1.绪论

- **数据生命周期：**数据获取->数据处理->数据管理->数据分析->数据挖掘->数据应用
- **数据是什么：**对客观事物的符号表示。是指那些未经过加工的事实或是着重对一种特定现象的客观描述。是客观事物性质、属性、位置以及相互关系的抽象表示。是指某一目标定性、定量描述的原始资料，包括数值、文字、符号、图形、图像以及它们能转换的数据等形式。
- **数据集类型：**记录；有序集；图和网络；空间、图像、多媒体数据
- **信息是什么：**信息是人们或者机器提供的关于现实世界新的事实的知识，是数据、消息中包含的意义，它不随载体的物理形式改变而改变。(信息是不确定性的消除)
- **数据与信息：**信息是有用的、经过加工的数据，信息是数据的内涵；数据是信息的表达，即数据是信息的载体。信息与数据不可分离。
- **空间数据：**指以地球表面空间位置为参照的自然、社会和人文经济景观数据，可以是图形、图像、文字表格和数字等，它所表达的信息是空间信息，反映了空间实体的位置及其各种相关附加属性的性质、关系、变化趋势和传播特性等的总和。
- **空间信息：**有关空间实体的性质、特征和运动状态的表征和一切有用的知识，是对空间数据的解释
- **地理学第一定律：**事物之间普遍存在联系，越是临近的事物关联性越强
- **第二空间特性---空间多样性：**不同的地方，地理数据的变化趋势是不一样的
- **数据挖掘：**从大量的数据中挖掘有趣的、有用的、隐含的、先前未知的和可能有用的模式或知识
- **四种主要技术促使数据挖掘的发展：**
  ① 超大规模的数据库
  ② 先进的计算机技术（以及庞大的算力）
  ③ 海量数据的快速访问
  ④ 统计方法的不断革新
- **数据挖掘的主要过程：**
  ① 数据清洗（data cleaning）：清除数据中的噪声和与挖掘主题明显无关的数据
  ② 数据集成（data integration）：将来自多数据源中的相关数据组合在一起
  ③ 数据转换（data transformation）：将数据转换为易于进行数据挖掘的数据存储形式
  ④ 数据挖掘（data mining）：利用智能方法挖掘数据模式或规律知识
  ⑤ 模式评估（pattern evaluation）：根据一定的评估标准从挖掘结果中筛选出有意义的相关知识
  ⑥ 知识表示（knowledge presentation）：利用可视化和知识表达技术向用户展示所挖掘的相关知识

2.数据

(1) 空间数据

- **空间数据的特征：**
  ① 空间特征：又称为几何特征，表示现象的空间位置或现在所处的地理位置。一般以坐标数据表示。

② 属性特征：表示实际现象或特征，例如变量、级别、数量特征和名称等等
③ 时间特征：指现象或物体随时间的变化，往往具有周期性变化

- **空间特征：**
  是空间数据区别于其他数据的根本特征；
  是由于地物或现象的空间分布所带来的；
  通常是通过特定空间参照系下的坐标直接表达；
  基于坐标的派生数据：面积、周长、质心、拓扑关系、方位关系等；
- **属性特征：**
  地物所固有的特征；
  专题属性特征通常以数字、符号、文本和图像的形式来表示；
  这类特征在其他类型的信息系统中均可存储和处理；
- **时间特征：**
  地物的生命周期、地物的移动、属性的时效性

• 尺度：研究某一现象或事件时采用的空间或时间单位。尺度与地理细节相关，是认识地理对象与地理空间的基础。不同尺度下，统一地理实体、现象可能有不同的表达形式、变化规律

• 可变面积单元 MAUP：由于空间单元大小和形状不同，会导致分析结果的不稳定性和偏差。

• 生态学谬误：由于生态学研究是由各个不同情况的个体“集合”而成的群体(组)为观察和分析的单位，以及存在的混杂因素等原因而造成的研究结果与真实情况不符（用整体去推断个体，认为个体具有整体的特征、属性）

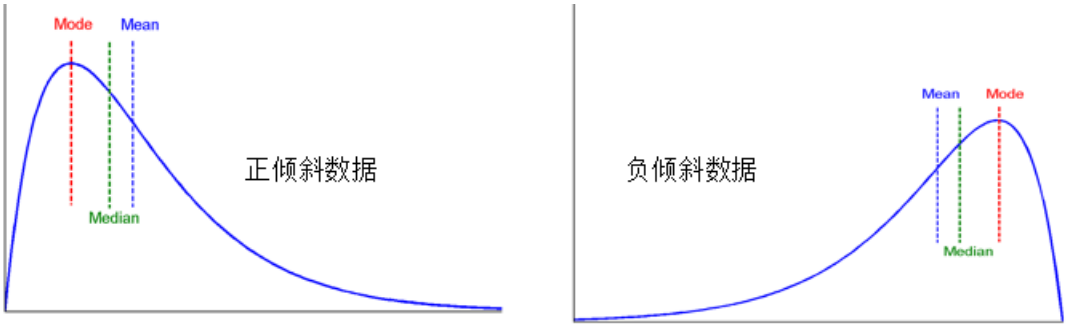
（2）数据对象与属性类型

- **数据对象：**
  数据集由数据对象组成，一个数据对象代表一个实体
  数据对象又称样本、实例、数据点、对象、数据元组
  数据对象由属性所描述

• 属性：属性（或纬度、特征、变量）是一个数据字段，表示数据对象的一个特征
  标称属性：类别、状态、事务的名称
  二元属性：只有两种状态的标称属性
  序数属性：值间具有有意义的排序
  区间标度属性：用相等的单位尺度度量，值有序，没有真正的零点
  比率标度属性：具有固定零点的数值属性
  离散属性：栅格
  连续属性：矢量

（3）属性的基本统计描述

- **集中趋势：**
  - 均值（Mean）：算术平均值、加权平均值、截尾平均值（去掉高低极端数据）
  - 中位数（Median）：有序集的中间值或中间两个值的平均
  - 众数（Mode）：集合中出现频率最高的值



- **离散趋势：**
  - 分位数：将数据从小到大排列，切分成等份的数据点
    - 四分位数：Q1=(n+1)*0.25;Q2=(n+1)*0.5;Q3=(n+1)*0.75
    - 四分位数极差（IQR）：Q3-Q1
    - 极差（range）：数据集中最大值和最小值的差
    - 盒图（五数概括：min、Q1、median、Q3、max）
    - 离群点：挑出落在至少高于第三个四分位数或低于第一个四分位数 1.5×IQR 处的值
  - 方差和标准差：衡量数据分布的离散程度

$$s^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n}[\sum_{i=1}^{n}x_i^2-\frac{1}{n}(\sum_{i=1}^{n}x_i)^2]$$

- **数据分布：**
  - 正态分布：中心极限定理说明，在适当的条件下，大量相互独立随机变量的均值经适当标准化后依分布收敛于正态分布，其中有三个要素：采样随机、因素独立与因素对结果的影响为相加

假设分布的均值为μ，方差为σ^2

- 在$\mu-\sigma$到$\mu+\sigma$的区间：大概68%的观测值落在该区间
- 在$\mu-2\sigma$到$\mu+2\sigma$的区间：大概95%的观测值落在该区间
- 在$\mu-3\sigma$到$\mu+3\sigma$的区间：大概99.7%的观测值落在该区间

  - 偏态系数：是数据平均值偏离状态的一种衡量

    一个对称的分布其中位数和均值应该接近或者相等。如果一个分布中位数和均值差的比较多，这样的分布就是有偏态的分布；

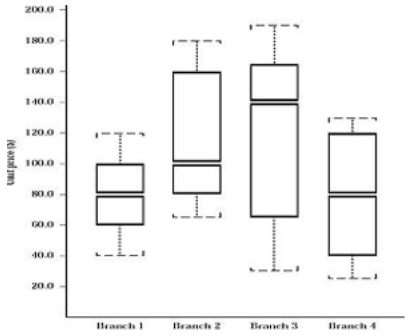
    如果偏态系数为正，就是正偏，峰值右偏；

    如果偏态系数为负，就是负偏，峰值左偏；

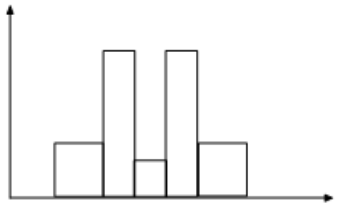
$$S=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{\frac{3}{2}}}$$

  - 峰态系数：峰态系数是数据分布集中强度的衡量。峰态系数越大，数据分布就越尖锐，峰态系数越小其分布就越缓
  - 对数正态分布：不是由事件相加决定，而是由事件相乘决定。表现为先急剧增加再缓慢减小
  - 幂率分布：不独立事件共同作用，有大量的极端事件。变现为从高处急剧下降后缓慢下降，长尾巴比对正态分布更长

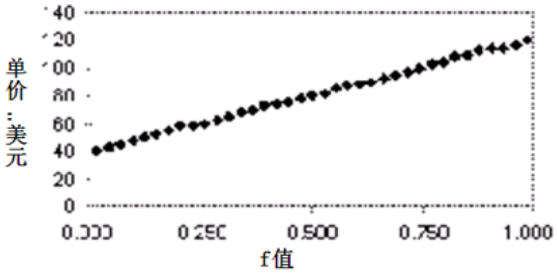
- **数据基本统计的图形显示：**
  - 盒图：五数概括



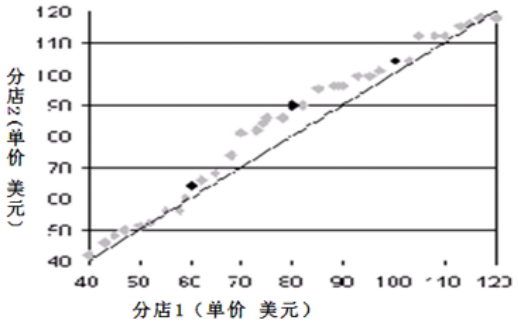
• 直方图：x 轴为值，y 轴为值出现的概率



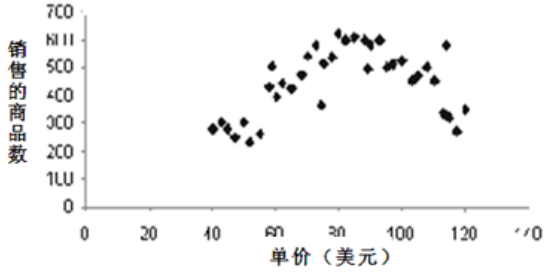
• 分位数图：每个值 xi 都有对应的值 fi，表示大概有 100*fi%的值是小于 xi 的



• 分位数-分位数图（q-q 图）：对着另一个单变量的分位数，绘制一个单变量分布的分位数



• 散点图：每个值对应一个代数坐标对，并作为一个点画在平面上



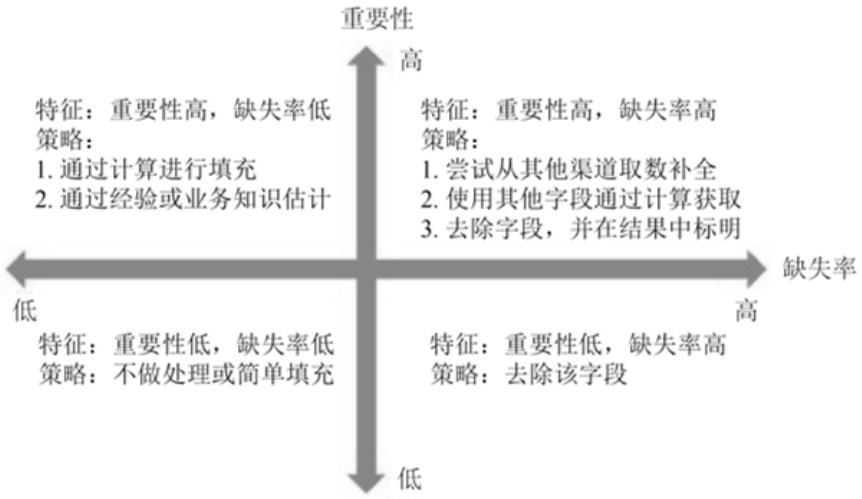
3.数据预处理

（1）数据预处理概述

- **为什么进行数据预处理**：现实世界数据是肮脏的（不完整、有噪声、数据不一致）
- **数据预处理的主要任务**：
  - 数据清理：填写缺失值、光滑噪声数据、识别并删除奇群点、解决不一致性
  - 数据集成：集成多个数据库、数据立方体或文件
  - 数据规约：维规约、数量规约、数据压缩
  - 数据变换：规范化、概念分层

