

14 聚类方法

没有标签时怎样组织相似样本

1 基本概念

聚类由数据表示、相似性规则和对“簇”的假设共同定义

■ 输入

- 样本特征 $x_1, \dots, x_N$, 没有已知类别标签

■ 输出

- 样本分组、层次结构、噪声标记或归属概率

■ 用途

- 探索数据结构、组织样本、形成后续分析的候选群体

■ 与分类的区别

- 分类学习预先定义类别；聚类依据表示与方法假设提出候选分组

■ 关键边界

- 算法得到的簇不自动等于现实中的真实类别

同一组样本可以存在多种合理划分

■ 按空间位置分组

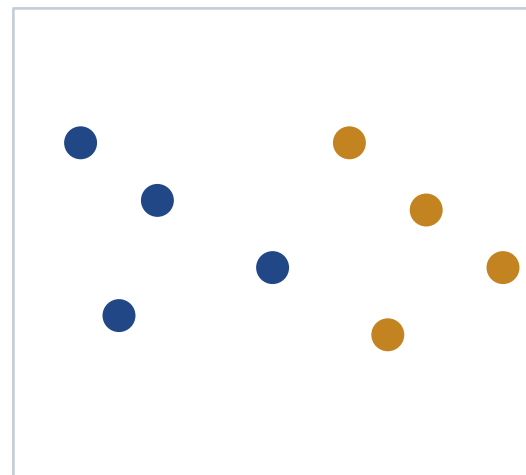
- 邻近点进入同一簇
- 适合寻找紧凑区域

■ 按方向或形状分组

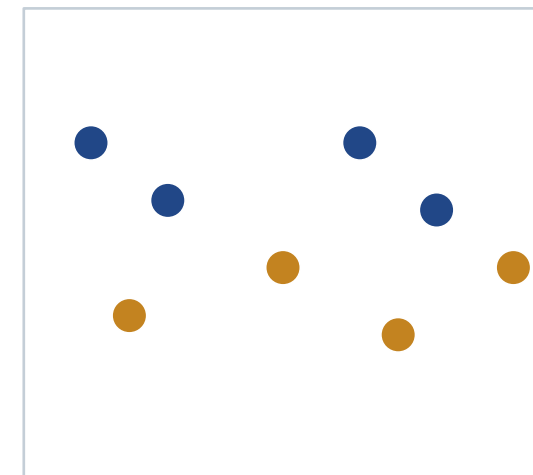
- 位于同一条带或曲线的点进入同一簇
- 适合寻找结构相似性

■ 判断依据

- 哪种表示和分组更能服务当前分析目的



按左右位置



按上下结构

点的位置完全相同，颜色表示两种分组解释

- 原始对象先被表示为特征向量
- 未进入表示的信息对聚类算法不可见
- 特征尺度会改变距离贡献
 - 例如 “收入以元计” 和 “年龄以年计” 可能由收入数值主导欧氏距离
- 标准化、编码与特征选择都可能改变聚类结果

- **欧氏距离**

- 直线距离，适合连续且尺度可比较的特征

- **曼哈顿距离**

- 各维绝对差之和，对坐标轴方向的差异更直接

- **余弦相似度**

- 比较方向而弱化整体长度，常用于稀疏文本向量

- **距离公式正确不表示它适合当前数据**

■ 硬聚类

- 每个样本只分到一个簇
- k 均值与常见层次聚类给出这种结果

■ 软聚类

- 为样本给出多个簇的归属程度或概率
 - 高斯混合模型可给出分量后验责任度
- 软归属表达边界模糊，不表示结果天然更正确

- 先问一个簇由中心、层次、密度还是概率分量定义
- 再看算法输出硬归属、软归属还是嵌套结构
- 方法地图用于比较假设，不表示算法排名

簇模型	怎样定义一个簇	典型方法
中心型	靠近代表中心	k 均值
层次型	属于嵌套结构	凝聚聚类
密度型	连接高密度邻域	DBSCAN
概率型	来自概率分量	GMM

2 k 均值

交替分配样本与更新质心，形成平坦的中心型划分

k 均值以质心定义簇

- 预先指定簇数 k
- 每个样本获得一个硬分配 $c_i \in \{1, \dots, k\}$
- 第 j 个质心 μ_j 是该簇样本的均值
- 目标是最小化簇内平方和
 - $J = \sum_{i=1}^N \|x_i - \mu_{c_i}\|_2^2$

