复习

1. 将词汇表中的词或短语映射为固定长度向量的技术被称为_____；

2. 简述在自然语言处理领域中，语言模型的基本概念（或基本任务）；

3. 以下不属于大语言模型关键技术的一项是（ ）：

   A. 大规模预训练模型　　B. 词汇网络构建

   C. 指令微调　　D. 基于人类反馈的强化学习

《神经网络与深度学习》

无监督学习

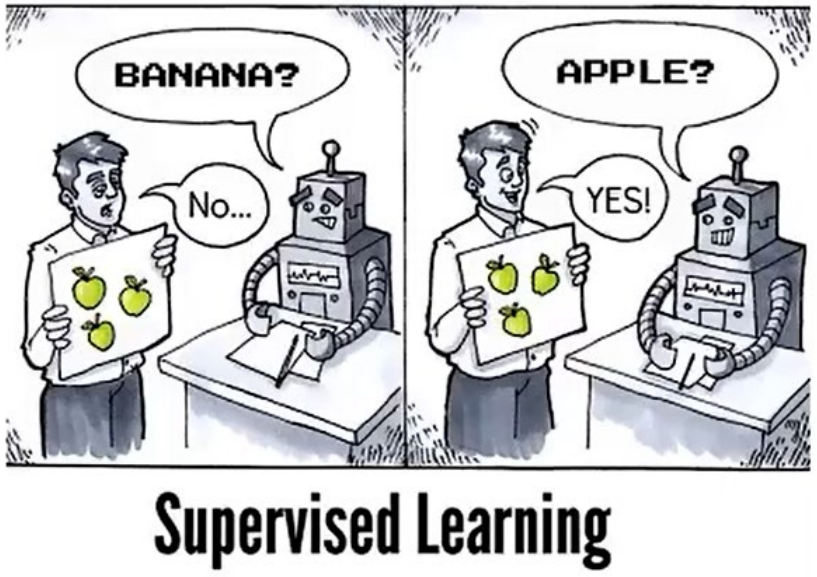
https://nndl.github.io/

课程大纲

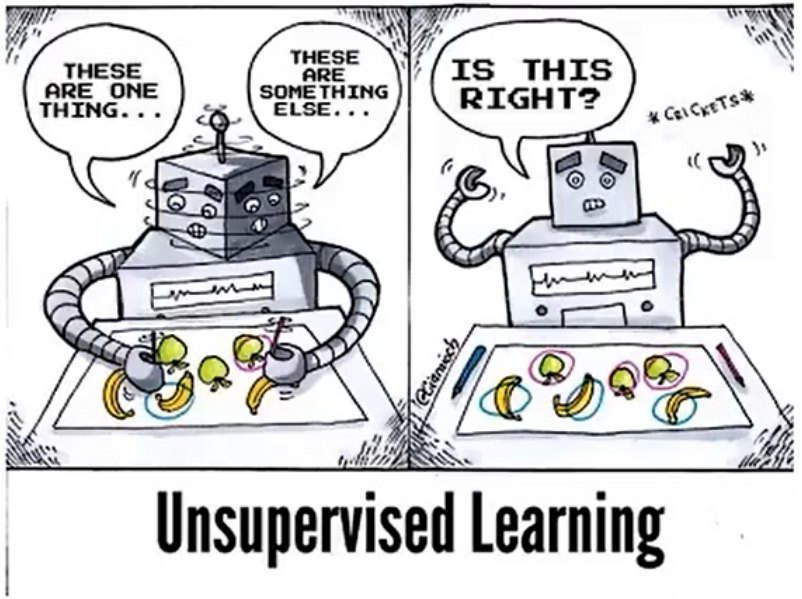
- 概述
- 基础网络模型
  - 前馈神经网络 √
  - 卷积神经网络 √
  - 循环神经网络 √
  - 网络优化与正则化 √
  - 记忆与注意力机制 √
  - 无监督学习 ○
- 深度学习计算
  - 深度学习框架与系统 √
- 进阶模型
  - 概率图模型 ○
  - 玻尔兹曼机
  - 深度信念网络 ○
  - 深度生成模型 ○
  - 深度强化学习
  - 序列生成模型 ○
- 深度学习应用
  - 计算机视觉 √
  - 自然语言处理 √

无监督学习（Unsupervised Learning）

- 监督学习
  - 建立映射关系 $f: x \rightarrow y$



- 无监督学习
  - 指从无标签的数据中学习出一些有用的模式。



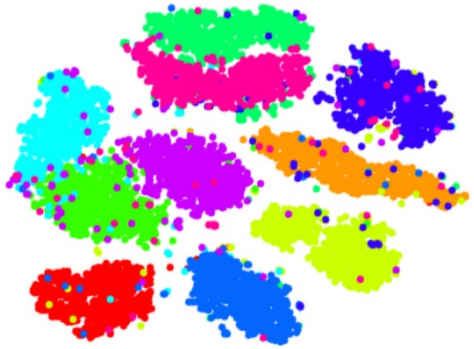
为什么要无监督学习?

