



consulting | technology | outsourcing

## 大数据分析--埃森哲

2015-7

High performance. Delivered.



# 目录

概述

数据分析框架

数据分析方法

数据理解&数据准备

分类与回归

聚类分析

关联分析

时序模型

结构优化

数据分析支撑工具

# 数据分析即从数据、信息到知识的过程，数据分析需要数学理论、行业经验以及计算机工具三者结合

- **工具支撑**

各种厂商开发了数据分析的工具、模块，将分析模型封装，使不了解技术的人也能够快捷的实现数学建模，快速响应分析需求。

- **机器学习**

不需要人过多干预，通过计算机自动学习，发现数据规律，但结论不易控制。

- **数据挖掘**

数据挖掘是挖掘数据背后隐藏的知识的重要手段

- **分析误区**

不了解分析模型的数学原理，会导致错误的使用模型，而得出错误的分析结论，影响业务决策，因此在选用分析模型时，要深入了解该模型的原理和使用限制

- **数学&统计学知识**

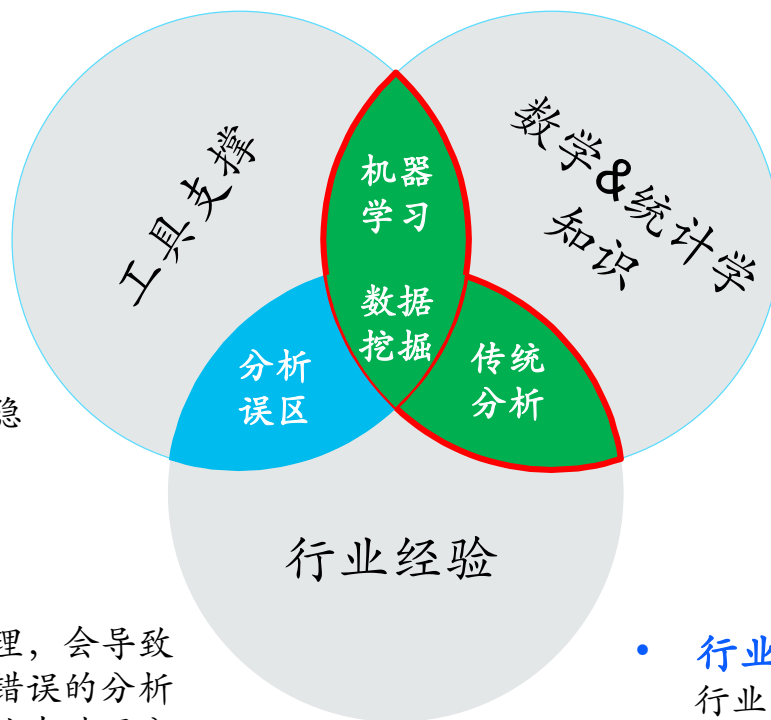
数据分析的基础，将整理、描述、预测数据的手段、过程抽象为数学模型的理论知识

- **传统分析**

在数据量较少时，传统的数据分析已能够发现数据中包含的知识，包括结构分析、杜邦分析等模型，方法成熟，应用广泛，本文不展开介绍

- **行业经验**

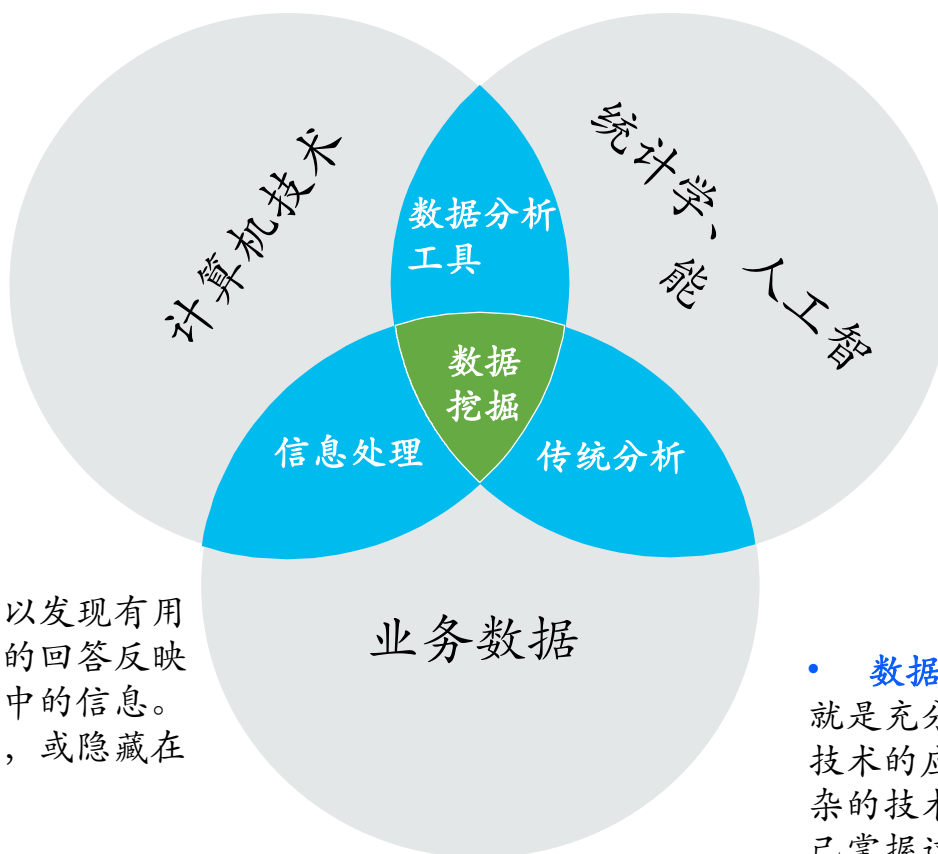
行业经验可在数据分析前确定分析需求，分析中检验方法是否合理，以及分析后指导应用，但行业特征不同，其应用也不同，因此本文不展开介绍



# 随着计算机技术发展和数据分析理论的更新，当前的数据分析逐步成为机器语言、统计知识两个学科的交集（备选）

- **数据分析工具**

各种厂商开发了数据分析的工具、模块，将分析模型封装，使不了解技术的人也能够快捷的实现数学建模，快速响应分析需求。



- **信息处理**

信息处理基于查询，可以发现有用的信息。但是这种查询的回答反映的是直接存放在数据库中的信息。它们不反映复杂的模式，或隐藏在数据库中的规律。

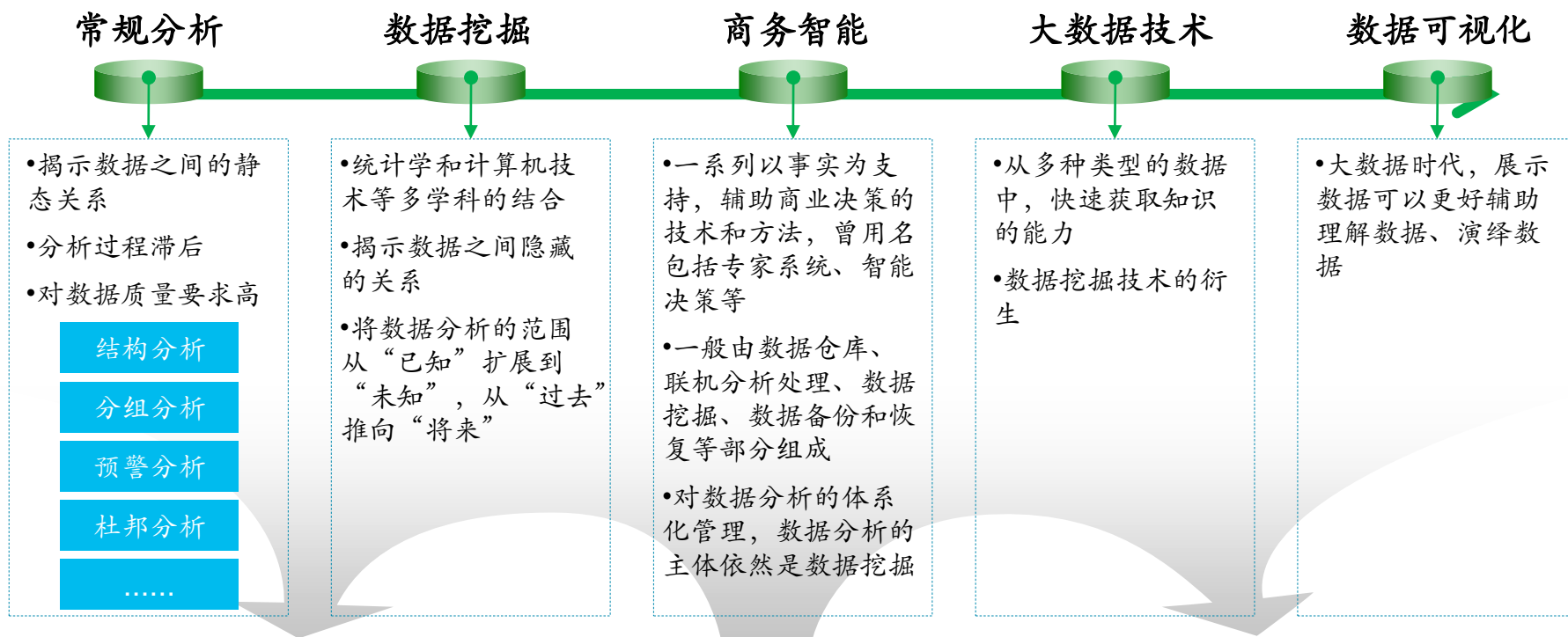
- **传统分析**

在数据量较少时，传统的数据分析已能够发现数据中包含的知识，包括结构分析、杜邦分析等模型，方法成熟，应用广泛，本文不展开介绍

- **数据挖掘**

就是充分利用了统计学和人工智能技术的应用程序，并把这些高深复杂的技术封装起来，使人们不用自己掌握这些技术也能完成同样的功能，并且更专注于自己所要解决的问题。

# 随着计算机科学的进步，数据挖掘、商务智能、大数据等概念的出现，数据分析的手段和方法更加丰富



## 数据分析

- 本文在描述数据分析的流程后，重点介绍通用的数据分析方法和主流的应用工具、软件。
- 随着数据量的不断扩大，数据分析理论正处于飞速发展期，因此本文的方法侧重于基础原理介绍。

# 目录

概述

数据分析框架

数据分析方法

数据理解&数据准备

分类与回归

聚类分析

关联分析

时序模型

结构优化

数据分析支撑工具

# 数据分析标准流程

CRISP-DM为90年代由SIG组织（当时）提出，已被业界广泛认可的数据分析流程。

## 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





# 数据分析框架

## 业务理解

### 理解业务背景， 评估分析需求

- **理解业务背景：**

数据分析的本质是服务于业务需求，如果没有业务理解，缺乏业务指导，会导致分析无法落地。

- **评估业务需求：**

判断分析需求是否可以转换为数据分析项目，某些需求是不能有效转换为数据分析项目的，比如不符合商业逻辑、数据不足、数据质量极差等。

## 数据理解

### 数据收集 数据清洗

- **数据收集：**

抽取的数据必须能够正确反映业务需求，否则分析结论会对业务造成误导。

- **数据清洗：**

原始数据中存在数据缺失和坏数据，如果不处理会导致模型失效，因此对数据通过过滤“去噪”从而提取出有效数据

## 数据准备

### 数据探索 数据转换

- **探索数据：**

运用统计方法对数据进行探索，发现数据内部规律。

- **数据转换：**

为了达到模型的输入数据要求，需要对数据进行转换，包括生成衍生变量、一致化、标准化等。

## 建立模型

### 选择方法、工具， 建立模型

- **建立模型：**

综合考虑业务需求、精度、数据情况、花费成本等因素，选择最合适的模型。在实践中对于一个分析目的，往往运用多个模型，然后通过后续的模型评估，进行优化、调整，以寻求最合适的模型。