（2）数据清理

- **数据清理的主要任务**：填写缺失值、光滑噪声数据、识别删除离群点、解决数据不一致性
- **如何处理缺失值**：
  ① 忽略元组（忽略有缺失值的这行数据）：当缺失较多时就不适用
  ② 人工填写：工作量大
  ③ 自动填写（填全局变量 none、填均值、填同类均值、使用回归等方法填写最可能值）



- **噪声**：一个测量变量中的随机错误或偏差
- **如何处理噪声**：
  ① 分箱：对数据进行排序，然后将他们分到等深的箱子当中，然后按照每个箱子的均值/边界/中值来对箱子内的数据进行平滑
  ② 回归：让数据适应回归函数来平滑数据
  ③ 聚类：通过聚类来检测并删除噪声
- **识别或删除离群点**：

检测方法	方法描述	方法评估
基于密度	将离群点视为在低密度区域中的对象	给出了对象是离群点的定量度量,即使数据具有不同区域也能很好处理,大数据集不适用
基于统计	需要构建一个概率分布模型,并计算对象符合该模型的概率,把具有低概率的对象视为离群点	前提是必须知道数据集服从什么分布,对于高维数据,检验效果可能很差
基于聚类	一种方法是丢弃远离其他簇的小簇,另一种更系统的方法,首先聚类所有对象,然后评估对象属于簇的程度(离群点得分)	可能是高度有效的;聚类算法产生的簇的质量对该算法产生的离群点的质量影响非常大.
基于邻近度	通过定义邻近性度量,把远离大部分点的对象视为离群点	简单,二维或三维数据可以做散点图观察,大数据集不适用,对于选择敏感,具有全局阈值,不能处理具有不同密度区域数据集

- **解决数据中的不一致性：**
  ① 格式内容清洗：时间日期格式、字符格式、不合理数据、矛盾内容
  ② 数据检验：元数据、检查规则

（3）数据集成

- **数据集成：**将不同来源、不同格式、不同特点的数据在逻辑上或物理上整绘
- **数据集成步骤：**
  ① 数据集成：将多个数据源中的数据整合到一个一致的存储当中
  ② 模式集成：整合不同数据源的原数据
  ③ 实体识别问题：匹配来自不同数据源的现实世界的实体
  ④ 检测并解决数据值的冲突

（4）数据规约

- **为什么要进行数据规约：**数据库和数据仓库当中往往有海量的数据，如果不进行处理而是在整个数据集上进行复杂的数据分析和挖掘，需要很长的时间
- **维规约：**去除不重要的属性(小波变换、主成分分析、流行学习、属性子集选择)
- **纬度诅咒：**当维度增加时，数据会变得愈加稀疏；数据点间的距离和密度对聚类非常重要，孤立点分析则没什么意义；子空间的可能组合会呈指数级增长
- **维规约的作用：**
  - 避免维度诅咒
  - 帮助去除不相关特征和减弱噪声
  - 减少数据挖掘中的时间和空间需求
  - 使得可视化更轻松
- **属性子集的选择：**剔除和挖掘主题不相关的维度（属性）
- **主成分分析（PCA）：**通过某种线性投影，将高维的数据映射到低维的空间中，并期望在所投影的维度上数据的信息量最大（方差最大），以此使用较少的数据维度同时保留较多的原始数据的特点。

- **主成分分析的优缺点：**

优点：

① 以方差衡量信息的无监督学习，不受样本标签的限制
② 各主成分之间正交，消除了原始数据成分间的相互影响
③ 因为这些主成分经过排序，可以减少指标选择的工作量
④ 计算方法简单，易于在计算机上实现

缺点：

① PCA 只适用于数值数据
② 其含义往往具有一定的模糊性
③ 贡献率小的主成分往往可能含有对样本差异的重要信息

- **因子分析：**是指研究从变量群众提取共性因子的统计技术
- **数量规约：**通过选择替代的、较小的数据表示形式来减少数据量
  - 有参方法：使用一个参数模型估计数据，最后只存储参数，不存储原始数据（线性回归、多元回归、对数线性模型）
  - 无参方法：不使用模型的方法存储数据（直方图、聚类、抽样）
- **直方图分析：**将数据分装于桶中，存储该桶中的均值或总和
- **抽样：**允许用数据的较小随机样本 s 表示较大的数据集 N
  - 抽样的关键在于：找出数据中具有代表性的子集作为样本
  - 简单随机抽样在数据倾斜是表现不佳，可以采用分成抽样
- **分层抽样：**

① 将抽样框架中的个体分为若干群，称为层(分层的标准是每层个体之间有相似的性质或者你对这些层有特别的兴趣)
② 每层采取简单随机抽样，把各层的样本合起来得到最终样本

(5) 数据变换

- **数据变换的目的：**为了研究数据特征之间潜在的规律，有时候还需要对特征运用某些函数进行变换，以便更容易地找到其中的规律
- **数据变换方法：**
  - 属性/特征构造：通过现有属性构造新的属性
  - 规范化：将数据按比例缩放，使之落入一个小的特定区间
  - 离散化
- **规范化：**
  - 最大最小规范化

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Z-Score 规范化：将数据转换为均值为 0、方差为 1，类似正态分布的数据

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - 小数标定规范化

$$v' = \frac{v}{10^j}$$ 其中j是令Max（|v'|）＜1的最小整数

• Box-Cox 变换：一种广义幂变换方法，用于连续的响应变量不满足正态分布的情况，其公式如下。这里 λ 是一个待定变换参数。对于不同的，所作的变换也不相同，所以 Box-Cox 变换是一族变换，通过参数的适当选择，使数据满足正态线性回归模型的所有假设条件。

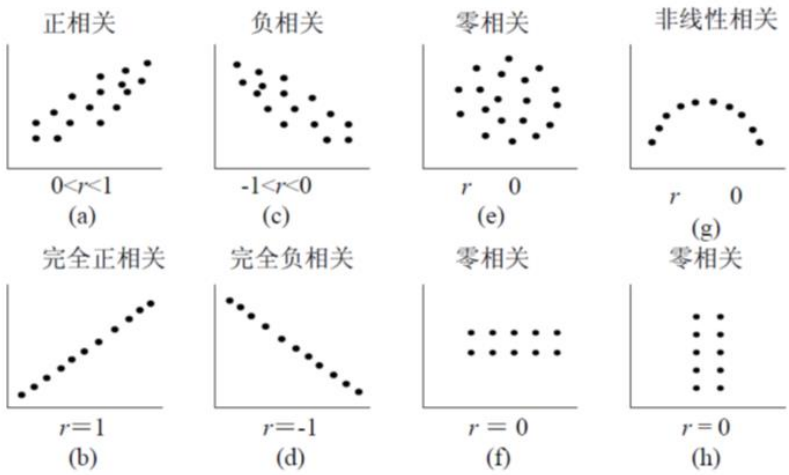
$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{(\lambda)} - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \ln y_i, & \text{if } \lambda = 0 \end{cases}$$

- **为什么要离散化：**
  - 一些算法如决策树、朴素贝叶斯、logistic 回归等要求变量必须离散
  - 离散化之后能降低利群数据的影响
  - 相对于连续型特征，离散型特征在计算速度、表达能力、模型稳定性等方面具有优势
- **离散化方法：**分箱、直方图分析、聚类分析、决策树分析、相关性分析
- **概念分层：**
  - **概念分层将概念分层进行组织，通常和数据仓库中的每一维都有关**
  - **通过概念分层可以将数据变换到多个粒度层来观察**
  - **概念分层的产生：通过将低层概念（例如年龄这样的数值）替换为高层概念（例如青少年、成年、老年等）来递归地减少数据**
  - **概念分层可以被专家或数据仓库设计者显式指定**

4.相关与回归

（1）相关系数

- **变量间的关系：**
  - 相关关系：两类现象在发展变化的方向与大小方面存在一定的关系
  - 共变关系：两事物本身之间没有直接的关系哦，但是它们都受第三种现象的影响而发生变化
  - 因果关系：一种现象是另外一种现象的原因，而另一种现象是结果
- **相关系数：**相关系数是用来表示变量间相关关系强度的指标
  - -1<=r<=1
  - 正负号表示相关的方向；取值大小表示相关的强弱程度
  - 相关系数不是等距量表值，更不是等比量表



- **皮尔逊相关系数：**也叫积差相关，是解释两个变量线性相关方向和程度最常用的方法
  - 要求成对的数值型数据
  - 正态双变量
  - 两列变量之间的关系应该是线性的，非线性不能用这个系数来计算相关性
  - n>=30
  - 相关性强不一定意味着皮尔逊相关系数高，还有可能是非线性相关

$$r=\frac{\sum(X-\overline{X})(Y-\overline{Y})}{NS_XS_Y}=\frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\cdot\sqrt{\sum(Y-\overline{Y})^2}}$$

- **斯皮尔曼相关系数：**在皮尔逊相关系数的基础上，在定义中把点坐标换成各自样本的等级
  - 适用于两列等级性质的变量（即可以是等级变量，也可以是连续变量赋以等级顺序转换而来）
  - 对数据的整体分布不做要求
  - 受异常值影响小
  - 首先对两个变量（X，Y）的数据进行排序，然后记下排序以后的位置（X’，Y’），（X’，Y’）的值就称为秩次，秩次的差值就是上面公式中的 d_i，n 就是变量中数据的个数，最后带入公式就可求解结果。

$$\rho_s=1-\frac{6\sum d_i^2}{n(n^2-1)}$$

- **肯德尔相关系数：**也是利用变量值出现的顺序。具体求取方法为：先将变量 X 进行升序排列，然后再对变量 Y 从第一个开始，依次往后进行两两比较，最后看随着 X 的增大变量 Y 增大的值有多少，降低的有多少，通过增大的个数和降低的个数的比较来判定两个变量的相关性。

（2）空间自相关

- **空间自相关：**描述空间中位置 S 处变量与其临近位置上同一变量的相关性
  - 如果数值接近就是正相关；否则则是负相关
  - 同时处理空间对象位置和属性的相似性
- **Moran 统计：**Moran 统计是最常见的空间自相关统计指标。Moran 指数为正，则该指标在

空间上正相关；反之则负相关。

- [-1, 1]
- 计算时必须先对每行进行标准化（对矩阵中的每一行，求和后，每个元素除以所在行元素之和）
- 一般大于 0.7 说明具有较强的正相关性

设研究区域中存在 n 个面积单元，第 i 个单元上的观测值记为 y_i，观测变量在 n 个单元中的均值记为 $\bar{y}$，则 Moran's I 定义为

$$I = \frac{n}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{W}_{ij}(y_i-\bar{y})(y_j-\bar{y})}{\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{W}_{ij}}$$

- **p 值：** 小于 0.05 说明本数据随机生成的概率小于 5%
- **z 值：** z 得分大于 1.96 说明这份数据呈现了明显的聚类特征

p值（p-value）是在假设检验中使用的一个概率值，表示观察到的数据或更极端情况下出现的概率。p值用于判断在给定假设下，观察到的数据是否具有统计显著性。通常，p值与预先设定的显著性水平进行比较（例如，通常使用的显著性水平是0.05），如果p值小于显著性水平，则可以拒绝原假设。