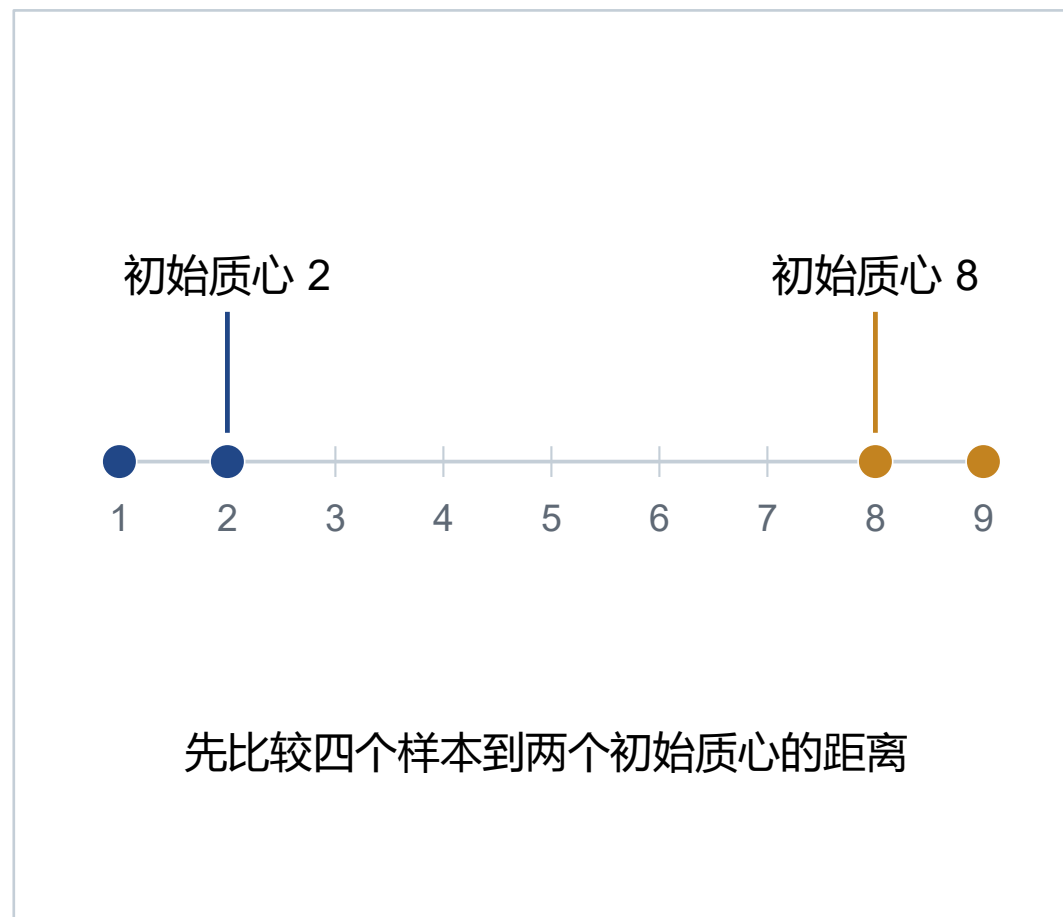
k 均值交替执行分配与质心更新

- **初始化**：选择 k 个起始质心
- **分配步骤**：每个样本分给最近的质心
- **更新步骤**：把每个质心移到所属样本的均值
- **重复**：直到分配不变、质心变化很小或达到迭代上限

