Machine Learning / Data Mining

Work based on the Crisp-DM

商業研究部

XSGames

XSGames商業研究部

ML/DM落地流程：Crisp-DM Model

XSGames

- CRISP-DM为90年代由SIG组织（当时）提出，已经成为被业界广泛认可的数据挖掘流程。

1.业务理解(business understanding)

确定目标、明确分析需求

2.数据理解（data understanding）

收集原始数据、描述数据、探索数据、检验数据质量

3.数据准备(data preparation)

选择数据、清洗数据、构造数据、整合数据、格式化数据

4.建立模型(modeling)

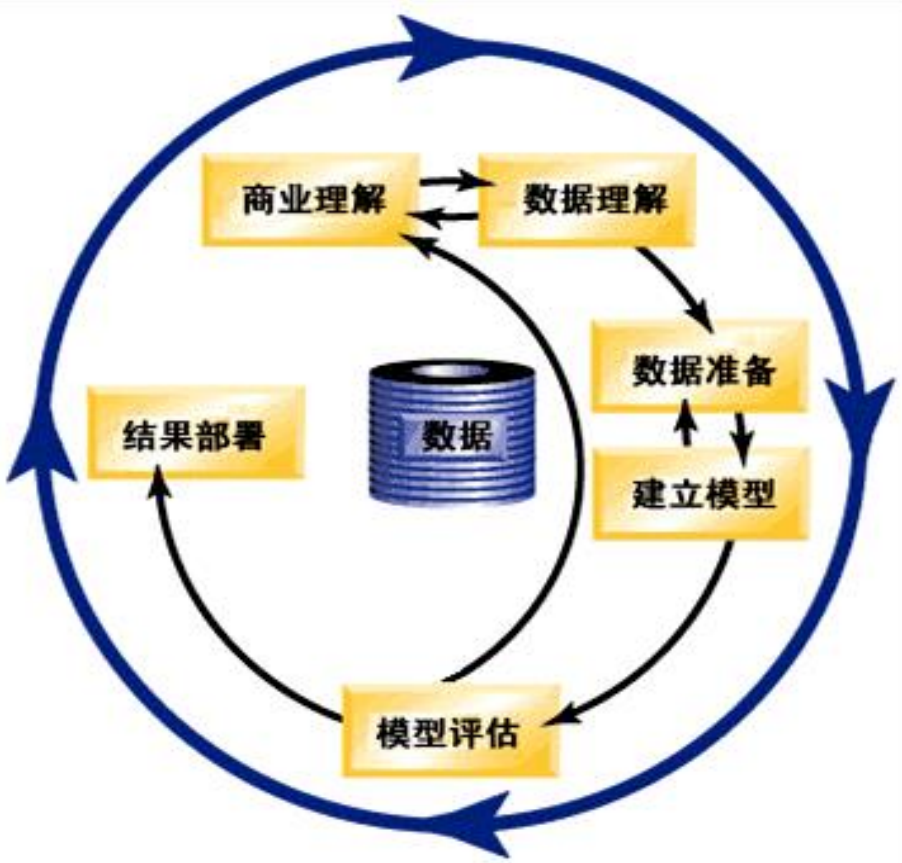
选择建模技术、参数调优、生成测试计划、构建模型

5.评估模型(evaluation)

对模型进行较为全面的评价，评价结果、重审过程

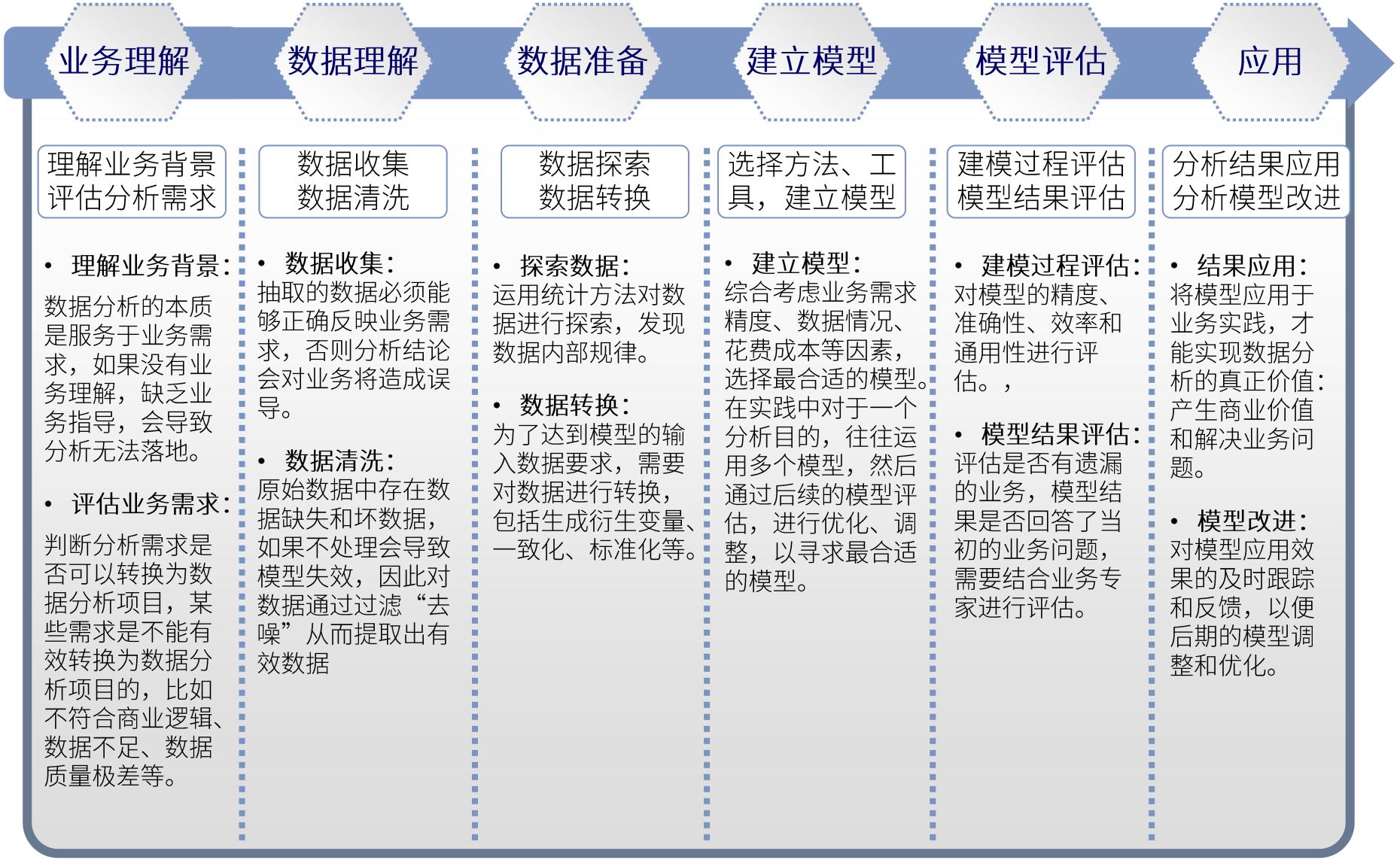
6.部署(deployment)

分析结果应用



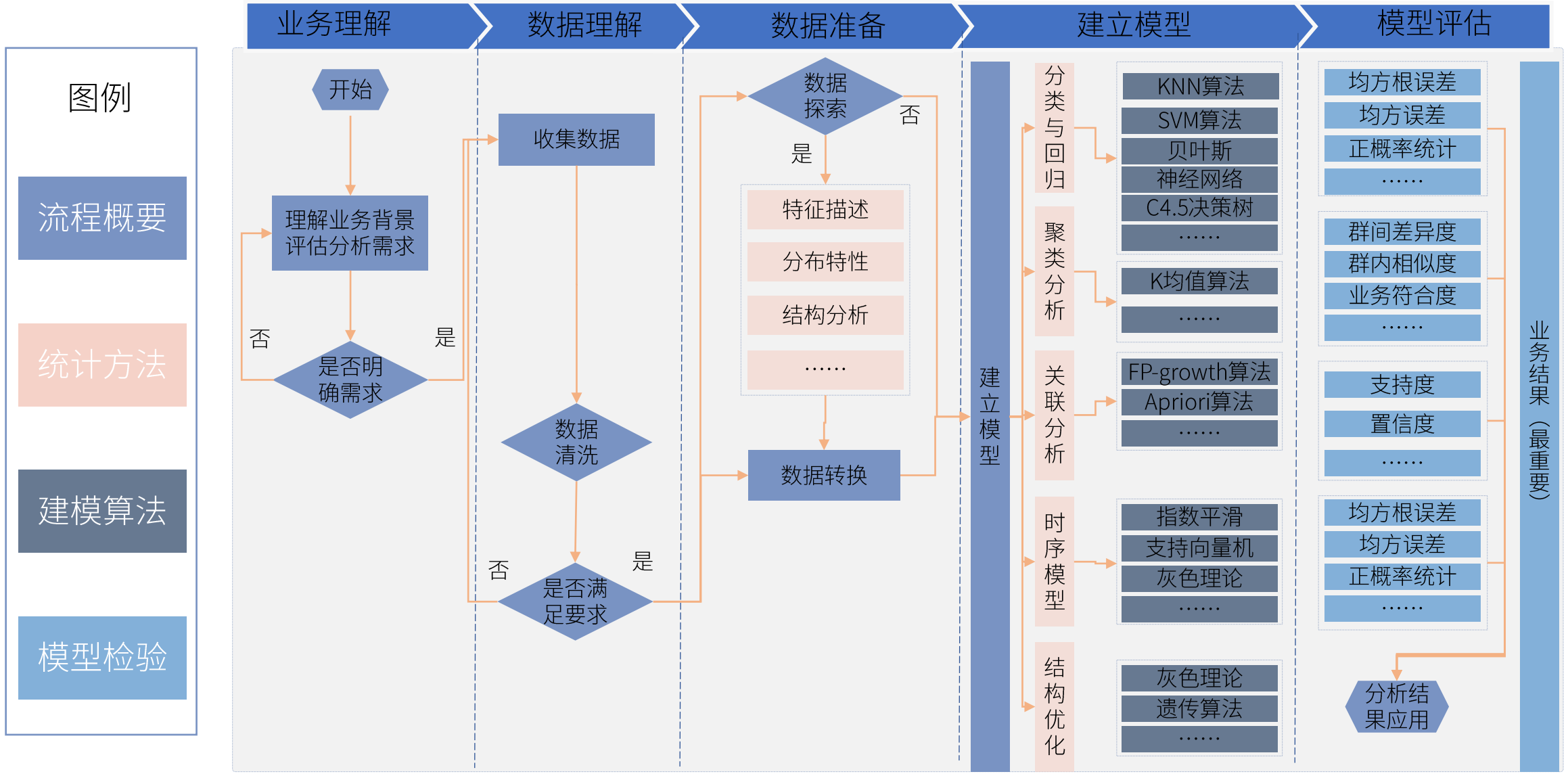
The workflow of Crisp-DM

XSGames



The workflow of Crisp-DM(continued)

