	3	8	2	3



聚类个数k=2

随机选择x3,x6作为中心

根据中心最近原则，
将其他对象分配到相应类中

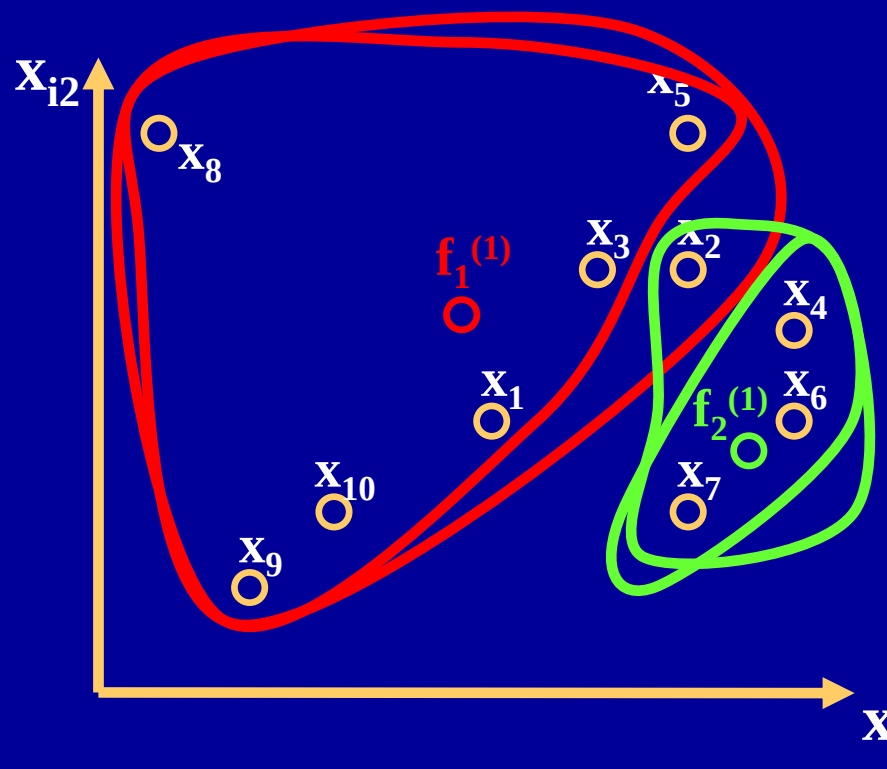
火龙果·整理
uml.org.cn

k-means算法步骤

3. 在完成对象的分配之后，针对每一个类，计算其所有对象的平均值，作为该类的新的“中心”。
4. 根据距离“中心”最近的原则，重新进行所有对象到各个相应类的分配。
5. 返回步骤（3），直到没有变化为止。

火龙果·整理
uml.org.cn

k-means算法例

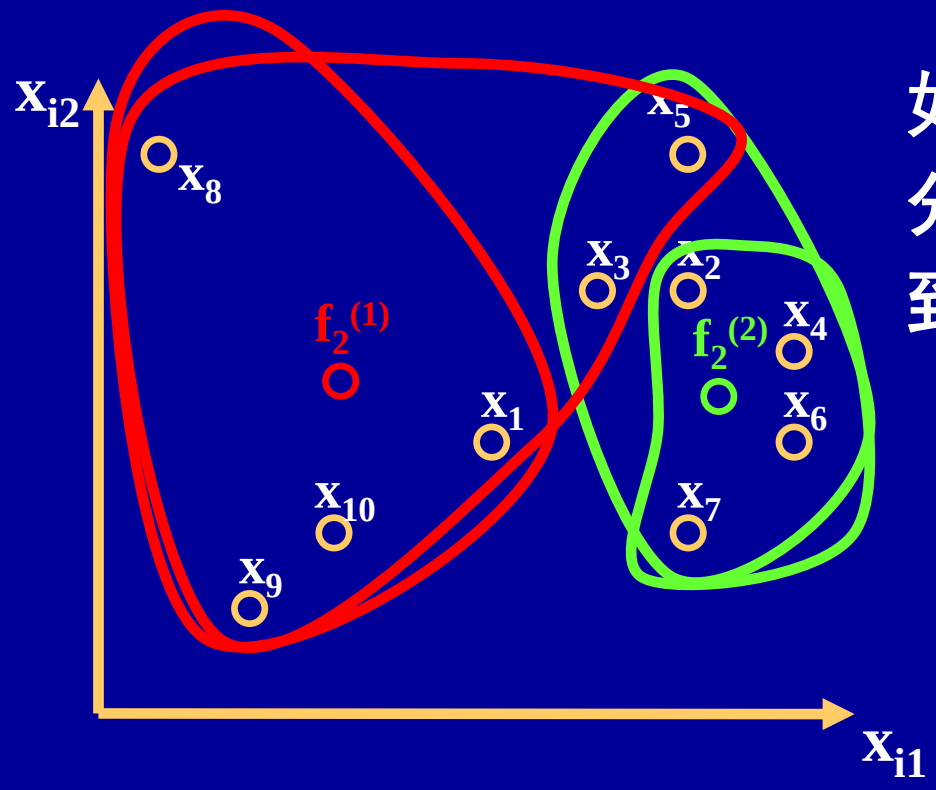


计算两个类的平均值分别为$f_1^{(1)}$、$f_2^{(1)}$。

根据距离“中心”最近的原则，重新分配所有对象。

火龙果·整理
uml.org.cn

k-means算法例



如此迭代计算，直至对象的分配不再发生变化，即可得到最终的聚类。

火龙果·整理
uml.org.cn

层次聚类方法概述

- 层次聚类方法（Hierarchical Clustering Method）是采用“自顶向下（Top-Down）”或“自底向上（Bottom-Up）”的方法在不同的层次上对对象进行分组，形成一种树形的聚类结构。
- 其包括分解型层次聚类法（自顶向下）和聚结型层次聚类法（自底向上）。

火龙果·整理
uml.org.cn

层次聚类方法思想

- 层次聚类方法按照一定的相似性判断标准，合并最相似的部分，或者分割最不相似的两个部分。
- 如果合并最相似的部分，从每一个对象作为一个类开始，逐层向上聚结，直到形成唯一的一个类 。
- 如果分割最不相似的两个部分，从所有的对象归属在唯一的一个类中开始，逐层向下分解，直到每一个对象形成一个类。

火龙果·整理
uml.org.cn

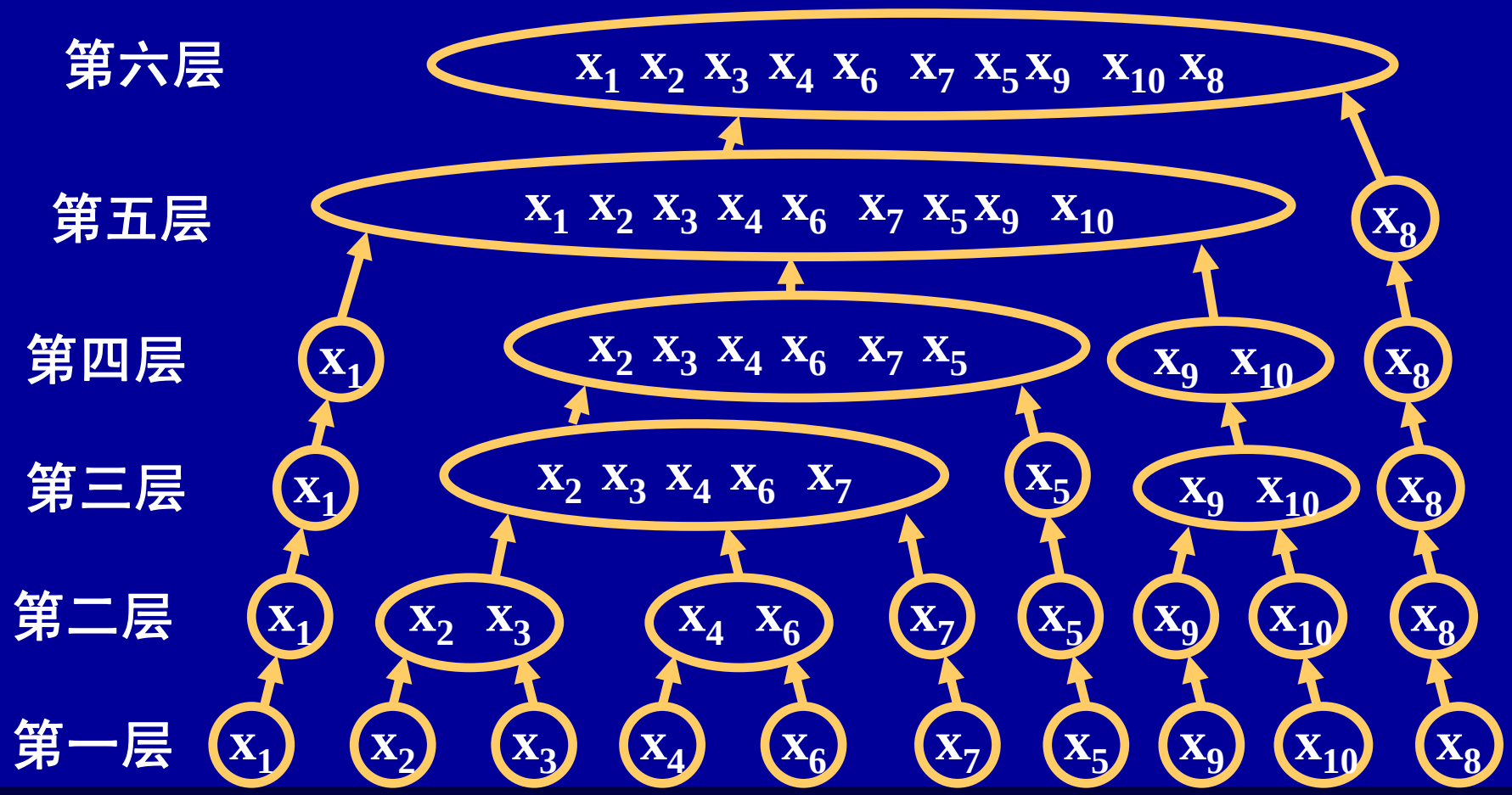
聚结层次聚类算法例

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_{i1}	5	7	6	8	7	8	7	1	2	3
x_{i2}	4	6	6	5	8	4	3	8	2	3