大脑有大约10^{14}个突触，我们只能活大约10^{9}秒。所以我们有比数据更多的参数。这启发了我们必须进行大量无监督学习的想法，因为感知输入（包括本体感受）是我们可以获得每秒10^{5}维约束的唯一途径。

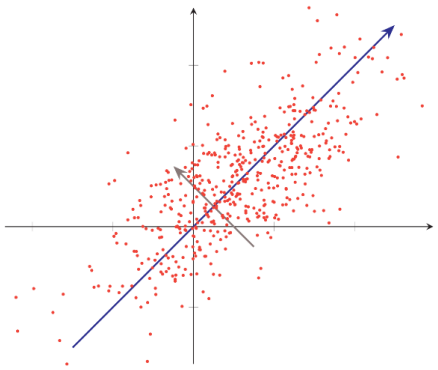
-- Geoffrey Hinton, 2014 AMA on Reddit

典型的无监督学习问题

聚类：发掘数据的纵向结构



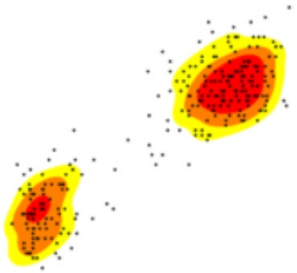
特征学习：
发掘数据的横向结构



$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	...	$x^{(n)}$

数据集

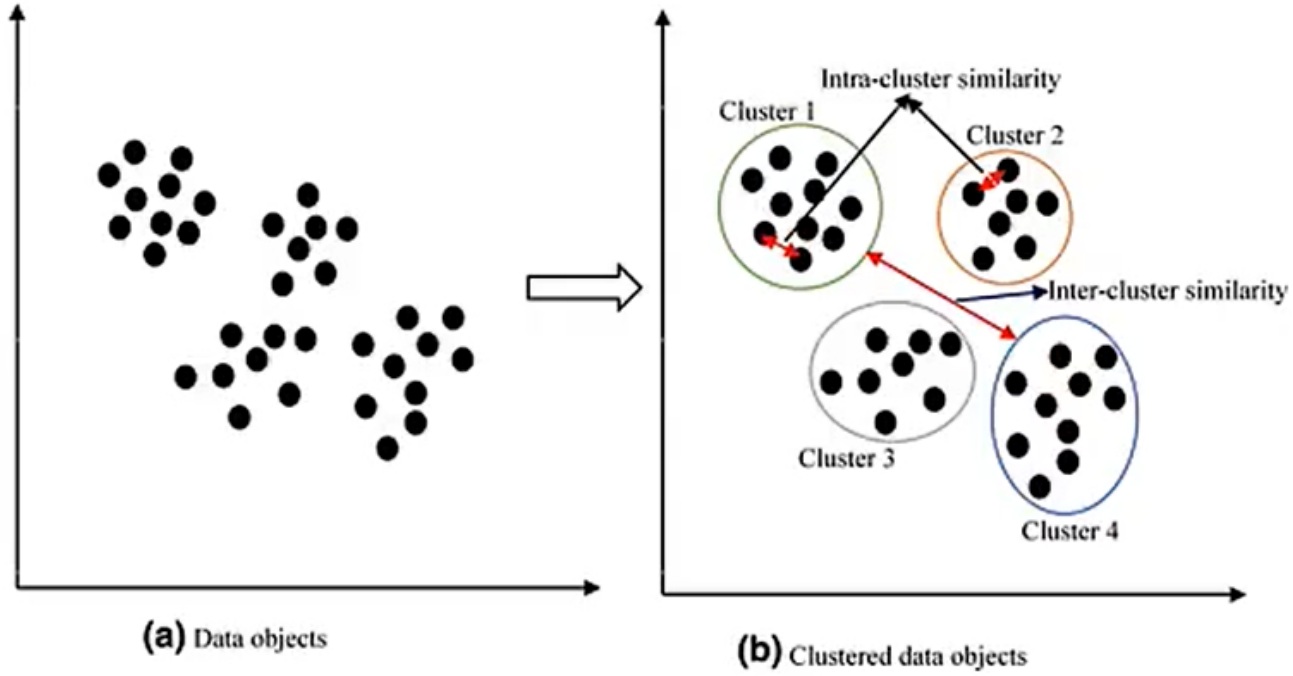
密度估计：
发掘底层数据分布



聚类

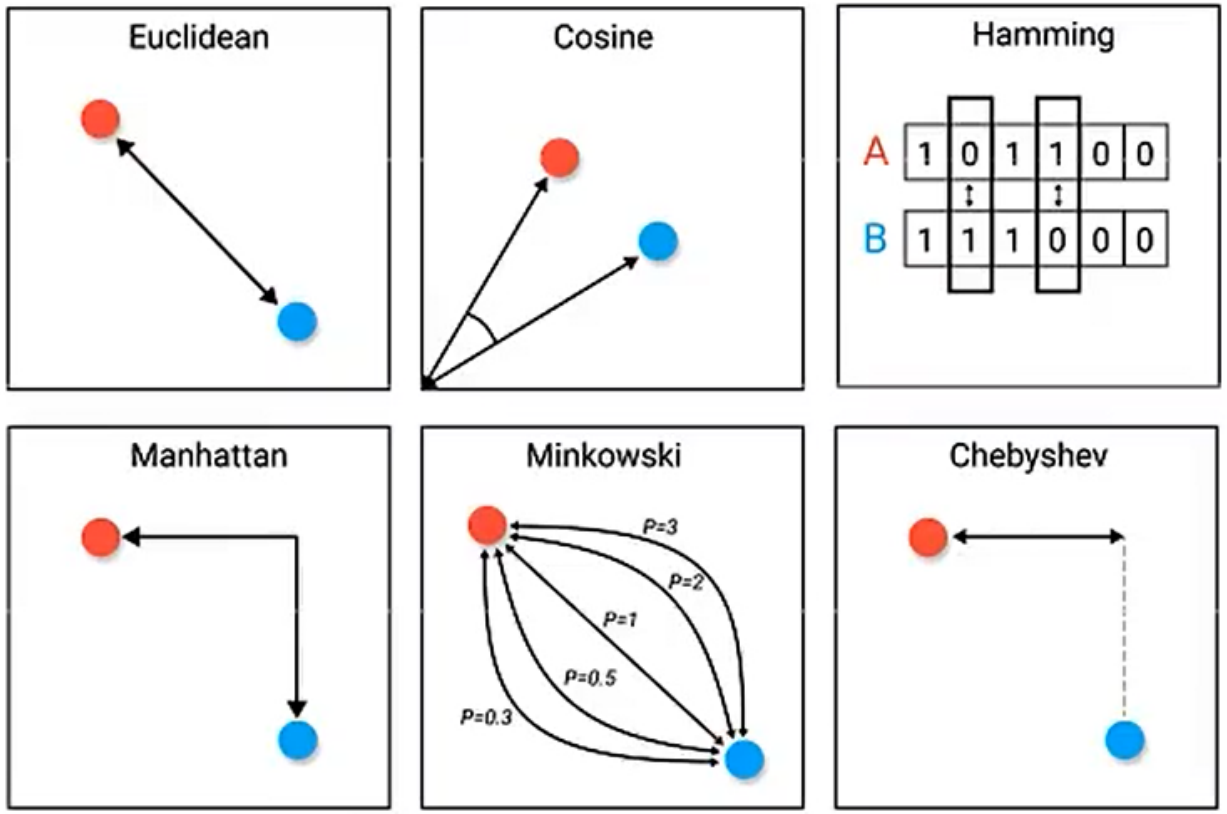
聚类

- 聚类（Clustering）
  - 将样本集合中相似的样本分配到相同的类/簇(cluster)，不相似的样本分配到不同的类/簇，使得类内样本间距较小而类间样本间距较大



