14 聚类方法

没有标签时怎样组织相似样本

本讲学习目标

- **分组依据**：区分聚类与分类
  - 说明数据表示、特征尺度、相似性和方法假设怎样影响分组
- **k 均值机制**：跟随一轮完整计算，解释分配与质心更新
  - 说明算法的目标、停止条件与主要局限
- **层次结构与其他输出**：看懂凝聚层次聚类的合并与树状图切割
  - 理解单链接的含义，简要认识密度分组与概率软归属
- **结果解释与验证**：结合消费行为分群示例，说明方法的候选理由与群体描述的依据
  - 知道内部质量、结果稳定性、业务解释和用途效果需要分别检查

1 基本概念

没有标签规定归属，分组依据需要明确选择

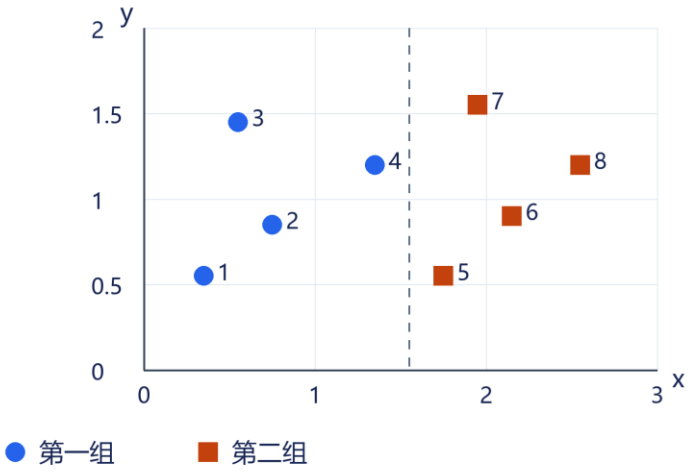
聚类与分类：标签的作用

- **聚类**：不使用已知类别标签来定义分组，而是依据数据表示、相似性和方法假设组织样本
  - 输入为样本特征 $x_1, \ldots, x_N$；按所选依据形成的样本组称为簇
- **与分类的区别**
  - 分类：使用已知类别标签，学习从样本特征到预先定义类别的判别规则
  - 聚类：不以已知类别标签规定归属，形成的组需要进一步解释其含义
- **用途**：探索结构、组织样本，为后续分析提供候选分组
  - 得到的簇不自动等于现实中的真实类别
- **分组依据**：没有标签规定归属时，需要选择依据什么信息、什么接近规则和什么结构形成簇

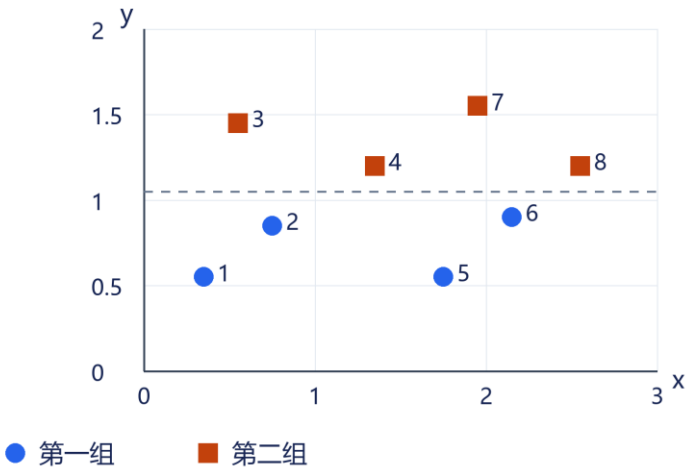
同一组样本的不同划分

- **样本**：八个样本各有两个特征，分别用横坐标和纵坐标表示
- **划分方式**
  - 左右划分：以横坐标 1.55 为界分成两组，用于比较横向位置差异
  - 上下划分：以纵坐标 1.05 为界分成两组，用于比较纵向位置差异
- **分组依据**：同一组样本可以依据不同维度划分，划分是否合理需要结合观察目的判断
  - 不同分组依据不意味着任意划分都合理；图中分组不代表已知真实类别