## 模型评估

### 建模过程评估 模型结果评估

- **建模过程评估：**

对模型的精度、准确性、效率和通用性进行评估。

- **模型结果评估：**

评估是否有遗漏的业务，模型结果是否回答了当初的业务问题，需要结合业务专家进行评估。

## 应用

### 分析结果应用 分析模型改进

- **结果应用：**

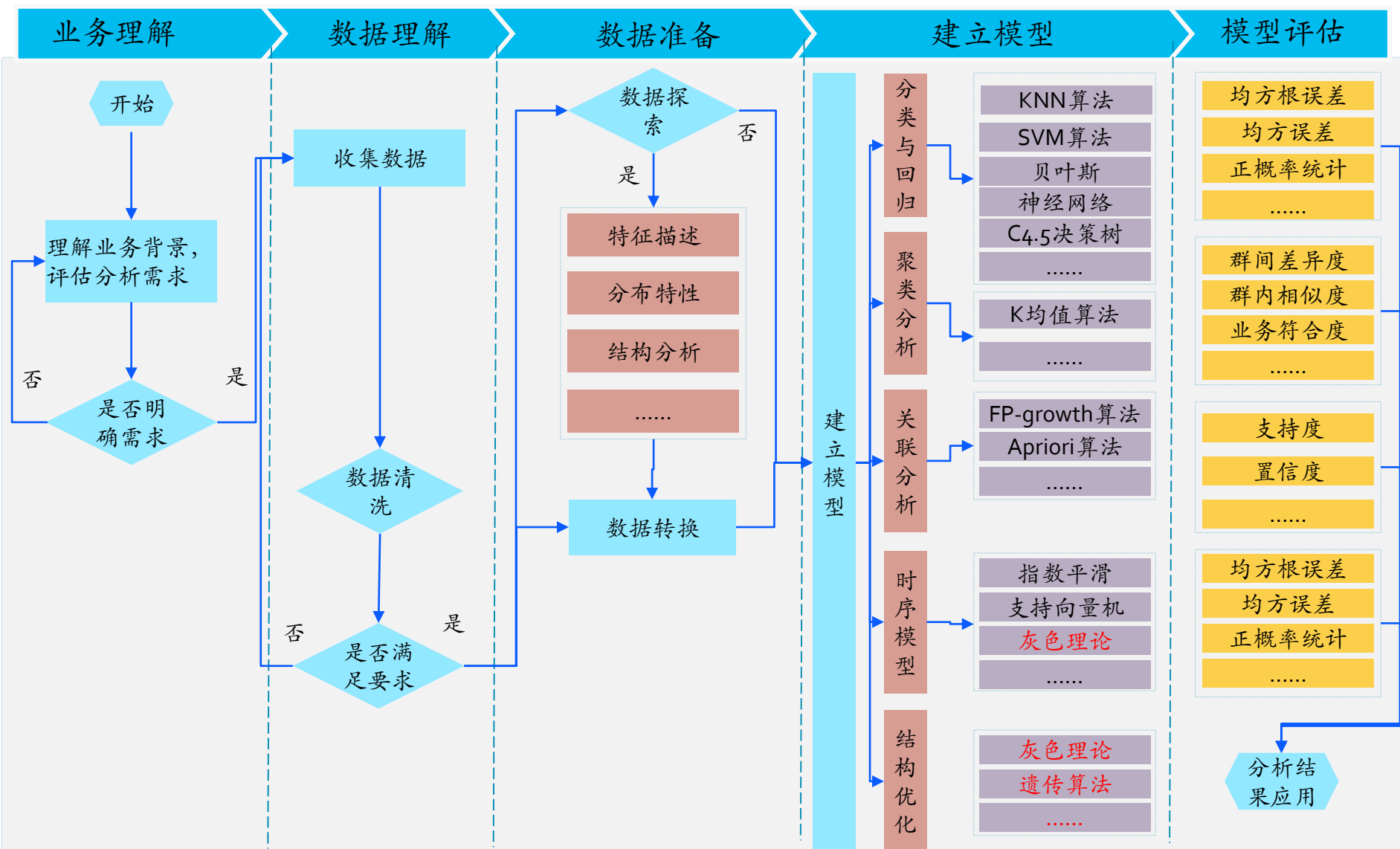
将模型应用于业务实践，才能实现数据分析的真正价值：产生商业价值和解决业务问题。

- **模型改进：**

对模型应用效果的及时跟踪和反馈，以便后期的模型调整和优化。



# 数据分析框架



# 目录

概述

数据分析框架

数据分析方法

数据理解&数据准备

分类与回归

聚类分析

关联分析

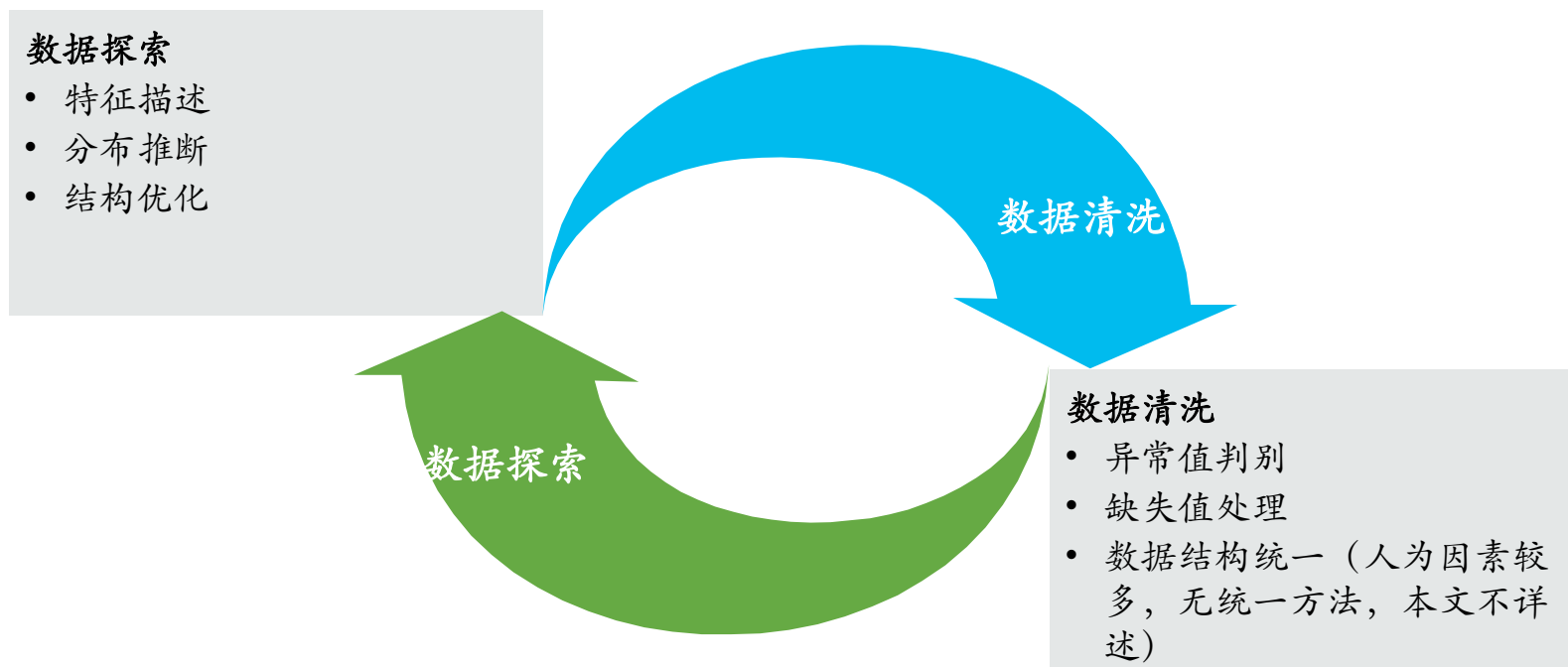
时序模型

结构优化

数据分析支撑工具

# 数据清洗&数据探索

数据收集的方法多种多样，本文不再详述。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。



- 数据清洗和数据探索通常交互进行
- 数据探索有助于选择数据清洗方法
- 数据清洗后可以更有效的进行数据探索

# 数据清洗：1.异常值判别

数据清洗的第一步是识别会影响分析结果的“异常”数据，然后判断是否剔除。目前常用的识别异常数据的方法有物理判别法和统计判别法

## 物理判别法

- 根据人们对客观事物、业务等已有的认识，判别由于外界干扰、人为误差等原因造成实测数据偏离正常结果，判断异常值。
- 比较困难

## 统计判别法

- 给定一个置信概率，并确定一个置信限，凡超过此限的误差，就认为它不属于随机误差范围，将其视为异常值。
- 常用的方法（数据来源于同一分布，且是正态的）：拉依达准则、肖维勒准则、格拉布斯准则、狄克逊准则、t检验。



### 注意

- **慎重对待删除异常值：**为减少犯错误的概率，可多种统计判别法结合使用，并尽力寻找异常值出现的原因；若有多个异常值，应逐个删除，即删除一个异常值后，需再行检验后方可再删除另一个异常值
- **检验方法以正态分布为前提，若数据偏离正态分布或样本较小时，则检验结果未必可靠，校验是否正态分布可借助W检验、D检验**

# 常见统计判别法

判别方法	判别公式	剔除范围	操作步骤	评价
拉依达准则 (3σ准则)	$p( x - u  > 3\sigma) \leq 0.003$	大于 $\mu + 3\sigma$ 小于 $\mu - 3\sigma$	求均值、标准差，进行边界检验，剔除一个异常数据，然后重复操作，逐一剔除	适合用于 $n > 185$ 时的样本判定
肖维勒准则 (等概率准则)	$ x_i - \bar{x}  > Z_c(n)\sigma$	大于 $\mu + Z_c(n)\sigma$ 小于 $\mu - Z_c(n)\sigma$	求均值、标准差，比对系数读取 $Z_c(n)$ 值，边界检验，剔除一个异常数据，然后重复操作，逐一剔除	实际中 $Z_c(n) < 3$ ，测算合理，当 $n$ 处于 $[25, 185]$ 时，判别效果较好
格拉布斯准则	$ x_i - \bar{x}  > T(n, \alpha)\sigma$	删除水平: $ x_i - \bar{x}  > T(n, \alpha_1)\sigma$ 异常检出水平: $T(n, \alpha_1)\sigma <  x_i - \bar{x}  < T(n, \alpha_2)\sigma$	逐一判别并删除达到删除水平的数据；针对达到异常值检出水平，但未及删除水平的数据，应尽量找到数据原因，给以修正，若不能修正，则比较删除与不删除的统计结论，根据是否符合客观情况做去留选择	$T(n, \alpha)$ 值与重复测量次数 $n$ 及置信概率 $\alpha$ 均有关，理论严密，概率意义明确。当 $n$ 处于 $[25, 185]$ 时 $\alpha = 0.05$ ，当 $n$ 处于 $[3, 25]$ 时 $\alpha = 0.01$ ，判别效果较好
狄克逊准则	$f_0 = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}}$ 或 $\frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}}$	$f_0 > f(n, \alpha)$ ，说明 $x_{(n)}$ 离群远，则判定该数据为异常数据	将数据由小到大排成顺序统计量，求极差，比对狄克逊判断表读取 $f(n, \alpha)$ 值，边界检验，剔除一个异常数据，然后重复操作，逐一剔除	异常值只有一个时，效果好；同侧两个数据接近，效果不好 当 $n$ 处于 $[3, 25]$ 时，判别效果较好
T检验	$ x_{(n)} - \bar{x}  > K(n, \alpha)\sigma$ 或 $ x_{(1)} - \bar{x}  > K(n, \alpha)\sigma$	最大、最小数据与均值差值大于 $K(n, \alpha)\sigma$	分别检验最大、最小数据，计算不含被检验最大或最小数据时的均值及标准差，逐一判断并删除异常值	异常值只有一个时，效果好；同侧两个极端数据接近时，效果不好；因而有时通过中位数代替平均数的调整方法可以有效消除同侧异常值的影响

# 数据清洗：2.缺失值处理

在数据缺失严重时，会对分析结果造成较大影响，因此对剔除的异常值以及缺失值，要采用合理的方法进行填补，常见的方法有平均值填充、K最近距离法、回归法、极大似线估计法等

## 平均值填充

取所有对象（或与该对象具有相同决策属性值的对象）的平均值来填充该缺失的属性值

## K最近距离法

先根据欧式距离或相关分析确定距离缺失数据样本最近的K个样本，将这K个值加权平均来估计缺失数据值

## 回归

基于完整的数据集，建立回归方程（模型），对于包含空值的对象，将已知属性值代入方程来估计未知属性值，以此估计值来进行填充；但当变量不是线性相关或预测变量高度相关时会导致估计偏差

## 极大似线估计

在给定完全数据和前一次迭代所得到的参数估计的情况下计算完全数据对应的对数似然函数的条件期望（E步），后用极大化对数似然函数以确定参数的值，并用于下步的迭代（M步）