火龙果·整理
uml.org.cn

聚结层次聚类算法例

选择最相似、也就是d_{min}最小的两个或多个类进行聚结



火龙果 · 整理
uml.org.cn

分类的目标

- 分类的目标是通过分析训练集中的数据，对类进行准确的描述或者建立模型，然后用它对数据库中的其它数据分类或者上升为分类规则。

火龙果 · 整理
uml.org.cn

分类发现的处理过程

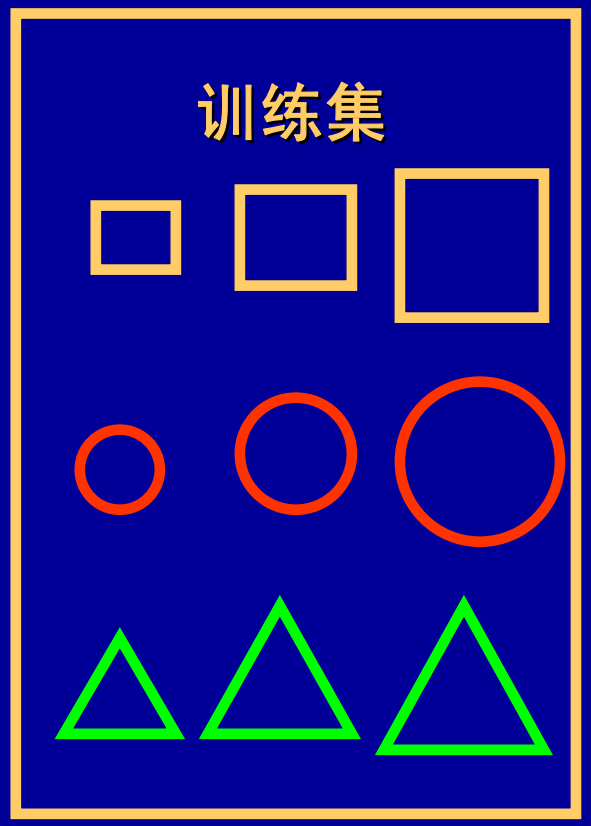
1. 分类模型的建立

   监督学习(Supervised Learning)

- 分类模型的建立是通过分析训练样本数据总结出一般性的分类规则，建立分类模型。
- 分类模型以分类规则、决策树或数学公式的形式给出。

火龙果·整理
uml.org.cn

分类发现的处理过程



分析

矩形是黄色
圆形是红色
三角是绿色

火龙果·整理
uml.org.cn

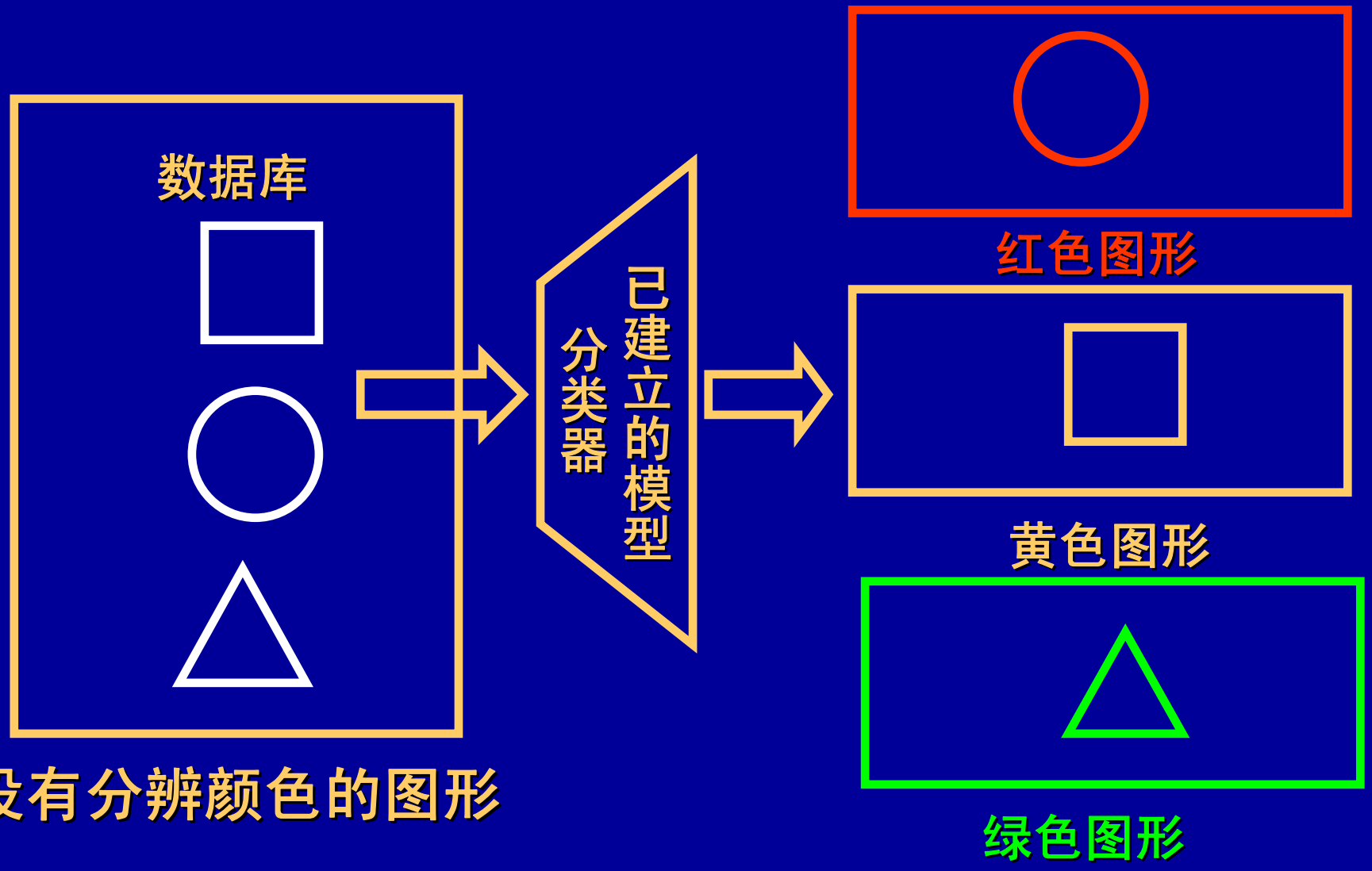
分类发现的处理过程

2. 分类模型的应用

- 在对建立的分类模型进行应用前，需要对建立的分类模型进行评估，在确保分类模型的准确性及精确度的情况下，才能运用该分类模型对未知其类别的数据样本进行分类处理。

火龙果·整理
uml.org.cn

分类发现的含义



火龙果 · 整理
uml.org.cn

分类发现的主要方法

1. 基于决策树模型的数据分类

   ——ID3算法

2. 基于统计模型的数据分类

   ——贝叶斯分类

3. 基于神经网络的数据分类

火龙果 · 整理
uml.org.cn

决策树生成过程

1. 用户根据实际需求以及所处理数据的特性，选择类别标识属性和决策树的决策属性集。

火龙果·整理
uml.org.cn

决策树生成过程例

销售时间	销售地区	销售产品	客户行业	销售渠道	利润	行数
2000年	华中	薄板	冶金	直销	高	200
200			冶金	直销	高	510
200			机电	分销	中	300
200	华南	线材	电力	直销	低	50
2000年	华东	线材	电力	分销	低	600
…	…	…	…	…	…	…
2000年	华中	薄板	电力	直销	中	80
2000年	华中	薄板	电力	直销	高	320

选择类别标示属性

选择决策属性集

火龙果 · 整理
uml.org.cn

决策树生成过程

2. 在决策属性集中选择最有分类标识能力的属性作为决策树的当前决策节点。

How?