左右划分：x = 1.55



上下划分：y = 1.05



表示与相似性决定接近程度

- **数据表示**：用选定特征组成向量，作为算法的样本输入
  - 特征选择需结合观察目的；未表示的信息不参与比较
- **特征尺度**：特征的单位和数值变化范围影响它在欧氏距离中的贡献
  - 收入以元计、年龄以年计时，收入的数值差异可能主导距离
  - 标准化可调整各维贡献，是否采用及具体方式需结合特征含义
- **相似性**：样本在当前数据表示和比较规则下的接近程度
  - 用距离或相似度衡量；规则选择需结合特征含义与分析目的
  - 欧氏距离为直线距离，曼哈顿距离为各维绝对差之和；距离越小，按该规则越接近
  - 余弦相似度比较非零向量的方向；数值越大，方向越接近
- **分组结果**：表示、尺度或比较规则改变，分组可能改变；比较样本之外，还需规定怎样组织簇

不同方法采用不同的簇结构规则

方法	簇结构与组织规则	结果形式	本讲学习范围
k 均值	质心周围的紧凑簇	硬归属的平坦分组	目标与完整迭代
凝聚层次聚类	按簇间距离合并	嵌套树及切割分组	合并与切割示例
DBSCAN	按密度连接形成区域	可弯曲的簇及噪声标记	认识输出形式
GMM	用多个高斯分量描述数据	分量后验概率	认识概率软归属

- **归属方式**：硬归属为每个已分簇样本指定一个簇；软归属给出对多个簇的归属程度，GMM 是概率型实例
  - 与第09讲联系：k 均值用硬分配与簇均值更新，GMM 的 EM 用概率责任度与加权参数更新；两者一般不等价
- **层次结构**：平坦分组只有一层，层次结构保存嵌套关系；与硬 / 软归属是不同维度；噪声指未分簇样本，不证明现实异常

2 k 均值

用目标解释交替更新、计算与停止边界

k 均值的簇内平方和目标

- **分组任务**：将无类别标签的样本分成预先给定的 k 个簇，用一个质心代表每个簇
  - 输入为 N 个样本向量 x_i；簇数 k 由分析者预先选择
- **待求量**
  - 分配 $c_i \in \{1, \ldots, k\}$：指定样本 x_i 属于哪个簇
  - 质心 μ_j：第 j 个簇的代表中心，均值更新后为该簇样本的算术均值，可以不是实际样本
- **建模思路**：用质心代表簇，以样本偏离其质心的平方距离衡量代表误差
  - 将各样本的误差相加，得到簇内平方和

$$J = \sum_{i=1}^{N} \| x_i - \mu_{c_i} \|_2^2$$

  - μ_{c_i} 是样本 x_i 所属簇的质心；$\|x_i - \mu_{c_i}\|_2$ 是两者的欧氏距离
- **优化目标**：在给定 k 下，调整分配与质心，使簇内平方和 J 最小

k 均值的分配与更新

分配与质心相互依赖，从初始质心出发交替求解同一簇内平方和目标。

- **分配步：固定质心，优化归属**
  - 对每个样本取 $c_i = argmin_j \|x_i - \mu_j\|_2^2$，逐项最小化平方误差
- **更新步：固定归属，优化质心**
  - C_j 为第 j 簇的样本索引集，$|C_j|$ 为样本数；对非空簇求梯度

$$\nabla_{\mu_j} \sum_{i \in C_j} \| x_i - \mu_j \|_2^2 = 2|C_j|\mu_j - 2 \sum_{i \in C_j} x_i$$