XSGames



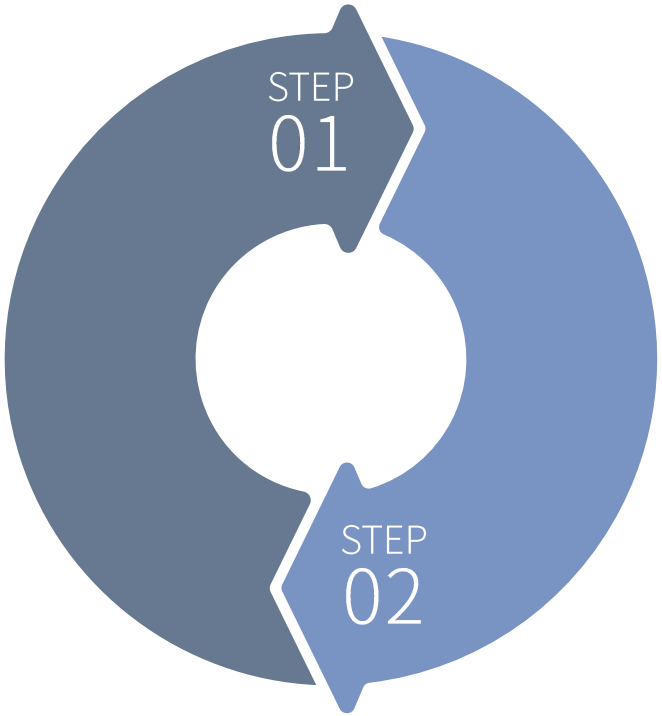
数据处理（特征工程）

XSGames

在对收集的数据进行分析前，要明确数据类型、规模，对数据有初步理解，同时要对数据中的“噪声”进行处理，以支持后续数据建模。

数据探索

- 特征描述
- 图形化观察/分布推断
- 结构优化



数据清洗

- 异常值/极端值/缺失值处理
- 单个变量处理
  - 连续变量/哑变量处理
- 多个变量处理
  - 降维、特征筛选/组合

数据探索和数据清洗通常交互进行

数据转换

XSGames

- 数据转换或统一成适合于挖掘的形式，通常的做法有数据泛化、标准化、属性构造等，本文详细介绍数据标准化的方法，即统一数据的量纲及数量级，将数据处理为统一的基准的方法。

基期标准化法

- 选择基期作为参照，
  各期标准化数据＝各期数据/基期数据

直线法

- 极值法：$x'_i = \frac{x_i}{\max(x_i)}, x'_i = \frac{\max(x_i) - x_i}{\max(x_i)}, x'_i = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)}$
- z-score法：$x'_i = \frac{x_i - \bar{x}}{s}$，其中$s = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$

折线法

- 某些数据在不同值范围，采用不同的标准化方法，通常用于综合评价

示例

$$x_i' = \begin{cases} 0 (x_i < a) \\ \frac{x_i - a}{b - a} (a \le x_i < b) \\ 1 (x_i \ge b) \end{cases}$$

曲线法

- Log函数法：$x' = \log(x_i) / \log(\max(x_i))$
- Arctan函数法：$x' = \arctan(x_i) \times 2/\pi$
- 对数函数法、模糊量化模式等

- 各方法都有缺点，要根据客观事物的特征及所选用的分析方法来确定，如聚类分析、关联分析等常用直线法，且聚类分析必须满足无量纲标准；而综合评价则折线和曲线方法用得较多
- **能简就简**，能用直线尽量不用曲线。

定义：

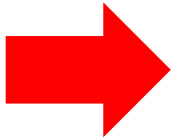
按照某种指定的属性特征将数据归类。需要确定类别的概念描述，并找出类判别准则。分类的目的是获得一个分类函数或分类模型（也常常称作分类器），该模型能把数据集合中的数据项映射到某一个给定类别。

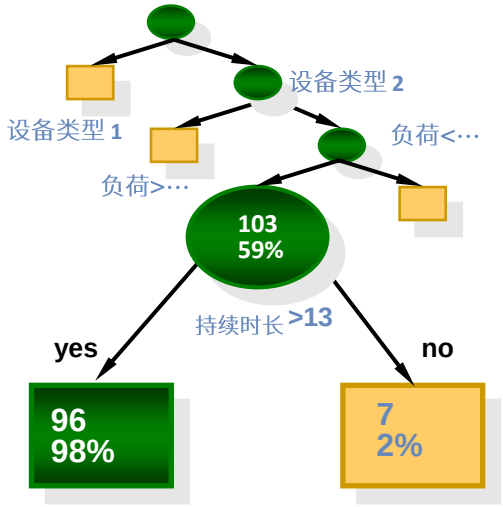
分类是利用训练数据集通过一定的算法而求得分类规则的。是模式识别的基础。

分类可用于提取描述重要数据类的模型或预测未来的数据趋势。

银行根据客户以往贷款记录情况，将客户分为低风险客户和高风险客户，学习得到分类器。对一个新来的申请者，根据分类器计算风险，决定接受或拒绝该申请

分析影响变压器正常运行的因素，预测变压器是否有故障，若有故障，故障为放电故障、过热故障、短路故障等的哪一种。

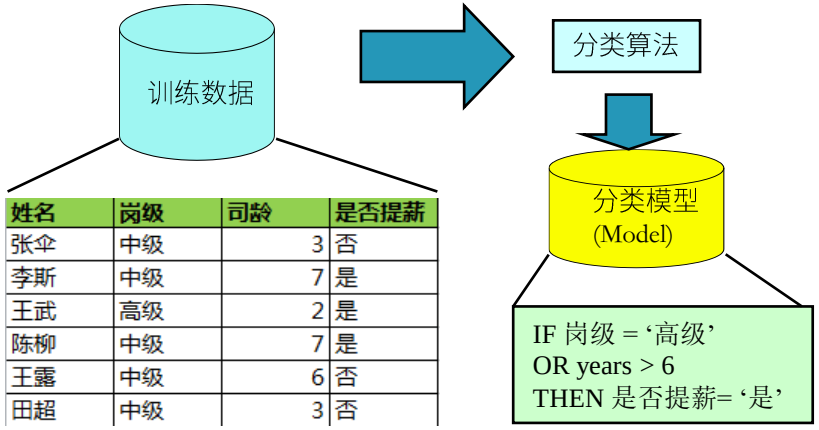




分类的实现：

模型的构建

- 对每个样本进行类别标记
- 训练集构成分类模型
- 分类模型可表示为：分类规则、决策树或数学公式

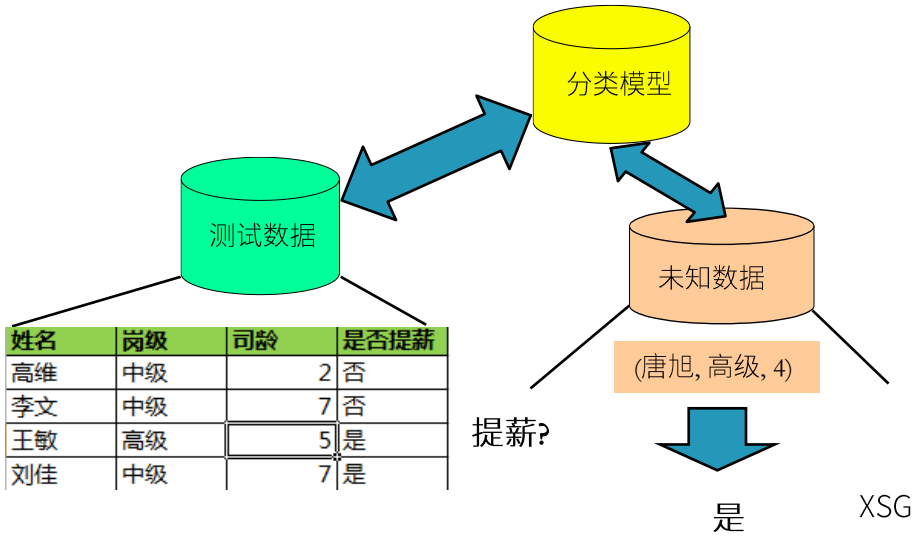


姓名	岗级	司龄	是否提薪
张伞	中级	3	否
李斯	中级	7	是
王武	高级	2	是
陈柳	中级	7	是
王露	中级	6	否
田超	中级	3	否

模型的使用

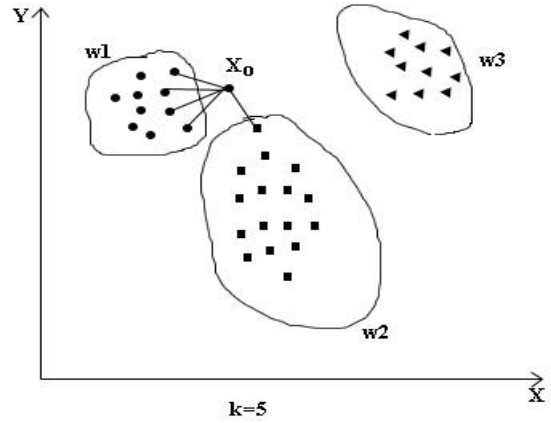
- 识别未知对象的所属类别
- 模型正确性的评价
  - ✓ 已标记分类的测试样本与模型的实际分类结果进行比较