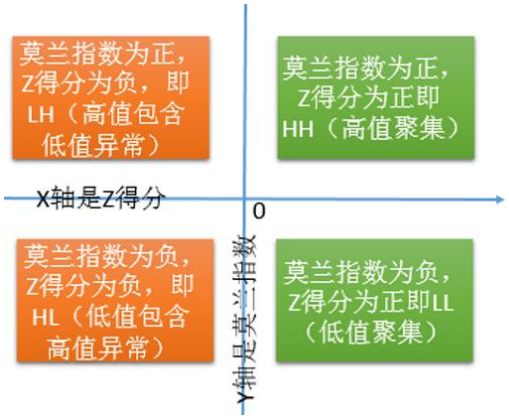
z值（z-score）是一个标准化的分数，用于测量一个观察值与其所属总体均值之间的距离（以标准差为单位）。z值可以将观察值与其所在总体进行比较，提供了一个相对的位置信息。计算z值的公式是：$z = (x - \mu) / \sigma$，其中x是观察值，μ是总体均值，σ是总体标准差。当原始数据近似符合正态分布时，z值可以告诉我们一个观察值相对于总体的位置，如处于平均值的左侧或右侧的多少个标准差。

- **广义 G 统计量：** Moran 统计不能区分不同类型的空间聚集模式，G 统计量可以

$$G(d) = \frac{\sum\sum w_{ij}(d)x_i x_j}{\sum\sum x_i x_j}$$

- 高，表示高数值的空间聚集，低则表示低数值的空间聚集。

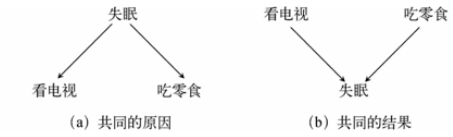
- 局部空间自相关：Moran 统计时全局统计量，但是有可能一个局部是正相关，另外一个局部是负相关



（3）相关与因果

- 相关系数是对称的，因果关系可能是不对称的
- 因果关系存在先后顺序

■ 如果A和B相关，存在多种可能：
- □A导致B
- □B导致A
- □C导致A和B
- □A和B互为因果
- □小样本引起的巧合



（4）回归

- **回归：**通过一个或几个变量的变化去解释另一变量的变化
- **判定系数：**估计的回归方程拟合优度的度量，表明Y的变异性能被估计的回归方程解释的部分所占比例。R^2越接近于1，说明回归拟合效果越好。

$$R^2 = \frac{\Sigma(y - y')^2}{\Sigma(y - \bar{y})^2}$$

- **多元回归**：一个因变量多个自变量的回归问题（使用最小二乘估计）
  - 自变量是非随机或者固定的，且相互不相关（无多重共线性）
  - 随机误差项具有0均值、同方差及不序列相关性
  - 自变量与随机误差项不相关
  - 随机误差性满足正态分布

设因变量y，k个自变量分别为x_1，x_2，…，x_k，描述因变量y如何依赖自变量x_1，x_2，…，x_k和误差项 ε 的方程，称为多元回归模型(multiple regression model)。多元回归模型一般形式为：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

其中b_0，b_1，b_2，…，b_k是参数
ε是被称为误差项的随机变量
y是x_1，x_2，…，x_k的线性函数加上误差项ε
ε包含在y里面但不能被k个自变量的线性关系所解释的变异性

- **多重判定系数：**回归平方和/总平方和。它是度量多元回归方程拟合程度的一个统计量，反映了在因变量y的变差中呗估计的回归方程所解释的比例。

- **多元回归中因变量离差平方和的分解：**
- **SST=SSR+SSE**

$$\sum(y_i-\bar{y})^2=\sum(\hat{y}_i-\bar{y})^2+\sum(y_i-\hat{y}_i)^2$$

总平方和=回归平方和+残差平方和

多重判定系数是多元回归中的回归平方和占总平方和的比例，它是度量多元回归方程拟合程度的一个统计量，反映了在因变量y的变差中被估计的回归方程所解释的比例。计算公式为

$$R^2=\frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\bar{y})^2}=\frac{SSR}{SST}=1-\frac{SSE}{SST}$$

- **多重共线性：**是指在回归模型中，两个或两个以上的自变量彼此相关
  - ① 变量之间高度相关时，可能会使回归的结果造成混乱，甚至会把分析引入歧途
  - ② 多重共线性可能对参数估计值的正负号产生影响，特别是各回归系数的正负号有可能和我们预期的相反
- **多重共线性的判别：**
  - 计算模型中各对自变量之间的相关系数，并对各相关系数进行显著性检验。如果有一个或多个相关系数是显著的，就表示模型中所用的自变量之间相关，因而存在着多重共线性问题
  - 如果休闲下列情况，则暗示存在多重共线性：
    - 模型中各对自变量之间显著相关
    - 当模型的线性关系检验显著（F 检验）时，几乎所有回归系数的 t 检验确不显著
    - 回归系数的正负号与预期相反
- **多重共线性的处理：**
  - 多重共线性问题带来的主要麻烦是对单个回归系数的解释和检验。如果仅仅是为了估计或预测，可以将所有自变量都保留在模型中
  - 在建立多元线性回归模型时，不要试图引入更多的自变量，除非有必要。特别是在社会科学的研究中
  - 将一个或多个相关的自变量从模型中剔除，使保留的自变量尽可能不相关
  - 使用主成分分析

（5）逻辑回归

- **逻辑回归：**解决二元分类。根据输入的特征变量来预测一个二元分类问题的输出结果

（6）地理加权回归

- **空间非平稳性：**在空间分析（Spatial analysis）中，变量的观测值（数据）一般都是按照某给定的地理单位为抽样单位得到的，随着地理位置的变化，变量间的关系或者结构会发生变化，这种因地理位置的变化而引起的变量间关系或结构的变化称之为空间非平稳性
- **地理加权回归：**将数据的空间位置嵌入到回归参数中，利用局部加权最小二乘方法进行逐点参数估计，其中权是回归点所在的地理空间位置到其他各观测点的地理空间位置之间的距离函数

- **空间权重函数：**
  - 距离阈值法：

    距离阈值法是最简单的权函数选取方法，它的关键是选取合适的距离阈值D，然后将数据点j与回归点i之间的距离d_{ij}与其比较，若大于该阈值则权重为0，否则为1，即

$$w_{ij}=\begin{cases}1 & d_{ij}\le D\\ 0 & d_{ij}>D\end{cases} \tag{2.37}$$

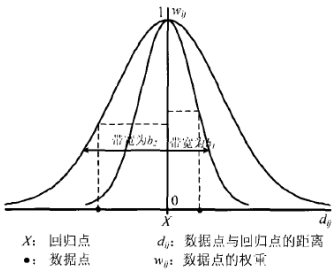
  - 距离反比法：

$$w_{ij}=1/d_{ij}^{\alpha} \tag{2.38}$$

    这里α为合适的常数，当α取值为 1 或 2 时，对应的是距离倒数和距离倒数的平方。这种方法简洁明了，但对于回归点本身也是样本数据点的情况，就会出现回归点观测值权重无穷大的情况，若要从样本数据中剔除却又会大大降低参数估计精度，所以距离反比法在地理加权回归模型参数估计中也不宜直接采用，需要对其进行修正。

  - 高斯函数法：

$$w_{ij}=\exp\left(-(d_{ij}/b)^2\right) \tag{2.39}$$



  - 截尾型函数法：

$$w_{ij}=\begin{cases}\left[1-\left(d_{ij}/b\right)^2\right]^2 & d_{ij}\le b\\ 0 & d_{ij}>b\end{cases} \tag{2.40}$$

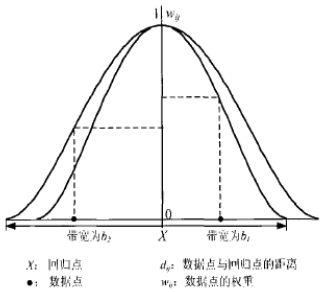


图 2.2 bi-square 空间权函数

- **地理加权回归模型：**

- **地理加权回归模型是对普通线性回归模型的扩展，将数据的地理位置嵌入到回归参数中，即：**

$$y_i = \beta_0(u_i,v_i) + \sum_{k=1}^{p} \beta_k(u_i,v_i)x_{ik} + \varepsilon_i \quad i = 1,2,\ldots,n \qquad (2.17)$$

这里的 (u_i,v_i) 为第i个采样点的坐标（如经纬度）$\beta_k(u_i,v_i)$ 是第i个采样点上的第k个回归参数，是地理位置的函数。$\varepsilon_i \sim N(0,\sigma^2)$，$Cov(\varepsilon_i,\varepsilon_j) = 0 (i \neq j)$

每个样本点计算的时候，其他的参与计算的样本都会根据与这个样本点不同的空间关系赋予不同的权值，这样最后就可以得出每个不同样本的相关回归系数

5.关联规则

- **关联规则挖掘：** 发现大量数据中项集之间有趣的关联或相关联系
- **支持度：** 对于一个规则（X，Y），其支持度表达所有所有事务中有多少事务同时完成 X、Y。支持度表示了关联规则的有用性、广泛性

$$Support(X,Y) = P(XY) = \frac{number(XY)}{num(AllSamples)}$$

- **置信度：** 对于一个规则（A=>B），置信度表示完成 A 的事务中有多少页完成了 B。置信度表示了关联规则的准确性

$$confidence(A => B) = P(B \mid A) = \frac{support\_count(A \cup B)}{support\_count(A)}$$

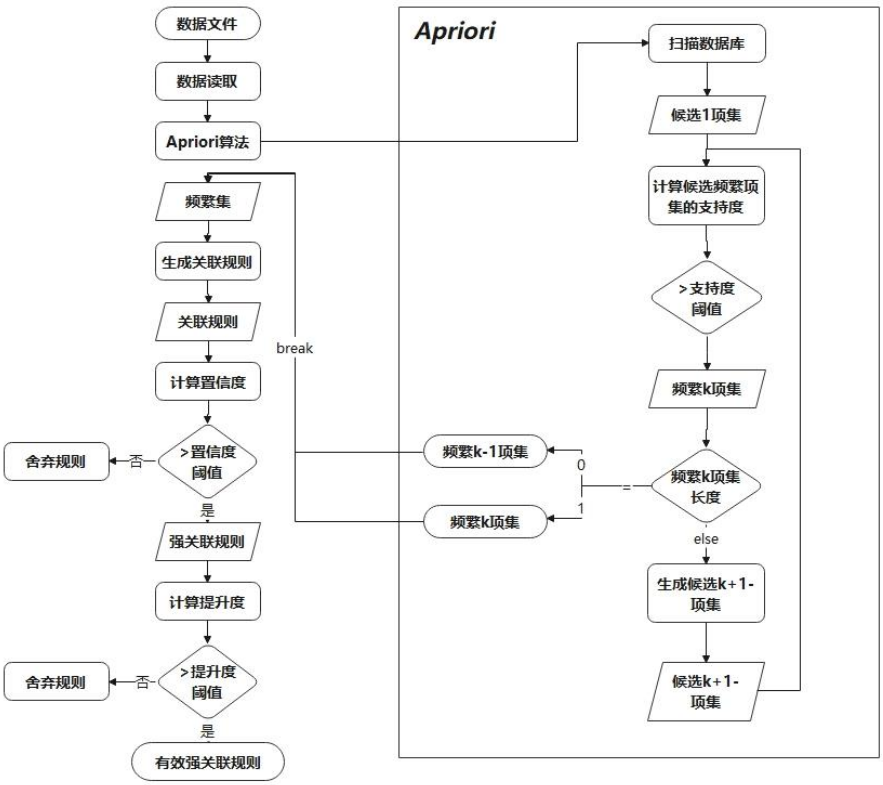
- **提升度：** 规则（A=>B），lift=1 说明 A、B 相互独立；如果 lift>1 则说明这个规则是一个有效的强关联规则。

$$Lift(X \Leftarrow Y) = P(X|Y)/P(X) = Confidence(X \Leftarrow Y)/P(X)$$

- **事务数据库、事务：**
  - **设I={ i_1，i_2，…，i_m }是一个项目集合，事务数据库D={ t_1，t_2，…，t_n }是由一系列具有唯一标识TID的事务组成，每个事务t_i（i=1，2，…，n）都对应I上的一个子集。**
- **项集：** 项的集合，包含 k 个项的项集为 k-项集
- **强规则：** 同时满足最小支持度阈值和最小置信度阈值的关联规则称为强规则
- **项集的频率：** 包含该项集的事务数
- **频繁项集：** 满足最小支持度的项集称为频繁项集
- **Apriori 算法：** 根据有关频繁项集性质的先验知识而命名。该算法适用一种逐层搜索的方法，利用 k-项集来探索 k+1-项集
- **Apriori 算法的性质：** 如果一个项集是频繁的，则这个项集的任意一个非空子集都是频繁
- **Apriori 算法步骤：**