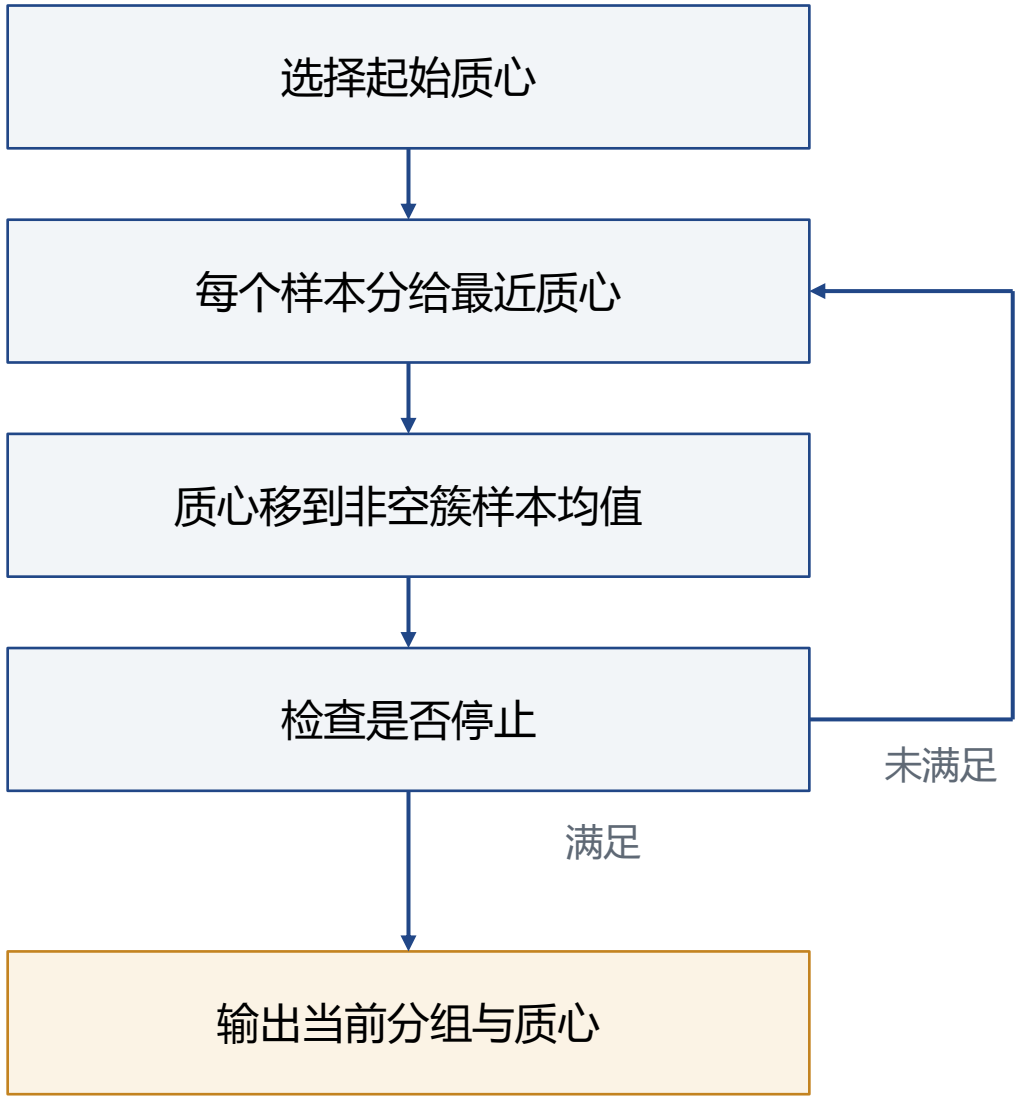
  - 令梯度为零，得到算术均值；非空簇的目标是严格凸二次函数，因此这是唯一最小点