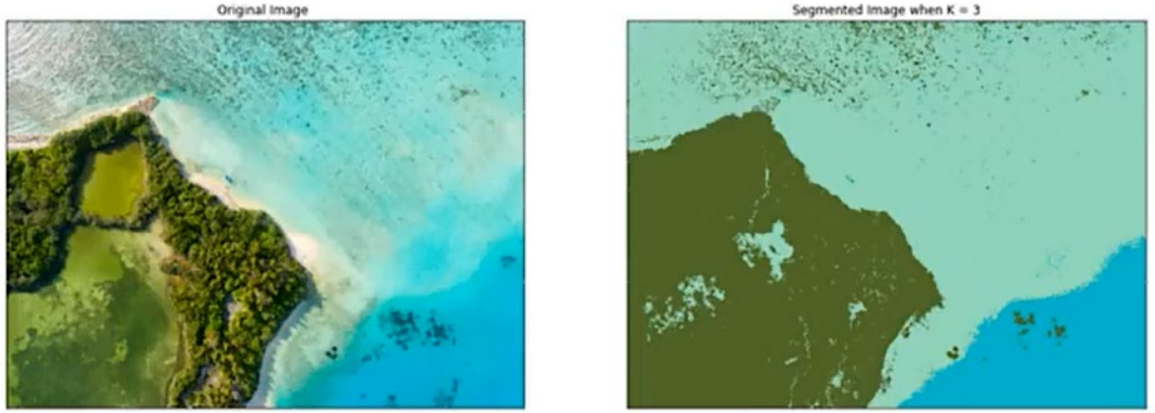
聚类

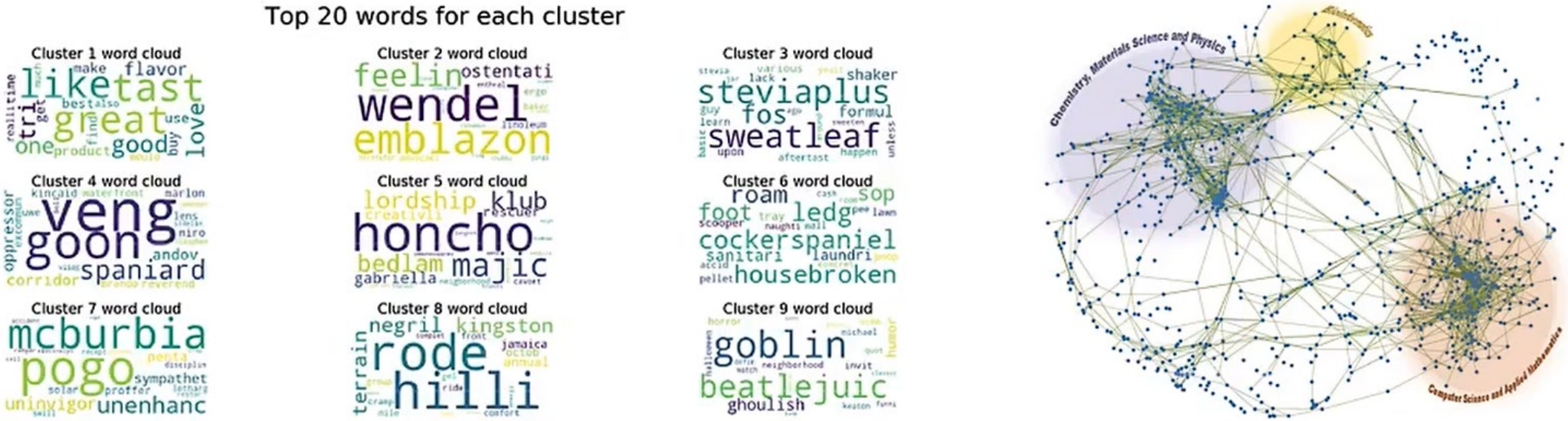
- 样本间距离/相似性
  - L1、L2距离
  - 余弦距离
  - 相关系数
  - 汉明距离