## 多重插补法

由包含m个插补值的向量代替每一个缺失值，然后对新产生的m个数据集使用相同的方法处理，得到处理结果后，综合结果，最终得到对目标变量的估计

随着数据量的增大，异常值和缺失值对整体分析结果的影响会逐渐变小，因此在“大数据”模式下，数据清洗可忽略异常值和缺失值的影响，而侧重对数据结构合理性的分析

# 数据探索

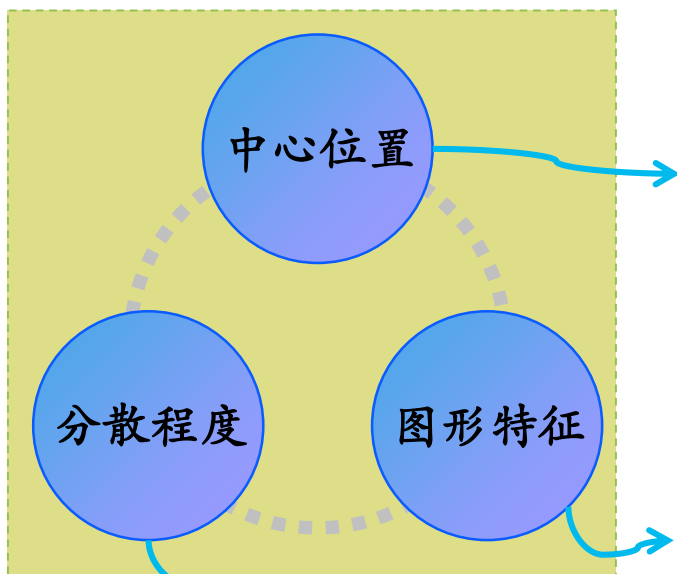
通过数据探索，初步发现数据特征、规律，为后续数据建模提供输入依据，常见的数据探索方法有数据特征描述、相关性分析、主成分分析等。

数据探索要遵循由浅入深、由易到难的步骤



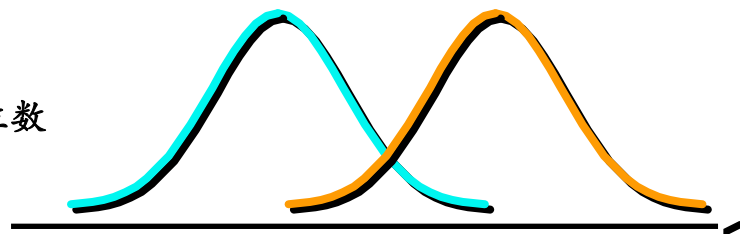


# 数据特征描述



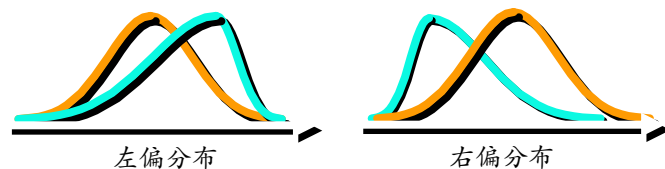
## 中心位置

- ❖ 众数
- ❖ 中位数/四分位数
- ❖ 均值

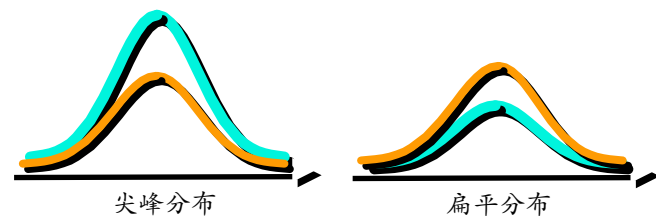


## 图形特征

- ❖ 偏度  
数据分布偏斜程度的测度

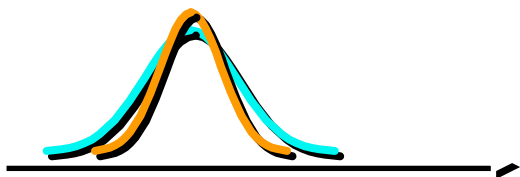


- ❖ 峰度  
数据分布扁平程度的测度



## 分散程度

- ❖ 方差和标准差
- ❖ 极差、四分位差
- ❖ 标准分数 z-score
- ❖ 离散系数



# 数据概率分布

概率分布可以表述随机变量取值的概率规律，是掌握数据变化趋势和范围的一个重要手段。

## 离散分布

### 均匀分布

离散型均匀分布是一个离散型概率分布，其中有限个数值拥有相同的概率

### 二项分布

- 1.在每次试验中只有两种可能的结果，而且是互相对立的；
- 2.每次实验是独立的，与其它各次试验结果无关；
- 3.结果事件发生的概率在整个系列试验中保持不变，则这一系列试验称为伯努力试验。

### 几何分布

以下两种离散型概率分布中的一种：

- 在伯努力试验中，得到一次成功所需要的试验次数 $X$ 。 $X$ 的值域是 $\{1, 2, 3, \dots\}$
- 在得到第一次成功之前所经历的失败次数 $Y = X - 1$ 。 $Y$ 的值域是 $\{0, 1, 2, 3, \dots\}$

### 泊松近似

泊松近似是二项分布的一种极限形式。其强调如下的试验前提：一次抽样的概率值相对很小，而抽取次数又相对很大。因此泊松分布又被称之为罕有事件分布。泊松分布指出，如果随机一次试验出现的概率为 $p$ ，那么在 $n$ 次试验中出现 $k$ 次的概率按照泊松分布应该为

### 正态分布

若随机变量 $X$ 服从一个数学期望为 $\mu$ 、方差为 $\sigma^2$ 的高斯分布，记为 $N(\mu, \sigma^2)$ 。其概率密度函数为正态分布的期望值 $\mu$ 决定了其位置，其标准差 $\sigma$ 决定了分布的幅度。因其曲线呈钟形，因此人们又经常称之为钟形曲线。我们通常所说的标准正态分布是 $\mu = 0, \sigma = 1$ 的正态分布

## 连续分布

### 均匀分布

如果连续型随机变量具有如下 $p=1/(b-a)$ 的概率密度函数,其中 $X[a, b]$ ，则称服从上的均匀分布

### 指数分布

指数分布可以用来表示独立随机事件发生的时间间隔，比如指数分布还用来描述大型复杂系统（如计算机）的平均故障间隔时间MTBF的失效分布

### 正态分布

# 数据分布初步推断

假设检验是数理统计学中根据一定假设条件由样本推断总体的一种方法，可以分为参数检验和非参数检验。

参数检验：数据的分布已知的情况下，对数据分布的参数是否落在相应范围内进行检验

检验方法名称	问题类型	假设	适用条件	抽样方法
单样本T—检验	判断一个总体平均数等于已知数	总体平均数等于A	总体服从正态分布	从总体中抽取一个样本
F—检验	判断两总体方差相等	两总体方差相等	总体服从正态分布	从两个总体中各抽取一个样本
独立样本 T—检验	判断两总体平均数相等	两总体平均数相等	1、总体服从正态分布 2、两总体方程相等	从两个总体中各抽取一个样本
配对样本T—检验	判断指标实验前后平均数相等	指标实验前后平均数相等	1、总体服从正态分布 2、两组数据是同一试验对象在试验前后的测试值	抽取一组试验对象，在试验前测得试验对象某指标的值，进行试验后再测得试验对象该指标的取值
二项分布假设检验	随机抽样实验的成功概率的检验	总体概率等于P	总体服从二项分布	从总体中抽取一个样本

非参数检验：一般是在不知道数据分布的前提下，检验数据的分布情况

检验方法名称	问题类型	假设
卡方检验	检测实际观测频数与理论频数之间是否存在差异	观测频数与理论频数无差异
K-S检验	检验变量取值是否为正态分布	服从正态分布
游程检验	检测一组观测值是否有明显变化趋势	无明显变化趋势
二项分布假设检验	通过样本数据检验样本来自的总体是否服从指定的概率为P的二项分布	服从概率为P的二项分布

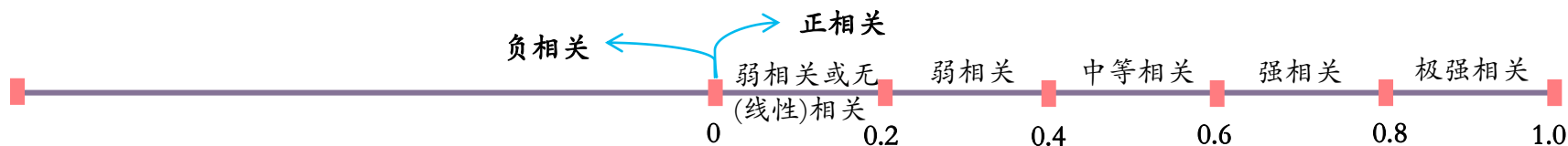
## 总结

- 1、参数检验是针对参数做的假设，非参数检验是针对总体分布情况做的假设。
- 2、二者的根本区别在于参数检验要利用到总体的信息，以总体分布和样本信息对总体参数作出推断；非参数检验不需要利用总体的信息。

# 结构优化——相关性分析

用于分析的多个变量间可能会存在较多的信息重复，若直接用来分析，会导致模型复杂，同时可能会引起模型较大误差，因此要初步探索数据间的相关性，剔除重复因素。

相关系数是考察变量之间的相关程度的变量，相关分析是优化数据结构的基础



## 二元变量相关分析

## 偏相关分析

## 距离相关分析

### Pearson相关系数

- 衡量两个变量**线性**相关性的强弱
- 在方差和协方差的基础上得到的，对异常值敏感
- 服从正态分布或接近正态的单峰分布
- 两个变量为连续数据

### Spearman秩相关系数

- 衡量两个变量之间联系（变化趋势）的强弱
- 在秩（排序）的相对大小基础上得到，对异常值更稳健
- 两个变量均为连续数据或等级数据

### Kendall相关系数

- 基于协同思想得到，衡量变量之间的协同趋势
- 对异常值稳健
- 两个变量均为连续数据或等级数据

- 研究两个变量之间线性相关关系时，控制可能对其产生影响的变量

- 对观测量之间或变量之间相似或不相似程度的一种测度

特点  
适用条件

# 结构优化——相关性分析

---

## 检验动机：

样本数据只是总体的一个实现，因此，根据现有数据计算出来的相关系数只是变量相关系数的一个观测值，又称为样本相关系数。欲根据这个样本相关系数来估计总体相关系数，必须进行显著性检验。其原假设：在总体中，两个变量的相关系数(总体相关系数)为零

## 检验意义：

计算在原假设成立的情况下(也就是在两个变量相关系数为零的情况下)，由于抽样的原因(收集样本数据的原因)得到当前的样本相关系数(可能这个系数并不为零，甚至还比较大)的概率。(p值越小说明越是小概率事件，不可能发生，拒绝原假设)

## 检验方法：

T检验（常用）：对于近似高斯分布的数据（如两个变量服从双变量正态分布），相关系数的样本分布近似地服从自由度为 $N - 2$ 的t分布；如果样本容量不是特别小（通常大于30），即使观测数据不服从正态分布，依然可使用t检验

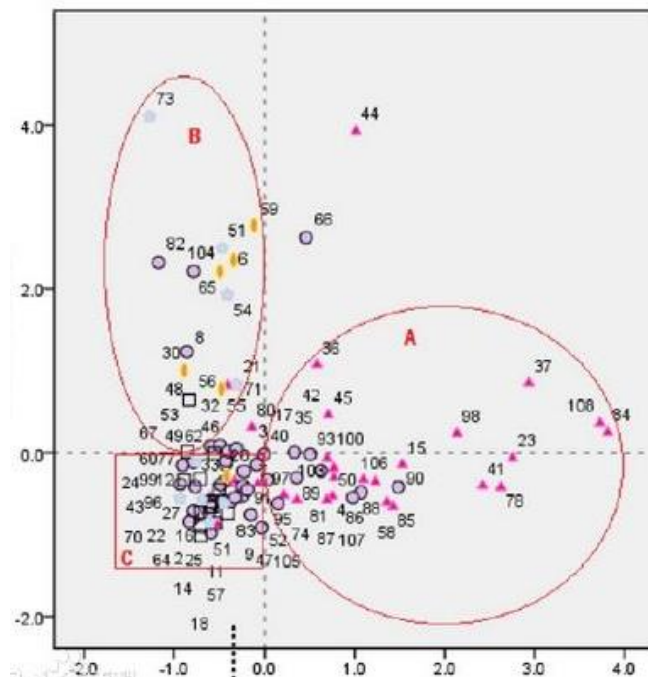
## 结构优化——主成分分析

Karl Pearson (1901) 探究如何通过少数几个主成分(principal component)来解释多个变量间的内部结构时提出主成分分析法, 旨在从原始变量中导出少数几个主分量, 使其尽可能多地保留原始变量的信息, 且彼此间互不相关

**内涵：**将彼此相关的一组指标变量转化为彼此独立的一组新的指标变量，并用其中较少的几个新指标变量就能综合反映原多个指标变量所包含主要信息的多元统计方法

**应用：**数据的压缩和解释，即常被用来寻找和简化判断事物或现象的综合指标，并对综合指标所包含的信息进行适当的解释

**原理：**设法将原来变量重新组合成一组新的互相无关的几个综合变量，同时根据实际需要从中可以取出几个较少的综合变量尽可能多地反映原来变量的信息的统计方法叫做主成分分析或称主分量分析，也是数学上用来降维的一种方法。



# 数据转换

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

## 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

$$x_i' = \begin{cases} 0(x_i < a) \\ \frac{x_i - a}{b - a}(a \leq x_i < b) \\ 1(x_i \geq b) \end{cases} \quad \text{示例}$$