$$\mu_j = \frac{1}{|C_j|} \sum_{i \in C_j} x_i$$

- **目标变化**：两步各自固定另一类待求量；精确执行时，每一步都不增加同一目标

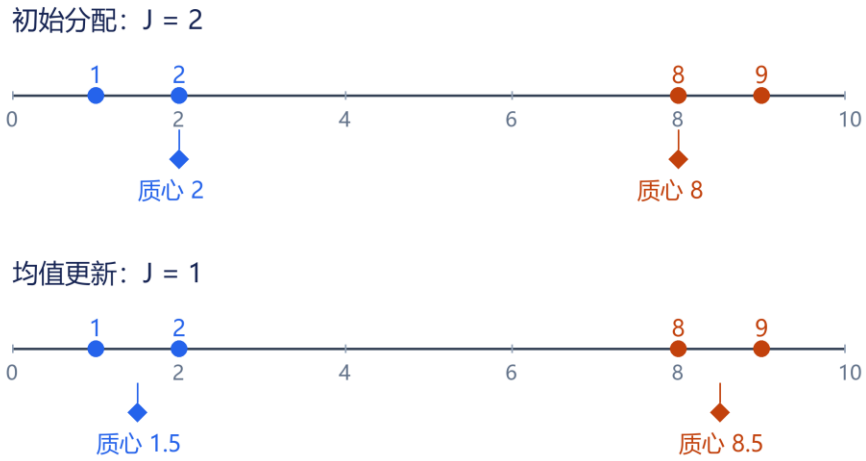
k 均值的迭代流程

- **初始化**：给定样本与簇数，选择起始质心
- **一次迭代**：按当前质心分配样本，再按新归属更新质心
- **循环出口**：满足停止条件时输出当前分组和质心；否则继续下一轮分配
  - 分配不变、质心变化很小或达到迭代上限，可作为停止规则
  - 容差或上限停止的性质需要另行判断，下一计算示例将核对实际稳定状态



示例：k 均值的一轮迭代

- **输入**：四个样本位于 1、2、8、9；$k = 2$，初始质心为 2、8
- **样本分配**：按最近质心分为 1、2 与 8、9 两簇
- **质心更新**：用各簇样本的均值更新质心
  - 第一簇：$(1 + 2)/2 = 1.5$；第二簇：$(8 + 9)/2 = 8.5$



- **目标变化**：质心更新后，簇内平方和由 2 下降到 1：$J^{(0)} = 2, J^{(1)} = 1$
  - 更新前的平方和为 $(1 - 2)^2 + (2 - 2)^2 + (8 - 8)^2 + (9 - 8)^2 = 2$
  - 更新后的四项误差均为 0.5^2，所以 $J^{(1)} = 4 \times 0.25 = 1$
- **交替作用**：先固定质心求分配，再固定分配求质心；每步只优化其中一类待求量

迭代停止与最优性

- **局部稳定**：当前分配与均值更新都不再改变状态
  - 本例以 1.5,8.5 为质心再次分配，样本归属不变；按相同归属更新质心，位置也不变，簇内平方和保持为 1
- **全局最优**：在给定 k 下，目标值在所有允许的分配与质心选择中最小
  - 局部稳定只说明当前两步更新无法继续改善，不保证得到全局最优
- **停止条件**：质心变化很小或达到迭代上限时也可停止，但不能据此认定分配已稳定
- **初始化影响**：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
- **特殊情况**：空簇和等距分配需明确处理规则；本例各簇非空且无等距分配

质心与平方距离的结构偏好

- **结构依据**：最近质心分配将空间分成围绕各中心的区域；平方距离惩罚样本偏离中心
  - 倾向形成围绕质心的紧凑划分；近似球形、尺度相近是常见适用情形，不是算法保证
  - 目标没有直接约束簇的形状、尺度或样本数，不保证各簇大小相近
  - 非凸形状、不同密度或显著不同尺度可能导致不理想划分；弯月形、环形可能被最近质心区域切开
- **影响因素**
  - 簇数 k 需要预先选择，决定平坦分组的数量
  - 初始化影响迭代起点，可能使算法通向不同结果
  - 特征尺度影响距离贡献，异常值可能拉动簇均值，进而影响划分
- **簇数选择**：结合稳定性、内部指标和实际用途判断
  - 稳定性关注改变初始化或样本后主要分组是否相似；内部指标按分组的距离、紧凑性等性质评分