聚类的应用

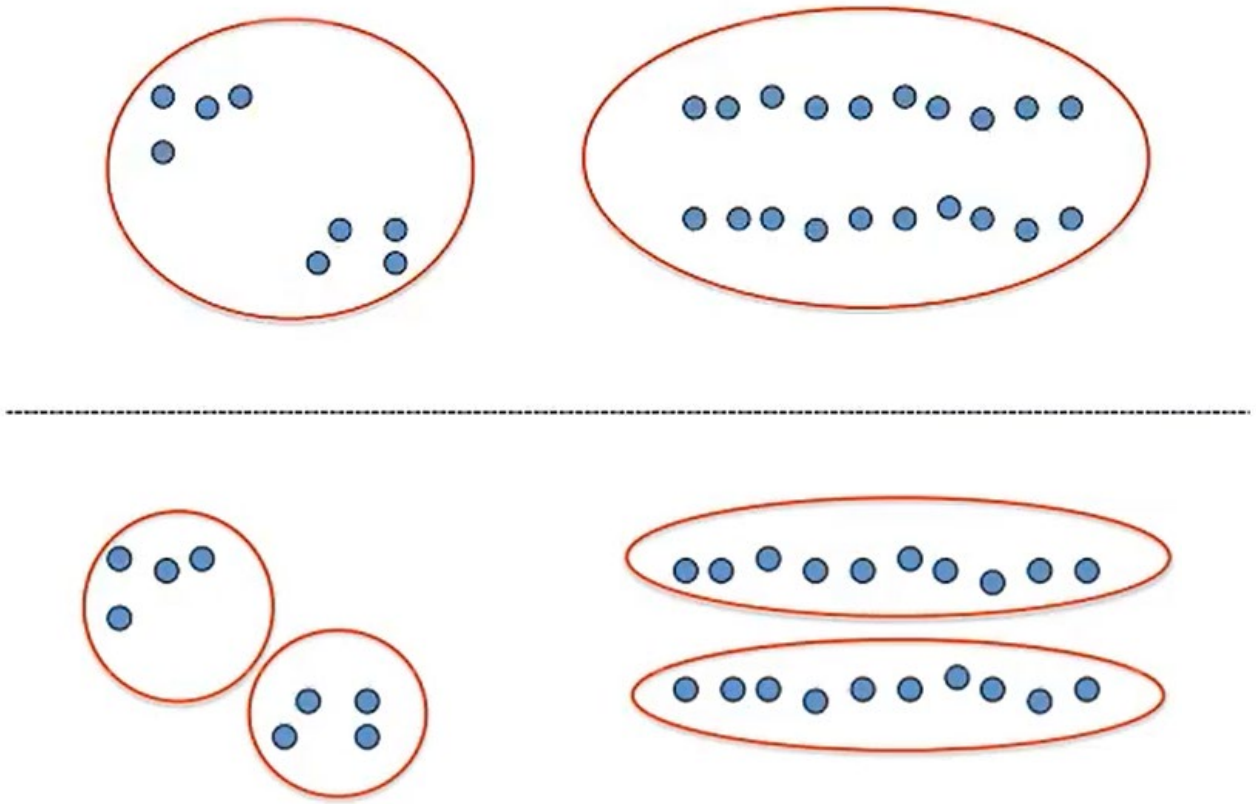
- 常见的聚类任务
  - 图像分割
  - 文本聚类
  - 社交网络分析





核心概念

- 类/簇(cluster)
  - 类没有一个严格的定义，可以理解为一组相似的样本



核心概念

- 类内间距
  - 样本间平均距离

$$avg(C) = \frac{2}{|C|(|C|-1)} \sum_{1 \le i < j \le |C|} d_{ij}$$

  - 样本间最大距离（直径）

$$dia(C) = \max_{1 \le i < j \le |C|} d_{ij}$$

- 类内间距
  - 样本间最短距离

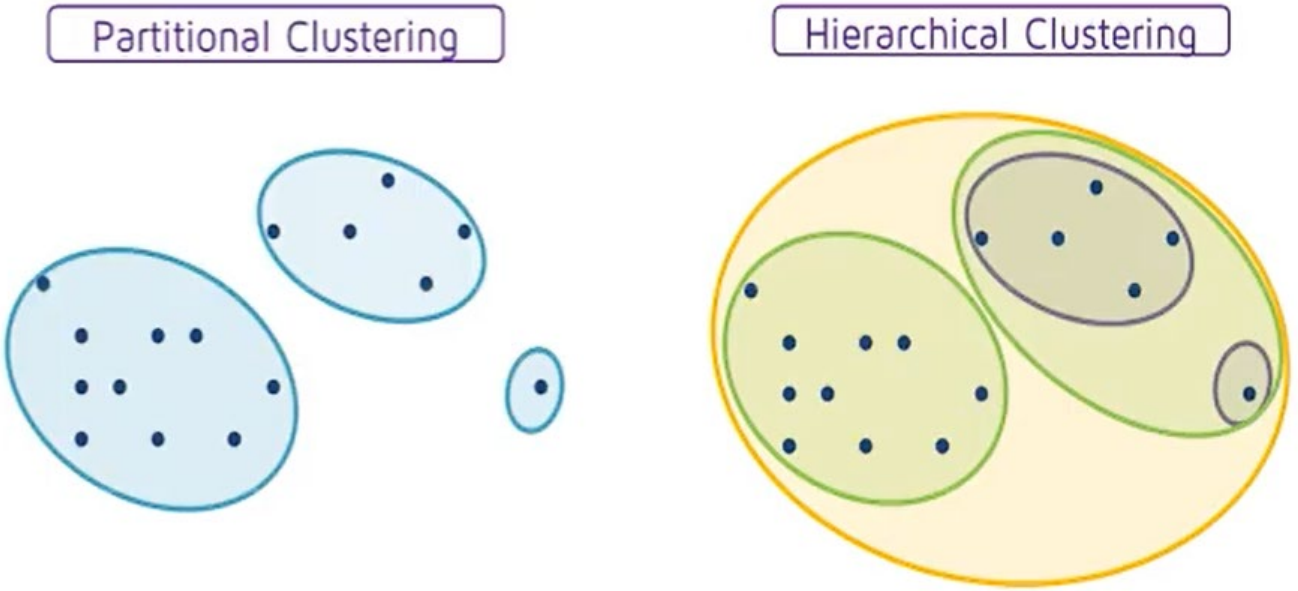