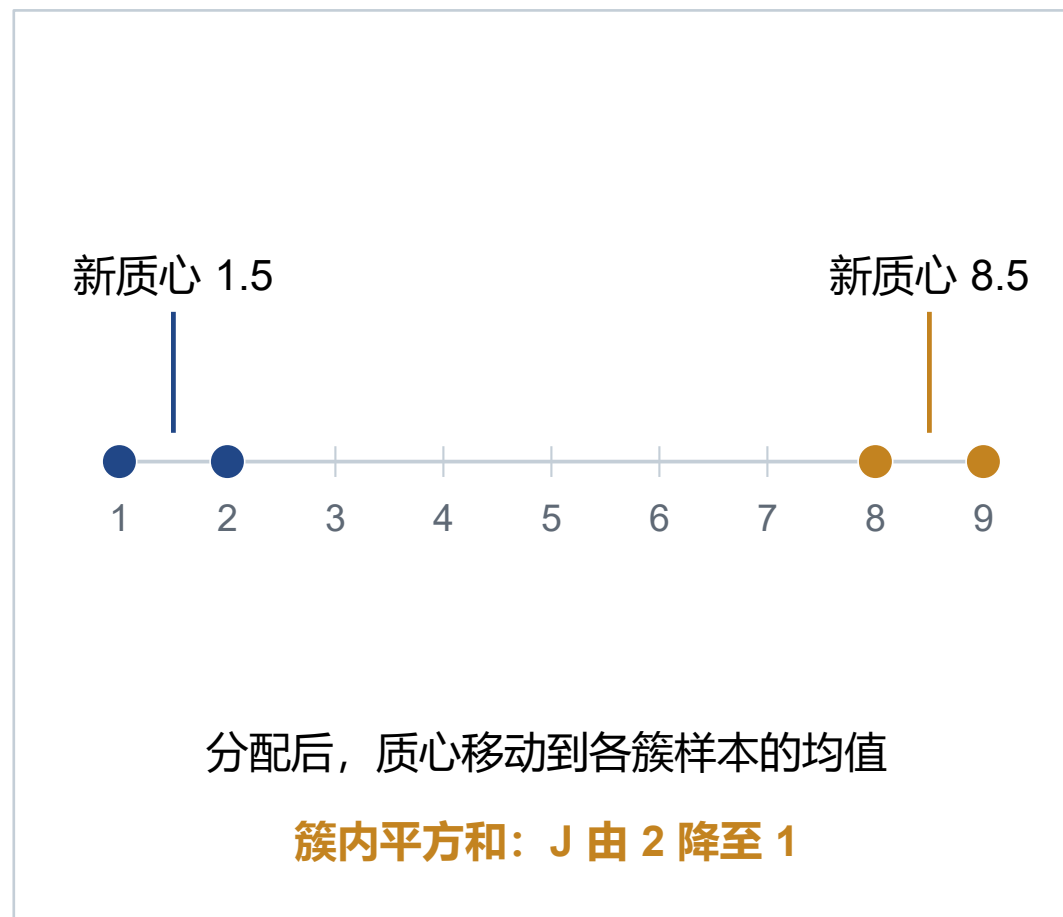


一维例子：四个样本与两个初始质心

- 四个样本位于数轴上的 1、2、8、9
- 设簇数为 2，初始质心位于 2 和 8
- 比较每个样本到两个质心的距离



- 最近质心分配：1、2 进入第一簇，8、9 进入第二簇
- 更新质心：第一簇均值为 1.5，第二簇均值为 8.5
- 新质心位于各簇样本的均值位置
- 质心更新后，簇内平方和由 2 下降到 1： $J^{(0)} = 2, J^{(1)} = 1$



分配不再改变时，当前迭代停止

- 以 1.5,8.5 为质心再次分配，两个簇保持不变
- 当前目标不再通过分配与均值更新下降
- 这表示达到一个局部稳定解
- 不同初始质心可能通向不同结果，因此实践中常进行多次初始化

k 均值适合围绕质心的紧凑簇

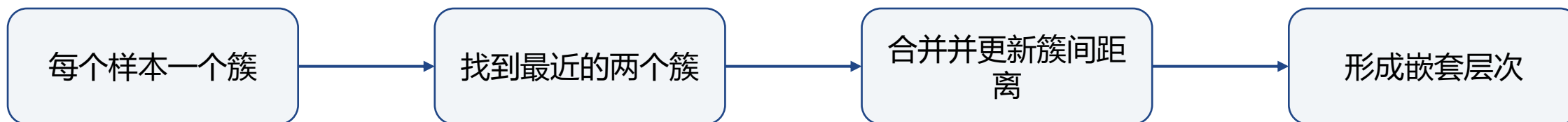
- 较适合尺度相近、近似球形且大小相近的簇
- 需要预先指定 k
- 对初始化、特征尺度和异常值敏感
- 对弯月形、环形或密度差异明显的的数据可能给出不自然划分
- 簇数选择需要结合稳定性、内部指标和实际用途