① 首先遍历事务数据库，将所有的项作为候选 1-项集，并统计各 1-项集支持度
② 筛选出满足最小支持度的频繁 1-项集
③ 由频繁 1-项集生成候选 2-项集（自由组合）
④ 计算候选 2-项集的支持度，得到满足最小支持度的频繁 2-项集
⑤ 由频繁 k-项集生成候选 k+1-项集，这里利用 Apriori 性质：频繁集的所有子集必须频繁，即某个候选 k+1-项集的某个 k-项子集不在频繁 k-项集中，则该候选 k+1-项集必然不是频繁的，剔除掉
⑥ 计算候选 k+1-项集支持度，得到满足最小支持度的频繁 k+1-项集
⑦ 若得出的频繁 k+1-项集个数为 1 则算法结束，结果为该 k+1-项集；若频繁 k+1-项集个数为 0 则算法结束，结果为所有的频繁 k-项集；否则，算法继续，调会第⑤步

- **Apriori 算法的瓶颈：**
  - 多次扫描事务数据库，需要很大的 I/O 负载
  - 可能产生庞大的候选集
- **Apriori 算法改进：**
  - 使用哈希表存储候选 k-项集的支持度计数
  - 移除不包含频繁项集的事务
  - 对数据采样
  - 划分数据
- **生成强关联规则：**通过 Apriori 算法得到所有频繁项目集后，对于每一个频繁项集生成其所有非空子集，对于每一个非空子集计算置信度，只要置信度满足最小置信度则其为强关联规则
- **关联规则挖掘的完整流程：**



6.聚类

聚类：聚类是一个数据集划分为若干类别的过程，使得同一个组内的数据具有高度相似性，而不同组的数据具有明显区别（无监督学习算法）

- 聚类生成的组称为簇，簇是数据对象的集合
- 同簇数据相似度高，异簇数据相似度低
- 数据对象间的距离是最常见的衡量指标

- **数据矩阵：**设有 n 个对象，可用 p 个变量（属性）描述每个对象，则 nxp 矩阵
- **相异度矩阵：**按 n 个对象两两间的相异度构建 n 阶矩阵（因为相异度矩阵是对称的，只需写出上三角或下三角即可（d 越接近 0 说明越相似）

$$\begin{pmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ \vdots & \vdots & \vdots & \\ d(n,1) & d(n,2) & \dots & \dots 0 \end{pmatrix}$$

- **数据对象间距离的计算：**
  - 闵可夫斯基距离：r=1，绝对值距离；r=2，欧式距离

$$d(x,y)=\left[\sum_{i=1}^{n}|x_i-y_i|^r\right]^{1/r}$$

  - 二次型距离：A=单位矩阵，欧式距离；A=对角阵，加权欧式距离

$$d(x,y)=\left((x-y)^T A(x-y)\right)^{1/2}$$

  - 余弦距离

$$d(x,y)=\frac{\sum_{i=1}^{n}x_i y_i}{\sqrt{\sum_{i=1}^{n}x_i^2\sum_{i=1}^{n}y_i^2}}$$

  - 二元特征样本的距离度量

    - **假定x和y分别是n维特征，x_i和y_i分别表示每维特征，且x_i和y_i的取值为二元类型数值{0，1}。则x和y的距离定义的常规方法是先求出如下几个参数，然后采用SMC、Jaccard系数或Rao系数。**
      - 设a，b，c和d分别是样本x和y中满足$x_i=y_i=1$，$x_i=1$且$y_i=0$，$x_i=0$且$y_i=1$和$x_i=y_i=0$的二元类型属性的数量，则
        - 简单匹配系数（Simple Match Coefficient,SMC）
        - $S_{SMC}(x,y)=\frac{a+b}{a+b+c+d}$
        - Jaccard系数 $S_J(x,y)=\frac{a}{a+b+c}$
        - Rao系数 $S_{Rao}(x,y)=\frac{a}{a+b+c+d}$

- **聚类分析方法分类：**
  - 划分法：基于一定标准构建数据的划分（k-means、k-modes、k-medoids、PAM）
  - 层次法：对给定数据对象集合进行层次分解

- 密度法：基于数据对象相连密度评价（DBSCAN）
- 网格法：将数据空间划分为有限个单元，基于网络结构进行聚类
- 模型法：给每个簇假定一个模型，然后去寻找能很好满足这个模型的数据对象

• 划分法：对于一个给定的 n 个对象或元组的数据库，采用目标函数最小化的策略，通过迭代把数据分成 k 个划分块，每个划分块为一个簇，这就是划分方法

- 每个分组至少一个数据对象
- 每个数据对象必属于且仅属于一个分组
- k-均值、k-中心点

• k-means 算法：

- 核心思想：通过迭代把数据对象划分到不同的簇中，以使目标函数最小化，使得生成的簇尽可能的紧凑和独立
- 步骤：
  a) 随机选取 k 个对象作为初始 k 个簇的质心
  b) 然后计算其余各个对象到各质心的距离，选择加入距离最近质心所代表的簇
  c) 计算每个簇的质心作为新质心
  d) 回到 b)，直至簇的质心不在变化
- 优点：
  - 聚类问题的经典算法，简单、快速
  - 处理大数据集，该算法是相对可伸缩和高效率的
  - 当结果簇密集时，其效果较好
- 缺点：
  - 在簇能够计算平均值的情况下才能使用，有些时候不适用
  - 必须事先给定聚类数目 k，而且对初始值敏感，如果给定不同的初始值可能导致不同结果
  - 不适于发现非凸面形状的簇或者大小差别很大的簇。而且其对噪声和孤立点敏感
- 改进：
  - 进行不止一次聚类，每次初始化不同的质心
  - 多次抽样，聚合中心点
  - 让初始化质心之间近距离很远
  - 和层次聚类相结合

• 轮廓系数：轮廓系数越接近 1，则说明聚类效果越好

- 使用内聚度和分离度来度量聚类效果的好坏

$$S(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$$

- 其中，a(i)代表样本点的内聚度 $a(i)=\frac{1}{n-1}\sum_{j\neq i}^{n}\text{distance}(i,j)$
  即点i到簇内点的平均距离
- b(i)是簇间不相似度，是点i到其他簇内点的平均距离的最小值
- 可以看出轮廓系数约接近1，聚类效果越好

$$S(i)=\begin{cases}1-\frac{a(i)}{b(i)} & a(i)<b(i)\\ 0 & a(i)=b(i)\\ \frac{b(i)}{a(i)}-1 & a(i)>b(i)\end{cases}$$

• K-中心点算法（PAM）：k-均值算法对孤立点是敏感的，如果具有极大值，就可能大幅度地

扭曲数据的分布。k-中心点算法是为消除这种敏感性提出的，它选择簇中位置最接近簇中心的对象（称为中心点）作为簇的代表点，目标函数仍然可以采用平方误差准则

- 优点：
  - 对属性类型没有局限性
  - 通过簇内主要点的位置来确定中心点，对孤立点的敏感性小

- **模糊 k-均值（FCM）：**
  - **对于每一个类别，对象有一个从属可能性在[0,1]的数值**
  - **一个对象对于所有类别的对应取值和应该为1**
  - **在簇中心点的更新过程中，这个数值反映了该对象对于这一类的贡献程度**
  - **这种一个对象可以属于多个类别的聚类方法被称为软聚类**
  - **其计算过程和k-means算法类似**

- **层次聚类：**层次聚类按数据分层建立簇，形成一棵以簇为节点的树，称为聚类图。
  - 自底向上分解：凝聚的层次聚类（凝聚的层次聚类采用自底向上的策略，开始时把每个对象作为一个单独的簇，然后逐次对各个簇进行适当合并，直到满足某个终止条件）
  - 自顶向下分解：分裂的层次聚类（分裂的层次聚类采用自顶向下的策略，与凝聚的层次聚类相反，开始时将所有对象置于同一个簇中，然后逐次将簇分裂为更小的簇，直到满足某个终止条件）
- **层次聚类的优缺点：**
  - 优点：可以在不同力度水平上对数据进行探测；更容易实现相似度量或距离度量
  - 缺点：终止条件模糊；执行分裂或者合并操作后不可更改，可能导致质量差；由于需要检查和估算大量对象或簇才能决定簇的合并或分裂，可扩展性比较差（一般将层次聚类和其他方法混用）
- **密度聚类算法：**只要一个区域的点的密度大于阈值，就把它加入到相近的聚类当中去（可以克服基于距离的算法只能发现“类圆形”的聚类的缺点，可以发现任意形状的聚类，且对噪声不敏感）
- **DBSCAN 算法：**
  - 主要思想：认为密度稠密的区域是一个聚类，各个聚类被密度稀疏的区域划分开来
  - 主要参数：Eps 和形成高密度区域所需要的最少点数 MinPts
  - 核心对象：对于任意一个点，如果其 Eps 邻域内有至少 MinPts 个样本点，则其为核心对象，否则为噪声
  - 直接密度可达：如果一个样本点 p 处于一个核心对象的 Eps 邻域内，则样本点 p 从核心对象 q 直接密度可达
  - 密度相连：对于样本点 p 和 q，如果存在核心对象 m，使得 p、q 均由 m 直接密度可达，则 p 和 q 密度相连
  - 聚类过程：先找到一个核心对象；找出它的直接密度可达对象；再从这些对象出发，寻找它们的直接密度可达对象；一致重复，直至最后没有可寻找的对象了，那么一个簇就完成了更新；重复上述过程，直至找不到新的簇
  - 优点：
    - 可以自主计算聚类数目，不需要认为指定
    - 不要求类的形状是凸的，可以是任意形状

- 对噪声数据不敏感
- 算法参数少（2 个）
- 聚类结果不依赖于遍历顺序

- 缺点：
  - 样本较多时，聚类收敛很慢
  - 聚类效果依赖于距离公式的选取
  - 不适合数据集中密度差异很大的情形

- **基于网格的聚类方法：**
  - 基于网格的方法首先将空间量化为有限数目的单元，然后在这个量化空间上进行所有的聚类操作
  - 这类方法的处理时间不受数据对象数目影响，仅依赖于量化空间中每一维上的单元数目，因此处理速度较快

- **基于模型的聚类方法：**基于模型的聚类方法建立在数据符合潜在的概率分布这一假设基础之上。该类方法试图优化给定数据与某些数学模型之间的拟合。主要有统计学方法和神经网络方法等。