模型的正确率是指测试集中被正确分类的样本数与样本总数的百分比。测试集与训练集相分离，否则将出现过拟合（over-fitting）现象



姓名	岗级	司龄	是否提薪
高维	中级	2	否
李文	中级	7	否
王敏	高级	5	是
刘佳	中级	7	是

分类的主要算法：KNN算法、决策树（CART、C4.5等）、SVM算法、贝叶斯算法、BP神经网络等



- 算法要点
  - 指导思想
    - KNN算法的指导思想是"近朱者赤，近墨者黑"，由你的邻居来推断出你的类别
    - 计算步骤
      1. 算距离：给定测试对象，计算它和训练集中每个对象的距离
      2. 找邻居：圈定距离最近的k个训练对象，作为测试对象的邻近
      3. 做分类：根据这k个近邻归属的主要类别，来对测试对象分类
  - 距离或相似度的衡量
    - 什么是合适的距离衡量？距离越近，应该意味着这两个点归属同一个分类的可能性越大
    - 常见的距离衡量包括欧式距离、夹角余弦等
    - 对于文本分类来说，使用夹角余弦（cosine）来计算相似度就比欧式（Euclidean）距离更合适
  - 类别的判定
    - 投票决定，少数服从多数，近邻中哪个类别的点最多就分到该类
    - 加权投票法：根据距离远近，对近邻的投票加权，距离越近则权重越大（权重为距离平方的倒数）

分类的主要算法：KNN算法、决策树（CART、C4.5等）、SVM算法、贝叶斯算法、BP神经网络等

算法介绍：

C4.5是一种类似二叉树或多叉树的树结构。树中的每个非叶结点（包括根结点）对应于训练样本集总一个非类属性的测试，非叶结点的每一个分支对应属性的一个测试结果，每个叶结点代表一个类或类分布。从根结点到叶子结点的一条路径形成一条分类规则。决策树可以很方便地转化为分类规则，一种非常直观的分类模型的表示形式。

C45属于一种归纳学习算法。归纳学习（Inductive Learning）旨在从大量经验数据中归纳抽取一般的判定规则和模式，它是机器学习（Machine Learning）中最核心、最成熟的一个分支。

根据有无导师指导，归纳学习又分为有导师学习（Supervised Learning，又称为示例学习）和无导师学习(Unsupervised Learning)。

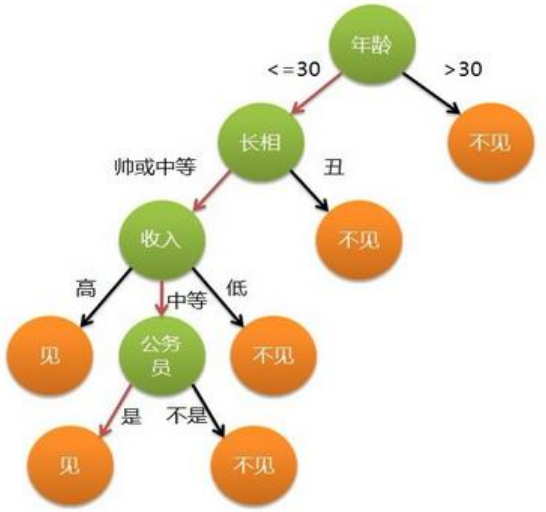
C45属于有导师的学习算法。

算法特点：

（1）模型直观清晰，分类规则易于解释；
（2）解决了连续数据值的学习问题；
（3）提供了将学习结果决策树到等价规则集的转换功能。

决策树示例：
套用俗语，决策树分类的思想类似于找对象。现想象一个女孩的母亲要给这个女孩介绍男朋友，于是有了下面的对话：

女儿：多大年纪了？
母亲：26。
女儿：长的帅不帅？
母亲：挺帅的。
女儿：收入高不？
母亲：不算很高，中等情况。
女儿：是公务员不？
母亲：是，在税务局上班呢。
女儿：那好，我去见见。



分类的主要算法：KNN算法、决策树（CART、C4.5等）、SVM算法、贝叶斯算法、BP神经网络等

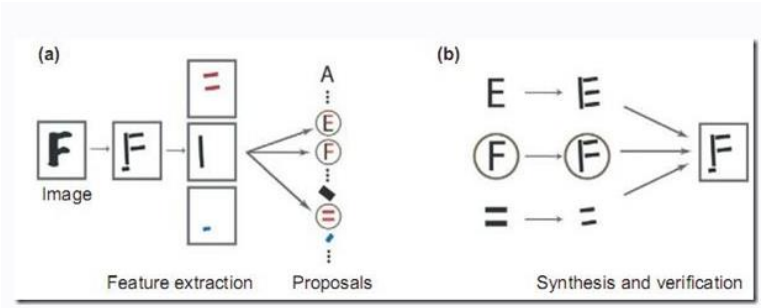
设每个数据样本用一个n维特征向量来描述n个属性的值，即：X={x1，x2，…，xn}，假定有m个类，分别用C1, C2,…，Cm表示。给定一个未知的数据样本X（即没有类标号），若朴素贝叶斯分类法将未知的样本X分配给类Ci，则一定是
P(Ci|X)>P(Cj|X) 1≤j≤m，j≠I

根据贝叶斯定理
由于P(X)对于所有类为常数，最大化后验概率P(Ci|X)可转化为最大化先验概率P(X|Ci)P(Ci)。如果训练数据集有许多属性和元组，计算P(X|Ci)的开销可能非常大，为此，通常假设各属性的取值互相独立，这样先验概率P(x1|Ci)，P(x2|Ci)，…，P(xn|Ci)可以从训练数据集求得。

根据此方法，对一个未知类别的样本X，可以先分别计算出X属于每一个类别Ci的概率P(X|Ci)P(Ci)，然后选择其中概率最大的类别作为其类别。

朴素贝叶斯算法成立的前提是各属性之间互相独立。当数据集满足这种独立性假设时,分类的准确度较高，否则可能较低。另外，该算法没有分类规则输出。

贝叶斯图像识别
贝叶斯方法是一个非常通用的推理框架。其核心理念可以描述成：Analysis by Synthesis （通过合成来分析）。06 年的认知科学新进展上有一篇论文就是讲用贝叶斯推理来解释视觉识别的，一图胜千言，下图就是摘自这篇论文：



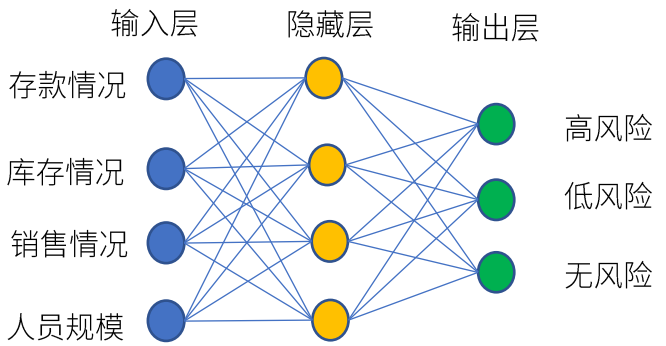
首先是视觉系统提取图形的边角特征，然后使用这些特征自底向上地激活高层的抽象概念（比如是 E 还是 F 还是等号），然后使用一个自顶向下的验证来比较到底哪个概念最佳地解释了观察到的图像

分类的主要算法：KNN算法、决策树（CART、C4.5等）、SVM算法、贝叶斯算法、BP神经网络等

BP（Back Propagation）网络是1986年由Rumelhart（鲁姆哈特）和McCelland（麦克利兰）为首的科学家小组提出，是一种按误差逆传播算法训练的多层前馈网络，是目前应用最广泛的神经网络模型之一。BP网络能学习和存贮大量的输入-输出模式映射关系，而无需事前揭示描述这种映射关系的数学方程。它的学习规则是使用最速下降法，通过反向传播来不断调整网络的权值和阈值，使网络的误差平方和最小。BP神经网络模型拓扑结构包括输入层（input）、隐层(hidden layer)和输出层(output layer)。

BP神经网络学习过程

- 正向传播：
  - 输入样本-----输入层------各隐藏层-----输出层
- 判断是否转入反向传播阶段
  - 若输出层的实际输出与期望输出不符
- 误差反传
  - 误差以某种形式在各层表示-----修正各层单元的权值
- 网络输出的误差减少到可接受的程度或达到预先设定的学习次数为止



BP神经网络的不足

首先，由于学习速率是固定的，因此网络的收敛速度慢，需要较长的训练时间。
其次，BP算法可以使权值收敛到某个值，但并不保证其为误差平面的全局最小值。
再次，网络隐含层的层数和单元数的选择尚无理论上的指导，一般是根据经验或者通过反复实验确定。
最后，网络的学习和记忆具有不稳定性。也就是说，如果增加了学习样本，训练好的网络就需要从头开始训练，对于以前的权值和阈值是没有记忆的。