3 层次与密度方法

层次方法组织多种粒度，密度方法连接局部高密度区域

凝聚层次聚类从单点簇开始逐步合并

- 初始时每个样本各自构成一个簇
- 每一步合并距离最近的两个簇
- 更新簇间距离，再继续合并
- 直到所有样本进入一棵层次树，或达到预设停止条件



- **单链接**

- 使用两个簇中最近点的距离，容易连接成长链

- **全链接**

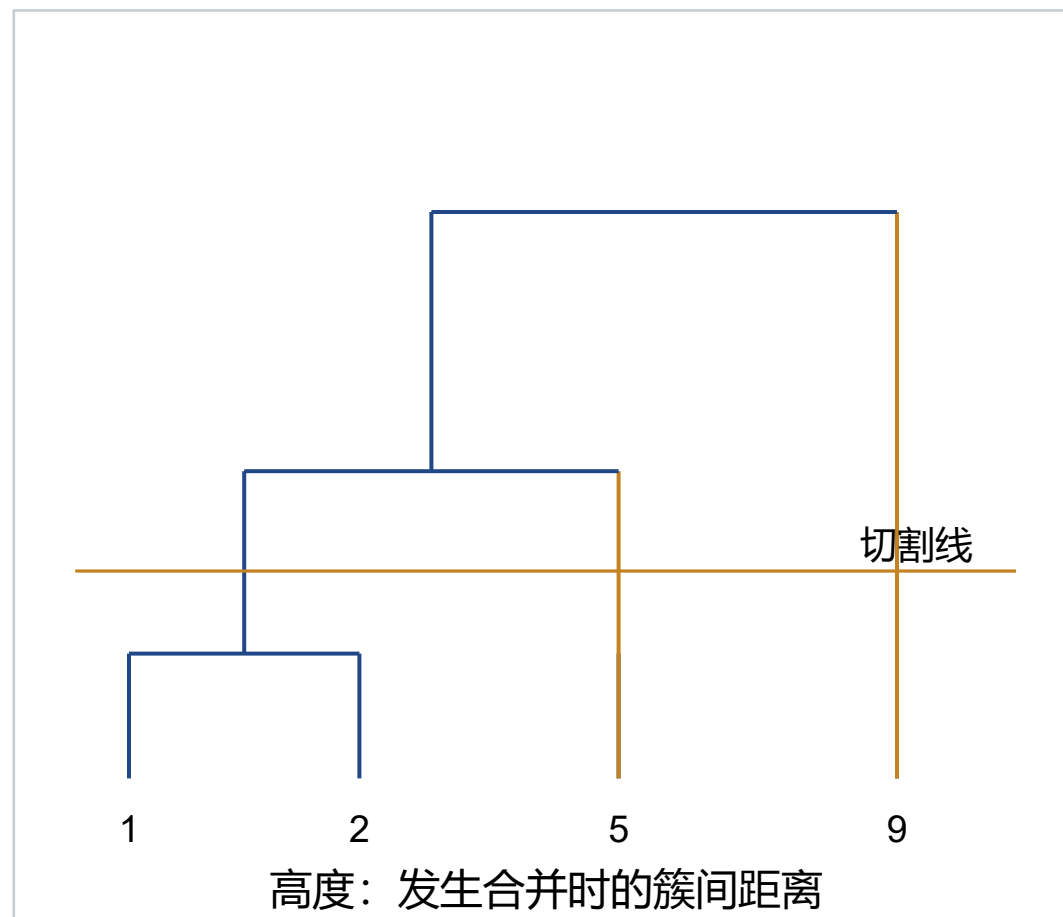
- 使用两个簇中最远点的距离，偏好紧凑簇

- **平均链接**

- 使用跨簇点对距离的平均值，折中描述整体接近程度

- **链接规则不同，合并顺序与最终层次可能不同**

- 四个一维点从四个单点簇开始，位置分别为 1、2、5、9
- 1 与 2 距离最近，因此先合并
- 纵轴高度表示按当前链接规则发生合并时的距离
- 在某个高度水平切割树，可得到相应粒度的平坦聚类
- 切割高度是分析选择，不是树状图自动给出的真值