$$D_{pq} = \min\{d_{ij} | x^{(i)} \in C_p, x^{(j)} \in C_q\}$$

  - 样本均值间距离

$$D_{pq} = d_{\mu_p \mu_q}$$

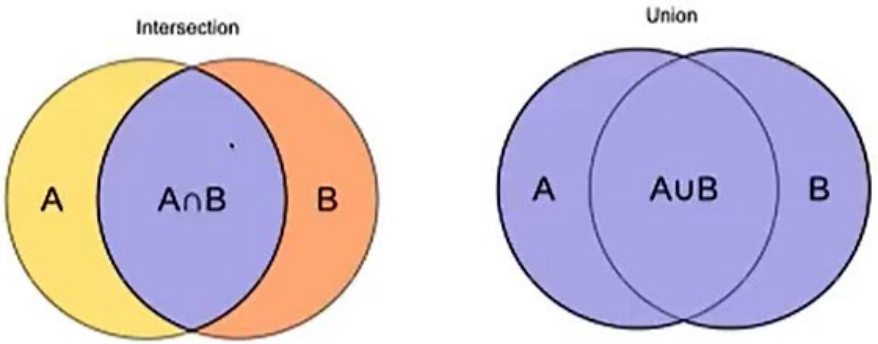
聚类方法

- 常见聚类方法
  - K均值聚类
  - 层次聚类
  - 密度聚类
  - 谱聚类



聚类效果评估

- 外部指标(external index)
  - 有外部参考聚类结果
    - Jaccard系数（Jaccard Coefficient）
    - FM指数（Fowlkes and Mallows Index）
    - Rand指数（Rand Index）