3 层次结构

通过合并与切割认识嵌套层次

层次聚类的嵌套结构与合并规则

k 均值给出一层分组；如果还想观察小簇怎样组成更大的簇，就需要保留不同粒度分组之间的关系。

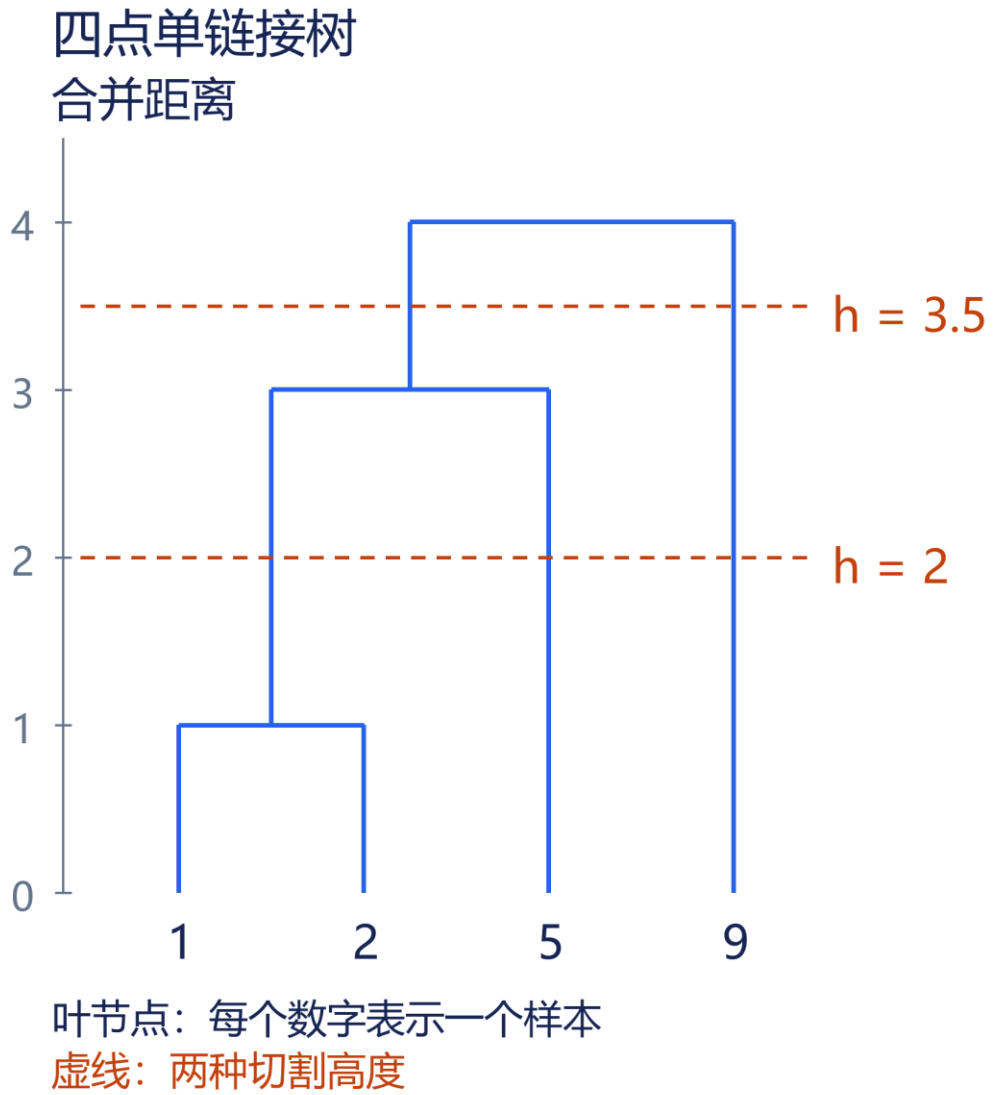
- **层次聚类**：将样本组织成多个粒度的嵌套分组，用层次结构保存簇之间的包含关系
  - 多个小簇组成一个大簇时，仍保留原来各小簇的分组关系，这就是嵌套
- **凝聚过程**：从每个样本各自成簇开始，从细到粗逐步合并；本讲采用这种方式
  - 按簇间距离选最近两簇合并，重算新簇到其余簇的距离，重复至全部样本成一簇
- **单链接规则**：比较分别来自两簇的样本点对，以最近点对的距离作为簇间距离
  - 全链接取最远点对距离，平均链接取点对距离的平均；规则改变时，层次可能改变
  - 单链接有链式效应：一串相近的桥接样本可能逐步连接两个整体相距较远的群体，形成一个大簇
- **树状图**：保存合并对象、距离与嵌套结构；按选定高度切割得到一层分组