3. 根据当前决策节点属性取值的不同，将训练样本数据集划分为若干子集。

火龙果·整理
uml.org.cn

决策树生成过程例

销售时间	销售地区	销售产品	客户行业	销售渠道	利润	行数
2000年	华中	薄板	冶金	直销	高	200
						10
						00
2000年	华南	线材	电力	直销	低	50
2000年	华东	线材	电力	分销	低	600
…	…	…	…	…	…	…
2000年	华中	薄板	电力	直销	中	80
2000年	华中	薄板	电力	直销	高	320

根据销售渠道的取值，将训练集分为两个子集

——————直销和分销

火龙果·整理
uml.org.cn

决策树生成过程例

可得：

火龙果·整理
uml.org.cn

决策树生成过程

4. 针对上一步中得到的每一个子集，重复进行上述的2、3两个步骤，直到最后的子集符合结束的三个条件之一。

火龙果·整理
uml.org.cn

三个条件

1. 子集中的所有元组都属于同一类；

2. 该子集是遍历了所有决策属性得到的；

3. 子集中的所有剩余决策属性取值完全相同，已不能根据这些决策属性进一步进行子集划分。

火龙果·整理
uml.org.cn

决策树生成过程

5. 根据符合条件的不同，生成叶子节点。

- 对满足“条件一”所产生的叶子节点，直接根据该子集的元组所属类别进行类别标识。
- 满足步骤“条件二”或“条件三”所产生的叶子节点，选取子集所含元组的代表性类别特征进行类别标识。

火龙果·整理
uml.org.cn

决策树生成过程例

在“销售渠道”为“分销”的分枝中进行同样的属性选择处理

销售时间	销售地区	销售产品	客户行业	销售渠道	利润	行数
2000年	华中	圆钢	机电	分销	中	300
…	…	…	…	…	…	…
2000年	华东	线材	电力	分销	低	600