7.分类

- **分类：**通过已有数据集（训练集）的学习，得到一个目标函数 f（模型），该模型能把每个属性集 x 映射到目标属性 y（类）
  - y 必须是离散的，y 如果是连续的则是回归算法
  - 步骤：①训练模型②分类
- **分类算法：**
  - K 近邻算法 KNN：如果一个样本在特征空间中的 k 个最相似（最邻近）的样本中大多数属于某个类别，则该样本也属于这个类别
  - 决策树 DT：根据特征集取值不同，将样本逐层划分并建立规则，直到某一个样本集合类的所有样本属于同一类
  - 朴素贝叶斯 NB：根据条件独立假设与贝叶斯公式，计算样本属于每个类的概率
- **KNN 算法：**
  - 基本思想：如果一个样本在特征空间中的 k 个最相似（最邻近）的样本中大多数属于某个类别，则该样本也属于这个类别
  - 算法步骤：
    1) 准备数据集：收集训练数据集，包括输入特征和对应的标签（类别或目标值）
    2) 选择 K 值：确定 K 的值，即选择在预测时要考虑的最近邻居的数量
    3) 计算距离：对于给定的测试样本，计算其与训练集中每个样本的距离。常见的距离度量方法包括欧氏距离、曼哈顿距离等
    4) 选择最近的 K 个样本：根据计算得到的距离，选择与测试样本最近的 K 个样本。
    5) 确定投票机制：对于分类问题，采用投票机制确定测试样本的类别。常见的投票机制包括多数表决，即根据 K 个最近邻中具有最多样本的类别确定测试样本的类别。对于回归问题，通常采用平均值或加权平均值来预测目标值
    6) 进行预测：根据投票结果或平均值，预测测试样本的类别或目标值
  - K 值的选择：交叉验证（将数据集划分成一定的比例，分别作为训练集和验证集，然后从一个较小的 k 开始逐渐增加 k 值，通过验证集计算模型误差，选择误差最小的模型

对应的 k 值）

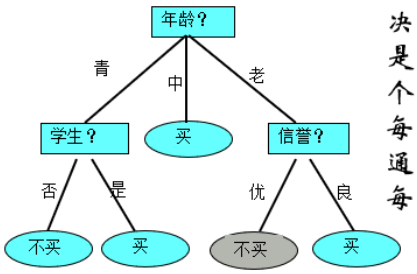
k折交叉验证

- 将数据集平均分割成K个等份
- 使用1份数据作为测试数据，其余作为训练数据
- 计算测试准确率
- 使用不同的测试集，重复2、3步骤
- 对测试准确率做平均，作为对未知数据预测准确率的估计

- 距离的度量：欧式距离
- 分类决策规则：
  - 按照数量最多的邻居进行分类
  - 加权 KNN 算法：给更近的邻居分配更大的权重（使用高斯函数加权，距离=0 权重为 1，距离越远权重减少，但不会为 0）
- 优点：
  - 简单易用
  - 训练时间快
  - 预测效果好，对异常值不敏感
- 缺点：
  - 无法给出数据的内在含义
  - 计算复杂度高，空间复杂性高
  - 无法处理样本不平衡问题

• 决策树：通过一系列规则对数据进行分类的过程，即根据特征集取值不同，将样本逐层划分并建立规则，直到某一个样本集合类的所有样本属于同一类

- 基本组成：决策节点、分支、叶子



决策树中最上面的结点称为根结点。是整个决策树的开始。每个分支是一个新的决策结点，或者是树的叶子。每个决策结点代表一个问题或者决策.通常对应待分类对象的属性。每个叶结点代表一种可能的分类结果

在沿着决策树从上到下的遍历过程中，在每个结点都有一个测试。对每个结点上问题的不同测试输出导致不同的分枝，最后会达到一个叶子结点。这一过程就是利用决策树进行分类的过程，利用若干个变量来判断属性的类别

- **熵：**刻画样本集的纯度，熵值越小，纯度越高，分类的不确定性越小
  - 假设某随机变量的概率分布为：

$$P(X=x_i)=p_i,\ i=1,2,...,n$$

  - 则它的信息熵计算公式为：

$$H(X)=-\sum_{i=1}^{n}p_i log p_i$$

- **信息增益：**划分前的信息熵-划分后的信息熵

假如决策树样本集为D，经过某属性划分后，样本集划分v个子集，D1,D2,...,Dv，划分前D的信息熵：

$$H(D)=-\sum_{i=1}^{y} p_i log p_i$$

pi表示第i类别的样本数占总样本数D的比例
划分后的子样本集的信息熵：

$$H(D^*)_{划分后}=\sum_{i=1}^{v}\frac{|D^i|}{|D|}H(D^i)$$

其中v是该属性的可取值的数量，Di表示该属性第i个值的样本数
信息增益：划分前的信息熵减去划分后的信息熵。

$$Gain(D,a) = H(D)-H(D^*)_{划分后}$$

其中a代表的此次划分中所使用的属性

- **ID3 算法：**
  - 经典的决策树学习算法
  - 基本思想：以信息熵为度量，用于决策树节点的属性选择
  - 目标：划分后使得信息熵降低，让整个系统变得更加确定
  - 过程：每次优先选择信息量最多的属性，即能使熵值变为最小的属性，以构造一颗熵值下降最快的决策树，到叶子节点处的熵值为 0
  - 具体步骤：

1 决定分类属性；
2 对目前的数据表，建立一个节点N
3 如果数据库中的数据都属于同一个类，N就是树叶，在树叶上标出所属的类
4 如果数据表中没有其他属性可以考虑，则N也是树叶，按照少数服从多数的原则在树叶上标出所属类别
5 否则，根据平均信息期望值E或GAIN值选出一个最佳属性作为节点N的测试属性
6 节点属性选定后，对于该属性中的每个值：
从N生成一个分支，并将数据表中与该分支有关的数据收集形成分支节点的数据表，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 算法偏向于取值数目较多的属性，如果一个属性是个体的唯一标识，那么通过 ID3 分类出每个分支中都只有一个样本，没有任何意义

- **过度拟合：**也称为过学习，指推出过多与训练集相一致的假设，过度拟合将导致做出的假设泛化能力过差
  - 出现原因：在创建决策树时，由于训练样本数量太少或数据中存在噪声和孤立点，许多分支反映的是训练样本集中的异常现象，建立的决策树会过度拟合训练样本集。
  - 防止办法：

① 先剪枝：该方法通过提前停止分支生成过程，即通过在当前结点上就判断是否需要继续划分该结点所含训练样本集来实现。一但停止分支，当前结点就成为一个叶结点。该叶结点中可能包含多个不同类别的训练样本。

② 后剪枝：由“完全生长”的树剪去分枝——对于树中的每个非树叶节点，计算该节点上的子树被剪枝可能出现的期望错误率

- **决策树优点：**
  - 推理过程容易理解

- 推理过程完全依赖于属性变量的取值特点，不需要使用者了解很多背景知识
- 可自动忽略对目标变量没有共享的属性变量，也为判断属性变量的重要性，减少变量的数目提供参考
- 能处理缺失数据

- **概率基础：**

■ 先验概率：由以往的数据分析得到的概率
- □P(A)是A的先验概率
- □P(B)是B的先验概率

■ 联合概率：两个事件共同发生的概率。
- □A与B的联合概率表示为P（AB）或者P（A∩B）

后验概率：得到"结果"的信息后重新修正的概率
- □P(A|B)是已知B发生后A的条件概率，也由于得自B的取值而被称为A的后验概率
- □P(B|A)是已知A发生后B的条件概率，也由于得自B的取值而被称为B的后验概率
- □通常，事件A在事件B（发生）的条件下的概率，与事件B在事件A的条件下的概率是不一样的

- **贝叶斯定理：**通过一个已知的结果，并结合一些经验性或统计性的信息来倒推出最可能产生该结果的原因

$$p(\text{类别}|\text{特征})=\frac{p(\text{特征}|\text{类别})p(\text{类别})}{p(\text{特征})}$$

$$P(A_i|B)=\frac{P(A_i)P(B|A_i)}{\sum_{j=1}^{n}P(A_j)P(B|A_j)},i=1,2,\cdots,n.$$

- **贝叶斯分类器：**通过某对象的先验概率，利用贝叶斯公式计算出其后验概率，即该对象属于某一类的概率，选择具有最大后验概率的类作为该对象所属的类
- **朴素贝叶斯：**在贝叶斯分类器的基础上假设各个特征之间相互独立

1、构建训练样本集。
2、统计得到在各类别下各个特征属性的条件概率估计。即
$P(a_1|y_1),P(a_2|y_1),\ldots,P(a_m|y_1);P(a_1|y_2),P(a_2|y_2),\ldots,P(a_m|y_2);\ldots;P(a_1|y_n),P(a_2|y_n),\ldots,P(a_m|y_n)$
3、如果各个特征属性是条件独立的，则根据贝叶斯定理有如下推导：

$$P(y_i|x)=\frac{P(x|y_i)P(y_i)}{P(x)}$$

因为分母对于所有类别为常数，因为我们只要将分子最大化皆可。
又因为各特征属性是条件独立的，所以有：

$$P(x|y_i)P(y_i)=P(a_1|y_i)P(a_2|y_i)\ldots P(a_m|y_i)P(y_i)=P(y_i)\prod_{j=1}^{m}P(a_j|y_i)$$

- 避免零概率问题：朴素贝叶斯方法需要每一个条件概率都是非零的，否则预测结果就会变为 0（矫正方法：拉普拉斯校正，给每一个计数都加一）
- 优点：容易实现；在大多情况下可以取得较好的结果
- 缺陷：类条件独立，缺乏准确性，实际上变量间是存在依赖关系的，这样的依赖关系

导致不能构建朴素贝叶斯分类器

• **贝叶斯网络**：一个有向无圈图，其中的节点代表随机变量，节点之间的边代表变量之间的直接依赖关系；每个节点都有一个条件概率表(Conditional Probability Table) P(Xi|Parents(Xi))，刻画了父变量对子变量的影响程度

1. 具有坚实的数学基础，长期的理论研究和实践应用，证明了其有效性和正确性。
2. 贝叶斯网络是有向无循环图，能够清晰和直观地显示变量之间的因果关系。
3. 贝叶斯网络可以图形化表示随机变量间的联合概率，利用概率理论能够处理各种不确定性信息。
4. 贝叶斯网络可以处理不完整和带噪音**的数据集。**

• **集成学习**：聚合一组分类器得到的结果通常比最好的单个预测器要好（原理：大数定理；前提：所有分类器相互独立）

• 好处：提升模型精度；可以处理过大、过小的数据集；可以形成很复杂的决策边界；适合处理多源异构数据

• **随机森林**：训练一组决策树分类器，每一棵树都基于训练集不同的随机子集进行训练。做出预测时，获得所有树各自的预测，然后给出得票最多的类别作为预测结果。这样一组决策树的集成被称为随机森林

• **bagging 和 pasting**：采样时如果将样本放回，这种方法叫作 bagging（bootstrap aggregating，也叫自举汇聚法）；采样时样本不放回，这种方法则叫用 pasting。

• **Boosting**：总体思路是循环训练预测器，，每一次迭代，都根据上一次迭代的结果，增加被错误分类的样本的权重。使模型在之后的迭代中更加注重难以分类的样本

• **分类器评估度量**：

• 几类样本：

（1）正样本：在二分类问题中，目标样本是正样本。如在人脸识别问题中，人脸样本就是正样本。

（2）负样本：非目标样本为负样本。如除了人脸样本之外的如窗户、门的样本就是负样本。

- 真正例（True Positive，TP）：被模型预测为正的正样本
- 假正例（False Positive，FP）：被模型预测为正的负样本
- 假负例（False Negative，FN）：被模型预测为负的正样本
- 真负例（True Negative，TN）：被模型预测为负的负样本

• 指标

度量	公式
准确率（识别率）	$\frac{TP+TN}{P+N}$
错误率（误分类率）	$\frac{FP+FN}{P+N}$
召回率（敏感度）	$\frac{TP}{P}$
特效性（真负例率）	$\frac{TN}{N}$
精度	$\frac{TP}{TP+FP}$
F_1值	$\frac{2\times 精度\times 召回率}{精度+召回率}$
F_β值	$\frac{(1+\beta^2)\times 精度\times 召回率}{(\beta^2\times 精度)+召回率}$

8.网络挖掘

- **社会网络：**社会行动者及其关系的集合

• 结构洞理论：当网络中某一个体处于结构相对稀疏地带，将两个不具有直接联系的行动者联结起来形成关系时，该网络个体作为纽带就处于结构洞的位置

根据结构洞理论，社会网络中的结构洞具有以下特征：

1. 桥接不同群体：结构洞位于不同社会群体之间，通过连接不同的群体，促进信息和资源的流动。这使得结构洞内的个体可以获得来自不同群体的多样化信息和资源。
2. 控制信息流动：结构洞的个体具有信息的控制权和中介地位，他们可以控制信息在网络中的流动。由于他们处于连接不同群体的位置，他们可以在不同群体之间传递信息，掌握信息的流向和内容。
3. 获取新颖信息：由于结构洞个体可以获取来自不同群体的信息，他们更有可能接触到新颖的信息和观点。这使得他们具有更广阔的知识和视野，并能够获得创新性的想法和机会。