## 直线法

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

## 曲线法

- Log函数法:  $x_i' = \log(x_i) / \log(\max(x_i))$
- Arctan函数法:  $x_i' = \arctan(x_i) \times 2/\pi$
- 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。



# 目录

概述

数据分析框架

数据分析方法

数据理解&数据准备

分类与回归

聚类分析

关联分析

时序模型

结构优化

数据分析支撑工具

# 分类

## 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。

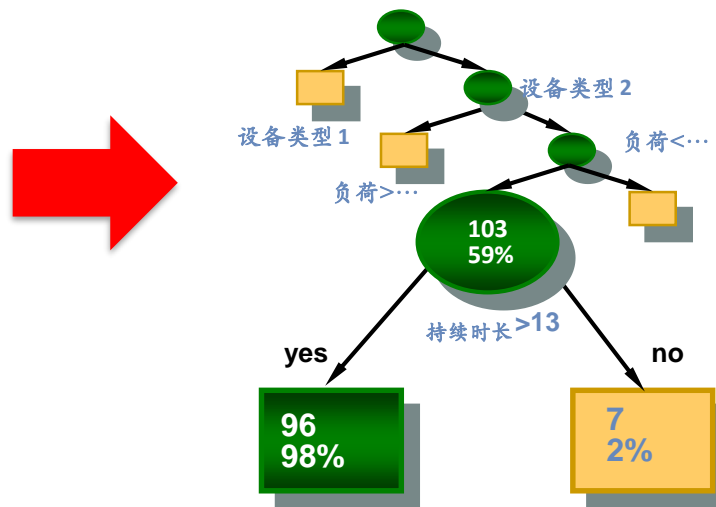
分类是利用训练数据集通过一定的算法而求得分类规则的。是模式识别的基础。

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。

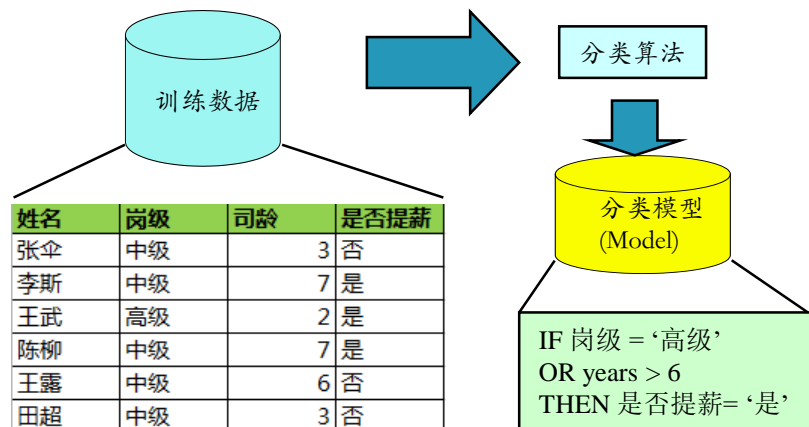


# 分类

## 分类的实现:

### 模型的构建

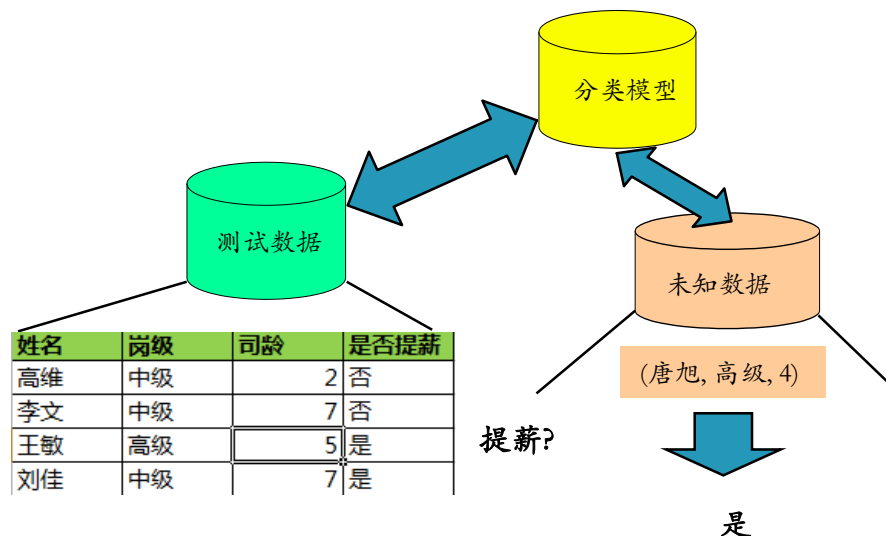
- 对每个样本进行类别标记
- 训练集构成分类模型
- 分类模型可表示为：分类规则、决策树或数学公式



### 模型的使用

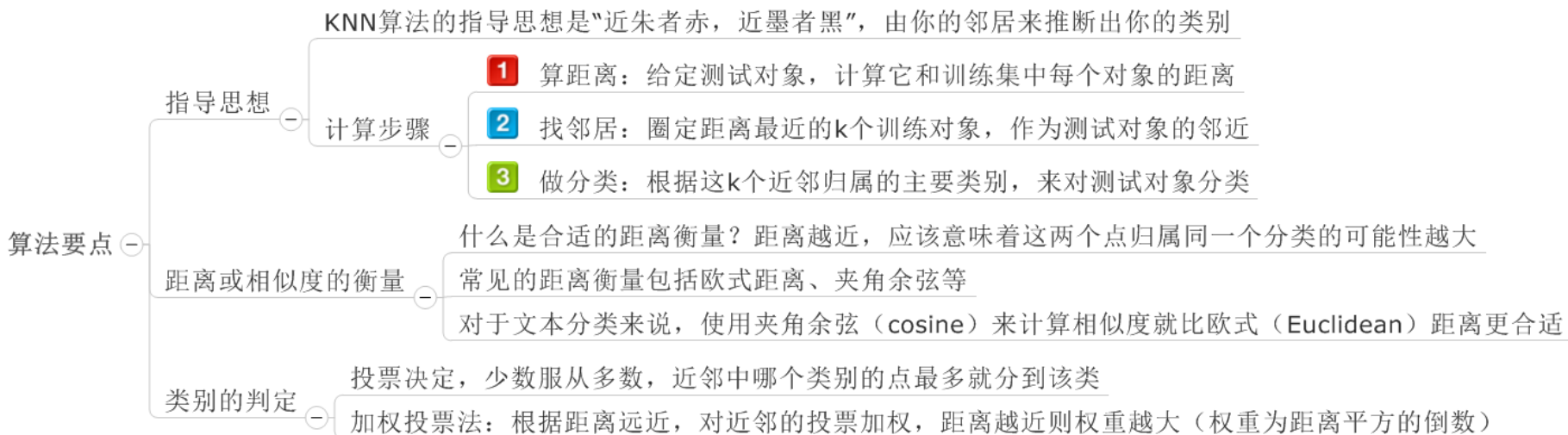
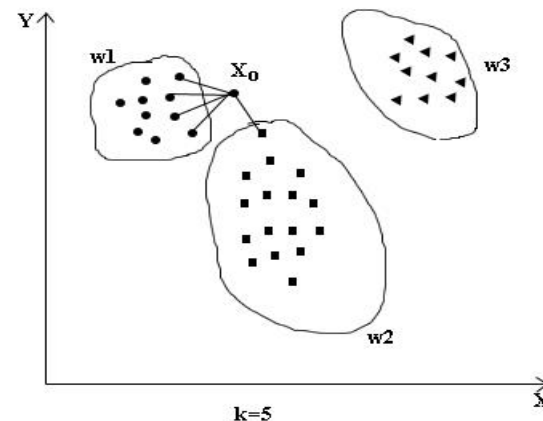
- 识别未知对象的所属类别
- 模型正确性的评价
  - ✓ 已标记分类的测试样本与模型的实际分类结果进行比较

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



# 分类

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



# 分类

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。

C4.5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）。

C4.5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。

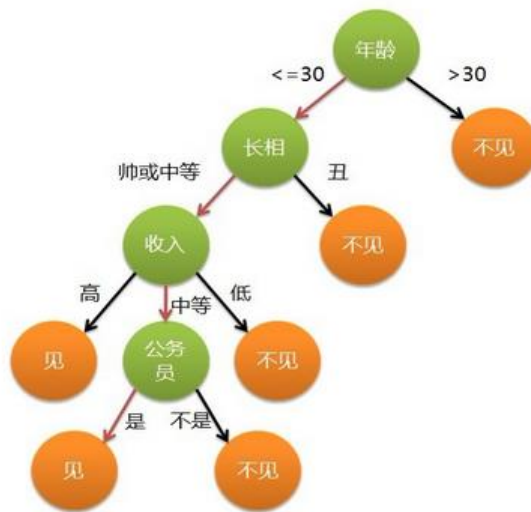
女儿：收入高不？

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

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： $X = \{x_1, x_2, \dots, x_n\}$ ，假定有 $m$ 个类，分别用 $C_1, C_2, \dots, C_m$ 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 $C_i$ ，则一定是 $P(C_i|X) > P(C_j|X) \quad 1 \leq j \leq m, j \neq i$

根据贝叶斯定理

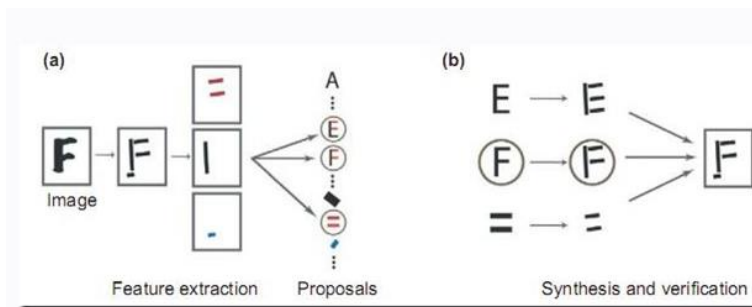
由于 $P(X)$ 对于所有类为常数，最大化后验概率 $P(C_i|X)$ 可转化为最大化先验概率 $P(X|C_i)P(C_i)$ 。如果训练数据集有许多属性和元组，计算 $P(X|C_i)$ 的开销可能非常大，为此，通常假设各属性的取值互相独立，这样先验概率 $P(x_1|C_i), P(x_2|C_i), \dots, P(x_n|C_i)$ 可以从训练数据集求得。

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 $C_i$ 的概率 $P(X|C_i)P(C_i)$ ，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

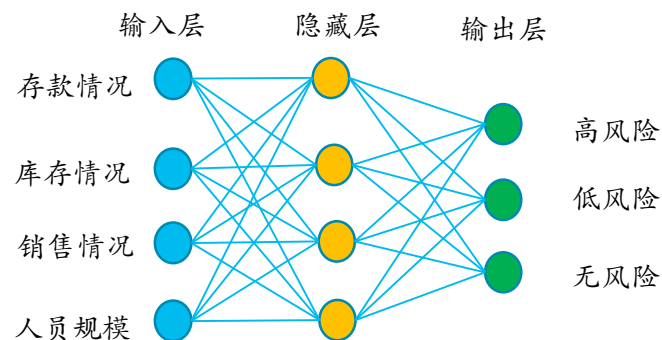
# 分类

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McClelland（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。



# 回归

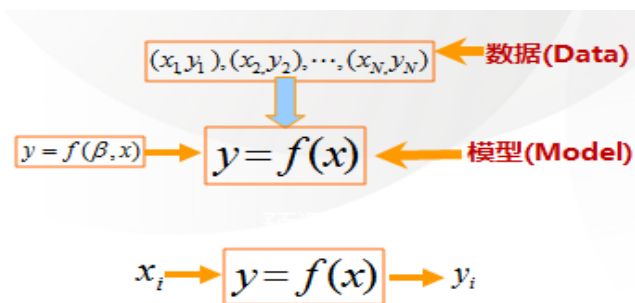
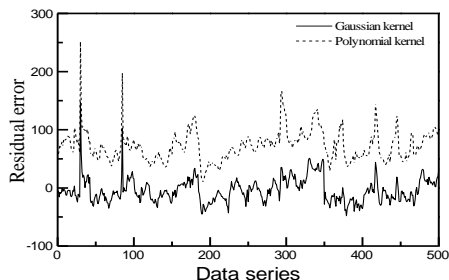
**产生：**英国统计学家F.GALTON（法兰西斯·高尔顿）(1822-1911)和其学生K.Pearson（卡尔·皮尔逊）(1856-1936)观察了1078对夫妇，以每对夫妇的平均身高为X，而取他们成年的儿子的身高为Y，得到如下经验方程：  
 $Y=33.73+0.516X$

**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。



$y$  是离散的，如 $\{-1,1\}$ ， $\{0,1,2\}$ 为分类问题

$y$  是连续值如温度，速度等为回归问题

变量间的关系

确定性关系或函数关系  $y=f(x)$

非确定性关系

人的身高和体重  
家庭的收入和消费  
商品的广告费和销售额  
粮食的产量和施肥量  
股票的价格和时间  
夏天气温与售电量...