火龙果·整理
uml.org.cn

决策树生成过程例

在“销售渠道”为“直销”的分枝中进行同样的属性选择处理

“客户行业”为“电力”，“销售地区”为“华中”这一分枝中，类别标识属性有两个不同的取值。

但已经没有可以继续分解的决策属性了。（第三种情况）

选择具有最大可能的类别标识属性值“高”作为其类别标识属性

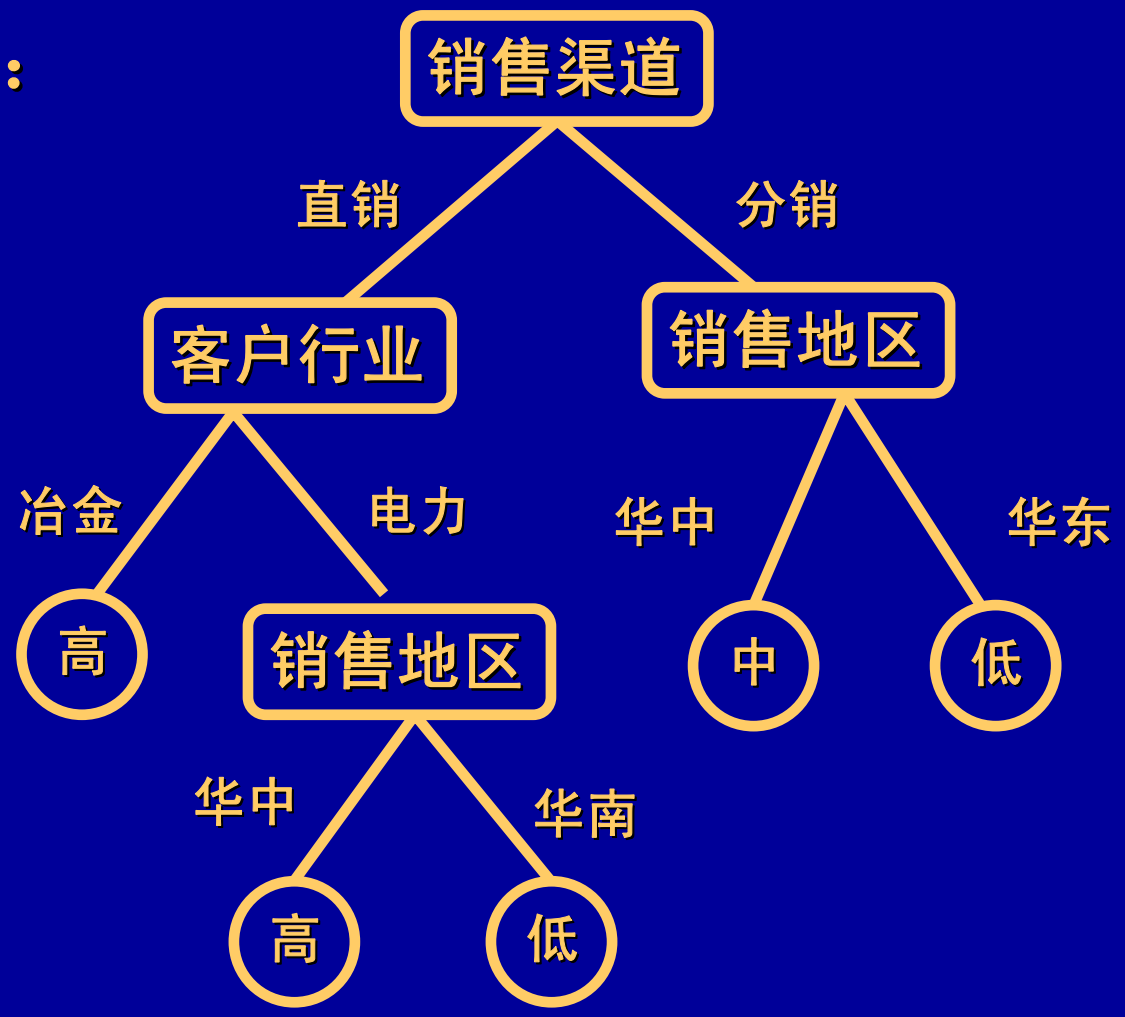
均属于“利润”为“高”的类，故该分枝终止进一步划分。（第一个条件）

道	利润	行数
消	高	200
…	…	…
直销	中	~~[illegible]~~
直销	高	320

火龙果·整理
uml.org.cn

决策树生成过程例

由此可得：



火龙果 · 整理
uml.org.cn

决策树剪枝

- 有决策树得到的初步规则中，有一些预测规则准确性较低，因此需要对上述得到的决策树进一步处理，这个进一步处理的过程由“剪枝”过程完成。

火龙果 · 整理
uml.org.cn

决策树剪枝

- 主要是采用新的样本数据集（称为测试数据集）中的数据检验决策树生成过程中产生的初步规则，将那些影响预测准确性的分枝剪除。

火龙果·整理
uml.org.cn

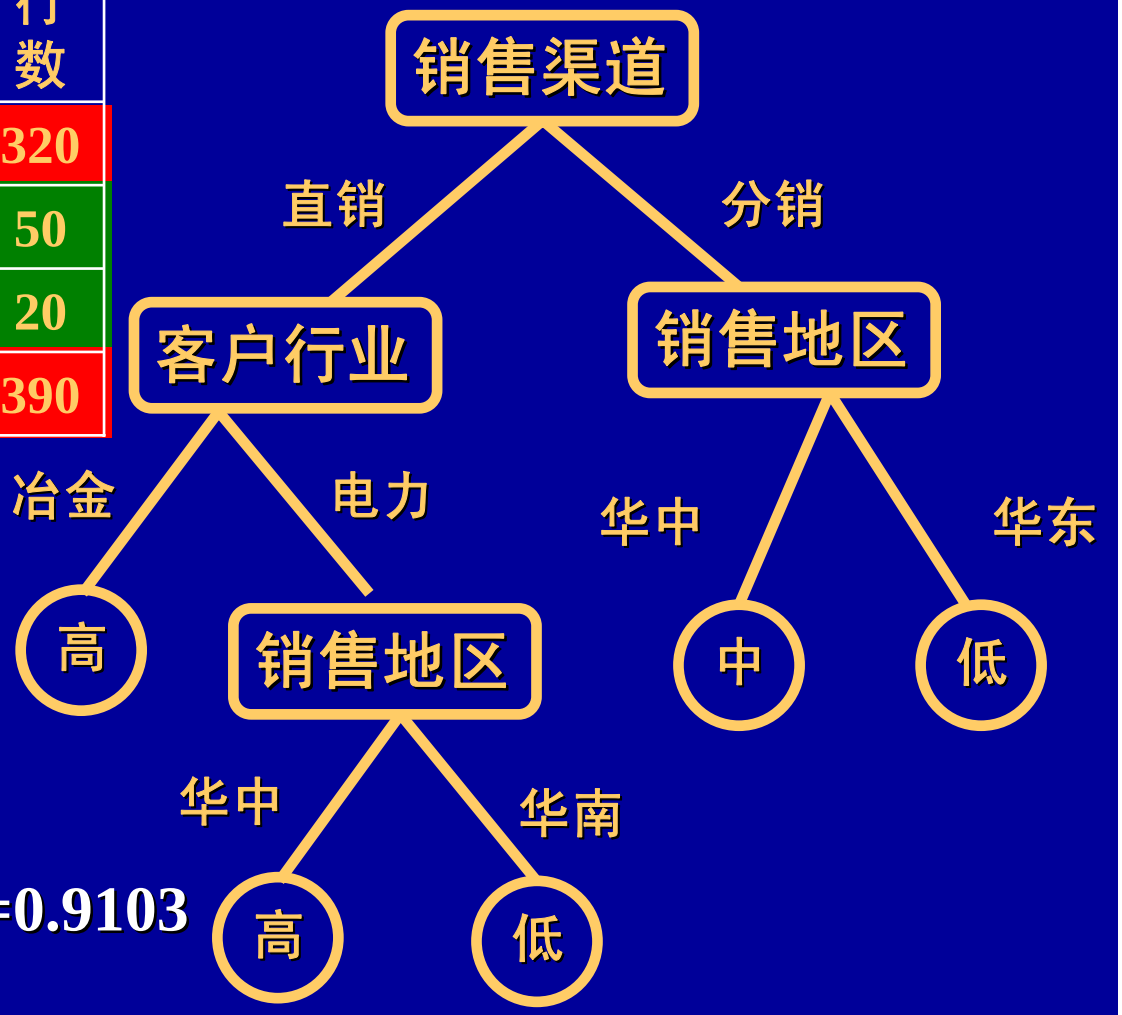
测试数据集

销售地区	客户行业	销售渠道	利润	行数	说明
华中	冶金	直销	高	320	这些元组符合：“销售渠道”为“直销”，“客户行业”为“冶金”的分枝（第一组）
华中	冶金	直销	中	50	
华中	冶金	直销	低	20	
华东	冶金	直销	高	390	
华南	电力	直销	高	20	这些元组符合：“销售渠道”为“直销”，“行业”为“电力”，“地区”为“华南”的分枝（第二组）
华南	电力	直销	低	50	
华南	电力	直销	中	150	
华中	机电	分销	中	300	略
…	…	…	…	…	…
华中	电力	直销	高	320	略