• 小世界网络：小世界网络对应于规则网络（所有节点仅与其最近邻居相连）和随机网络（所有节点随机相连）之间的中间状态。小世界网络被认为是有效的网络体系结构，平衡了网络集成和隔离的过程

• 无标度网络：少数的节点拥有大量的连接，大部分节点的连接很少，这种复合幂律分布的复杂网络称为无标度网络

- **图的基本概念：**
  - 定义

定义1 一个有序二元组(V, E)称为一个图，记为$G=(V, E)$，其中

① V称为G的顶点集，$V\neq\phi$，其元素称为顶点或结点，简称点；（社会网络中的行动者）

② E称为G的边集，其元素称为边，它联结V中的两个点，如果这两个点是无序的，则称该边为无向边，否则，称为有向边.（社会网络中的关系）

  - 邻域：与某个特定的点响铃的那些点的集合，用 N（v）表示
  - 度数：图 G 中与顶点 v 关联的边的数目，用 d（v）表示
    - 入度：有向图中以顶点 v 为终点的边的数目
    - 出度：有向图中以顶点 v 为起点的边的数目
    - 孤立点：d=0；悬挂点：d=1；奇点：d=奇数；偶点：d=偶数
  - 相邻边：有一个公共端点的边
  - 链：由两两相邻的点及其相关联的边构成的点边序
  - 割点：图中去掉该点后，原来向量的图就分割成两个不相连的子图的结点
  - 桥：图中去掉该边后，原来相连的图就分割成两个不相连的子图的边
  - 块：最大不可分子图
  - 图的连通性：

□通路（walks）、轨迹（Trails）和路径（paths）

□定义：设$G = (V, E)$是一个图，$v_0, v_1, \ldots, v_k \in V$，且$\forall 1 \le i \le k$，$v_{i-1}v_i \in E$，则称$v_0\ v_1\ \ldots\ v_k$是$G$的一条通路. 如果通路中没有相同的边，则称此通路为轨迹（道路）. 始点和终点相同的道路称为圈或回路. 如果通路中既没有相同的边，又没有相同的顶点，则称此通路为路径，简称路.

- 旅程：每条线至少过一次
- 路径长度：一条路径经过的边的数量
- 距离：最短路径长度
- 直径：最大路径长度

- **图的整体属性：**
  - 网络规模=顶点数或边数目
  - 网络密度=实际连线数/最大可能连线数
  - 平均节点度=所有结点的平均度数
  - 网络平均距离=所有点对之间的最短路径的算术平均值
  - 集群系数=网络平均度/网络规模（反映网络的群集程度）
  - 明星：网络中突出的结点
- **中心性度量：**
  - 两种度量方法：中心度（一个节点在网络中处于核心地位的程度）和中心势（描述整个图的紧密程度或一致性，即一个图的中心度）

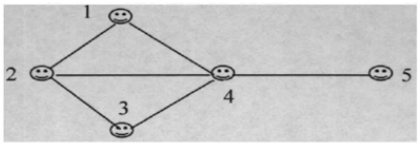
① 点度中心度：与该点有直接关系的点的数目

② 点度中心势：对于一个网络来说，它的中心势指数由如下思想给出：首先找到图中的最大中心度数值；然后计算该值与任何其他点的中心度的差，从而得到多个“差值”；再计算这些“差值”的总和；最后用这个总和除以各个差值总和的最大可能值。

$$C = \frac{\sum_{i=1}^{n}(C_{\max} - C_i)}{\max\left[\sum_{i=1}^{n}(C_{\max} - C_i)\right]}$$

③ 中间中心度：其他节点通过该节点的最短路径的数量（很高说明能控制其他人交往）

1－4－5是一个连接1和5的测地线，1和5之间的测地线仅此一条，4的中间中心度为1。
2－4－5是一个连接2和5的测地线，2和5之间的测地线仅此一条，4的中间中心度多了1。
3－4－5是一个连接3和5的测地线，3和5之间的测地线仅此一条，4的中间中心度又多了1。
1－4－3是一个连接1和3的测地线，1和3之间的测地线有2条（1－4－3和1－2－3），4的中间中心度赋予1/2。所以，行动者4的中间中心度为：1+1+1+1/2=3.5，记作CB(4)=3.5



④ 中间中心势：网络中中间中心性最高的节点的中间中心性与其他节点的中间中心性

的差距。该节点与别的节点的差距越大，则网络的中间中心势越高，表示该网络中的节点可能分为多个小团体而且过于依赖某一个节点传递关系，该节点在网络中处于极其重要的地位

$$C_B = \frac{\sum_{i=1}^{n}(C_{ABmax} - C_{ABi})}{n^3 - 4n^2 + 5n - 2} = \frac{\sum_{i=1}^{n}(C_{BBmax} - C_{RBi})}{n - 1}$$

其中，C_{ABmax}是点的绝对中间中心度，C_{RBmax}是点的相对中间中心度。

⑤ 接近中心度：节点与其他节点之间平均最短路径的倒数（评价的是一个节点传播信息时不依赖他人的程度，如果很高则说明其为网络的重心）

- **社区发现：**
  - 社区：内部连接比较紧密的节点自己和对应的子图叫做社区
  - 社区发现：给定一个网络图，找出其社区结构的过程叫做社区发现
  - 常用社区发现算法-Louvain 算法：一种基于模块度的社区发现算法。
    - 基本思想：网络中节点尝试遍历所有邻居的社区标签，并选择最大化模块度增量的社区标签
    - 模块度：评估一个社区网络划分好坏的度量方法。它的取值范围是[-0.5,1)。可以简单的理解为社区内部所有边权重和减去社区相连的边权重和。
    - 算法步骤：
      1、初始时将每个顶点当作一个社区，社区个数与顶点个数相同。
      2、依次将每个顶点与之相邻顶点合并在一起，计算它们最大的模块度增益是否大于 0,如果大于 0,就将该结点放入模块度增量最大的相邻结点所在社区。
      3、迭代第二步，直至算法稳定，即所有顶点所属社区不再变化。
      4、将各个社区所有节点压缩成为一个结点，社区内点的权重转化为新结点环的权重，社区间权重转化为新结点边的权重。
      5、重复步骤 1-3，直至算法稳定
    - 效率和效果上都表现较好，并且能够发现层次性的社区结构
- **页面重要性的评价方法-PageRank 算法：**
  - 基本假设：许多优质的网页链接的网页必定是优质网页；互联网是一个有向图，每一个网页是图的一个顶点，网页间的每一个超链接是图的一个有向边
  - PageRank 算法中一个网页的重要性由 1 导入该网页链接的数量和这些导入链接的重要性来衡量
- **社会网络中的信息传播：**
  - 传播的三个要素：传播者；接受者；传播媒介
  - 主要的信息传播方式：羊群效应、信息级联、创新扩散、流行病

① 羊群效应：个体在观察到其他个体动作后做出的一致行为
   a) 传播时不一定需要社会网络
   b) 可利用贝叶斯或群体智能建模

② 信息级联：指在社交网络中，由于个体之间的相互影响和信息传递，一个人或一组人的决策或行为会引发其他人的决策或行为，从而形成一系列连锁反应或传播效应的过程。

③ 创新扩散：指新的想法、产品、技术或实践从最初的创造者或创新者开始传播到广大群体和市场的过程

④ 流行病：模拟传染病传播的传播方式

9.文本挖掘

• **文本挖掘（Text Mining）**：将数据挖掘的成果用于分析以自然语言描述的文本，这种方法称为文本挖掘

• **自然语言处理（Natural Language Processing，NLP）**：研究的是如何通过机器学习等技术让计算机学会处理人类语言，乃至最终理解人类语言

• **文本预处理**：句子切分、词语切分、文本清洗、删除特殊字符、扩展缩写词、大小写转换、词语校正、删除停用词、词干提取等

• **停用词**：指文档中出现的连词、介词、冠词等无太大意义的词

• **分词**：即将文本中的句子划分为一个一个词语

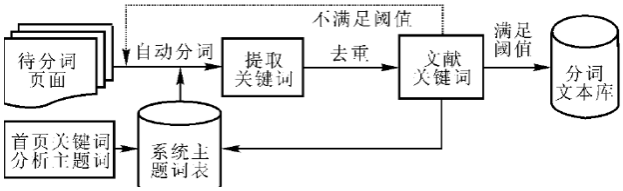
• 词典分词法：简单、易于实现，广泛用于实际工程。匹配速度慢，词典功能能越强词典中词条的数目就越大

• 正向最大匹配法（MM），逆向最大匹配法（RMM），双向最大匹配法（BM）

• MM 和 RMM 都是先切分出 i 个字符（i 为词典中最长词条的长度），然后和词库进行比对，如果在词库中存在，则记录下来；否则减少一个字符，继续比较，一直到剩下单个字符则终止

• MM 为从左到右缩减字符，RMM 为从右到左缩减，RMM 法更加精确

• BM 法是将 MM 法和 RMM 法的结果进行比较，选择较短的结果作为最终分词



• 基于统计的分词法：统计分词认为词是稳定的组合，因此在上下文中，相邻的字同时出现的次数越多，就越有可能构成一个词

• 该方法又称无词典分词，分词效果依赖于训练预料的质量，计算量比词典分词高得多

• 可以对训练文本中相邻出现的各个字的组合的频度进行统计，当组合频率高于某一个阈值时，便可以认为此字组可能构成了一个词。统计方法主要有隐马尔可夫模型 HMM，条件随机场 CRF

• 混合分词法

• **词性标注**：标注动词、名词、形容词等

• **命名实体识别**：主要是识别语料中的人名、地方名、组织机构名等一些命名实体

• **词袋模型**：首先根据所有文本中出现的词语组件一个词库，然后统计具体一个文本中出现的单词及其出现次数

• 缺点：纬度灾难、无法保留词序信息、存在语义鸿沟

假如文本集一共有如下两个文本
1.John likes to watch movies. Mary likes movies too.
2.John also likes to watch football games.
基于这两个文本构建的词库如下：
["John","likes","to","watch","movies","also","football","games","Mary","too"]
词库中共有10个词，两个文本利用词袋模型分别可以表示为：
1.[1, 2, 1, 1, 2, 0, 0, 0, 1, 1]
2.[1, 1, 1, 1, 0, 1, 1, 1, 0, 0]

- **TF-IDF=TF*IDF**
  - TF-Term Frequency 词频

$$\mathrm{tf}_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}}$$

    - 式子分子是该词在文件中的出现次数，而分母则是在文件中所有字词的出现次数之和
  - IDF-Inverse Document Frequency 逆向文件频率

某一特定词语的IDF，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取对数得到

$$\mathrm{idf}_i=\log\frac{|D|}{|\{j:t_i\in d_j\}|}$$

  - 如果一个词语在一篇文章中的词频高，且在其他文章中出现少，则认为这个词语具有很好的类别区分能力。可以用来分类
- **权重的计算：**
  - 最简单-根据频率进行赋权 pi＝mi / n
  - 为了防止权为 0 可以 pi＝（λ＋mi） / n
  - 根据标签信息等乘以加权指数，即对粗体标注的活标题等特殊信息加权
  - 布尔权重：a_{ij}=1(TF_{ij}>0) or (TF_{ij}=0)0
  - TF-IDF 权重
  - 基于熵概念的权重：该值越大，说明分布越均匀，越有可能出现在较多的类别中；该值越小，说明分布越倾斜，词可能出现在较少的类别中
- **文本表示：**词频矩阵
  - 行对应关键词 t，列对应文档向量 d
  - 将每一个文档视为空间向量 v
  - 向量值反映单词 t 与文档 d 的关联度
  - 矩阵元素可以是词频，也可以是布尔值