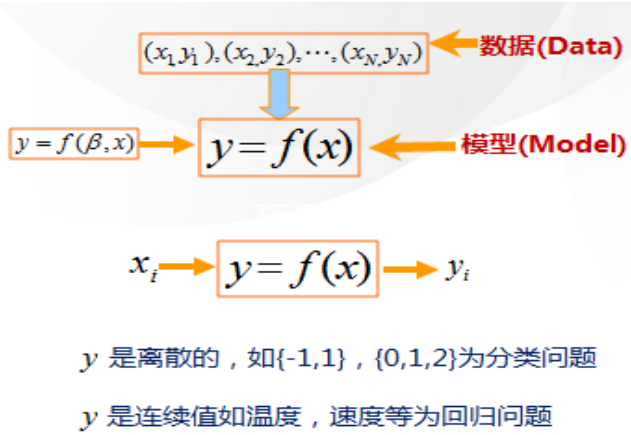
产生：英国统计学家F.GALTON（法兰西斯·高尔顿）(1822-1911)和其学生K.Pearson（卡尔.皮尔逊）(1856-1936)观察了1078对夫妇，以每对夫妇的平均身高为X，而取他们成年的儿子的身高为Y，得到如下经验方程：Y=33.73+0.516X

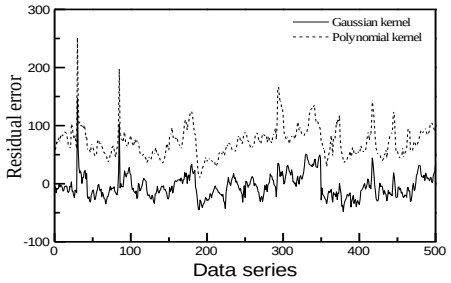
定义：

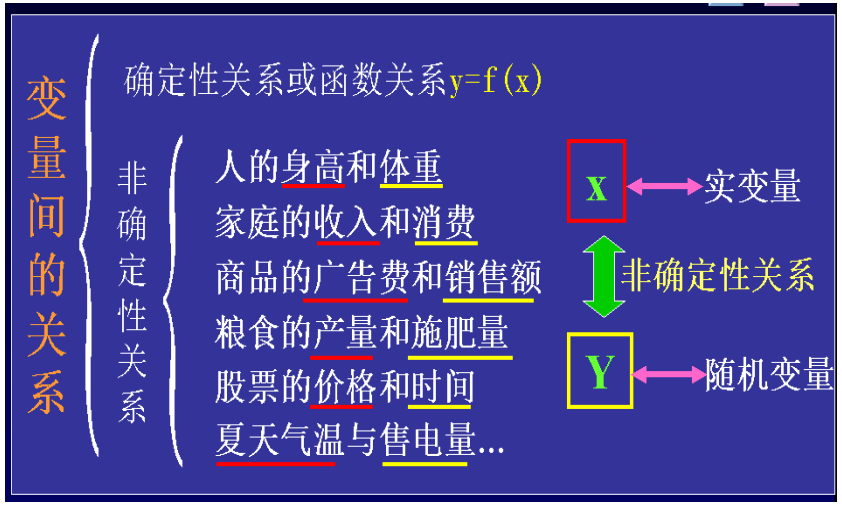
假定同一个或多个独立变量存在相关关系，寻找相关关系的模型。不同于时间序列法的是：模型的因变量是随机变量，而自变量是可控变量。分为线性回归和非线性回归，通常指连续要素之间的模型关系，是因果关系分析的基础。（回归研究的是数据之间的非确定性关系）

线性回归算法寻找属性与预测目标之间的线性关系。通过属性选择与去掉相关性，去掉与问题无关的变量或存在线性相关性的变量。

在建立回归模型之前，可先进行主成分分析，消除属性之间的相关性。最后通过最小二乘法，算法得到各属性与目标之间的线性系数。







线性回归

XSGames

分类：

一元线性回归

只有一个变量X与因变量Y有关，X与Y都是连续型变量，因变量Y或其残差必须服从正态分布

多元线性回归

分析多个变量与因变量Y的关系， X与Y都是连续型变量，因变量Y或其残差必须服从正态分布

LOGISTIC线性回归

分析多个变量与因变量Y的关系， Y通常是离散型或定性变量，该模型对因变量Y的分布无要求

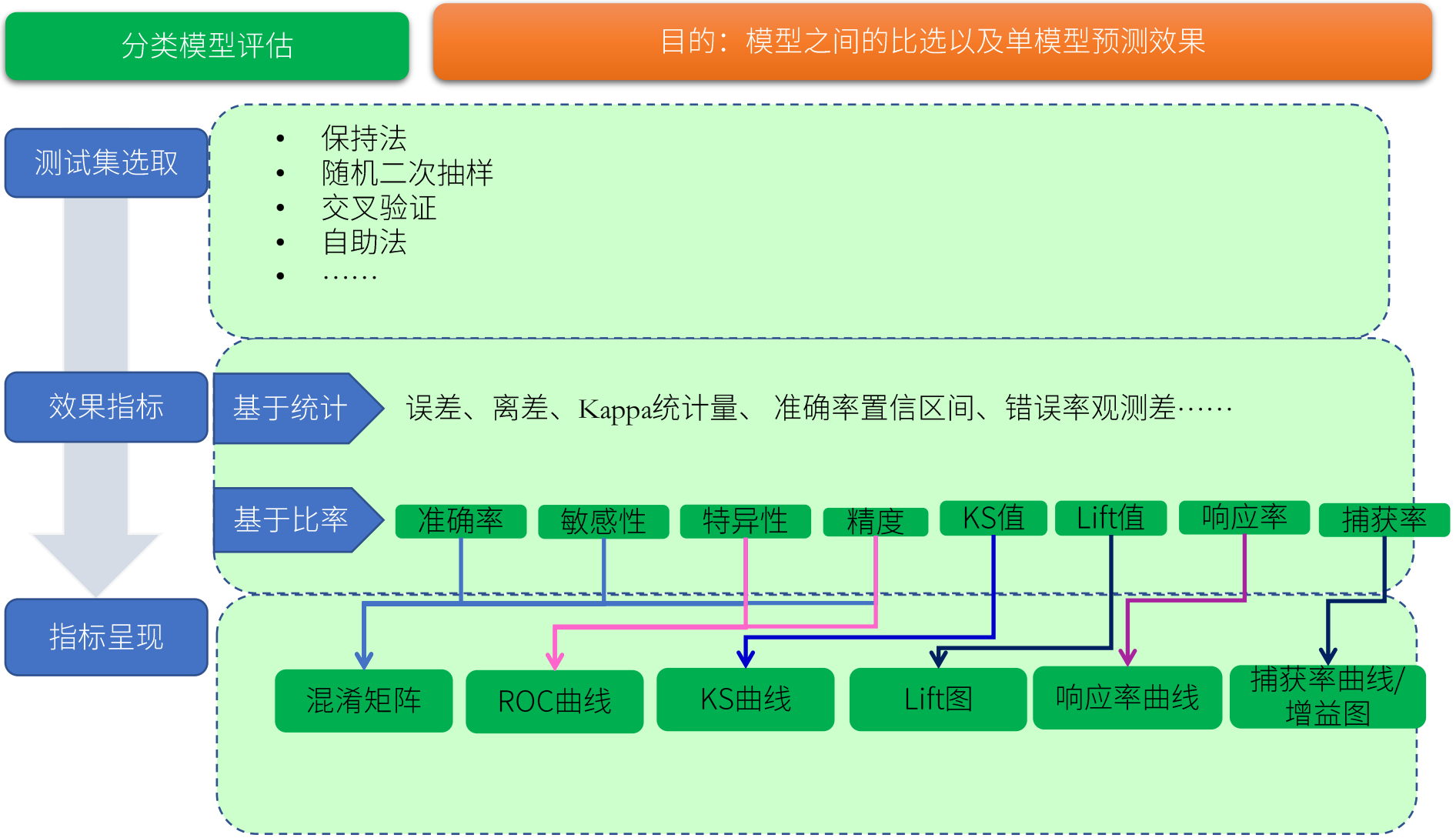
前提：

- 正态性假设：总体误差项需服从正态分布，反之则最小二乘估计不再是最佳无偏估计，不能进行区间估计和假设检验
- 零均值性假设：在自变量取一定值的条件下，其总体各误差项的条件平均值为零，反之无法得到无偏估计
- 等方差性假设：在自变量取一定值的条件下，其总体各误差项的条件方差为一常数，反之无法得到无偏估计
- 独立性假设：误差项之间相互独立（不相关）,误差项与自变量之间应相互独立，否则最小二乘估计不再是有效估计

检验：

回归模型
- 样本容量：$n \geqslant 30$或$n \geqslant 3(p+1)$；其中n为样本容量，p为自变量数目
- 模型显著性：F检验，读取p值并同置信度比较，判断回归模型显著性
- 系数显著性：t检验，读取p值并同置信度比较，判断该自变量是否显著影响因变量
- 拟合优度
  - 残差检验：绘制残差图（标准化残差直方图及散点图），检验残差的正态性及独立性，若所描绘的点都在以0为横轴的直线上下随机散布，则回归直线对各个观测值的拟合情况良好
  - 决定系数：R^2 度量自变量对因变量变化的解释程度，愈接近1则线性关系愈强