X

实变量

非确定性关系

Y

随机变量

# 回归-线性回归

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

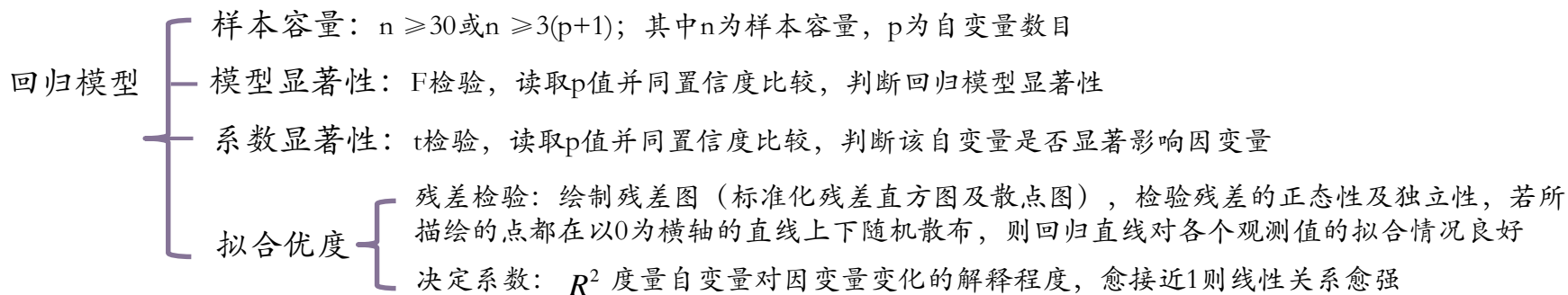
## LOGISTIC线性回归

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:



# 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

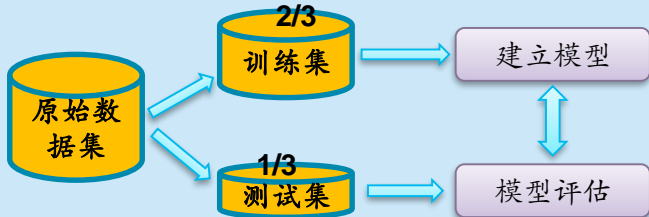
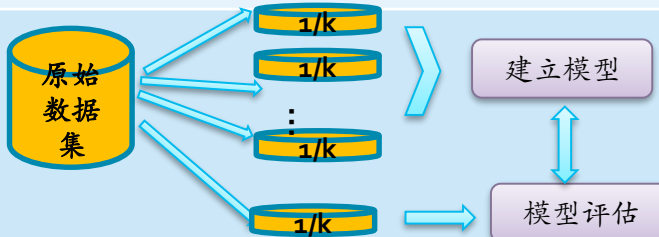
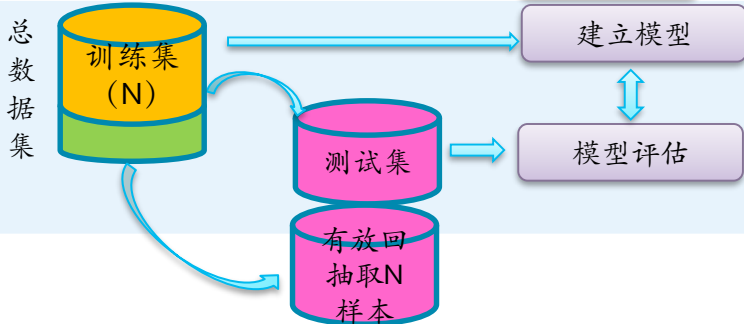
Lift图

响应率曲线

捕获率曲线/  
增益图

# 分类模型评估

## 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

## 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。

预测类 实际类	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

预测类 实际类	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%。

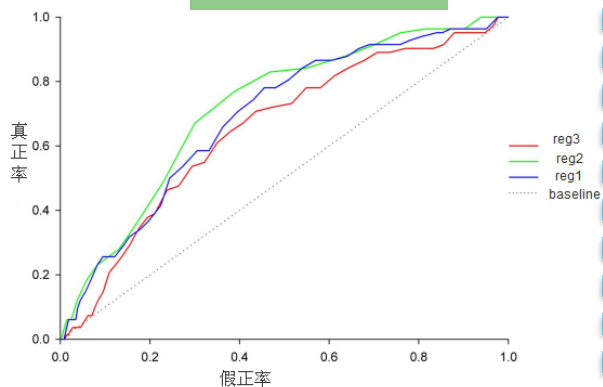
精度 $= a/(a+c)$

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。

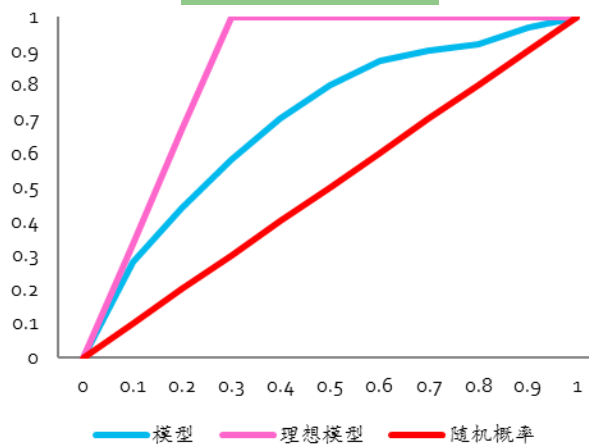
# 分类模型评估

## ROC曲线



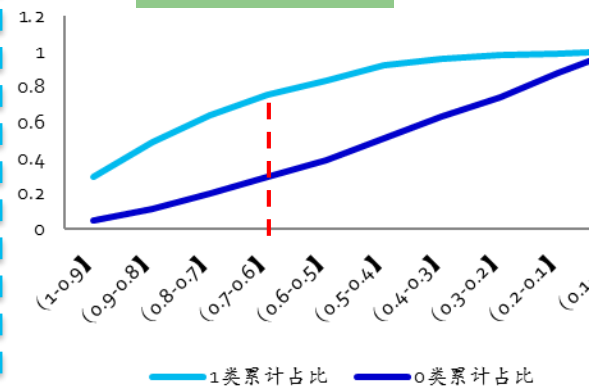
以真正率及敏感性为纵轴，假正率=1-特异性为横轴做图。给定一个二类问题，我们可以对检验集的不同部分，显示模型可以正确识别正样本的比例与模型将负样本错误标识为正样本的比例之间的比较评定。敏感性的增加以错误正例的增加为代价。

## 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。

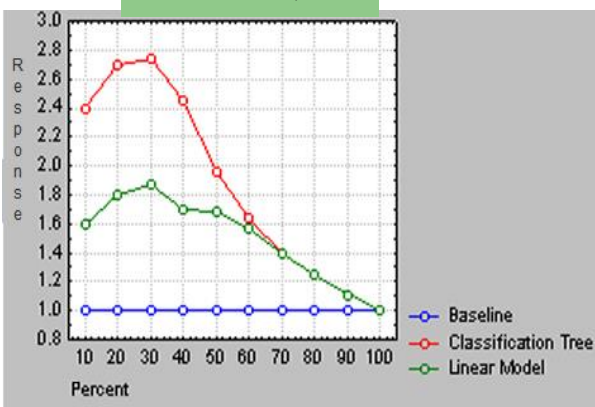
## KS曲线



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。

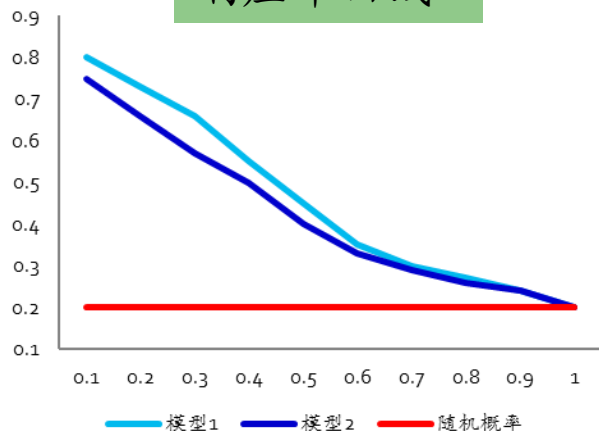
# 分类模型评估

## Lift图



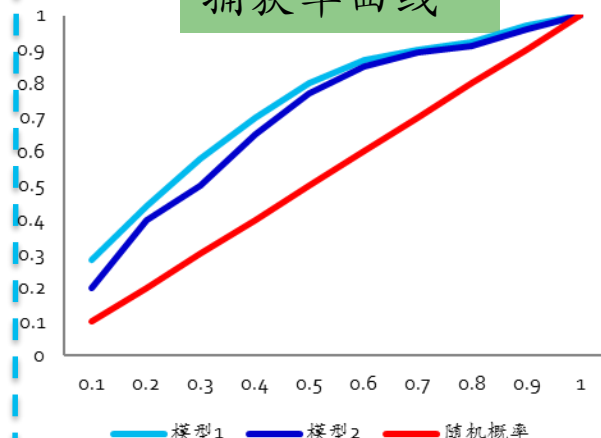
Lift值=响应率/随机概率。比如对10000名潜在客户进行概率打分，预测其购买商品的可能性，若实际中有900人会购买，则9%为随机概率。抽取概率排名前10%的人数，即1000人，预测600人购买，则前10%的响应率为600/1000=60%，则Lift值=60%/9%=6.67。

## 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。



# 目录

概述

数据分析框架

数据分析方法

数据理解&数据准备

分类与回归

聚类分析

关联分析

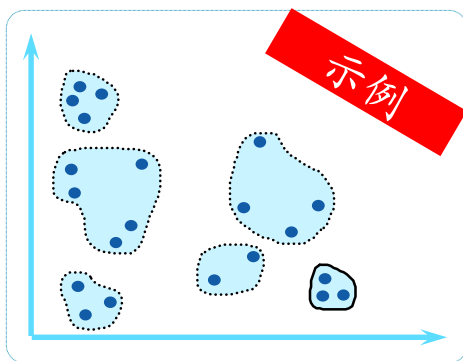
时序模型

结构优化

数据分析支撑工具

# 聚类分析

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

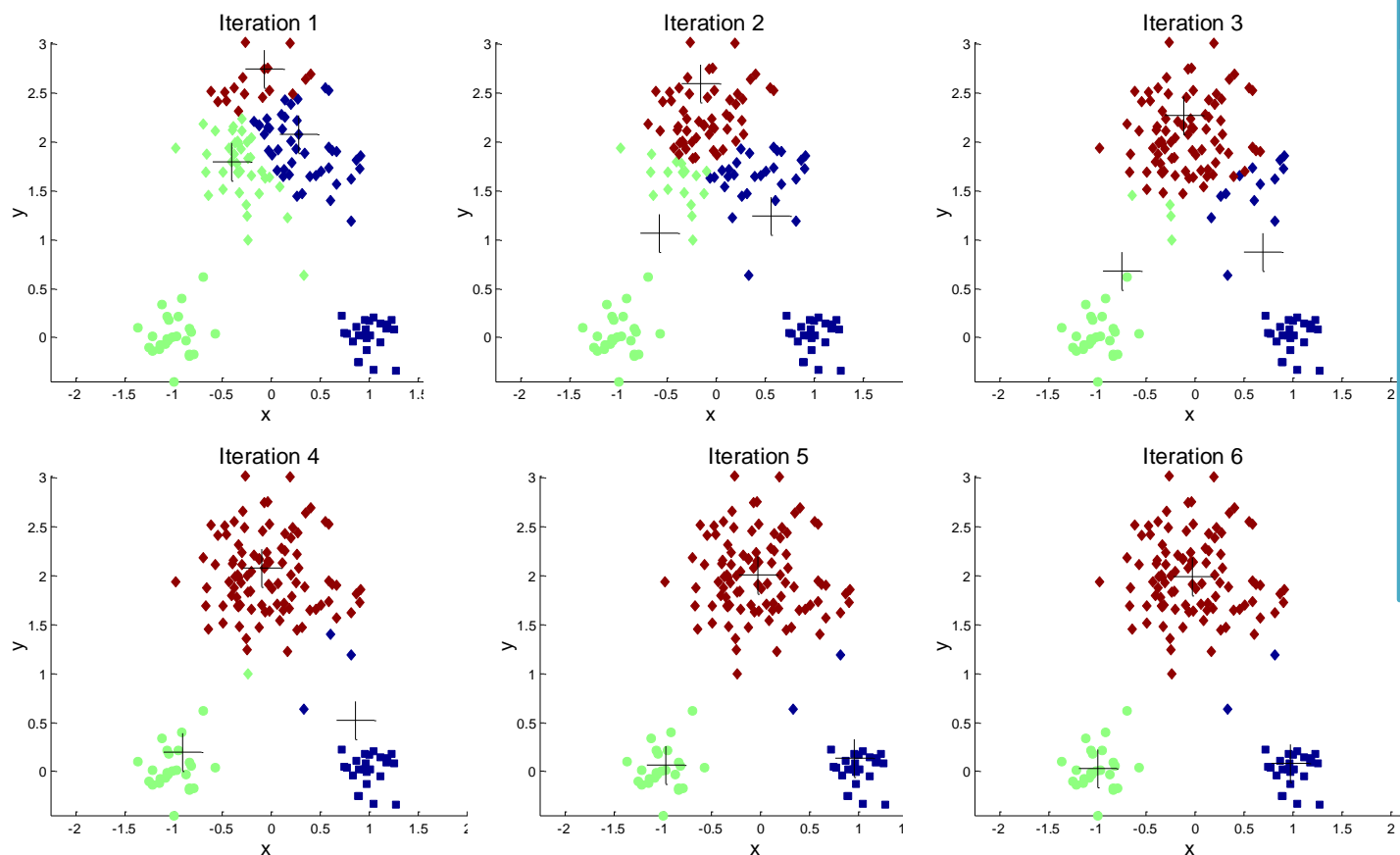
- 首先将问题空间量化为有限数目的单元，形成一个空间网格结构，随后聚类在这些网格之间进行。
- 典型算法：STING、WareCluster和CLIQUE等。