表示文档词频的词频矩阵

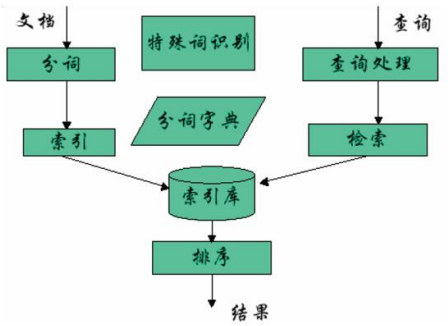
	d_1	d_2	d_3	d_4	d_5	d_6
t_1	322	85	35	69	15	320
t_2	361	90	76	57	13	370
t_3	25	33	160	48	221	26
t_4	30	140	70	201	16	35

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

- **文档相似度：**向量空间模型文档表示为一个矢量，文档相似度计算可以采用向量夹角的余弦

$$Sim(x,y)=\frac{x\bullet y}{|x|\cdot|y|}=\frac{\sum_{k=1}^{t}(x_k\cdot y_k)}{\sqrt{\sum_{k=1}^{t}x_k^{\,2}}\cdot\sqrt{\sum_{k=1}^{t}y_k^{\,2}}}$$

- **信息检索：**反之用户从包含各种信息的文档中查找所需要信息或知识的过程
  - 主要问题：根据用户查询（描述所需信息的关键词）在文档中定位相关文档
  - 主要过程：



• 正向索引：以词为单位，记录每个关键词的词频、格式、位置等权重信息，把页面转换为一个关键词组成的集合，无法满足迅速返回查询结果的要求

• 倒排索引：把文件对应到关键词的映射转换为关键词到文件的映射；方便查询但不利于构建，特别不利于删除

文档表（document_table）

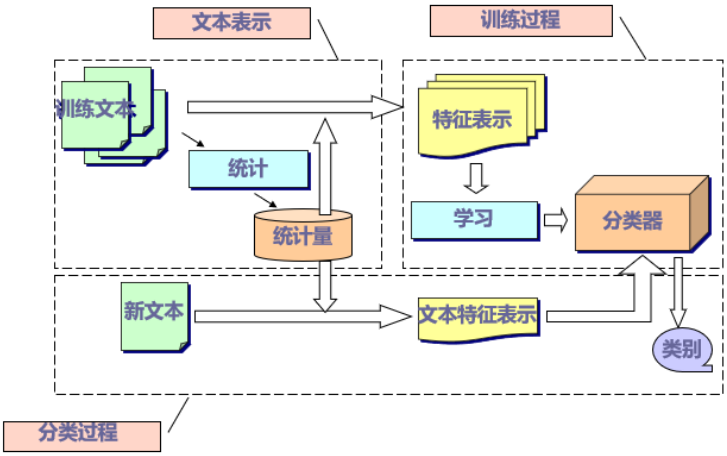
doc_ID	posting_list
Doc_1	t_1_1, ... ,t_1_n
Doc_2	t_2_1, ... ,t_2_n
⋮	⋮
Doc_n	t_n_1, ... ,t_n_n

词表(term_table)

term_ID	posting_list
Term_1	doc_1, ... , doc_i
Term_2	doc_1, ... , doc_ j
⋮	⋮
Term_n	doc_1, ... , doc_n

• **主题分析**：主题模型是以非监督学习的方式对文集的隐含语义结构进行聚类的统计模型（LDA，潜在狄利克雷分布）

• **文本分类**：给定分类体系，将文本分到某个或几个类别当中



• **情感分析**：对文本进行褒义、贬义、中性的判断

10. 时间序列

• **时间序列**：按照时间先后顺序取得的一系列观测值（反映客观现象的观测值+所属时间）

• **时间序列的基本指标**：

  • 频率：采样点的时间间隔

  • 时间跨度：如果没有时间缺口，时间跨度=观测数目*频率

  • 均值：

$$\hat{\mu} = \frac{1}{n}\sum_{t=1}^{n} x_t$$

- 方差：

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{t=1}^{n}(x_t - \hat{\mu})^2$$

- 自协方差和自相关系数：描述同一事物在两个不同使其之间的相关性

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \quad r(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var[X]Var[Y]}}$$

- **时间序列的分类：**
  - 平稳序列：基本不存在趋势的序列，各观察值基本在某个固定水平波动
  - 非平稳序列
    - 有趋势序列：线性的、非线性的
    - 复合型序列：有趋势、季节性、周期性的复合型序列
- **平稳性时间序列：**
  - 概率函数不随时间的平移而变化
  - 期望值、方差和自协方差是不依赖于时间的常数
  - 随机时间序列模型是以时间序列的平稳性为基础建立的
  - 所有观测值围绕某一水平随机波动
  - 现实生活中大多数时间序列都是非平稳的
  - 非平稳时间序列可以通过一次或多次差分的方式变成平稳时间序列
- **时间序列的成分：**
  - 趋势：持续向上或持续向下的状态或规律
  - 季节性：时间序列在一年内重复出现的周期性波动
  - 周期性：围绕长期趋势的一种波浪形或震荡式变动
  - 随机性：除去趋势、周期性、季节性之后的偶然波动
- **STL-时间序列分解算法：**基于局部加权回归将某时刻的数据分解为趋势分量、季节周期分量、残差分量（随机分量）
  - 季节性分量：显示时间序列数据在每个季节周期中的重复模式。
  - 趋势分量：显示时间序列数据中的长期趋势或变化方向。
  - 残差分量：表示无法由季节性和趋势性解释的剩余部分，可以视为随机波动或噪声。
- **时间序列预测方法：**
  - 朴素预测法：

$$y_{t+1} = y_t$$

  - 简单平均法：

$$\hat{y}_{x+1} = 1/x\sum_{i=1}^{x} y_i$$

  - 滑动窗口法

□取前面p个点的平均值作为预测值

$$\hat{y_i} = \frac{1}{p}(y_{i-1} + y_{i-2} + y_{i-3} \cdots\cdots + y_{i-p})$$

  - 指数平滑法

- 指数平滑的原理为：当利用过去观测值的加权平均来预测未来的观测值时（这个过程称为平滑）离得越近的观测值要给以更多的权。
- 而“指数”意味着：按照已有观测值“老”的程度，其上的权数按指数速度递减。

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1-\alpha)y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + \cdots$$

- 双指数平滑法在单指数平滑法基础上增加趋势信息
- 三次指数平滑法增加了一个季节分量

22

•ARIMA 模型：整合自回归移动平均模型(Autoregressive Integrated Moving Average)

- 基础为自回归模型（Autoregressive）和移动平均模型（Moving Average）
- 一套比较成熟的时间序列建模方案
- 必须满足平稳性假设
- 在短期预测和小样本情况下表现良好

它由两个特殊模型发展而成，一个特例是自回归模型或AR (Autoregressive) 模型。假定时间序列用$X_1, X_2, \ldots, X_t$表示，则一个纯粹的AR (p)模型意味着变量的一个观测值由其以前的p个观测值的线性组合加上随机误差项a_t（该误差为独立无关的）而得：

$$X_t = \varphi_1 X_{t-1} + \cdots + \varphi_p X_{t-p} + a_t$$

ARMA 模型的另一个特例为移动平均模型或MA (Moving Average) 模型，一个纯粹的MA (q)模型意味着变量的一个观测值由目前的和先前的q个随机误差的线性的组合：

$$X_t = a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$$

ARMA(p,q)模型为AR (p)模型和MA(q)模型的组合：

$$X_t = \varphi_1 X_{t-1} + \cdots + \varphi_p X_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$$

要想拟合ARIMA模型，必须先把它利用差分变成ARMA(p,q)模型，并确定是否平稳，然后确定参数p,q。

•LSTM：长短期记忆神经网络（Long-Short Term Memory）是一种特殊的递归神经网络RNN，其适合于处理和预测时间序列中间隔和延迟非常长的事件

11. 空间统计与空间数据挖掘

- **变异函数：**
  - 概念：又称变异矩，是地质（空间）统计分析所特有的基本工具

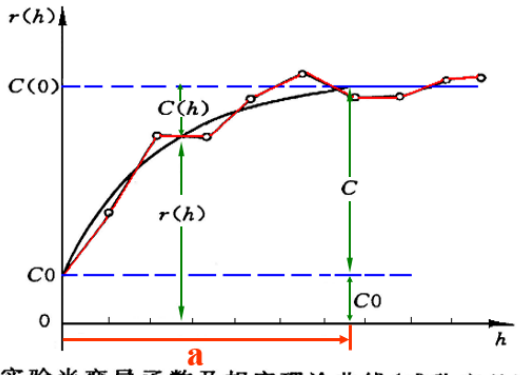
在一维条件下变异函数定义为：当空间点Z在一维x轴上变化时，区域化变量$Z(x)$在点x和$x+h$处的值$Z(x)$与$Z(x+h)$差的方差的一半为变量$Z(x)$在x轴方向上的变异函数，记为$\gamma(h)$，即：

$$\gamma(x,h) = \frac{1}{2}\mathrm{Var}[Z(x) - Z(x+h)]$$

$$= \frac{1}{2}E[Z(x) - Z(x+h)]^2 - \frac{1}{2}\{E[Z(x)] - E[Z(x+h)]\}^2$$

•参数：基台值（Sill）、变程（Range）、块金值（Nugget）、分维数（Fractal Dimension）

•这些参数决定变异函数的形状，反映自然现象空间分布的结构或空间相关的类型



实验半变异函数及相应理论曲线（或称变差）图

a—变程；*C*(0)—有限方差或基台值；*h*—样品间距或滞后；*C*0—块金常数；*C*—拱高，或结构随机变化极大值；*r*(*h*)—在 *h* 点上的半变异函数值；*C*(*h*)—在 *h* 点上的方差

① 基台值：当变异函数随着间隔距离 *h* 的增大，从非零值达到一个相对稳定的常数时，该常数称为基台值 C_0+C，它是系统或系统属性中最大的变异。基台值意味着在对应（或大于）距离的样点之间没有空间相关性，因为方差不再随距离变化。

② 变程：曲线从较低的方差值升高，到一定的间隔值时到达基台值，这一间隔称为变程。它描述了该间隔内样点的空间相关特征。在变程内，样点越接近，两点之间相似性、即空间上的相关性越强。很明显，如果某点与已知点距离大于变程，那么该点数据不能用于数据内插（或外推），因为空间上的自相关性不复存在。

③ 块金值：理论方差函数曲线不穿过原点，而是存在一个最小的方差值。当不计分析误差时，一个样品自身的品位误差应等于 0，当两个样品即使相距很近，品位间仍存在差异时，这种现象称为“块金效应”，C_0 反映了区域化变量内部随机性的可能程度。块金是在间隔距离小于采样间距时的测量误差或空间变异，或者是二者的和。测量误差是由仪器的内在误差引起的，空间变异是自然现象在一定空间范围内的变化。小于采样间距的微观尺度上空间变异是块金的一部分。

④ 分维数：用于表示变异函数的特性，由变异函数 γ (h) 和间隔距离 h 之间的关系确定。分维数 D 的大小，表示变异函数曲线的曲率，可以作为随机变异的量度。

$$2\gamma(h) = h^{(4-2D)}$$

12. 大数据挖掘

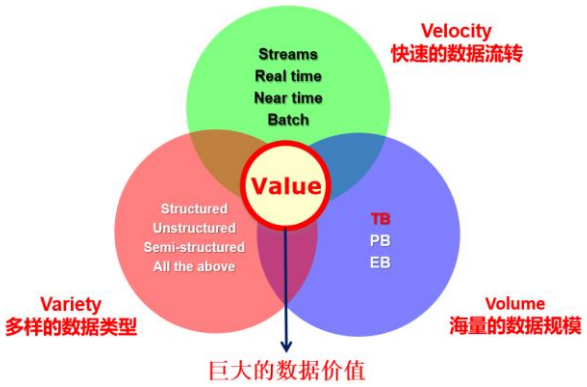
• 大数据的 4V 特点：

① 海量的数据规模（Volume）

② 多样的数据类型（Variety）

③ 快速的数据流转（Velocity）

④ 巨大的数据价值（Value）



- **时空大数据的特征：**
  ① 位置（点线面体的三维坐标、拓扑、方向）
  ② 时间（位置和属性随时间变化而变化）
  ③ 属性（数量、质量、说明信息等属性）
  ④ 尺度（比例尺）
  ⑤ 多源异构（空间基准）
  ⑥ 多维