示例：单链接的合并距离

- **样本与规则**：四个一维样本的坐标为 1、2、5、9，采用欧氏距离与单链接
- **第一次合并**：最近的样本为 1、2，距离为 $2-1=1$，合并为 $\{1,2\}$
- **第二次合并**：更新簇间距离后，最近两簇为 $\{1,2\}$、$\{5\}$
  - 跨簇最近点对为 2、5，距离为 $5-2=3$，合并为 $\{1,2,5\}$
- **第三次合并**：$\{1,2,5\}$ 与 $\{9\}$ 的最近点对为 5、9
  - 距离为 $9-5=4$，全部样本合为一簇
- **记录结果**：三次合并高度为 1、3、4；这些距离与合并对象共同构成下一页的树状图

示例：树状图的高度与切割

- **叶节点**：底部数字表示原样本坐标；横向等距排列不表示样本间距离
- **合并高度**：从下向上读三条横线，分别记录上一页距离为 1、3、4 的簇合并
- **切割规则**：在选定高度画水平线，只保留线下的合并；已连接的叶节点属于同一簇
  - 高度 2：得到 $\{1,2\},\{5\},\{9\}$ 三个簇
  - 高度 3.5：得到 $\{1,2,5\},\{9\}$ 两个簇
- **分组粒度**：升高切割线后分组更粗；高度由分析需要决定，不自动代表真实类别