火龙果·整理
uml.org.cn

第一组测试数据

销售地区	客户行业	销售渠道	利润	行数
华中	冶金	直销	高	320
华中	冶金	直销	中	50
华中	冶金	直销	低	20
华东	冶金	直销	高	390



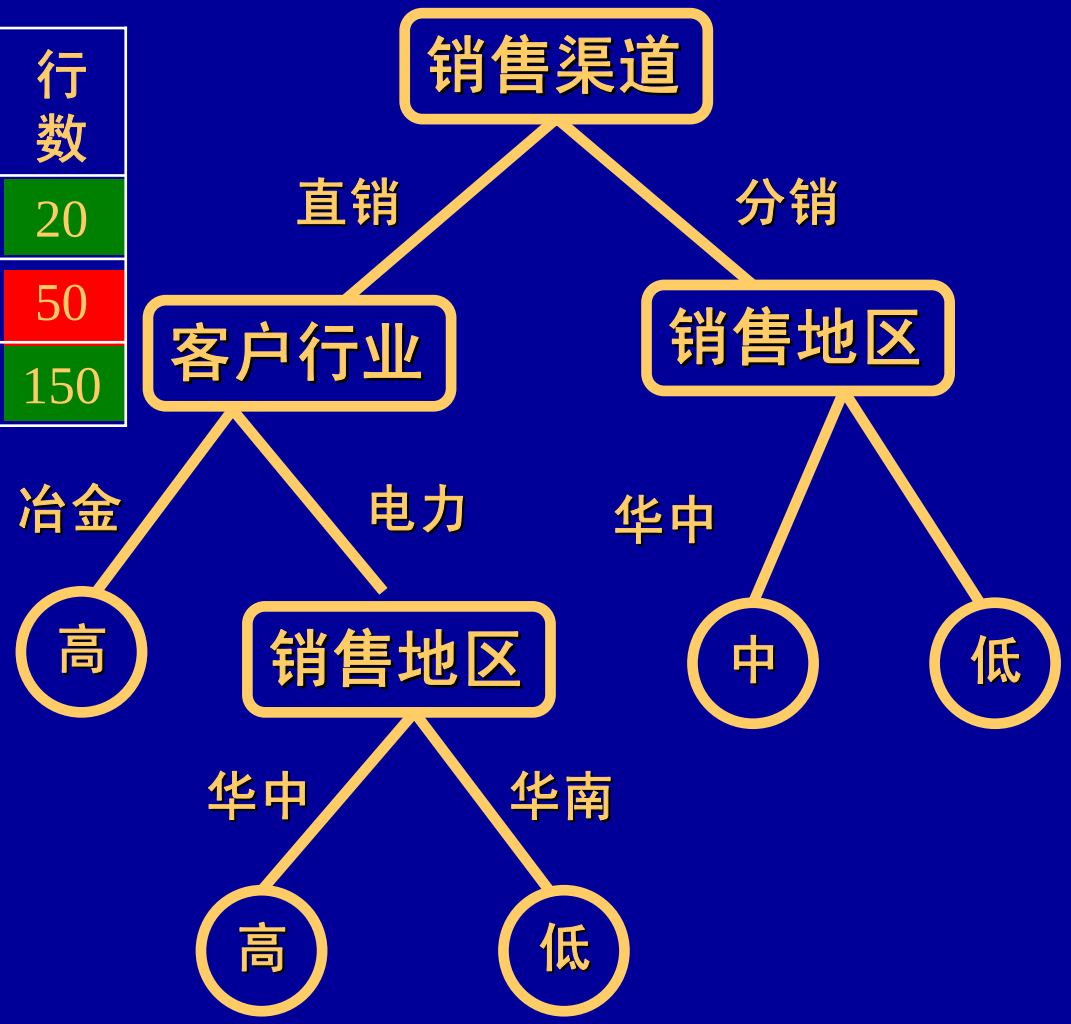
其准确率为

(320+390)/(320+50+20+390)=0.9103

火龙果·整理
uml.org.cn

第一组测试数据

销售地区	客户行业	销售渠道	利润	行数
华南	电力	直销	高	20
华南	电力	直销	低	50
华南	电力	直销	中	150



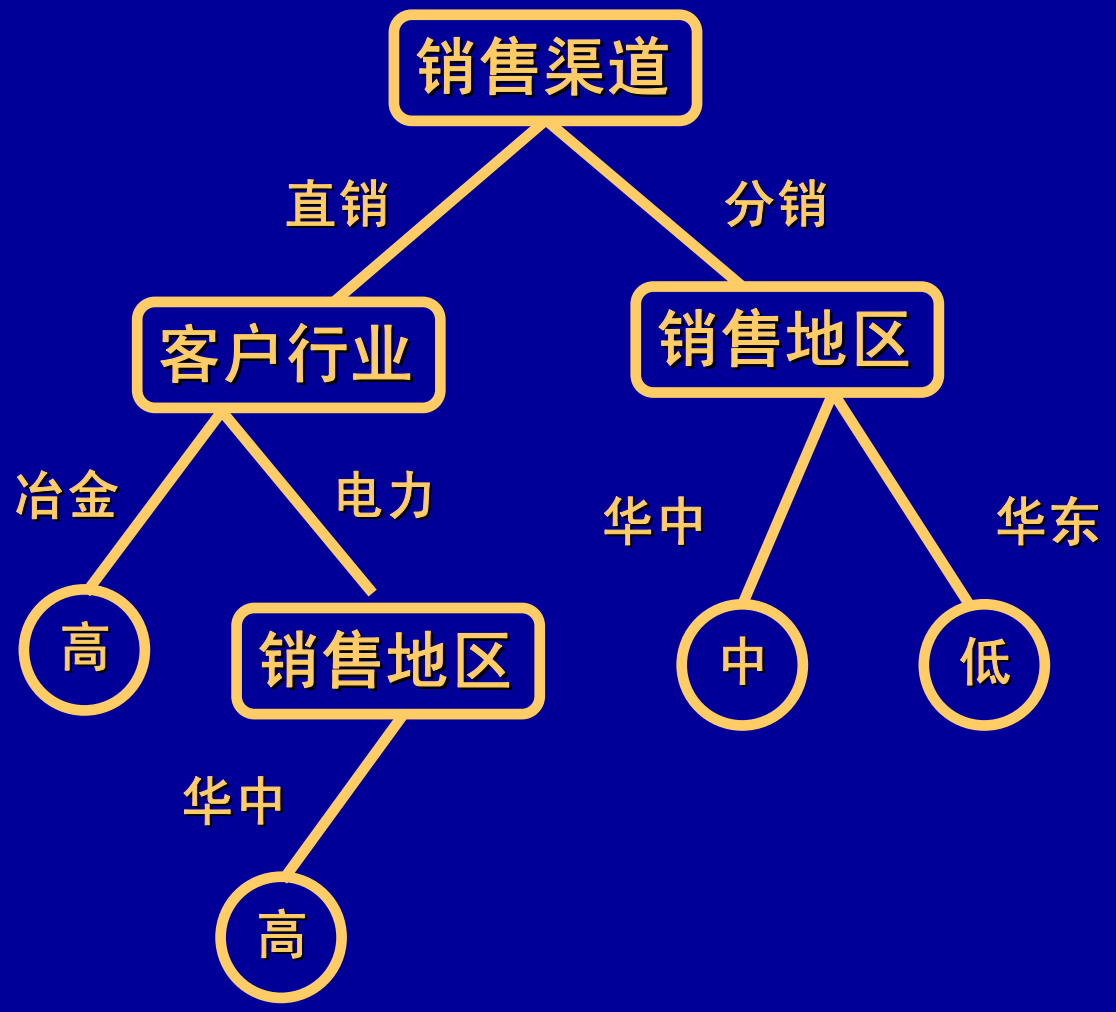
其准确率为

50/(20+50+150)=0.2272

其准确度较低，可以考虑从决策树中剪除

火龙果·整理
uml.org.cn

最终决策树



火龙果·整理
uml.org.cn

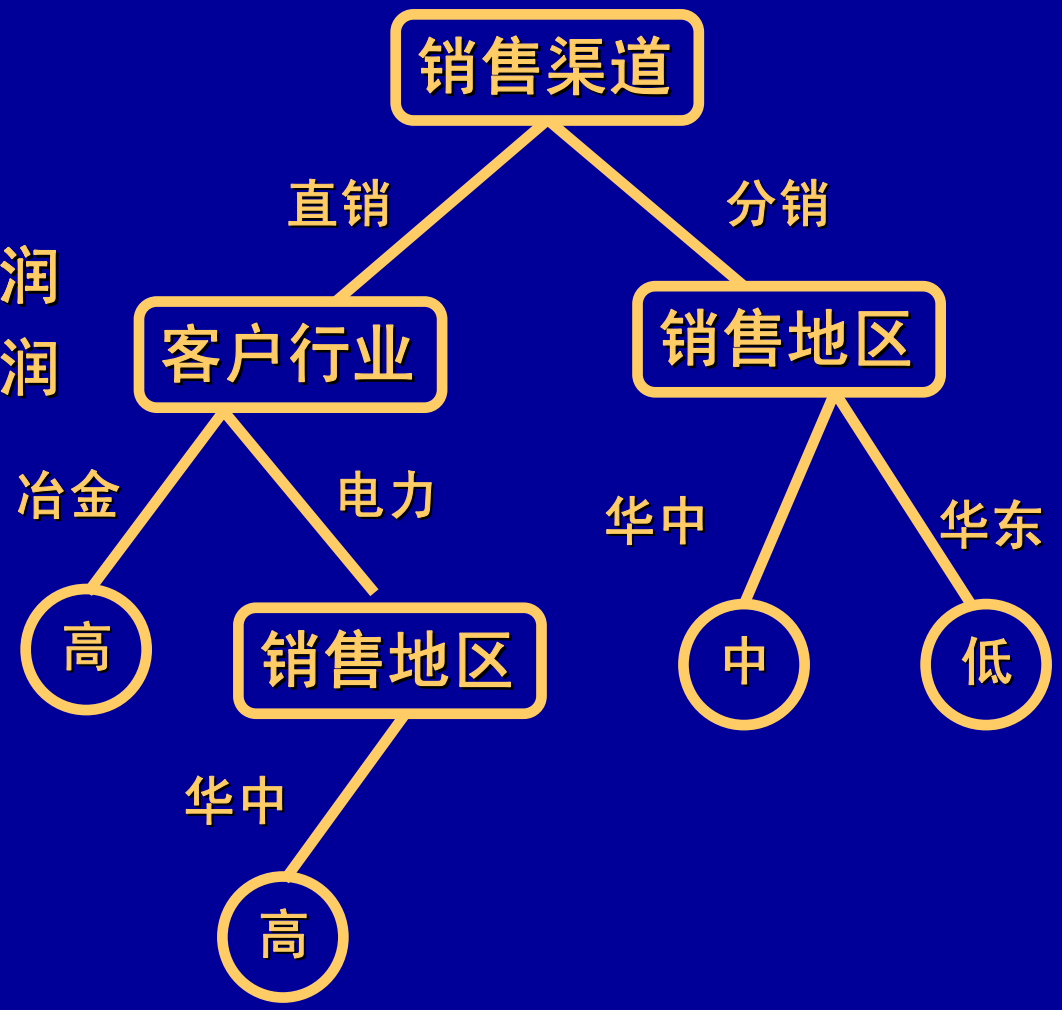
IF – THEN

规则4：

IF 销售渠道=“分销”

and 销售地区=“华东”

THEN 销售业务属于“低”利润



火龙果·整理
uml.org.cn

贝叶斯原理

- X为未知其类标识的训练样本数据；
- H表示作出的一些假设(例如，假设训练样本数据X属于某一特定类C)；
- 我们想得到P(H|X)，即该假设成立的可能性。
- P(H|X)被称为假设H在训练样本数据X的基础上的后验概率。

火龙果 · 整理
uml.org.cn

简单贝叶斯分类例

部门	职位	年龄	工资
销售部	高级	31…40	41K…55K C2
销售部	低级	21…30	26K…40K C3
销售部	低级	31…40	26K…40K C3
系统部	低级	21…30	41K…55K C2
系统部	高级	31…40	56K…70K C1
系统部	低级	21…30	41K…55K C2
系统部	高级	41…50	56K…70K C1
市场部	高级	31…40	41K…55K C2
市场部	低级	31…40	41K…55K C2
秘书处	高级	41…50	26K…40K C3
秘书处	低级	21…30	26K…40K C3

X= (部门 = '系统部'，职位 = '高级'，年龄 = '21…30')

火龙果·整理
uml.org.cn

简单贝叶斯分类例

1. 计算$P(c_i)$。

部门	职位	年龄	工资
		31…40	41K…55K C2
		21…30	26K…40K C3
		31…40	26K…40K C3
		21…30	41K…55K C2
		31…40	56K…70K C1
		21…30	41K…55K C2
系统部	高级	41…50	56K…70K C1
市场部	高级	31…40	41K…55K C2
市场部	低级	31…40	41K…55K C2
秘书处	高级	41…50	26K…40K C3
秘书处	低级	21…30	26K…40K C3

$P(c_1) = 2/11 = 0.1818$

$P(c_2) = 5/11 = 0.4545$

$P(c_3) = 4/11 = 0.3636$

11个元组

火龙果·整理
uml.org.cn

简单贝叶斯分类例

2. 计算$P(X|c_1)$。

当部门为系统部时，
有两个工资位C_1的元组

部门	职位	年龄	工资
			41K…55K C2
			26K…40K C3
			26K…40K C3
系统部	低级	21…30	41K…55K C2
系统部	高级	31…40	56K…70K C1
系统部	低级	21…30	41K…55K C2
系统部	高级	41…50	56K…70K C1
[illegible]	[illegible]	[illegible]	41K…55K C2
			41K…55K C2
			26K…40K C3
[illegible]	低级	21…30	26K…40K C3

2个元组工资为C_1

P(部门 = '系统部' | c_1) =2/2= 1

P(职位 = '高级' | $c1$) = 2/2= 1

P(年龄 = '21…30' | $c1$) = 0

火龙果 · 整理
uml.org.cn

简单贝叶斯分类例

由此可得：

$$P(X|c_1) = P(\text{部门} = \text{'系统部'} \mid c_1) \times P(\text{职位} = \text{'高级'} \mid c_1) \times P(\text{年龄} = \text{'21…30'} \mid c_1) = 0$$

$$P(X|c_1)P(c_1) = 0$$

火龙果 · 整理
uml.org.cn

简单贝叶斯分类例

同理可得：

$$P(X|c_2) = P(\text{部门} = \text{'系统部'} \mid c_2) \times P(\text{职位} = \text{'高级'} \mid c_2) \times P(\text{年龄} = \text{'21…30'} \mid c_2)$$

$$= 0.4 \times 0.4 \times 0.4$$

$$= 0.064$$

$$P(X|c_2)P(c_2) = 0.064 \times 0.4545 = 0.029$$

火龙果 · 整理
uml.org.cn

简单贝叶斯分类例

同理可得：

$$P(X|c_3) = P(\text{部门} = \text{'系统部'} \mid c_3) \times P(\text{职位} = \text{'高级'} \mid c_3) \times P(\text{年龄} = \text{'21…30'} \mid c_3)$$

$$= 0 \times 0.5 \times 0.4$$

$$= 0$$

$$P(X|c_3)P(c_3) = 0$$

火龙果 · 整理
uml.org.cn

简单贝叶斯分类例

- $P(X|c_2)P(c_2) > P(X|c_3)P(c_3) = P(X|c_1)P(c_1)$
- 因此，未知训练样本数据最有可能属于C_2类，该结果表示对于年龄在21~30岁之间、所属部门是“系统部”同时其地位属于“高级”的员工的工资水平最有可能在41K~55K之间。