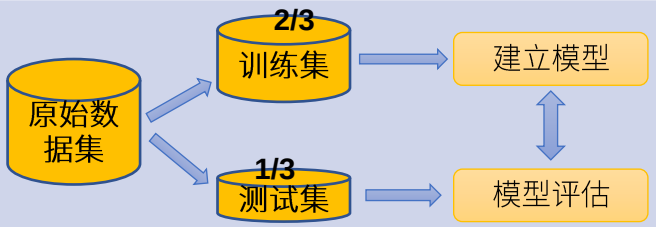
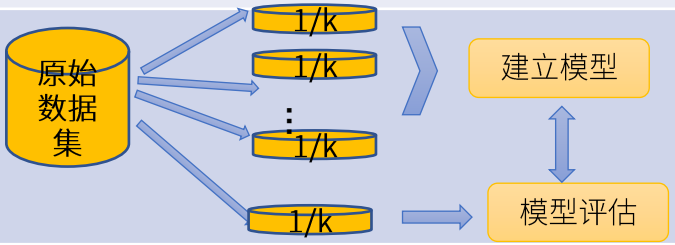
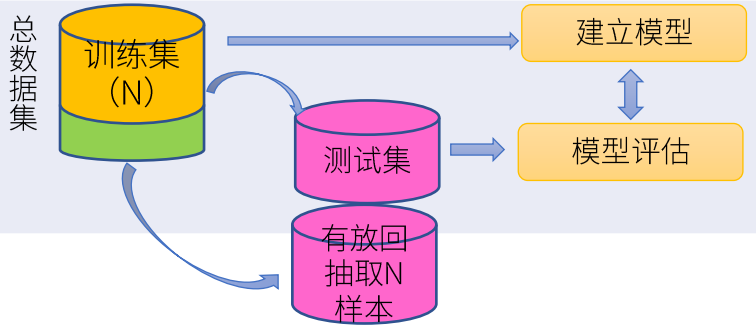
分类模型评估


分类模型评估
目的：模型之间的比选以及单模型预测效果
测试集选取
保持法
随机二次抽样
交叉验证
自助法
……
效果指标
基于统计
误差、离差、Kappa统计量、 准确率置信区间、错误率观测差……
基于比率
准确率
敏感性
特异性
精度
KS值
Lift值
响应率
捕获率
指标呈现
混淆矩阵
ROC曲线
KS曲线
Lift图
响应率曲线
捕获率曲线/
增益图

分类模型评估

XSGames

测试集选取方法

方法	描述	图示
保持法	将原始数据集随机地划分到两个独立的集合:训练集和检验集。通常，三分之二的数据分配到训练集，其余三分之一分配到检验集。模型的效果指标如准确率、误差等由训练集导出。	
随机二次抽样	多次重复使用保持法，得到一组准确率等效果指标。	
交叉验证	最常用的是k-折交叉法，将原始数据分成k份，每次用其中一份为测试集，其余为训练集运行，总共运行k次，记录误差。	
自助法	有放回抽样。训练集的样本为N，放回原数据集，重新有放回地均匀抽取N个样本后，剩余的数据集作为测试集。	

分类模型评估

XSGames

效果指标—基于比率

以二分类为例，说明几个重要效果指标概念。下图为混淆矩阵。通过银行办理信用卡的例子做指标的业务解释。

实际类 \ 预测类	1	0	合计
1	a	b	a+b
0	c	d	c+d
合计	a+c	b+d	a+b+c+d

示例

实际类 \ 预测类	违约	不违约	合计
违约	80	120	200
不违约	20	980	1000
合计	100	1100	1200

准确率 =(a+d)/(a+b+c+d)

最常用的评估指标，用以评价模型分类是否正确。但是，对于不平衡问题（即0类的占大多数），准确率去评价就不够。例如银行办理信用卡，模型只用一条规则“所有人不违约”，结果准确率达到1000/1200=83.3%。但这样的模型毫无意义。准确率适合于平衡问题。

敏感性=a/(a+b)

正确识别正元组的百分比。如例中，敏感性为80/200=40%，因此该模型正确标识真元组（稀有类）的能力还是比较差的，但是还是高于违约的总占比200/1200=16.7%

特异性=d/(c+d)

正确识别负元组的百分比。例子中为98%。

精度=a/(a+c)

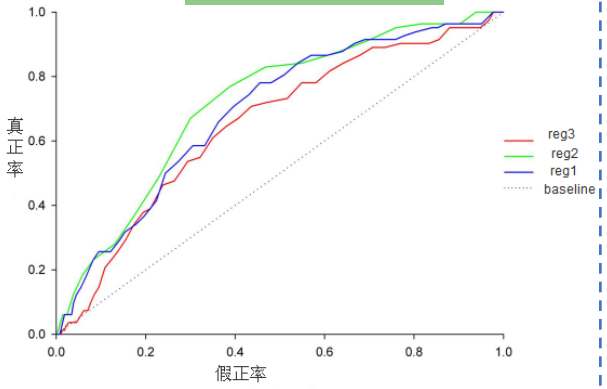
预测为正元类中实际为正元类所占的百分比。衡量预测类1的精确性。例子中为80%。

该案例中模型对于违约的人群，可以识别40%；如果一个人通过模型判断为违约类，则80%可能该人为违约的。敏感性和精度是两个重要指标，可以综合这两个指标，如F等。

分类模型评估

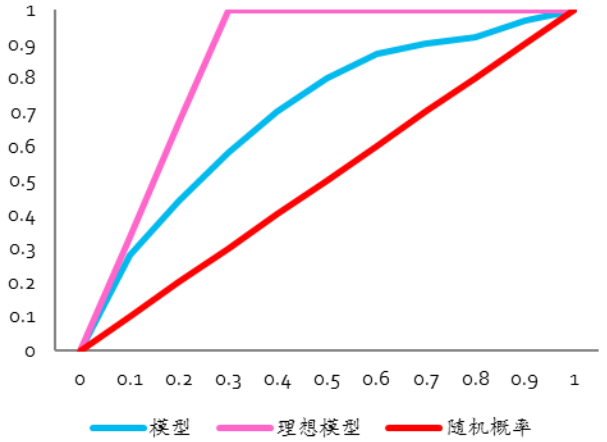
XSGames

ROC曲线



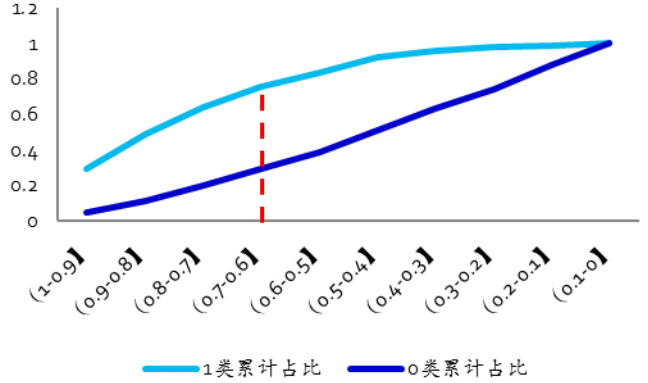
以真正率及敏感性为纵轴，假正率=1-特异性为横轴做图。给定一个二类问题，我们可以对检验集的不同部分，显示模型可以正确识别正样本的比例与模型将负样本错误标识为正样本的比例之间的比较评定。敏感性的增加以错误正例的增加为代价。

增益图



和捕获率曲线是一样的，详见捕获率曲线。
理想模型：100%预测正确下的曲线。这里假设1类占总数为30%。
模型的曲线越靠近理想曲线，预测水平越高。可用Gini系数衡量。
Gini系数=模型曲线与随机曲线之间的面积/理想模型曲线与随机曲线之间的面积。越接近1越好。

KS曲线

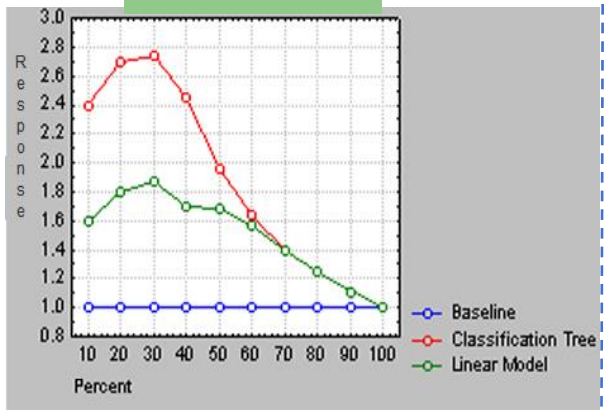


模型预测为概率值，即为1类的概率为多少，为0类的概率为多少。将1类、0类的概率 按照大小由高到底排列，并将各自的累计百分比画在一个图里。纵坐标代表累计百分比，横坐标为预测的概率区间。
0、1曲线的最大距离为KS值，反映模型区分0、1类的能力，越大代表模型将0、1分开程度越大。一般大于0.2较好。如图KS=0.47.

分类模型评估

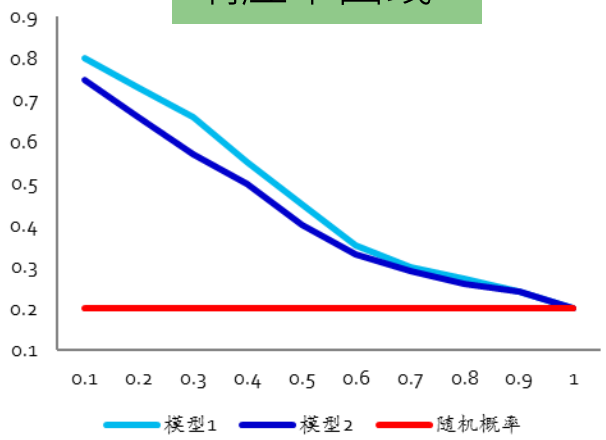
XSGames

Lift图



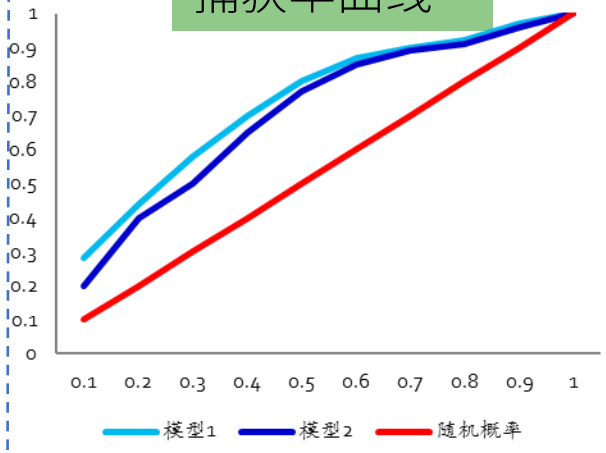
Lift值=响应率/随机概率。比如对10000名浅在顾客进行概率打分，预测其购买商品的可能性，若实际中有900人会购买，则9%为随机概率。抽取概率排名前10%的人数，即1000人，预测600人购买，则前10%的响应率为600/1000=60%，则Lift值=60%/9%=6.67.