混淆矩阵		预测结果	
		同类	不同类
真实结果	同类	#TP	#FN
	不同类	#FP	#TN

$$JC = \frac{\#TP}{\#TP + \#FN + \#FP}$$

Jaccard系数

数据集中每组数据$(x^{(i)}, x^{(j)})$在预测/参考中结果

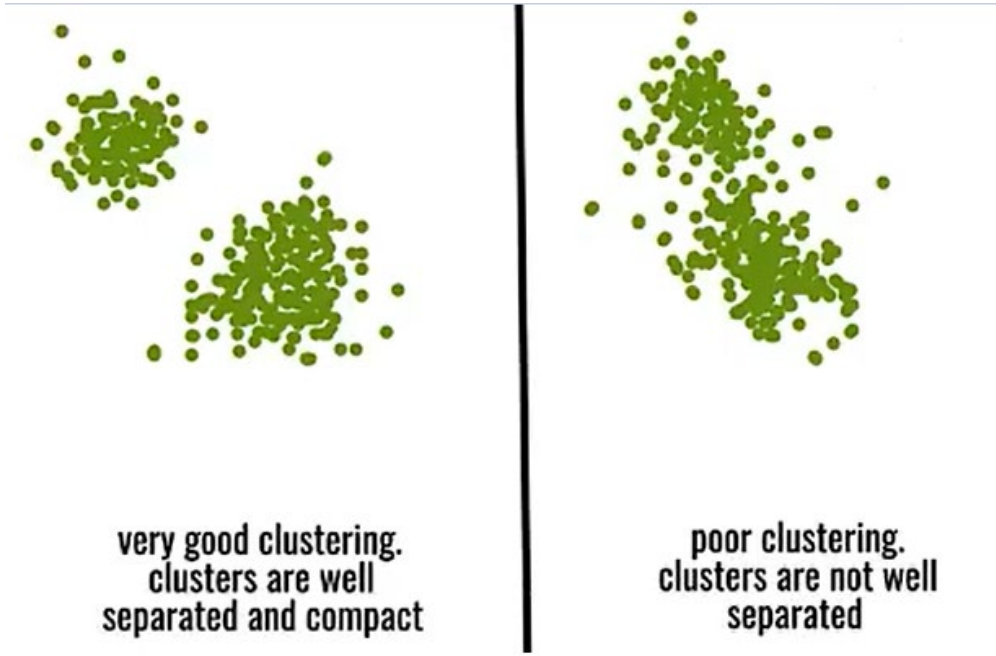
聚类效果评估

- 内部指标(internal index)
  - 无外部参考聚类结果
    - DB指数（Davies-Bouldin Index）
    - Dunn指数（Dunn Index）

类内间距 类内间距

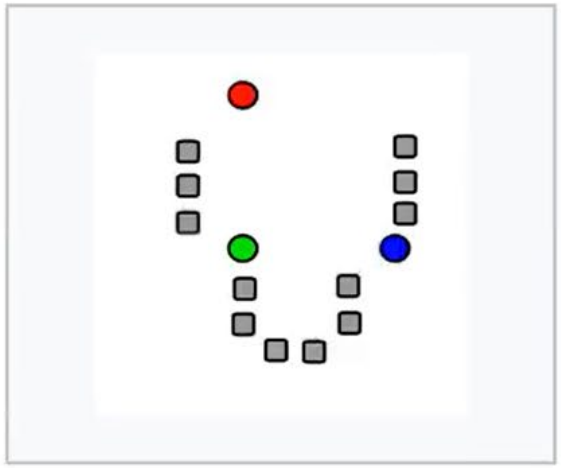
$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i}\left(\frac{avg(C_i) + avg(C_j)}{d_{\mu_i \mu_j}}\right)$$