火龙果 · 整理
uml.org.cn

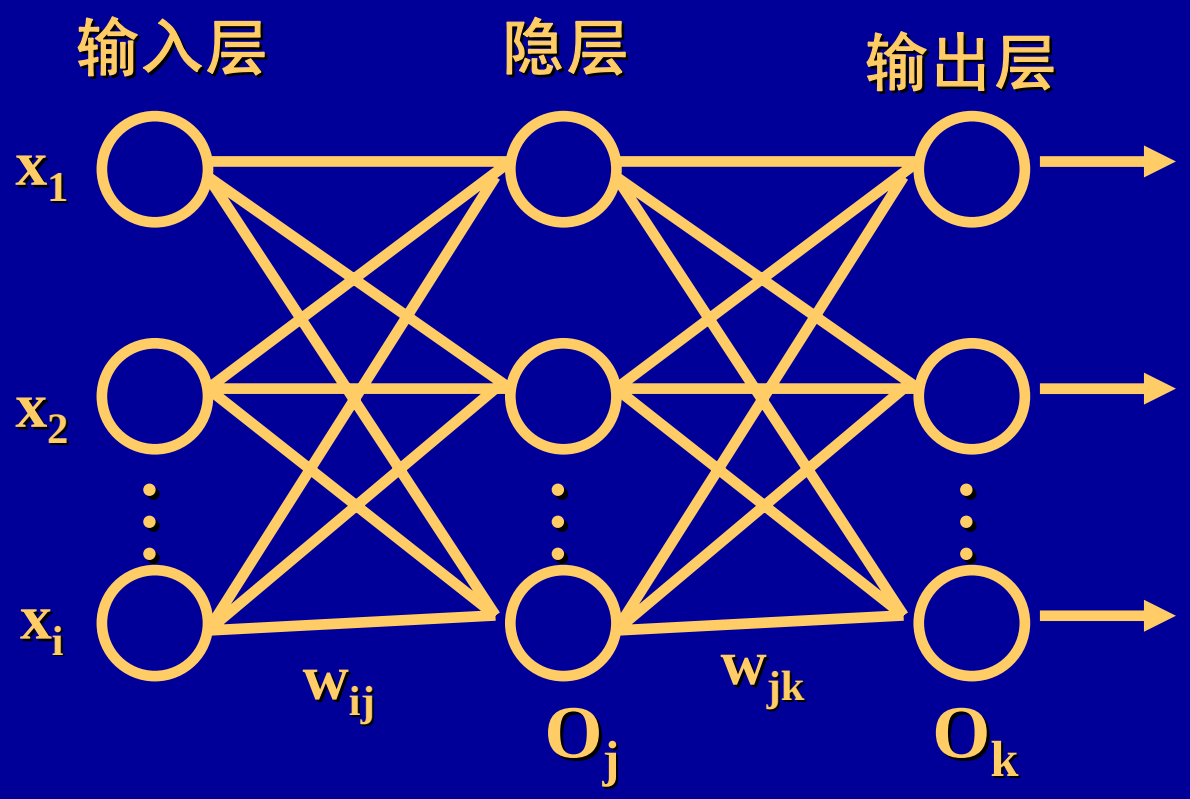
神经网络

- 神经网络由一组相关的输入及输出单元组成，对于每一个关联都有一个权重与之相对应。
- 在网络构建阶段，神经网络通过调整权重以达到能正确预测输入的训练样本数据的类别归属，由于神经网络学习是一种在各个单元之间的关联，因而又称之为关联学习。

火龙果·整理
uml.org.cn

多层前馈神经网络

输入节点由衡量训练样本的非类别标识属性组成



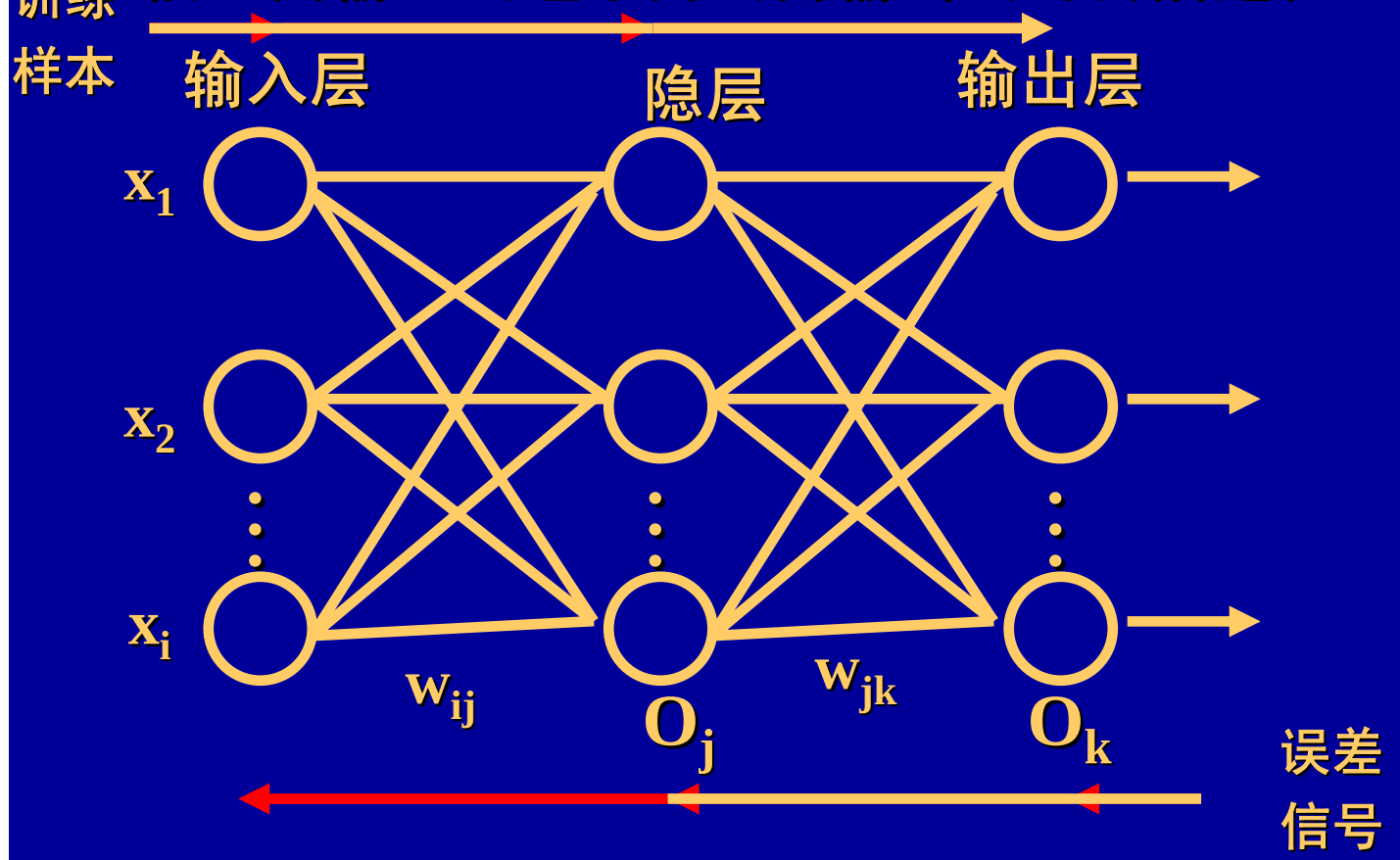
隐层及输出层中的单元统称为输出单元

由训练样本的已知类别标识属性组成，表示对给定训练样本的网络预测

火龙果·整理
uml.org.cn

BP算法

模式顺传播过程是指将一组训练样本作为输入信息由输入层经中间层向输出层的传播过程。



误差逆传播过程是指网络的希望输出与网络的实际输出之间的误差信号由输出层经中间层向输入层逐层修正连接权的过程

火龙果 · 整理
uml.org.cn

关联规则

- 关联规则是指大量数据中项集之间有趣的关联或相关联系。
- 关联规则发现最初的形式是零售商的货篮分析。

火龙果 · 整理
uml.org.cn

支持度

1. 支持度(Support)