4 结果解释与验证

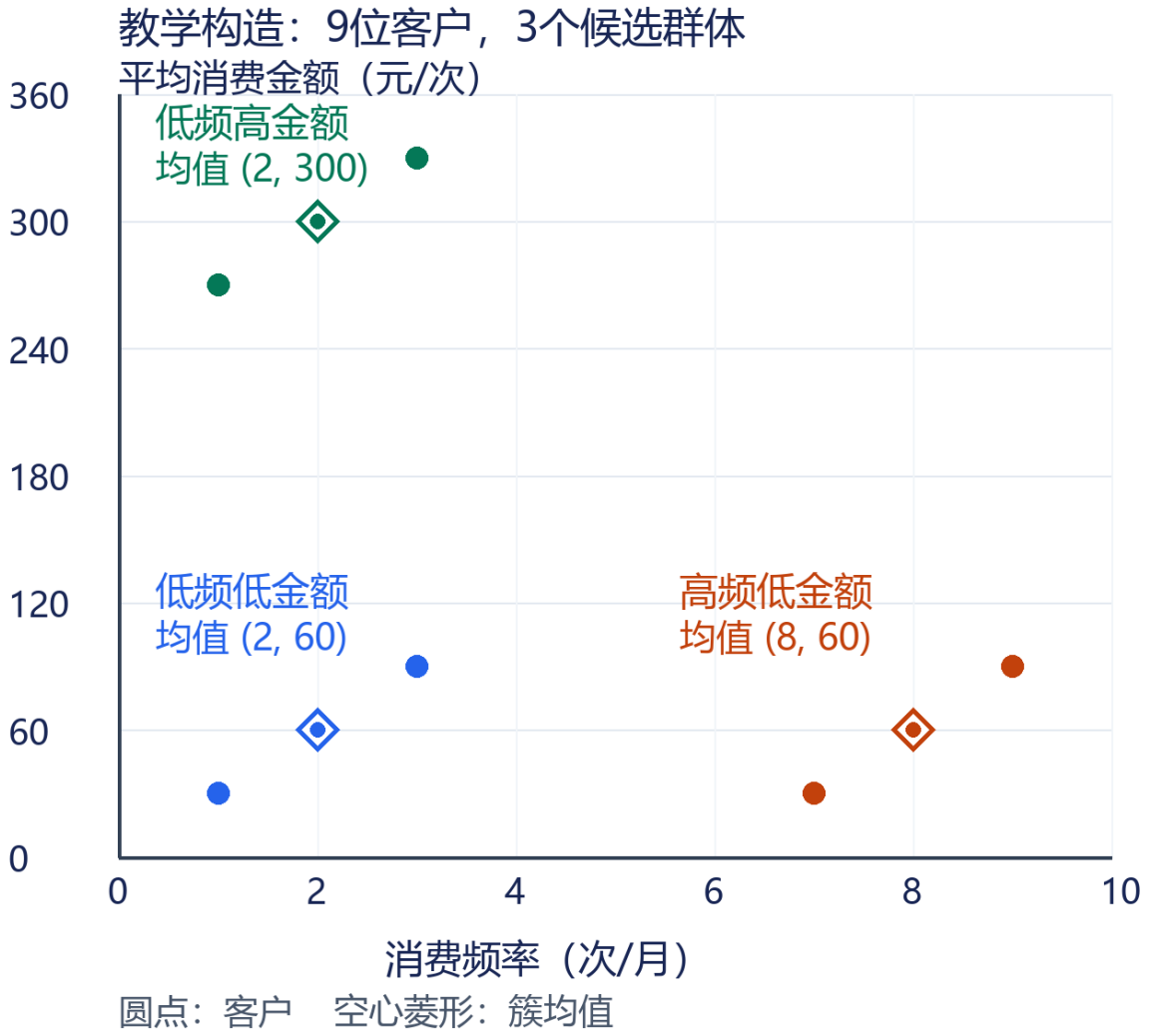
从消费特征解释分组，以不同证据判断结果与用途

示例：消费行为分群

教学构造：9位客户，3个候选群体。

- **表示**：消费频率与平均金额；按两维标准差缩放后比较距离
- **候选方法**：$k = 3$ 的中心型平坦分组；弯曲密度区域或嵌套层次需另选规则
- **输出**：客户归属与质心；图中质心换回原单位
- **解释**：比较簇均值，形成图中三个消费特征名称
  - 高低相对本例其他簇而言；不说明真实类别、满意度或留存

后续用途还需哪些证据？



聚类验证：不同问题需要不同证据

检查维度	回答的问题	证据与条件
内部质量	是否符合几何准则?	固定数据、表示和 k，簇内平方和下降支持中心型紧凑性
结果稳定性	分组能否重复出现?	固定特征、距离与 k，改变初始化或样本，比较分组及特征差异
业务解释	业务判断是否有依据?	消费特征支持描述；业务判断需领域知识与未参与聚类的满意度、留存等变量
用途效果	是否帮助后续任务?	按分组设计调查或服务方案后，检查任务是否实际受益

- **证据职责**：四类证据分别回答不同问题；目标下降或结果稳定，不能单独证明真实类别；稳定性和解释不能替代用途效果证据

本讲总结

- **分组依据**：聚类依据表示、相似性和方法规则组织无标签样本；用途引导这些选择，结果通常没有唯一答案
- **k 均值机制**：给定簇数 k，以减小簇内平方和为目标，形成中心型平坦分组
  - 从初始质心出发，交替执行最近质心分配与簇均值更新
  - 精确分配与非空簇均值更新不增加目标；分配与质心不再改变表示当前稳定，停止不保证全局最优
- **层次结构与其他输出**
  - 凝聚层次聚类逐步合并簇，保存嵌套关系；切割树状图，得到所选粒度的一层分组
  - DBSCAN 按密度规则形成簇，并保留未分簇样本的噪声标记
  - GMM 给出样本对各分量的后验概率，表达概率软归属
- **结果解释与验证**：依据特征描述分组，分别检查内部质量、结果稳定性、业务解释与用途效果
  - 算法输出不自动等于真实类别；目标下降或结果稳定，不能单独证明用途有效