类间间距

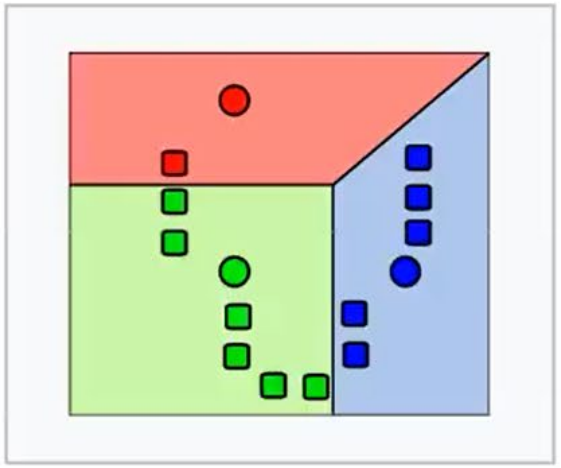


聚类
K均值聚类方法

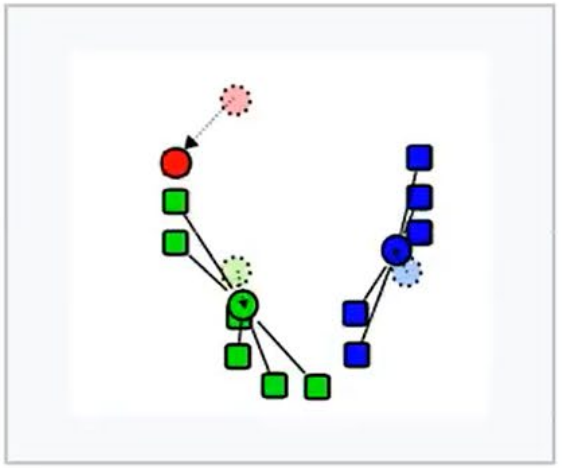
K-means方法



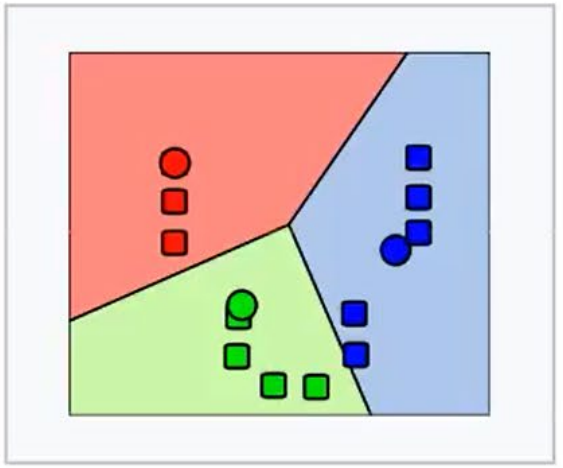
1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).



2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.



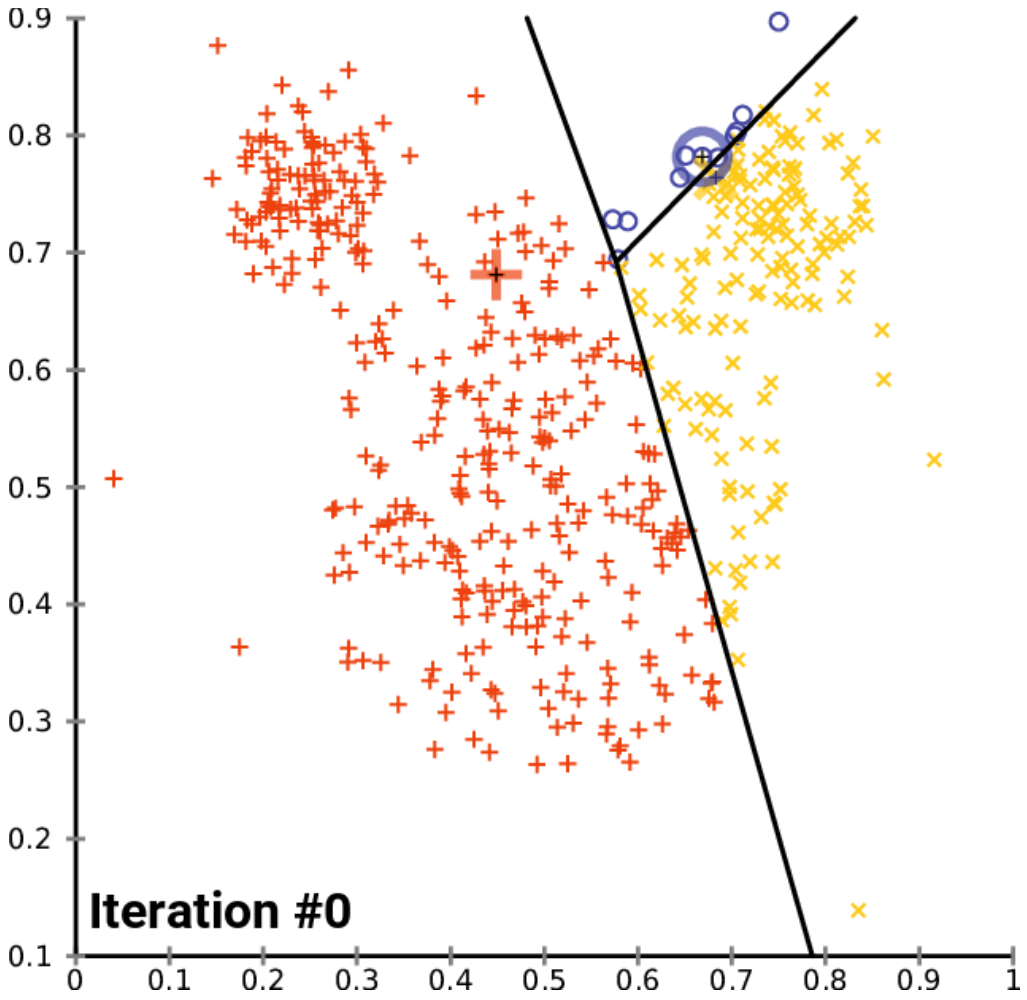
3. The centroid of each of the *k* clusters becomes the new mean.