■ ε 邻域

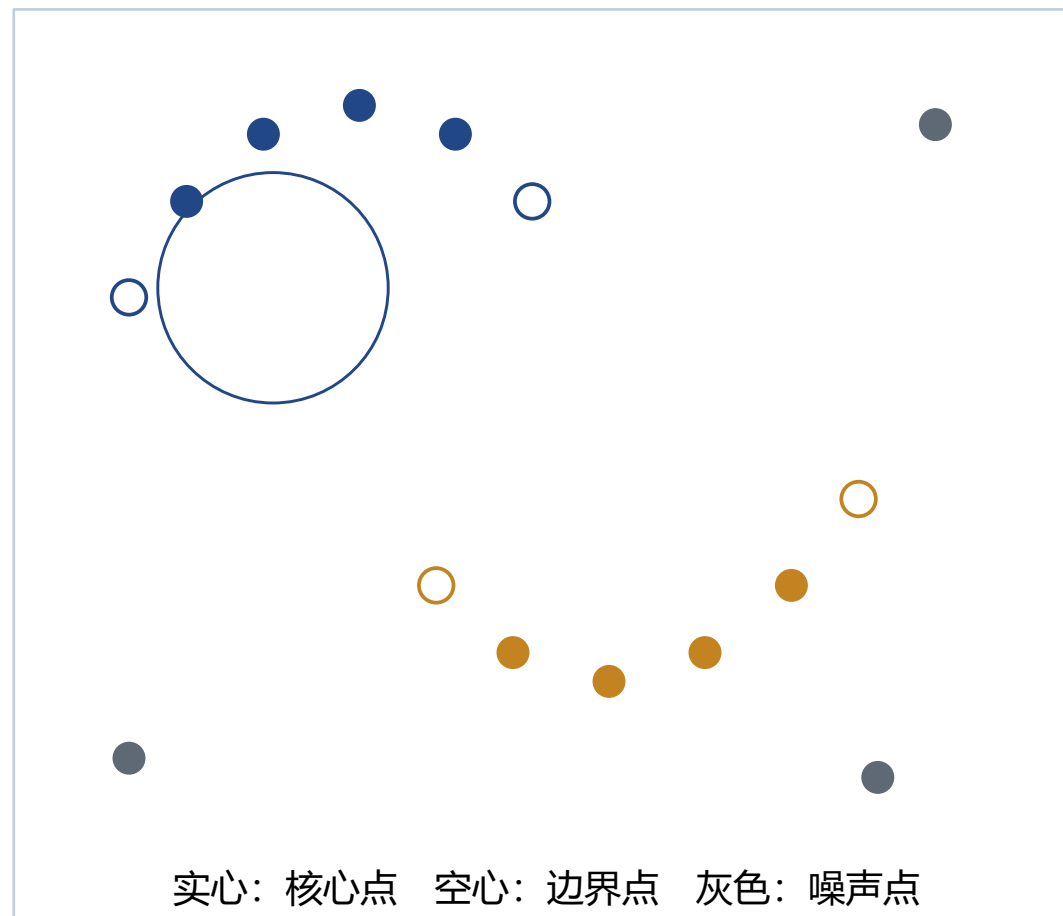
- $N_\varepsilon(x) = \{x_j \mid d(x, x_j) \leq \varepsilon\}$