响应率曲线



在每个区间里进行计算，1类的累计数占该区间累计的总数比例作为**响应率**。比如在排序前10%中，模型1得出1类样本占比80%，模型2为73%。响应率越高越好，改图显示模型1较模型2更好。

捕获率曲线



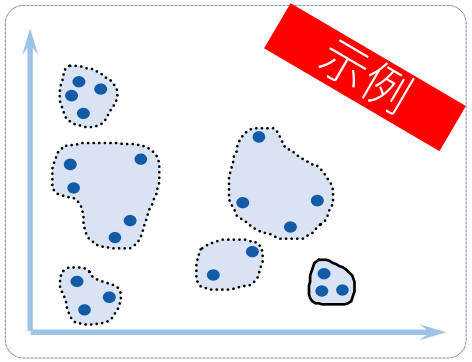
是在每个区间段，计算1类的累计值占总体1类的百分比作为**捕获率**。衡量的是某累计区间抓住1类的对象占总体的比例。

随机概率：不用模型随机抽取数据得到的比率。比如响应率，总数据中1类占比20%，那抽取10%数据理论占比应该还是20%。
横坐标：按照模型结果概率得分从高到底排序，分成10个区间。适合于模型输出值为概率得分，如贝叶斯分类、后向传播等。

三个指标在实际业务中使用比较多，因为其直观、通俗易懂；同时有利于划分不同的目标人群，前10%？、20%？根据业务需要挑选受众规模。

聚类

聚类分析对具有共同趋势或结构的数据进行分组，将数据项分组成多个簇（类），簇之间的数据差别应尽可能大，簇内的数据差别应尽可能小，即“最小化簇间的相似性，最大化簇内的相似性”。



基于划分的聚类

- 对给定的数据集合，事先指定划分为k个类别。
- **典型算法：k-均值法**和k-中心点算法等。

基于层次的聚类

- 对给定的数据集合进行层次分解，不需要预先给定聚类数，但要给定终止条件，包括凝聚法和分裂法两类。
- **典型算法：** CURE、Chameleon、BIRCH、Agglomerative

基于密度的聚类

- 只要某簇邻近区域的密度超过设定的阈值，则扩大簇的范围，继续聚类。这类算法可以获得任意形状的簇。
- **典型算法：**DBSCAN、OPTICS和DENCLUE等

基于网格的聚类

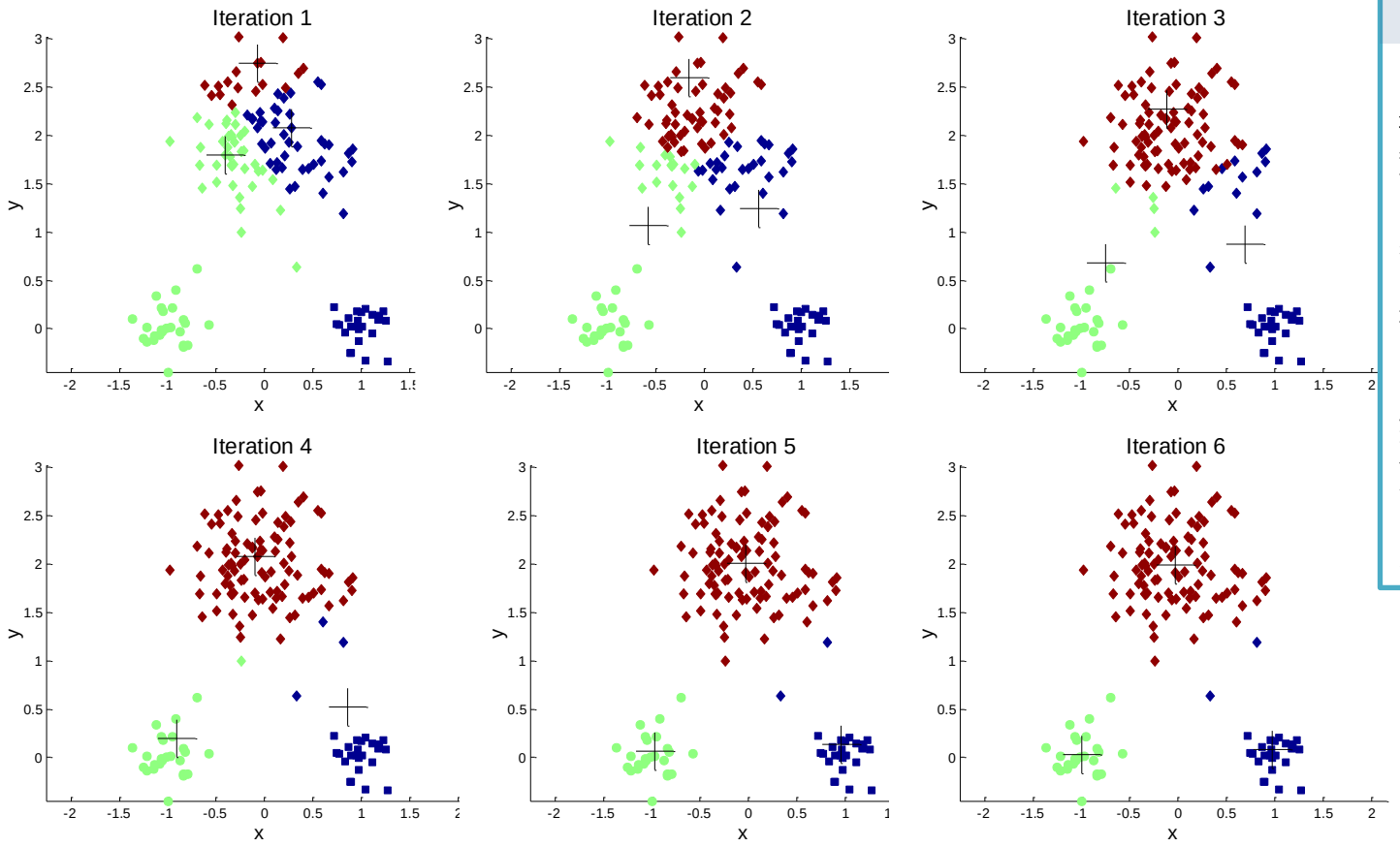
- 首先将问题空间量化为有限数目的单元，形成一个空间网格结构，随后聚类在这些网格之间进行。
- **典型算法：**STING、WareCluster和CLIQUE等 。

基于模型的聚类

- 为每个簇假定一个模型，寻找数据对模型的最佳拟合。所基于的假设是：数据是根据潜在的概率分布生成的。
- **典型算法：**COBWEB和神经网络算法等。

K-Means聚类

K-Means算法，也被称为K-平均或K-均值，是一种得到最广泛使用的聚类算法。主要思想是：首先将各个聚类子集内的所有数据样本的均值作为该聚类的代表点，然后把每个数据点划分到最近的类别中，使得评价聚类性能的准则函数达到最优，从而使同一个类中的对象相似度较高，而不同类之间的对象的相似度较小。



应用实例

利用K-means聚类算法，把原始数据聚成三个不同的簇的应用实例如左图示（K=3）。

基本思路：

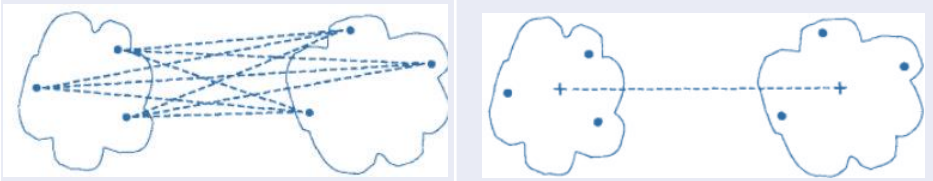
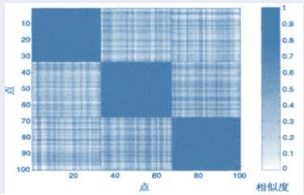
（1）首先，随机选择k个数据点做为聚类中心；

（2）然后，计算其它点到这些聚类中心点的距离，通过对簇中距离平均值的计算，不断改变这些聚类中心的位置，直到这些聚类中心不再变化为止。

聚类模型评估

聚类

目的：评估聚类效果、确定合适的分类数量、聚类模型的选择

评估指标	公式定义	图示定义
凝聚度	衡量一个族内对象凝聚情况	
分离度	衡量族与族之间的差异	
轮廓系数	综合了凝聚度和分离度	
相似度矩阵	通过与理想相似矩阵比较，看聚类效果	
共性分类相关系数	衡量共性分类矩阵与原相异度矩阵之间的相关度，用以评估哪种层次聚类方法最好。	

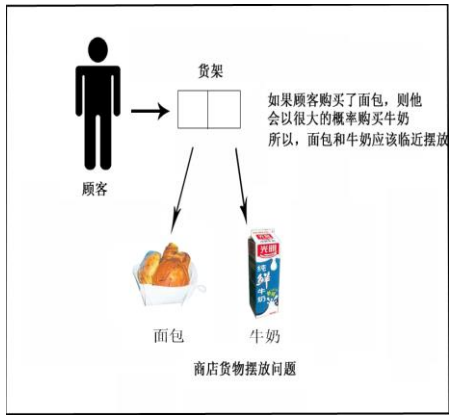
关联规则

XSGames

定义：

自然界中某种事物发生时其他事物也会发生，则这种联系称之为关联。反映事件之间依赖或关联的知识称为关联型知识（又称依赖关系）。要求找出描述这种关联的规则,并用以预测或识别。

关联分析的目的是找出数据集合中隐藏的关联网，是离散变量因果分析的基础。

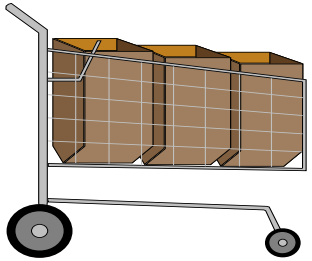


举例：

通过发现顾客放入其购物篮中不同商品之间的联系，分析顾客的购买习惯。通过了解哪些商品频繁地被顾客同时购买，这种关联的发现可以帮助零售商制定营销策略。例如，在同一次购物中，如果顾客购买牛奶的同时，也购买面包（和什么类型的面包）的可能性有多大?