## 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。

# 聚类分析——K均值聚类

**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：


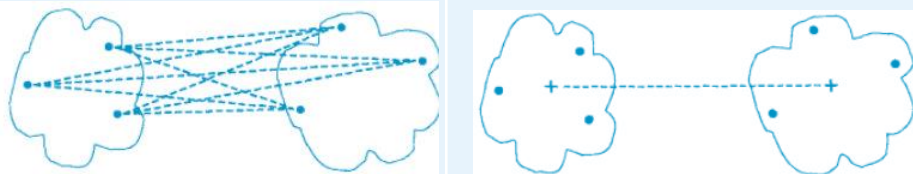
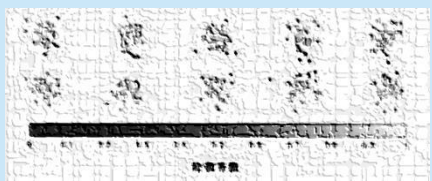
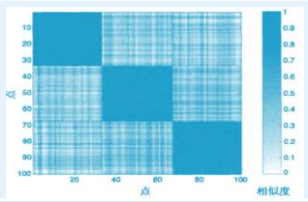
(1) 首先，随机选择 $k$ 个数据点做为聚类中心；

(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。	

# 目录

概述

数据分析框架

数据分析方法

数据理解&数据准备

分类与回归

聚类分析

关联分析

时序模型

结构优化

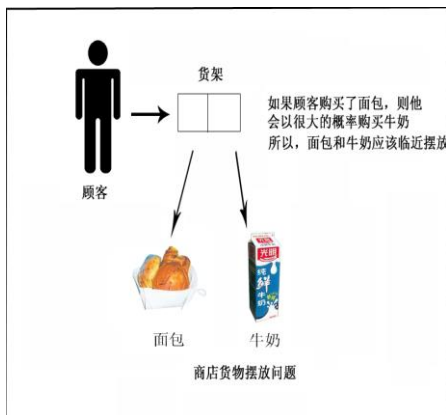
数据分析支撑工具

# 关联规则

## 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。

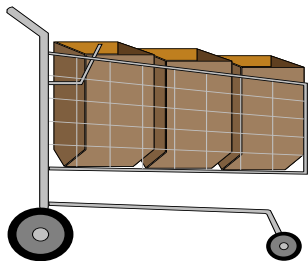


## 举例：

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

## 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

支持度（{喝茶}  $\rightarrow$  {喝咖啡}）= $150/1000=15\%$ ；  
置信度（{喝茶}  $\rightarrow$  {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}  $\rightarrow$  {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。

## 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



# 关联规则

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。

核心思想是项集的反单调性：“如果一个项集是非频繁的，那么它的超集（superset）也一定是非频繁的”

所谓频繁项集是指发生频率超过最小支持度的项集

算法要点

计算步骤

迭代

1 首先扫描交易数据库，找出项数为1的频繁项集（即频繁的单项集）此时 $k=1$

2 连接步：从 $k$ 频繁项集中生成 $k+1$ 候选频繁项集

3 扫描数据集，计算出每个候选频繁项集的支持度

4 修剪步：根据最小支持度要求，从中筛选出 $k+1$ 频繁项集

5 直到 $k+1$ 达到用户指定的最大项数，或者 $k+1$ 频繁项集为空

如果指定的最大项数为 $K_{max}$ ，则Apriori算法最多扫描数据集 $K_{max}+1$ 次

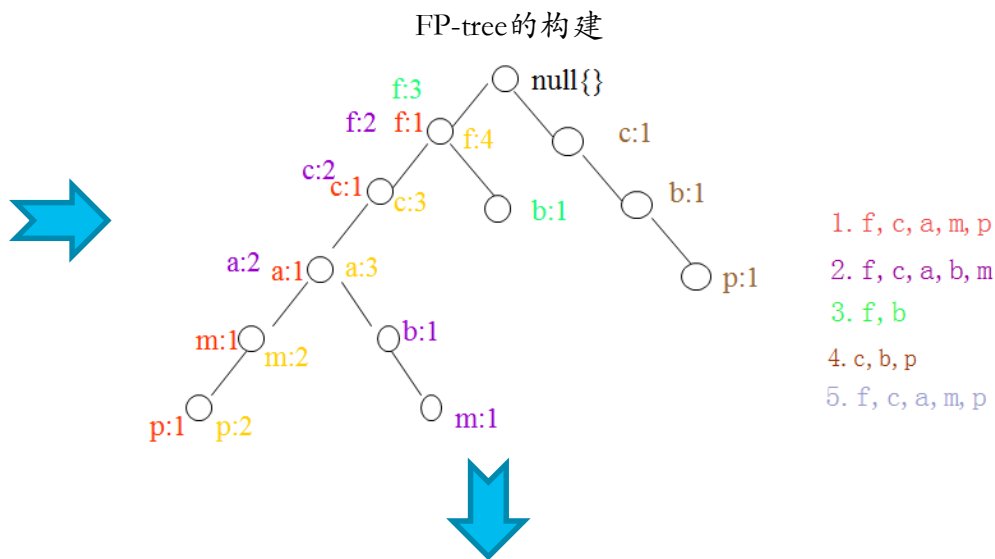
# 关联规则

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



f, c, b组合满足条件

# 关联规则模型评估

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
- .....

# 目录

概述

数据分析框架

数据分析方法

数据理解&数据准备

分类与回归

聚类分析

关联分析

时序模型

结构优化

数据分析支撑工具

# 时间序列分析

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

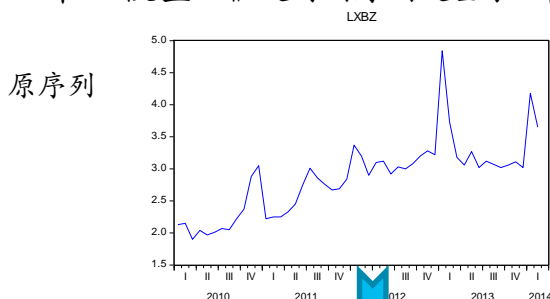
循环变动 (C) : 沿着趋势线如钟摆般地循环变动，又称景气循环变动

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

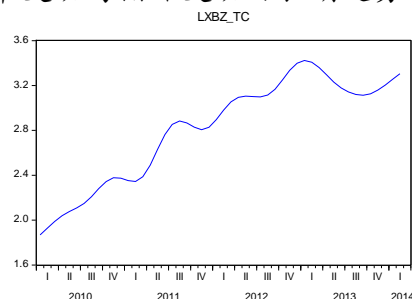
时间  
序列

组合模型：

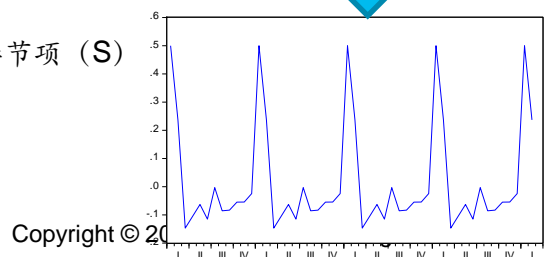
- 加法模型：假定时间序列是基于4种成份相加而成的。长期趋势并不影响季节变动； $Y=T+S+C+I$
- 乘法模型：假定时间序列是基于4种成份相乘而成的。假定季节变动与循环变动为长期趋势的函数； $Y=T \times S \times C \times I$



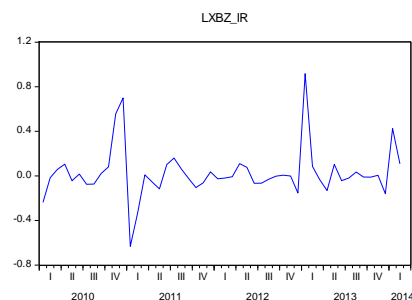
趋势循环项 (TC)



季节项 (S)



随机扰动项 (I)



# 时间序列分析

## 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

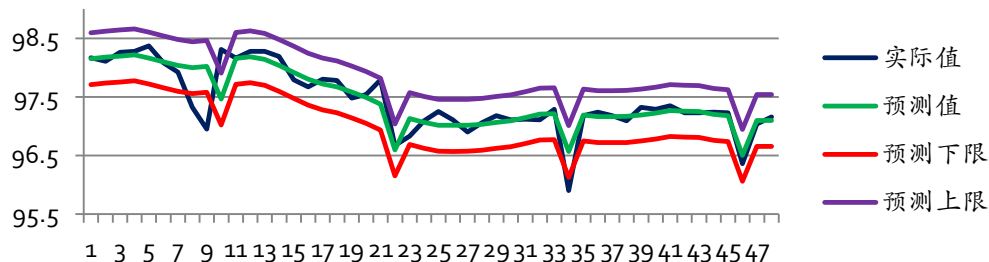
采用季节拆分建模

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

# 时间序列算法介绍

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



# 时间序列算法介绍-ARIMA

ARIMA又称自回归求积移动平均模型，是存在序列相关的非平稳时间序列建模方法。

## 建模前提：

### 1、序列平稳性

平稳序列是指均值和方差在时间过程中保持常数。非平稳时间序列要么均值随时间而变化，要么方差随时间而变化，或者二者同时在发生变化。

- 对于一个平稳的时间序列可以通过过去时间点上的信息，建立模型拟合过去信息，进而预测未来的信息。而非平稳时间序列在各个时间点上的随机规律是不同的，难以通过序列已知的信息去掌握时间序列整体上的随机性。因此，对于一个非平稳序列去建模，预测是困难的。
- 时间序列建模依赖于序列自身所表现的自相关，有时候，自相关是由于时间序列非平稳所引起的。

### 2、序列相关

如果不同的样本点之间不是完全相互独立的，而是存在某种相关性，则认为出现了序列相关性。序列相关分为自相关和偏自相关，序列相关的表现为协方差不为0

$$\text{cov}(u_t, u_{t-s}) \neq 0 \quad s \neq 0, t = 1, 2, \dots, T$$

# 时间序列算法介绍-ARIMA

- AR(p)是p阶自回归模型，AR(p)模型适用于具有p阶偏自相关的序列。

$$u_t = c + \phi_1 u_{t-1} + \phi_2 u_{t-2} + \cdots + \phi_p u_{t-p} + \varepsilon_t$$

对于AR(p)模型，求出滞后k阶自相关系数p(k)时，实际上得到并不是u(t)与u(t-k)之间单纯的相关关系。因为u(t)同时还会受到中间k-1个随机变量u(t-1)、u(t-2)、……、u(t-k+1)的影响，而这k-1个随机变量又都和x(t-k)具有相关关系，所以自相关系数p(k)里实际掺杂了其他变量对u(t)与u(t-k)的影响。

- MA(q)是q阶移动平均模型，MA(q)适用于具有q阶自相关的序列。

$$u_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

- ARMA(p,q)是p阶自回归模型和q阶移动平均模型的组合，适合于具有p阶偏自相关和q阶自相关的序列。

$$u_t = c + \phi_1 u_{t-1} + \cdots + \phi_p u_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