- 设事务集D中有s%的事务同时支持项集X和Y，s%称为关联规则X⇒Y的支持度。
- 支持度s%描述了X和Y这两个项集的并集X∪Y在所有事务中出现的概率。

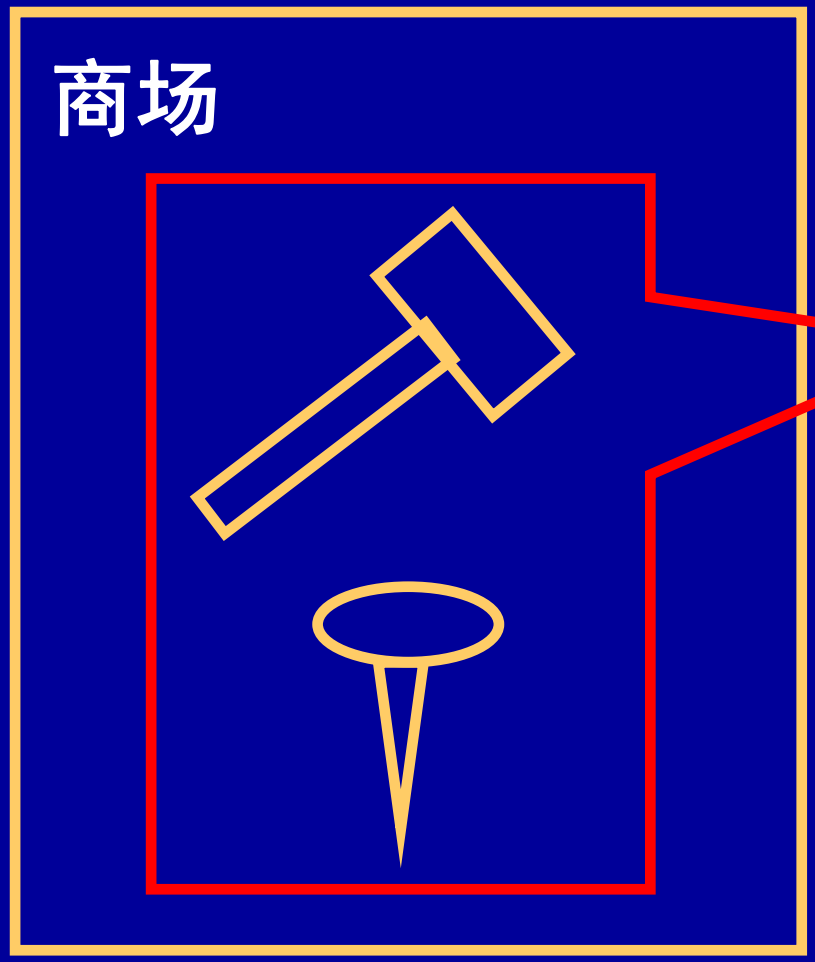
火龙果 · 整理
uml.org.cn

支持度

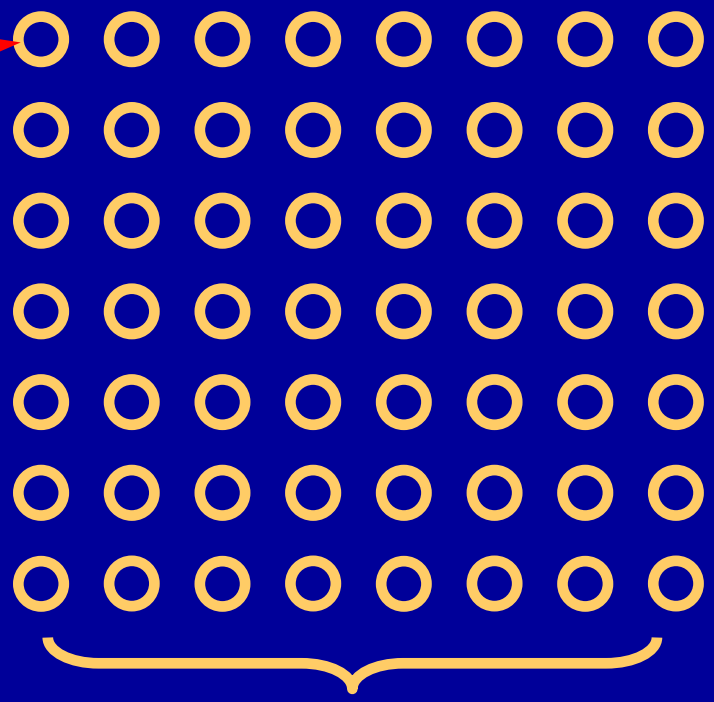
项集X ⇒ 项集Y

顾客购买锤子⇒顾客购买钉子

的关联规则的支持度为10%



100个顾客同时购买锤子和钉子



事务集D ⟵ 1000个顾客去商场购买东西

火龙果·整理
uml.org.cn

支持度

- 项集X在D中的支持度：

$$support(X) = \frac{|\{T \mid T \in D \text{ and } X \subseteq T\}|}{|D|}$$

- 关联规则X⇒Y的支持度：

$$support(X \Rightarrow Y) = support(X \cup Y) = \frac{|\{T \mid T \in D \text{ and } (X \cup Y) \subseteq T\}|}{|D|}$$

其中“|　|”表示集合中的元素个数。

火龙果 · 整理
uml.org.cn

可信度

2. 可信度 (Confidence)

- 设事务集D中支持项集X的事务中，有c%的事务同时也支持项集Y，c%称为关联规则X⇒Y的可信度。
- 可信度就是指在出现了项集X的事务中，项集Y也同时出现的概率有多大。

火龙果 · 整理
uml.org.cn

可信度

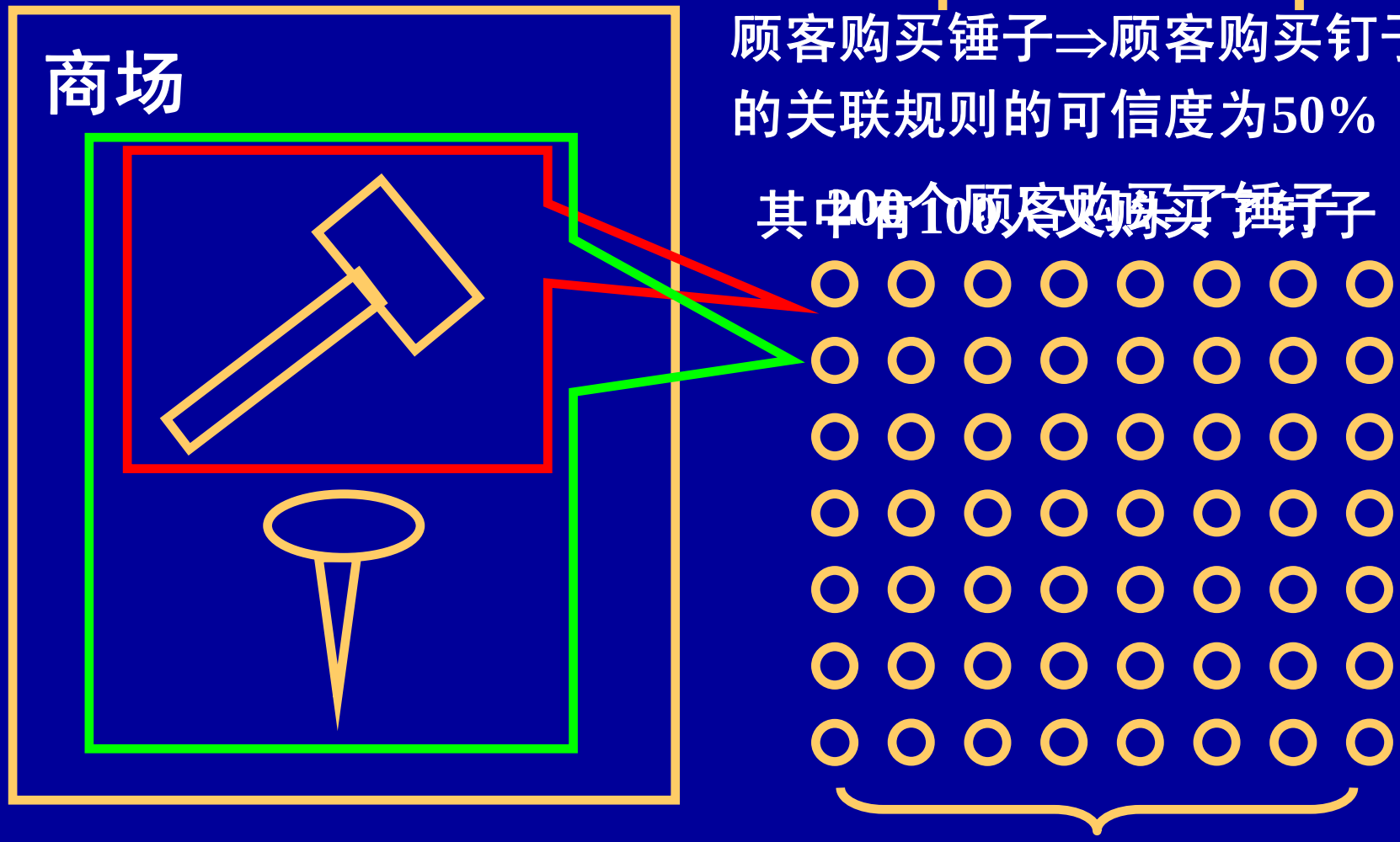
项集X ⇒ 项集Y

顾客购买锤子⇒顾客购买钉子

的关联规则的可信度为50%

其中有100人购买了锤子

200个顾客购买了钉子



事务集D ⟵ 1000个顾客去商场购买东西

火龙果 · 整理
uml.org.cn

可信度

关联规则X⇒Y的可信度

$$confidence(X \Rightarrow Y) = \frac{|\{T \mid T \in D \text{ and } (X \cup Y) \subseteq T\}|}{|\{T \mid T \in D \text{ and } X \subseteq T\}|} = \frac{support(X \cup Y)}{support(X)}$$

火龙果·整理
uml.org.cn

期望可信度

3. 期望可信度 (Expected Confidence)