13. 数据挖掘应用

- **推荐方法：**基于内容的推荐、协同过滤、基于潜在因素的推荐
- **基于内容的推荐：**根据物品的内容特征与用户的偏好进行匹配来进行推荐的方法
  - 步骤：
    ① 内容特征提取：对于每个物品，需要从中提取有关其内容的特征信息。例如，对于电影推荐，可以使用电影的类型、演员、导演、剧情等作为特征。
    ② 用户偏好建模：了解用户的喜好和偏好是基于内容的推荐的关键。可以通过用户的历史行为、评分、喜欢的物品等来建模用户的偏好。
    ③ 特征匹配：通过计算物品的特征与用户偏好之间的相似度或相关性，来确定推荐的物品。常用的计算相似度的方法包括余弦相似度、欧几里德距离、皮尔逊相关系数等。
    ④ 推荐生成：根据物品与用户之间的特征匹配程度，选取相似度高的物品作为推荐结果。
  - 优点：
    ① 不需要其他用户的数据（没有稀疏性的问题）
    ② 能给品味一致的用户推荐
    ③ 解决推荐系统中的用户冷启动和物品冷启动问题（能给新用户、新项目推荐）
    ④ 能够提供解释（能给推荐的项目作出内容特征的描述）
  - 缺点：
    ① 找到适当的特征时困难的
    ② 过度集中（人们有多方面的兴趣，但系统不会推荐用户内容偏好之外的项目；且不能利用其它用户的优质判断）
    ③ 对新用户的推荐（如何给新用户进行偏好建模）
- **协同过滤：**基于用户的历史行为数据和与其他用户的比较，通过发现用户之间的共同兴趣和行为模式来进行推荐

- 基于用户的协同过滤：找到与用户 x 有最相似评分的用户集合 N，根据 N 中用户的评分估计用户 x 的评分

  - r_x ： 为用户 **x**的评分矢量
  - ***N***： 为对项目 ***i***的评分与用户**x**最相似的 ***k*** 个用户的集合
  - 用户**x**对项目 **s**的评分预测

$$r_{xi} = \frac{1}{k}\sum_{y\in N} r_{yi}$$

$$r_{xi} = \frac{\sum_{y\in N} s_{xy}\cdot r_{yi}}{\sum_{y\in N} s_{xy}} \qquad s_{xy} = sim(x,y)$$

- 基于项目的协同过滤：寻找与项目 i 最相似的项目集合 N，用 N 的评分估计 i 的评分

$$r_{xi} = \frac{\sum_{j\in N(i;x)} s_{ij}\cdot r_{xj}}{\sum_{j\in N(i;x)} s_{ij}}$$

s_{ij}... similarity of items *i* and *j*
r_{xj}...rating of user *u* on item *j*
N(i;x)... set items rated by *x* similar to *i*

- 基于模型的协同过滤
- User CF ： Item CF（一般 item cf 用的多一点，因为简单）

  - **个性化程度**
    - UserCF推荐结果着重于和用户兴趣相似的小群体的热点，推荐更加社会化，反映了用户所在兴趣群体中物品的热门程度
    - ItemCF着重于维护用户的历史兴趣，推荐更加个性化，反映了用户自己的兴趣传承，可以用于长尾物品丰富的领域
  - **可解释性**
    - ItemCF可以利用用户的历史行为给推荐结果提供推荐解释，UserCF很难对推荐结果作出解释
  - **相似度矩阵**
    - UserCF要计算用户相似度矩阵，适用于用户个数远少于物品个数的场景，ItemCF要计算物品相似度矩阵，适用于物品个数远少于用户个数的场景
    - 互联网中物品的相似度相对于用户的兴趣一般比较稳定，UserCF中用户相似度矩阵相比ItemCF中物品相似度矩阵更新更频繁

  - **物品冷启动**
    - UserCF对物品冷启动不敏感， 对于新加入系统的物品，一旦有用户对该新物品产生行为，UserCF可以将该新物品推荐给和该用户相似的人群，而ItemCF无法很好地推荐新加入的物品
  - **用户冷启动**
    - ItemCF对用户冷启动不敏感，对于新加入的用户，一旦新用户对某物品产生行为，ItemCF可以给该新用户推荐和该物品相似的物品，而UserCF无法很好地给新加入的用户产生推荐实际在互联网中使用ItemCF比较多

- **基于潜在因素的推荐：**

每个用户都有自己的偏好，同时每个物品也包含所有用户的偏好信息；那么可以认为用户对于物品的高评分体现的是物品中所包含的偏好信息恰好就是用户喜好的信息。得到用户-潜在因子矩阵Q和物品-潜在因子矩阵P就可以计算出用户对于物品的评分信息。

$$\hat{R} = QP^T$$

- 根据已有的评分情况，分析出评分者对各个物品因子的喜好程度，以及各个物品对于这些因子的包含程度，最后再反过来根据分析结果预测评分。通过SVD的方式可以找出影响评分的显示因子和隐藏因子

$$R_{n*m} = U_{n*n} * \Sigma_{n*m} * V_{m*m} \qquad R_{n*m} \approx U_{n*k} * \Sigma_{k*k} * V^{T}_{k*m}$$

•冷启动：在推荐系统中，冷启动（Cold Start）指的是面对新用户或新物品时的一种挑战。

① 用户冷启动：当系统中没有关于新用户的足够历史数据或用户特征信息时，需要在没有个性化偏好信息的情况下进行推荐。这意味着系统无法准确了解新用户的兴趣和喜好，从而难以生成个性化推荐。解决用户冷启动问题的方法包括基于用户属性的推荐（如年龄、性别、地理位置等）、基于项目的协同过滤、基于内容的推荐等。

② 物品冷启动：当系统中没有足够的历史交互数据或物品特征信息时，需要推荐给用户新添加的物品。在没有足够的用户反馈或相关特征的情况下，很难准确预测新物品的兴趣和适合性。解决物品冷启动问题的方法包括基于内容的推荐、基于用户协同过滤的推荐、使用专家知识进行标注等。

• 解决冷启动问题的关键是收集和利用尽可能多的相关信息，如用户属性、物品特征、上下文信息等。通过这些信息，可以使用不同的技术和算法来进行推荐，以提供更准确和个性化的推荐结果。此外，可以采用主动学习、探索性推荐、引导性推荐等策略来引导新用户或新物品与系统进行交互，以逐渐积累数据并改善推荐效果。

- **推荐系统的构成：**前台展示页面、后台日志系统、推荐算法、用户参与
- **离线实验：**

- **离线实验的方法一般由如下几个步骤构成**
  - 通过日志系统获得用户行为数据,并按照一定格式生成一个标准的数据集;
  - 将数据集按照一定的规则分成训练集和测试集;
  - 在训练集上训练用户兴趣模型,在测试集上进行预测;
  - 通过事先定义的离线指标评测算法在测试集上的预测结果。特征选择过程必须确保不丢失重要特征

- **用户调查：**
  - 用户调查是推荐系统评测的一个重要工具，很多离线时没有办法评测的与用户主观感受有关的指标都可以通过用户调查获得
  - 用户调查的成本很高
  - 测试用户必须认真填写才有效

• AB 测试：一种常用的在线评测推荐算法的实验方法。通过一定的规则把用户随机分为几组，并对不同组的用户采用不同的算法，然后通过统计不同组用户的各种不同的评测指标来比较不同的算法

- **推荐系统评测指标：**
  - 用户满意度：最重要指标；只能通过用户调查、在线实验方式获取，不能离线获取；用户调查获取用户满意度的方式主要是调查问卷
  - 预测准确度：度量一个推荐系统或推荐算法预测用户行为的能力（最重要离线指标）

- 在计算该指标时需要有一个离线的数据集,该数据集包含用户的历史行为记录
- 然后,将该数据集通过时间分成训练集和测试集
- 最后,通过在训练集上建立用户的行为和兴趣模型预测户在测试集上的行为,并计算预测行为和测试集上实际行为的重合度作为测准确度。

• 覆盖率：描述一个推荐系统对物品长尾的发掘能力。简单的定义为推荐系统能够推荐出来的物品占总物品集合的比例

• 多样性：描述了推荐列表中物品两两之间的不相似性。为了满足用户广泛的兴趣，推荐列表需要能够覆盖用户不同的兴趣领域，即推荐结果需要具有多样性

• 信任度：用户对推荐结果的信任程度

提高推荐系统的信任度主要有两种方法

- 需要增加推荐系统的透明度 而增加推荐系统透明度的主要办法是提供推荐解释。只有让用户了解推荐系统的运行机制,让用户认同推荐系统的运行机制,才会提高用户对推荐系统的信任度。
- 其次是考虑用户的社交网络信息,利用用户的好友信息给用户做推荐,并且用好友进行推荐解释。

• 实时性：物品(新闻、微博等)具有很强的时效性，这样的推荐系统需要考虑实时性

- 推荐系统需要实时地更新推荐列表来满足用户新行为变化。
  - 比如,当一个用户购买了 Phone,如果推荐系统能够立即给他推荐相关配件,那定比第二天再给用户推荐相关配件更有价值。
  - 很多推荐系统都会在离线状态每天计算一次用存列表,然后于在线期间将推荐列表展示给用户
  - 与用所为相应的实时性,可以通过推荐列表的变化速率来评测。如果推荐列表在用户有行为后变化或者没有变化,说明推荐系统的实时性不高。
- 推荐系统需要能够将新加入系统的物品推荐给用户
  - 这主要考验了荐系统处理物品冷启动的能力。我们可以利用用户推荐列表中有多大比例的物品是当天新新加的来评测。

• 健壮性：衡量了一个推荐系统抗击作弊的能力（如注入攻击、恶意刷分等）

算法健壮性的评测主要利用模拟攻击

- 给定一个数据集和一个算法,可以用这个算给这个数据集中的用户生成推荐列表。
- 用常用的攻击方法向数据集中注入噪声数据
- 利用算法在注入噪声后的数据集上再次给用户生成推荐列表
- 最后,通过比较攻击前后推荐列的相似度评测算法的健壮性
- 如果攻击后的推荐列表相对于攻击前没有发生大的变化,就说明算法比较健壮

• 今日头条的推荐算法：

• 输入：

1） 内容特征（图文、视频、问答、评论等内容特征的提取）

2） 用户特征（兴趣标签、职业、年龄、性别、机型等参数中提取用户特征）

3） 环境特征（位置、时间、场景等环境不同用户的偏好不同，要提取环境特征）

• 点击率、阅读时间、点赞数、评论数、转发量等可以量化的内容，不能完全由指标去估计，需要具体内容的干预（如国家政策形势等）

• 没有通用的模型、各种算法的复杂组合

• 平衡计算成本和效果

• 采用流式计算框架

• 面对复杂情况的处理

1）过滤噪声：过滤停留时间短的点击，打击标题党；

2）惩罚热点：用户在热门文章上的动作做降权处理；

3）时间衰减：随着用户动作的增加，老的特征权重会随时间衰减，新动作贡献的特征权重会更大；

4）惩罚展现：如果一篇推荐给用户的文章没有被点击，相关特征（类别、关键词、来源）权重会被惩罚；

5）考虑全局背景：考虑给定特征的人均点击比例

- 想要达到比较好的效果，算法需要解决这样的问题
  1） 相关性特征：解决内容和用户匹配的问题
  2） 环境特征：解决基础特征和匹配
  3） 热度特征：在冷启动上很有效
  4） 协同特征：考虑相识用户的兴趣，在一定程度上解决算法越推越窄的问题