- ARIMA(p,d,q)是经过d次差分后满足平稳性条件后，建立ARMA(p,q)的建模方法。

因为大多数时间序列都在一定的序列相关性，使得ARIMA建模方法的预测比平滑法在应用中更为有效。

# 目录

概述

数据分析框架

数据分析方法

数据理解&数据准备

分类与回归

聚类分析

关联分析

时序模型

结构优化

数据分析支撑工具

# 结构优化-遗传算法

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。

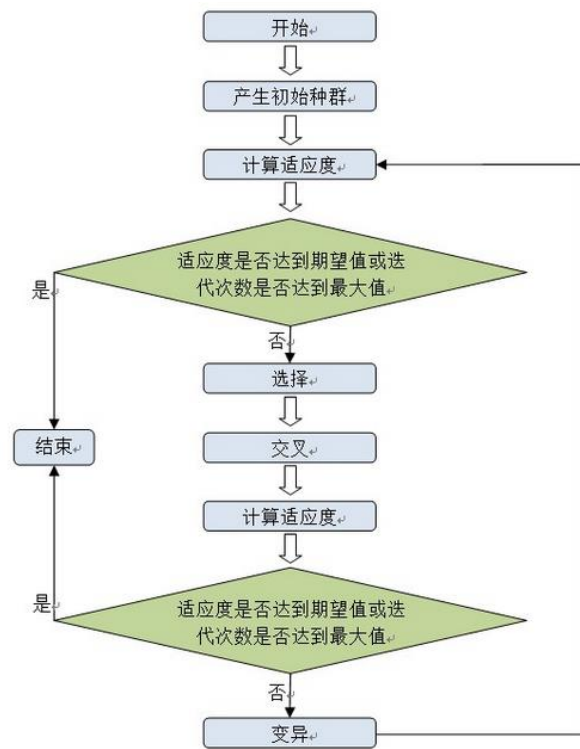
## 算法特点：

**(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

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。

# 目录

概述

数据分析框架

数据分析方法

数据理解&数据准备

分类与回归

聚类分析

关联分析

时序模型

结构优化

数据分析支撑工具

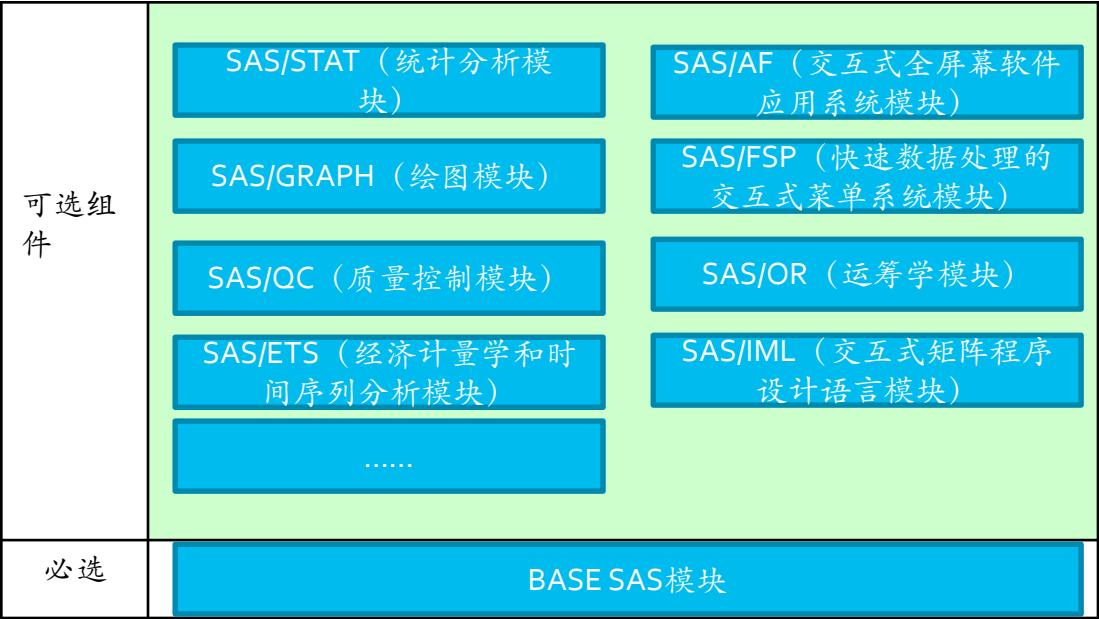
# 常用的数据分析工具

	操作		编程			
	Eviews	SPSS	SAS	Stata	Matlab	R
主导优势	时间序列分析	多元横截面数据	数据管理及挖掘	面板数据处理	数值分析, 复杂模型	算法及绘图
应用领域	经济	通信, 政府, 金融, 制造, 医药, 教育等	市场调研, 医药研发, 能源公共事业, 金融管理等	经济	建筑工程	学术研究, 医药研发, IT
处理功能	推断统计	推断及多元统计	批量数据集		统计预测, 优化建模	统计分析, 数据挖掘
界面设计	直观, 可视化	简易, 可视化	语言机械规范化	可视, 代码灵活	偏向底层	语言丰富灵活
数据安全	软件稳定	大数据易丢失	软件稳定	软件稳定	软件稳定	软件稳定
处理效率	高, 稳定	低, 不适宜大数据	高, 稳定	高, 稳定	高, 稳定	极适合大量数据
结合形式	Excel, SAS, SPSS	Excel	Excel, txt,	txt	All	All



# 数据挖掘工具-SAS

SAS 系统全称为Statistics Analysis System，最早由北卡罗来纳大学的两位生物统计学研究生编制，并于1976年成立了SAS软件研究所，正式推出了SAS软件。SAS是用于决策支持的大型集成信息系统，SAS 是由大型机系统发展而来，其核心操作方式就是程序驱动，经过多年的发展，现在已成为一套完整的计算机语言，其用户界面也充分体现了这一特点：它采用MDI（多文档界面），用户在PGM视窗中输入程序，分析结果以文本的形式在OUTPUT视窗中输出。使用程序方式，用户可以完成所有需要做的工作，包括统计分析、预测、建模和模拟抽样等。但是，这使得初学者在使用SAS时必须学习SAS语言，入门比较困难。

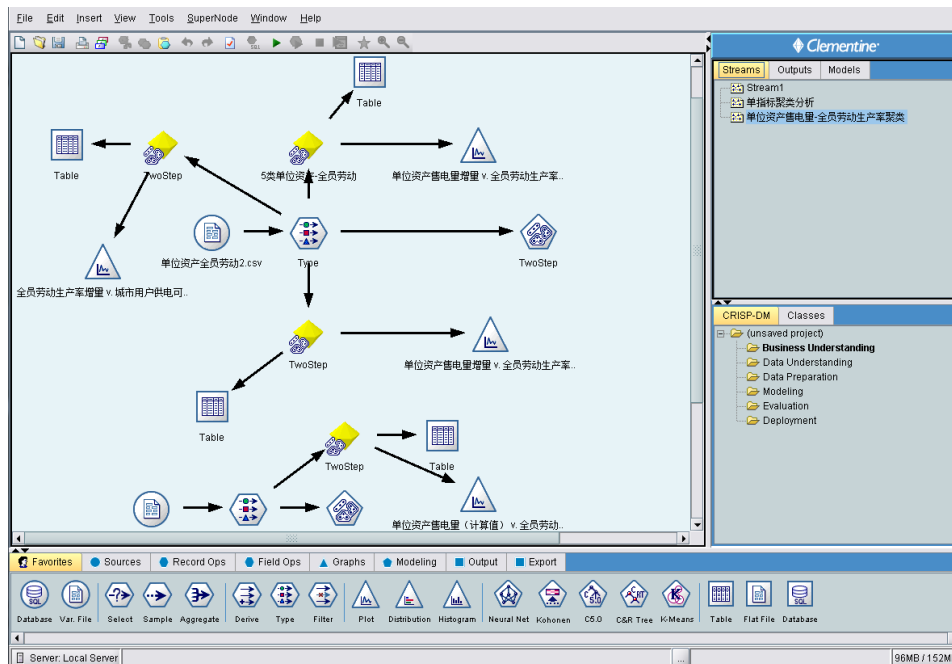


目前SAS已在全球100多个国家和地区拥有29000多个客户群，直接用户超过300万人。在我国，国家信息中心，国家统计局，卫生部，中国科学院等都是SAS系统的大用户。SAS已被广泛应用于政府行政管理，科研，教育，生产和金融等不同领域，并且发挥着愈来愈重要的作用。

# 数据挖掘工具- SPSS Clementine (现已更名为: PASW Modeler)

Clementine是ISL(Integral Solutions Limited)公司开发的数据挖掘工具平台。1999年SPSS公司收购了ISL公司,对Clementine产品进行重新整合和开发,现在Clementine已经成为SPSS公司的又一亮点。

Clementine的图形化操作界面,使得分析人员能够可视化数据挖掘过程的每一步。通过与数据流的交互,分析人员和业务人员可以合作,将业务知识融入到数据挖掘过程中。这样数据挖掘人员就可以把注意力集中于知识发现,而不是陷入技术任务,例如写代码,所以他们可以尝试更多的分析思路,更深入地探索数据,揭示更多的隐含关系。



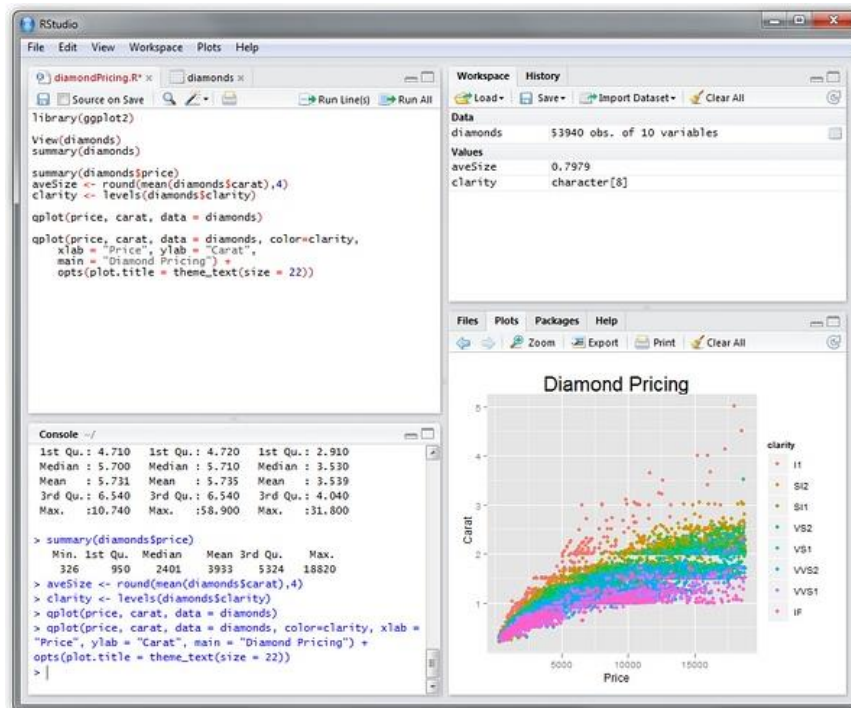
使用全面整合到Clementine的Text Mining,您可以从任何类型的文本—例如内部报告、呼叫中心记录、客户的邮件、媒体或者杂志文章、博客等中抽取内容和评论。使用WebMining for Clementine®,您可以发现访问者网上行为模式。直接获取Dimension产品的调查数据,您可以把人口统计信息、态度和行为信息用于模型—更深入地理解您的客户。Clementine还提供大量的应用模板:

- CRM CAT--针对客户的获取和增长,提高反馈率并减少客户流失;
- Web CAT--点击顺序分析和访问行为分析;
- cTelco CAT--客户保持和增加交叉销售;
- Crime CAT--犯罪分析及其特征描述,确定事故高发区,联合研究相关犯罪行为;
- Fraud CAT--发现金融交易和索赔中的欺诈和异常行为;
- Microarray CAT--研究和疾病相关的基因序列并找到治愈手段

# 数据挖掘工具- R

R语言，一种自由软件编程语言与操作环境，主要用于统计分析、绘图、数据挖掘。R本来是由来自新西兰奥克兰大学的Ross Ihaka和Robert Gentleman。R主要是以命令行操作，同时有人开发了几种图形用户界面。开发（也因此称为R），现在由“R开发核心团队”负责开发。

- R内置多种统计学及数字分析功能。R的功能也可以通过安装包（Packages，用户撰写的功能）增强。因为S的血缘，R比其他统计学或数学专用的编程语言有更强的面向对象（面向对象程序设计）功能。
- R的另一强项是绘图功能，制图具有印刷的素质，也可加入数学符号。
- 虽然R主要用于统计分析或者开发统计相关的软件，但也有人用作矩阵计算。其分析速度可媲美专用于矩阵计算的自由软件GNU Octave和商业软件MATLAB。