- 设事务集D中有e%的事务支持项集Y，称为关联规则X⇒Y的期望可信度。
- 期望可信度描述了在没有任何条件影响时，项集Y在所有事务中出现的概率。

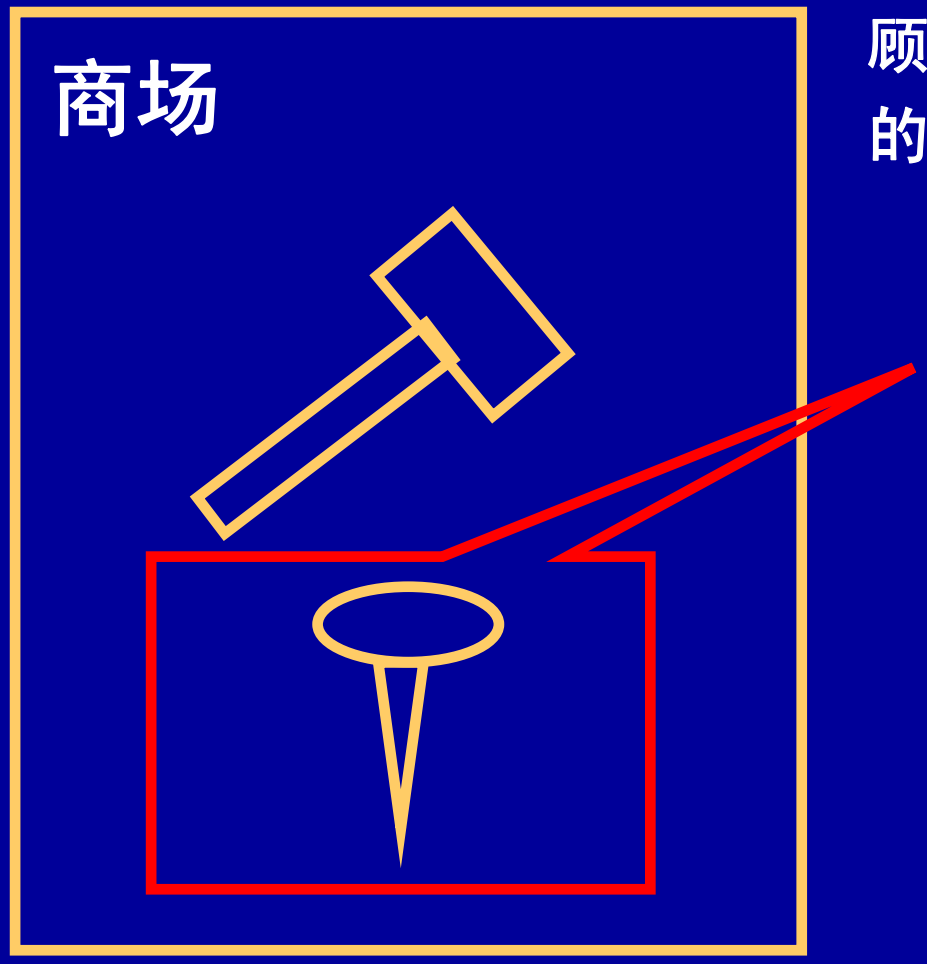
火龙果·整理
uml.org.cn

期望可信度

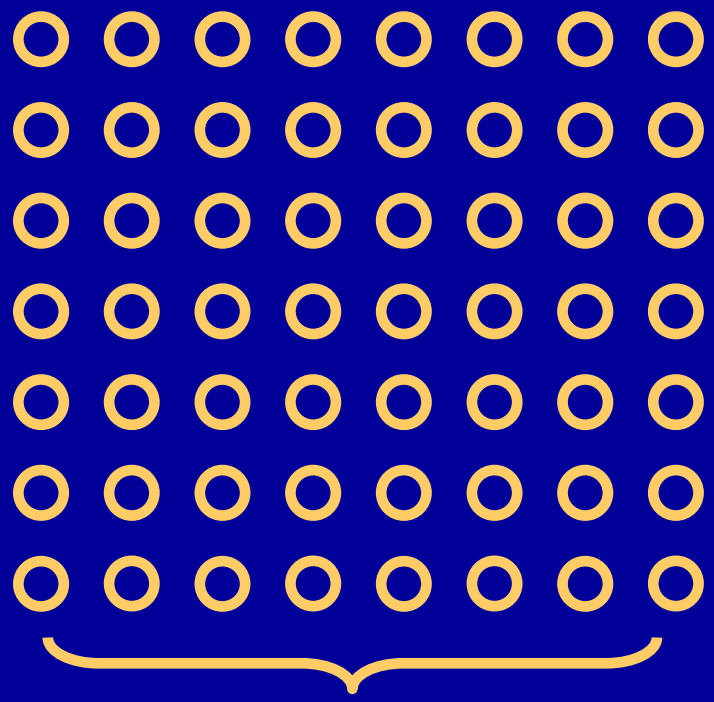
项集X ⇒ 项集Y

顾客购买锤子⇒顾客购买钉子

的关联规则的期望可信度为20%



200个顾客购买钉子



事务集D ⟵ 1000个顾客去商场购买东西

火龙果·整理
uml.org.cn

期望可信度

关联规则X⇒Y的期望可信度

$$expectedconfidence(X \Rightarrow Y) = \frac{|\{T \mid T \in D \text{ and } Y \subseteq T\}|}{|D|}$$

火龙果 · 整理
uml.org.cn

作用度

4. 作用度 (Lift)

- 作用度是可信度和期望可信度之间的比值。
- 作用度描述了项集X的出现对项集Y的出现有多大的影响。

火龙果·整理
uml.org.cn

作用度

- 项集Y在所有事务中出现的概率是期望可信度
- 而项集Y在项集X出现的事务中出现的概率是可信度
- 通过可信度和期望可信度的比值反映了在加入“项集X出现”这个条件后，项集Y的出现概率发生了多大的变化。

火龙果 · 整理
uml.org.cn

作用度

关联规则X⇒Y的作用度

$$lift(X \Rightarrow Y) = \frac{confidence(X \Rightarrow Y)}{expectedconfidence(X \Rightarrow Y)}$$

火龙果·整理
uml.org.cn

关联规则四个参数

名称	描述	公式
支持度	项集X和Y同时出现的概率	$P(X \cup Y)$
可信度	在项集X出现的前提下，Y出现的概率	$P(Y\|X)$
期望可信度	项集Y出现的概率	$P(Y)$
作用度	可信度与期望可信度的比值	$P(Y\|X)/P(Y)$

火龙果·整理
uml.org.cn

关联规则四个参数意义

- 支持度是对关联规则重要性的衡量
- 可信度是对关联规则的准确度的衡量
- 期望可信度描述了在没有项集的作用下，项集本身的支持度
- 作用度描述了项集对项集的影响力的大小，有用的关联规则的作用度都应该大于1

火龙果·整理
uml.org.cn

强规则（Strong Rules）

- 最小支持度（minsup）

  用户规定的关联规则必须满足的最小支持度。
- 最小可信度（minconf）

  用户规定的关联规则必须满足的最小可信度。
- 一般称满足最小支持度和可信度要求的规则为强规则。

火龙果·整理
uml.org.cn

Apriori算法举例

输入：

事务数据库D

TID	项列表
T1	1，2，5
T2	1，2
T3	2，4
T4	1，2，4
T5	1，3
T6	1，2，3，5
T7	1，2，3
T8	2，5
T9	2，3，4
T10	3，4

最小支持度minsup=20%，从而最小支持计数为10×20%=2。

火龙果 · 整理
uml.org.cn

Apriori算法举例

1. 初始化。
2. 寻找频繁1-项集。

C_1

项集	支持计数
{1}	6
{2}	8
{3}	5
{4}	4
{5}	3

与最小支持计数比较

计算支持计数

TID	项列表
T1	1，2，5
T2	1，2
T3	2，4
T4	1，2，4
T5	1，3
T6	1，2，3，5
T7	1，2，3
T8	2，5
T9	2，3，4
T10	3，4

火龙果 · 整理
uml.org.cn

Apriori算法举例

3. 寻找频繁2-项集

连接产生候选2-项集的集合

与最小支持数相比

C_2

项集	支持计数
{1，2}	5
{1，3}	1
{1，4}	1
{1，5}	2
{2，3}	3
{2，4}	3
{2，5}	3
{3，4}	2
{3，5}	1
{4，5}	0

L_1

项集	支持计数
{1}	6
{2}	8
{3}	5
{4}	4
{5}	3

TID	项列表
T1	1，2，5
T2	1，2
T3	2，4
T4	1，2，4
T5	1，3
T6	1，2，3，5
T7	1，2，3
T8	2，5
T9	2，3，4
T10	3，4

火龙果 · 整理
uml.org.cn

Apriori算法举例

4. 寻找频繁3-项集

L_3

项集	支持计数
{1,2,3}	2
{1,2,5}	2
{1,3,5}	
{2,3,4}	1
{2,3,5}	
{2,4,5}	
{3，4}	

使用L 产生候选3-项集的集合

删除在D中支持计数相比

于最小支持计数相比

TID	项列表
T1	1，2，5
T2	1，2
T3	2，4
T4	1，2，4
T5	1，3
T6	1，2，3，5
T7	1，2，3
T8	2，5
T9	2，3，4
T10	3，4

火龙果·整理
uml.org.cn

Apriori算法举例

5. 寻找频繁4-项集

删除存在不频繁子集的项集

找出了所有频繁集。

C_4

L_3

项集
{1，2，3，5}
{1，2，5}

火龙果·整理
uml.org.cn

提取关联规则过程

1. 对于每个频繁集l，求出其所有非空子集。

2. 对于频繁集l的每个非空子集s，计算规则s⇒(l-s)的可信度，即

$$confidence = \frac{l\text{的支持计数}}{s\text{的支持计数}}$$

若*confidence*≥*minconf*则产生关联规则s⇒(l-s)。

火龙果·整理
uml.org.cn

提取关联规例

以频繁集{1，2，3}为例

1.生成非空子集　2.计算s⇒(l-s)可信度　3.和最小可信度65%比较

$1 \wedge 3 \Rightarrow 2$

$2 \wedge 3 \Rightarrow 1$

$$confidence = \frac{l\text{的支持计数}}{s\text{的支持计数}}$$