这种信息可以引导销售，可以帮助零售商有选择地经销和安排货架。例如，将牛奶和面包尽可能放近一些，可以进一步刺激一次去商店同时购买这些商品。

关联分析 Association



- 市场组合分析
- 套装产品分析
- 目录设计
- 交叉销售

关联规则

XSGames

基本概念

设关联规则：$A \rightarrow B$，｛A｝或｛B｝为项集，支持度=$\{A \cap B\}/\{A\}+\{B\}$，表示同时包含A、B事务占总事务的百分比；置信度=$\{A \cap B\}/\{A\}$，是预测性指标，表示A事务发生B事务发生的可能性。显然支持度为对称指标，即$A \rightarrow B$或$B \rightarrow A$都一样，而置信度为非对称指标，二者不同。我们以茶和咖啡的案例做指标说明。

	A	¬A	合计
B	F11	F10	F1+
¬B	F01	F00	F0+
合计	F+1	F+0	F

示例

	喝咖啡（A）	不喝咖啡（¬A）	合计
喝茶（B）	150	50	200
不喝茶（¬B）	650	150	800
合计	800	200	1000

支持度（｛喝茶｝→｛喝咖啡｝）=150/1000=15%；
置信度（｛喝茶｝→｛喝咖啡｝）=150/200=75%。即一个人喝茶那么他75%可能喝咖啡。
再看，不管一个人是否喝茶，其喝咖啡的比例为800/1000=80%>75%。即一个人喝茶其喝咖啡的可能性由80%降低到75%，因此｛喝茶｝→｛喝咖啡｝的高置信度实际上是一个误导，其忽略了喝咖啡的支持度。因此，支持度-置信度的评估框架是不完善的。

兴趣因子

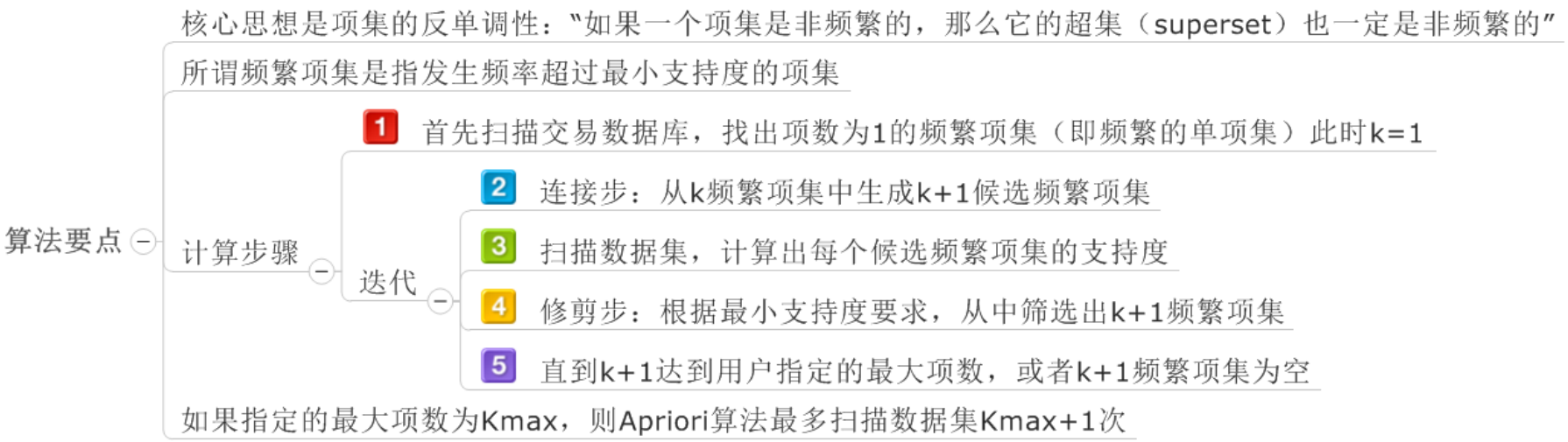
置信度除以喝咖啡的支持度，即75%/80%=0.94。大于1表示正相关，而且越大相关性越强；等于1表示相互独立；小于1表示负相关。

相关性

对于连续变量相关性用pearson相关系数，Pearson相关系数用来衡量两个数据集合是否在一条线上面，它用来衡量定距变量间的线性关系。如衡量国民收入和居民储蓄存款、身高和体重、高中成绩和高考成绩等变量间的线性相关关系。

主要的关联算法：Apriori关联算法、FP-growth关联算法等；

Apriori算法是最基本的一种关联规则算法，它采用布尔关联规则的挖掘频繁项集的算法，利用逐层搜索的方法挖掘频繁项集。



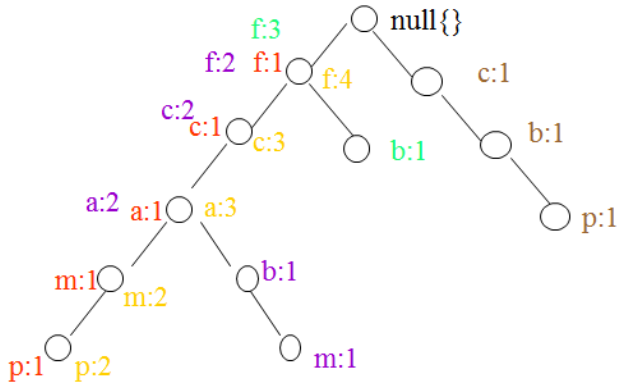
主要的关联算法：Apriori关联算法、FP-growth关联算法等；

FP-Growth算法不产生候选集而直接生成频繁集的频繁模式增长算法，该算法采用分而治之的策略：在第一次扫描数据库之后，把数据库中的频繁项目集压缩到一棵频繁模式树中，形成投影数据库，同时保留其中的关联信息，随后继续将FP-tree分成一些条件树，对这些条件树分别进行挖掘。

交易编号	所有购物项	（排序后的）频繁项
100	f, a, c, d, g, i, m, p	f, c, a, m, p
200	a, b, c, f, l, m, o	f, c, a, b, m
300	b, f, h, j, o	f, b
400	b, c, k, s, p	c, b, p
500	a, f, c, e, l, p, m, n	f, c, a, m, p

其中，最小支持度阈值为3

FP-tree的构建



1. f, c, a, m, p
2. f, c, a, b, m
3. f, b
4. c, b, p
5. f, c, a, m, p

f，c，b组合满足条件

关联模型评估


关联规则
目的：识别有意义（有价值）的关联模式
客观度量
评价项集：对称度量指标
支持度
相关性
兴趣因子
余弦
Jaccard
集体强度
……
评价关联规则：非对称客观度量
置信度
J度量
Gini指标
可信度因子
互信息
信任度
……
主观度量
可视化
基于主观模板的度量
基于主观兴趣的度量
……