■ 核心点

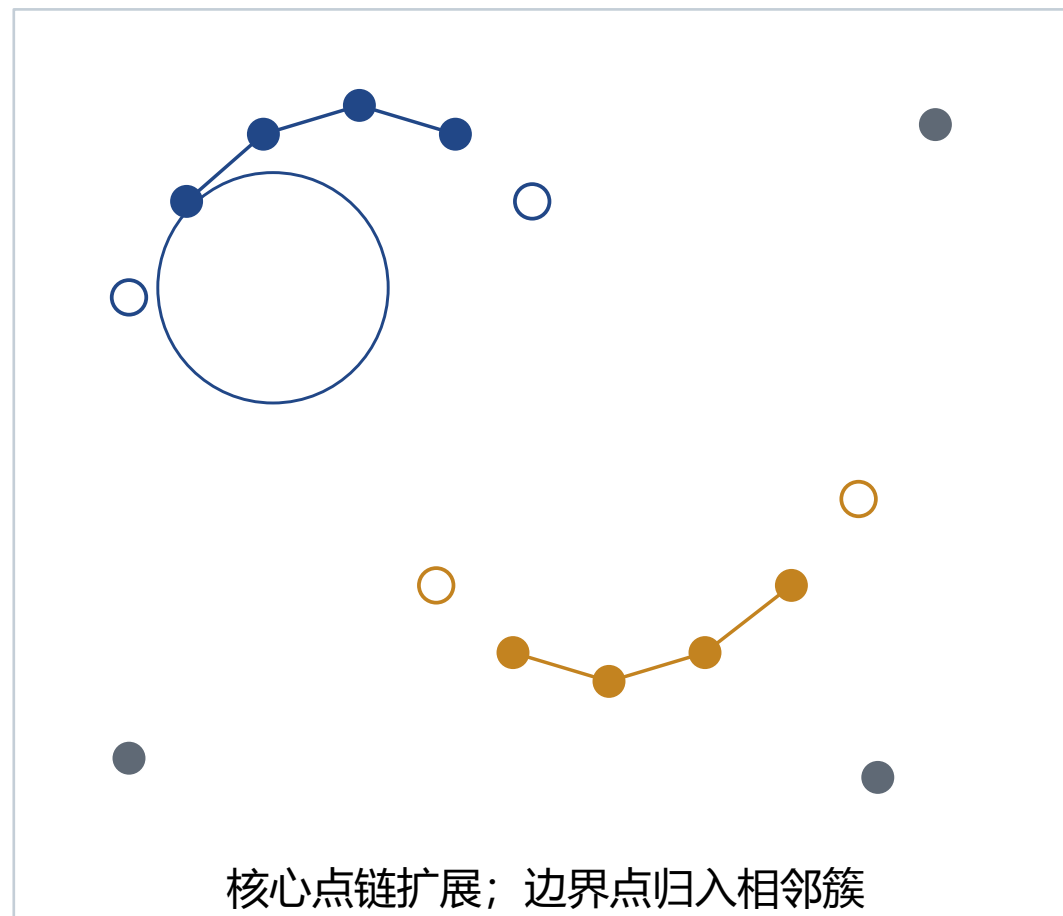
- $|N_\varepsilon(x)| \geq \text{MinPts}$
- 本课件约定邻域计数包含样本自身

■ 边界点：可归入相邻核心点形成的簇，但不能独立扩展

■ 噪声点：当前参数下未被任何簇覆盖，不自动等于异常



- 从一个核心点开始，寻找邻域中的其他核心点
- 相邻核心点继续扩展，形成密度连通的区域
- 边界点随相邻核心区域归入簇，噪声点留在簇外
- 这种扩展能够形成弯曲的非球形簇
- 密度差异较大时，一组参数可能无法同时描述所有簇



