lec14.3 Divisive clustering

分裂层次聚类（Divisive Clustering）

分裂（Top-Down）方法的主要思想

1. 自顶向下的方式：使用自顶向下的方法，将数据对象逐步划分成树状结构。
2. 平面聚类算法的使用：
    - 在每一步划分中，可以使用任意平面聚类算法（例如 k-means 算法）来进行划分
    - 这种算法 $\mathscr{A}$ 可以是任意聚类算法，不一定是基于距离的算法

分裂层次聚类的策略：
分裂方法提供了灵活性，可以在树结构的平衡性和每个簇中的对象数量的平衡之间进行选择。以下是两种常见策略：

- 策略 1：拆分最重的节点：
    - 拆分包含对象数量最多的簇（最重的节点）
    - 结果：叶节点（簇）中将包含相似数量的对象
- 策略 2：将每个簇分割成相同数量的子簇：
    - 将每个簇分割成相同数量的子簇
    - 结果：树结构是平衡的，每个节点有相同数量的子节点，但叶节点（最细粒度的簇）将包含不同数量的对象

一般分裂聚类算法

1. input：
    - 数据集 $\mathscr{D}$：待聚类的数据集
    - 平面算法 $\mathscr{A}$：用于每一步划分的聚类算法，可以是任意聚类算法，不一定是基于距离的算法
2. init：初始化树 $\mathscr{T}$，包含一个根节点，该节点代表整个数据集 $\mathscr{D}$
3. repeat：
    1. 在树 $\mathscr{T}$ 上选择一个叶节点 L
        - 需要定义一个标准来选择在每一步进行划分的叶节点。例如，可以选择包含对象最多的节点或其他特定标准
    2. 使用指定的平面聚类算法 $\mathscr{A}$ 将选择的叶节点 L 划分成若干子簇 $L_1, \ldots, L_k$
    3. 将子簇 $L_1, \ldots, L_k$ 作为 L 的子节点添加到树 $\mathscr{T}$ 中
4. until：持续进行步骤 3 的操作，直到达到预定义的终止条件
    - 需要定义一个明确的终止条件，决定何时停止进一步的划分。例如，可以是达到预定的簇数量或每个簇中的对象数量少于某个阈值等
5. return：返回最终的聚类结果或构建的层次聚类结构

Bisecting k-Means

1. input：
    - 数据集 $\mathscr{D}$：待聚类的数据集
    - 簇的数量 s：希望得到的簇的数量

2. init：初始化树 $\mathcal{T}$，包含一个根节点，该节点代表整个数据集 $\mathcal{D}$
3. repeat：
    1. 在树 $\mathcal{T}$ 上选择一个叶节点 L，使其具有最大的平方和距离

$$\sum_{X,Y \in L} \text{dist}(X,Y)^2$$

    2. 使用 k-means 将选择的叶节点 L 划分成两个簇 L_1、L_2
    3. 将 L_1、L_2 作为 L 的子节点添加到树 $\mathcal{T}$ 中
4. until：$\mathcal{T}$ 中的叶节点数量达到 s
5. return：返回当前的叶节点（簇），即最终的聚类结果