时间序列

XSGames

时间序列：是按时间顺序的一组数字

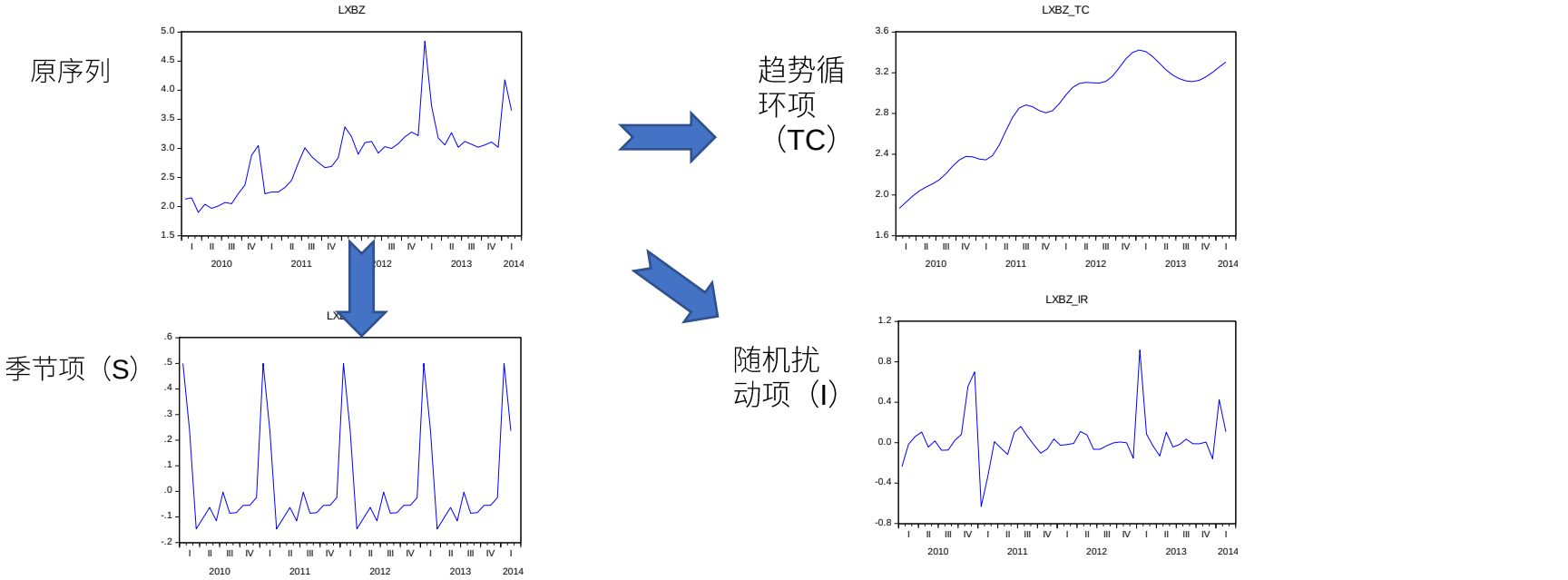
序列构成：

- **长期趋势（*T*）**:时间序列随时间的变化而逐渐增加或减少的长期变化的趋势
- **季节变动（*S*）**:时间序列在一年中或固定时间内，呈现出的固定规则的变动
- **循环变动（C）**:沿着趋势线如钟摆般地循环变动，又称景气循环变动
- **不规则变动（I）**:在时间序列中由于随机因素影响所引起的变动

→ 时间序列

组合模型：

- 加法模型：假定时间序列是基于4种成份相加而成的。长期趋势并不影响季节变动；Y=T+S+C+I
- 乘法模型：假定时间序列是基于4种成份相乘而成的。假定季节变动与循环变动为长期趋势的函数；$Y = T \times S \times C \times I$



时间序列

XSGames

建模步骤：

- 用观测、调查、统计、抽样等方法取得被观测系统时间序列动态数据

- 根据动态数据作相关图，进行相关分析，求自相关函数
- 相关图能显示出变化的趋势和周期，并能发现跳点和拐点（跳点是指与其他数据不一致的观测值，拐点则是指时间序列从上升趋势突然变为下降趋势的点）

- 辨识合适的随机模型,进行曲线拟合,即用通用随机模型去拟合时间序列的观测数
- 短的或简单的时间序列，可用趋势模型和季节模型加上误差来进行拟合；平稳时间序列，可用通用ARMA模型及其特殊情况的自回归模型、滑动平均模型或组合-ARMA模型等来进行拟合，当观测值多于50个时一般采用ARMA模型；非平稳时间序列则要先经差分运算化为平稳时间序列，再用适当模型去拟合这个差分序列

举例：成本费用收入比单指标（累计值）预测

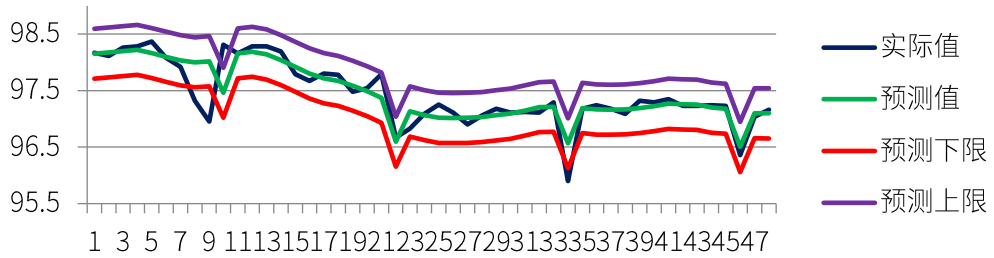
采用季节拆分建模

拟合优度：0.7628

平均绝对误差：0.15

平均相对误差： 0.00156

标准误差：0.2211



	实际值	预测值	下限值	上限值
2014年1月	96.36	96.503303	96.0609034	96.9457034
2014年2月	97.04	97.098057	96.6556572	97.5404572
2014年3月	97.16	97.097295	96.6548955	97.5396955

时间序列算法

XSGames

时间序列预测方法分为平滑法预测和ARIMA模型预测，平滑法是通过时间序列的发展趋势来进行预测，而ARIMA模型是通过时间序列的自相关性来预测。两类方法的适用范围和特点为：

	预测方法	适用范围	特点
平滑法	简单移动平均	没有明显的趋势和季节性	
	加权移动平均	没有明显的趋势和季节性	考虑了不同时刻对预测值影响权重不同
	单指数平滑	适用于无线性趋势，无季节因素的序列	考虑了各期数据对预测值的影响
	双指数平滑	适用于有线性趋势，无季节因素的序列	加入了线性趋势项
	Winter无季节	适用于有线性趋势，无季节因素的序列	与双指数平滑类似，双指数平滑法只用了一个参数，Winters无季节用了两个参数
	Winter加法	适用于有线性趋势和不变季节因素的序列	加入了季节变动的因素
	Winter乘法	适用于有线性趋势和变化季节因素的序列	加入了季节变动的因素
ARIMA	AR(p)	适用于具有p阶偏自相关的序列	通过自回归来预测
	MA(q)	适用于具有q阶自相关的序列	通过随机扰动项的移动平均来预测
	ARMA(p,q)	适用于具有p阶偏自相关和q阶自相关的序列	综合考虑了自回归和随机扰动项的移动平均
	ARIMA(p,d,q)	适用于具有p阶偏自相关和q阶自相关，且d阶差分后平稳的序列	可以对非平稳时间序列建模

结构优化——遗传算法

XSGames

遗传算法是计算机科学人工智能领域中用于解决最优化的一种搜索启发式算法，是进化算法的一种。这种启发式通常用来生成有用的解决方案来优化和搜索问题。进化算法最初是借鉴了进化生物学中的一些现象而发展起来的，这些现象包括遗传、突变、自然选择以及杂交等。

遗传算法广泛应用在生物信息学、系统发生学、计算科学、工程学、经济学、化学、制造、数学、物理、药物测量学和其他领域之中。

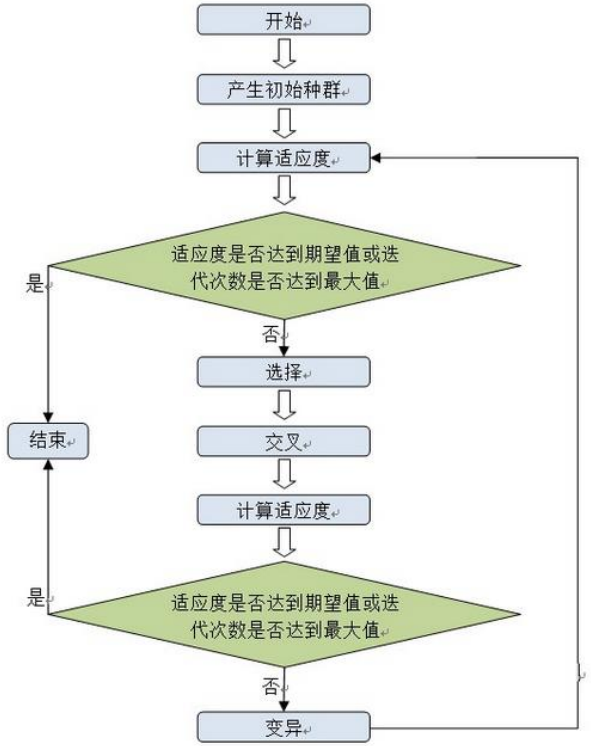
算法特点：

(1)遗传算法从问题解的串集开始搜索，而不是从单个解开始。这是遗传算法与传统优化算法的极大区别。传统优化算法是从单个初始值迭代求最优解的；容易误入局部最优解。遗传算法从串集开始搜索，覆盖面大，利于全局择优。

(2)遗传算法同时处理群体中的多个个体，即对搜索空间中的多个解进行评估，减少了陷入局部最优解的风险，同时算法本身易于实现并行化。

(3)遗传算法不是采用确定性规则，而是采用概率的变迁规则来指导他的搜索方向。

(4)具有自组织、自适应和自学习性。遗传算法利用进化过程获得的信息自行组织搜索时，适应度大的个体具有较高的生存概率，并获得更适应环境的基因结构。



结构优化——灰色理论

XSGames

灰色系统是指“部分信息已知，部分信息未知”的“小样本”，“贫信息”的不确定性系统。它通过对“部分”已知信息的生成、开发去了解、认识现实世界，实现对系统运行行为和演化规律的正确把握和描述。

严格来说，灰色系统是绝对的，而白色与黑色系统是相对的。社会、经济、农业等系统的预测都属于特征性灰色系统的预测。

灰色系统认为：尽管客观系统表象复杂，数据离散，但它们总是有整体功能的，总是有序的。因此，它必然潜藏着某种内在规律。关键在于要用适当方式去挖掘它，然后利用它。

应用：

(1)数列预测：即用观察到的反映预测对象特征的时间序列来构造灰色预测模型，预测未来某一时刻的特征量，或达到某一特征量的时间。

(2)灾变与异常值预测：即通过灰色模型预测异常值出现的时刻，预测异常值什么时候出现在特定时区内。

(3)季节灾变与异常值预测：通过灰色模型预测灾变值发生在一年内某个特定的时区或季节的灾变预测。

(4)拓扑预测：将原始数据作曲线，在曲线上按定值寻找该定值发生的所有时点，并以该定点为框架构成时点序列，然后建立模型预测该定值所发生的时点

(5)系统预测：通过对系统行为特征指标建立一组相关联的灰色模型，预测系统中众多变量间的相互协调关系的变化。

Thank You