■ k 均值

- 中心型结构，输出硬平坦划分
- k 控制粒度；对尺度、初始化和异常值敏感

■ 层次聚类

- 嵌套结构，输出树状图
- 切割高度控制粒度；链接规则改变结果

■ DBSCAN

- 密度连通结构，输出簇与噪声标记
- 密度参数控制粒度；对密度差异敏感

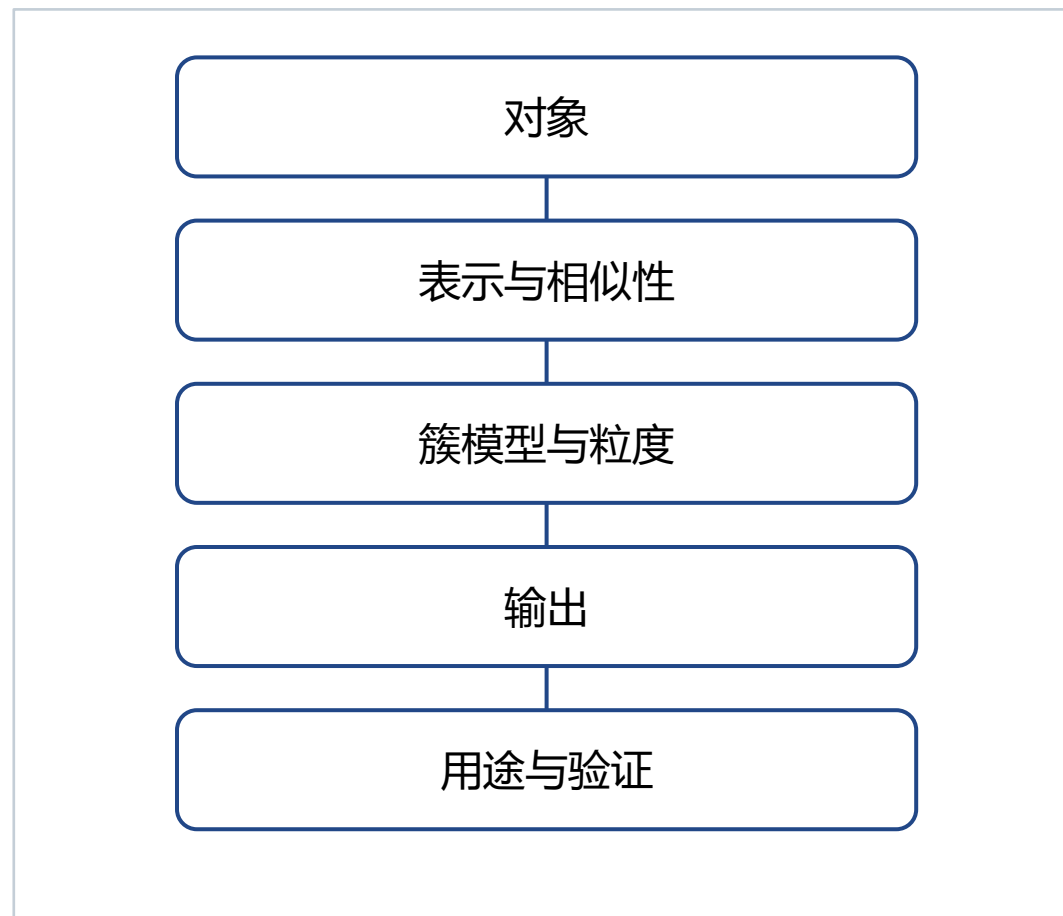
■ GMM

- 概率分量结构，输出软归属
- 分量数控制粒度；需要分布假设

4 应用与验证

聚类结果只有经过解释、稳定性检查和使用验证才具有实际价值

- **表示与相似性**：算法比较什么
- **簇模型与粒度**：寻找什么结构、观察多细
- **输出与验证**：怎样使用、证据何在



- **对象**：客户、用户或其他待分析样本
- **表示与相似性**：用频率、数量、偏好或测量指标形成特征，并处理量纲差异
- **结构与粒度**：若需要固定数量且便于描述的平坦群体，可比较不同 k 的中心型划分
- **输出与验证**：得到候选群体及其代表性特征，再检查跨样本和跨时间的稳定性
- **用途**：辅助描述群体、设计后续调查或比较方案

- **对象与表示**：把搜索结果、新闻或论文表示为词项权重或语义向量
- **相似性**：余弦相似度比较文档向量的方向
- **结构与输出**：按浏览任务形成平坦主题组或可逐层展开的层次，并提取代表内容
- **用途与验证**：支持结果浏览与语料探索；检查主题连贯性和用户任务效果

- **对象**：像素或先形成的局部图像区域
- **颜色量化**：按颜色聚类像素，用少量代表色近似原图
- **图像分割**：加入位置或局部区域特征，形成具有空间意义的区域
- **结构与粒度**：代表色数量或区域尺度决定结果保留多少细节
- **验证区别**：量化检查视觉信息与压缩效果；分割检查任务所需边界

- **结构粒度**： k 、切割高度和密度参数控制分组粗细
- **内部质量**：例如轮廓系数；它偏好簇内紧凑、簇间分离的几何结构
- **稳定性**：改变样本、初始化或合理参数后，主要结构是否仍出现
- **外部证据**：用留出变量、领域知识和后续任务效果解释结果

- **对象与表示**：先确定什么被聚类，以及算法能够看见哪些特征
- **相似性与簇模型**：距离、尺度和中心、层次、密度或概率假设共同定义簇
- **结构粒度与输出**：参数控制观察粗细，算法给出分组、树、软归属或噪声标记
- **聚类验证**：检查内部质量、稳定性、领域解释和后续任务价值