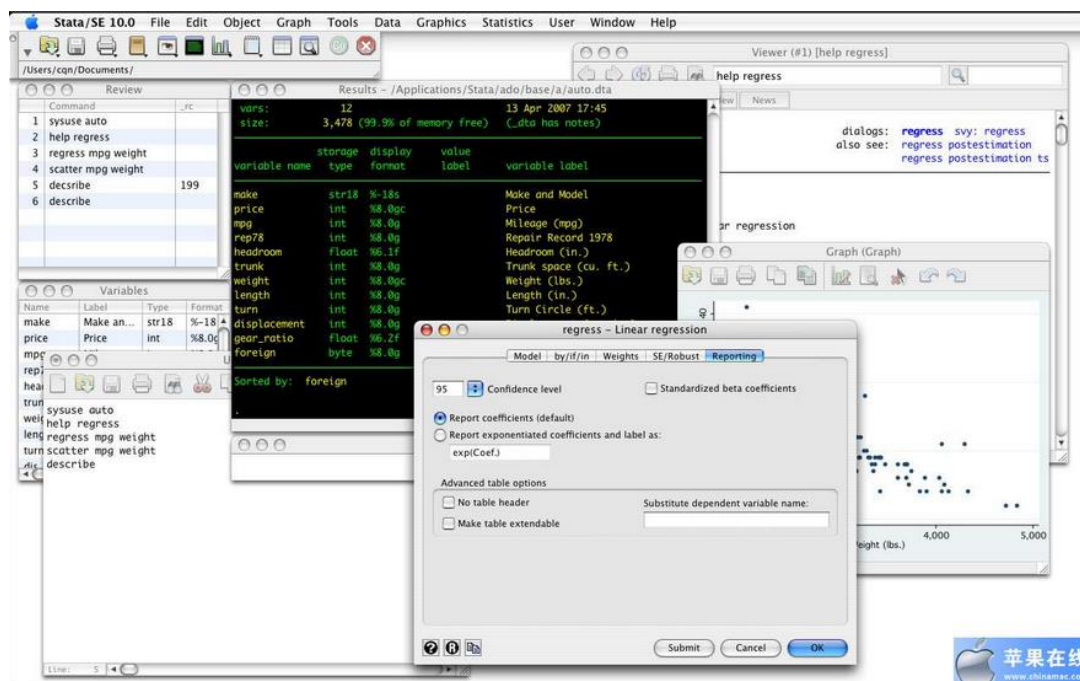


# 数据挖掘工具- Stata

Stata是Statacorp于1985年开发出来的统计程序，在全球范围内被广泛应用于企业和学术机构中。许多使用者工作在研究领域，特别是在经济学、社会学、政治学及流行病学领域。

作为一个小型的统计软件，其统计分析能力远远超过了SPSS，在许多方面也超过了SAS！由于Stata在分析时是将数据全部读入内存，在计算全部完成后才和磁盘交换数据，因此计算速度极快（一般来说，SAS的运算速度要比SPSS至少快一个数量级，而Stata的某些模块和执行同样功能的SAS模块比，其速度又比SAS快将近一个数量级！）Stata也是采用命令行方式来操作，但使用上远比SAS简单。其生存数据分析、纵向数据（重复测量数据）分析等模块的功能甚至超过了SAS。用Stata绘制的统计图形相当精美，很有特色。在长远趋势上，Stata有超越SAS的可能（据消息灵通人士透露：在SAS的老家——北卡，真正搞生物统计的人青睐的反而是Stata！）

Stata最大的缺点应该是数据接口太简单，实际上只能读入文本格式的数据文件；其数据管理界面也过于单调



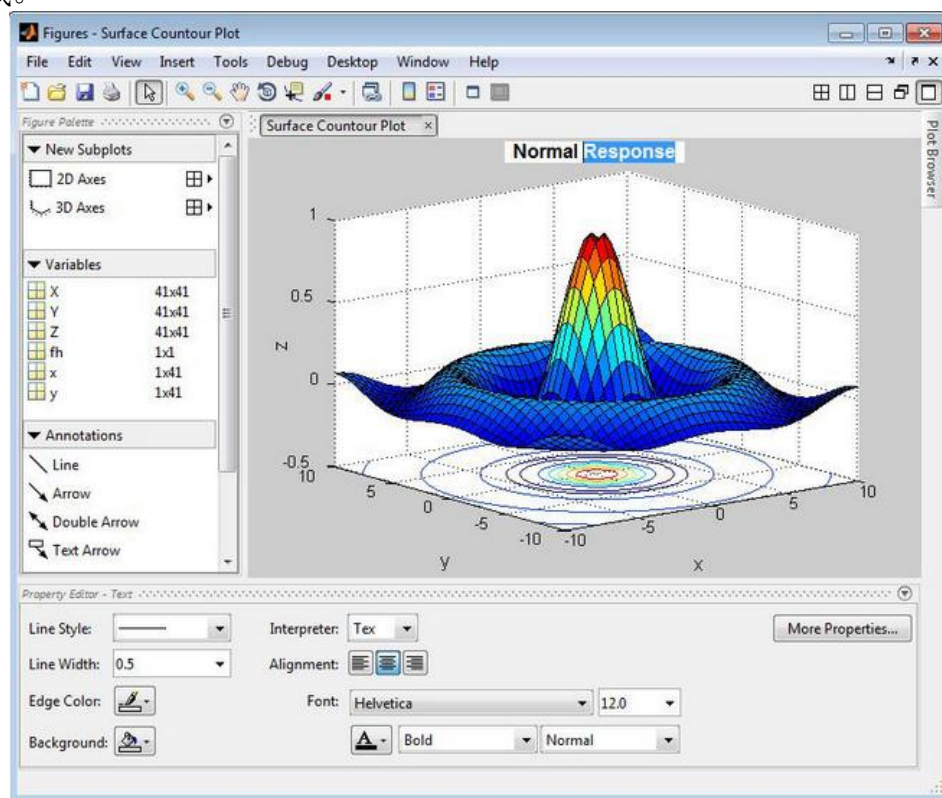
# 数据挖掘工具- MATLAB

MATLAB（矩阵实验室）是MATrix LABoratory的缩写，是一款由美国The MathWorks公司出品的商业数学软件。MATLAB是一种用于算法开发、数据可视化、数据分析以及数值计算的高级技术计算语言和交互式环境。除了矩阵运算、绘制函数/数据图像等常用功能外，MATLAB还可以用来创建用户界面及与调用其它语言（包括C，C++和FORTRAN）编写的程序。

MATLAB和Mathematica、Maple并称为三大数学软件。它在数学类科技应用软件中在数值计算方面首屈一指。MATLAB可以进行矩阵运算、绘制函数和数据、实现算法、创建用户界面、连接其他编程语言的程序等，主要应用于工程计算、控制设计、信号处理与通讯、图像处理、信号检测、金融建模设计与分析等领域。

软件特点：

- 1) 高效的数值计算及符号计算功能，能使用户从繁杂的数学运算分析中解脱出来；
- 2) 具有完备的图形处理功能,实现计算结果和编程的可视化;
- 3) 友好的用户界面及接近数学表达式的自然化语言，使学者易于学习和掌握；
- 4) 功能丰富的应用工具箱(如信号处理工具箱、通信工具箱等),为用户提供了大量方便实用的处理工具



# 数据挖掘工具- 其他

## EViews

是美国GMS公司1981年发行第1版的Micro TSP的Windows版本，通常称为计量经济学软件包。EViews是Econometrics Views的缩写，它的本意是对社会经济关系与经济活动的数量规律，采用计量经济学方法与技术进行“观察”。计量经济学的核心是设计模型、收集资料、估计模型、检验模型、运用模型进行预测、求解模型和运用模型。正是由于EViews等计量经济学软件包的出现，使计量经济学取得了长足的进步，发展成为实用与严谨的经济学科。使用 EViews软件包可以对时间序列和非时间序列的数据进行分析，建立序列（变量）间的统计关系式，并用该关系式进行预测、模拟等等。

## Minitab

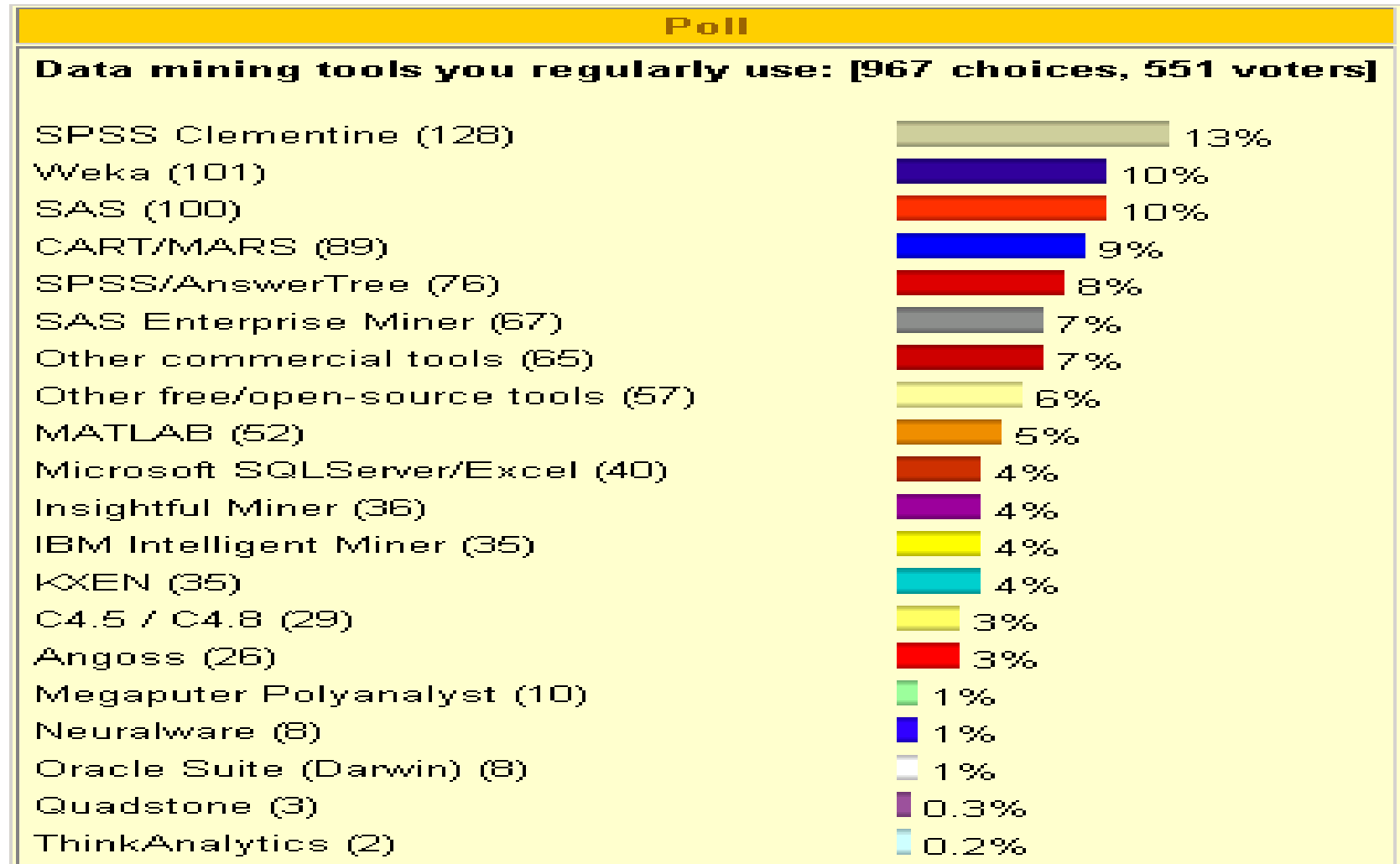
同样是国际上流行的一个统计软件包，其特点是简单易懂，在国外大学统计学系开设的统计软件课程中，Minitab与SAS、BMDP并列，根本没有 SPSS的份，甚至有的学术研究机构专门教授Minitab之概念及其使用。MiniTab for Windows统计软件比SAS、SPSS等小得多，但其功能并不弱，特别是它的试验设计及质量控制等功能。MiniTab提供了对存储在二维工作表中的数据进行分析的多种功能，包括：基本统计分析、回归分析、方差分析、多元分析、非参数分析、时间序列分析、试验设计、质量控制、模拟、绘制高质量三维图形等，从功能来看，Minitab除各种统计模型外，还具有许多统计软件不具备的功能——矩阵运算。

## WEKA

WEKA的全名是怀卡托智能分析环境（Waikato Environment for Knowledge Analysis），同时weka也是新西兰的一种鸟名，而WEKA的主要开发者来自新西兰。WEKA作为一个公开的数据挖掘工作平台，集合了大量能承担数据挖掘任务的机器学习算法，包括对数据进行预处理，分类，回归、聚类、关联规则以及在新的交互式界面上的可视化。



# 数据挖掘的工具及软件





# 数据分析发展历程

## 从数据到信息的进化