4. Steps 2 and 3 are repeated until convergence has been reached.

K-means方法

- 聚类过程
  - 1) 选择K个点作为聚类中心
    - a) 根据与聚类中心的距离对每个样本点进行聚类
    - b) 求每类样本的平均值作为该类别新的聚类中心
  - 2) 不断迭代a)、b)步骤，直至收敛（例如每个样本的类别不再改变）



K-means方法

- K均值聚类的目标函数

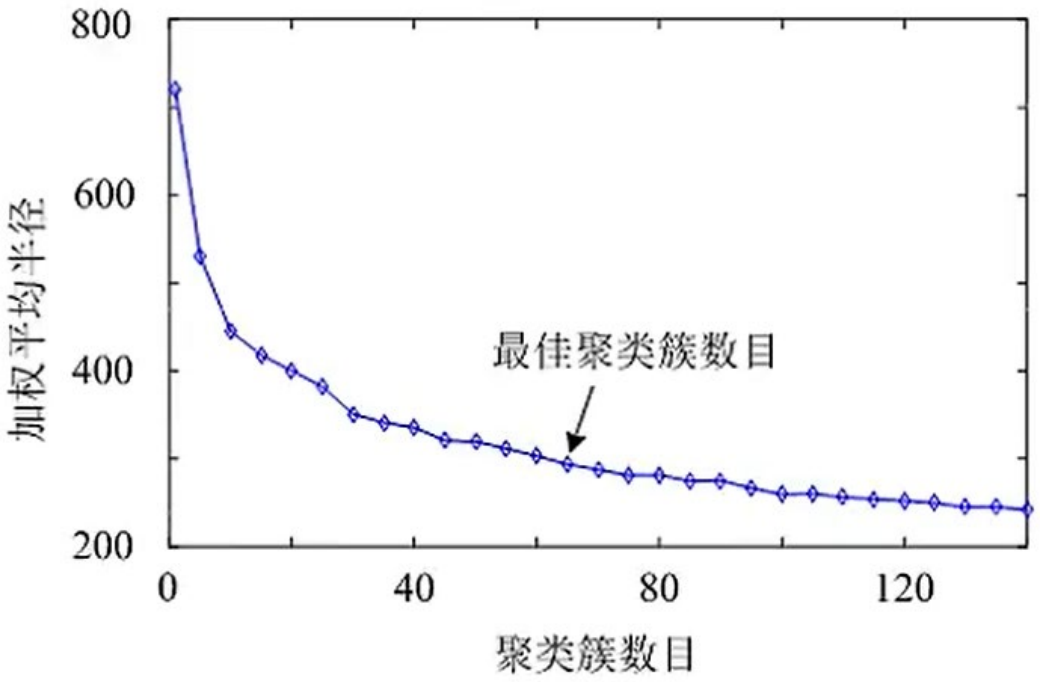
$$E = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||_2^2$$

  - 其中μ_i是第i个簇C_i的均值向量
  - E值刻画了簇内样本围绕簇均值向量的紧密程度

- 收敛性
  - NP-Hard问题，迭代优化：
    - 固定均值向量，优化划分→a)步
    - 固定划分，优化均值向量→b)步

K-means超参数选择

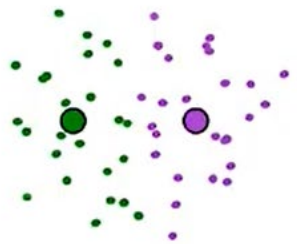
- K的选择

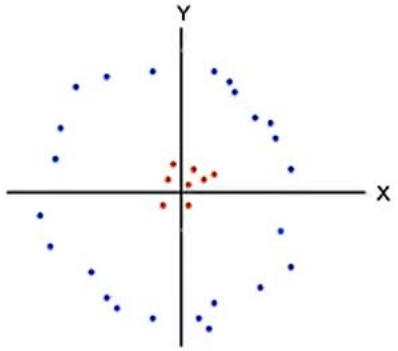


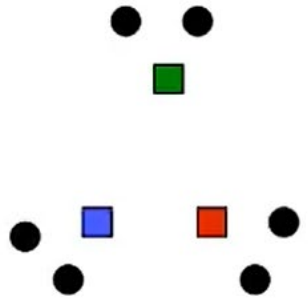
- 类中心初始化
  - 大于最小间距的随机点/样本点
  - K个相互距离最远的样本点
  - K个等距网格点

K-means方法

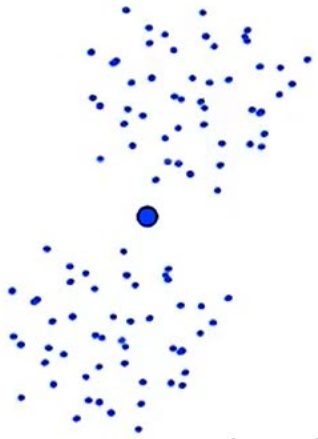
- 优点
  - 实现简单
  - 时间复杂度低——O(N)

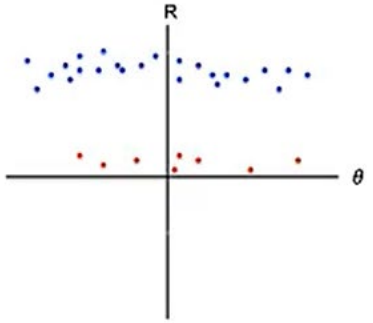


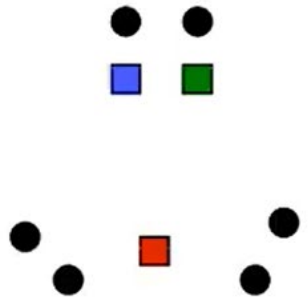




- 缺点
  - K值选择
  - 主要适合凸集问题
  - 初始值影响较大



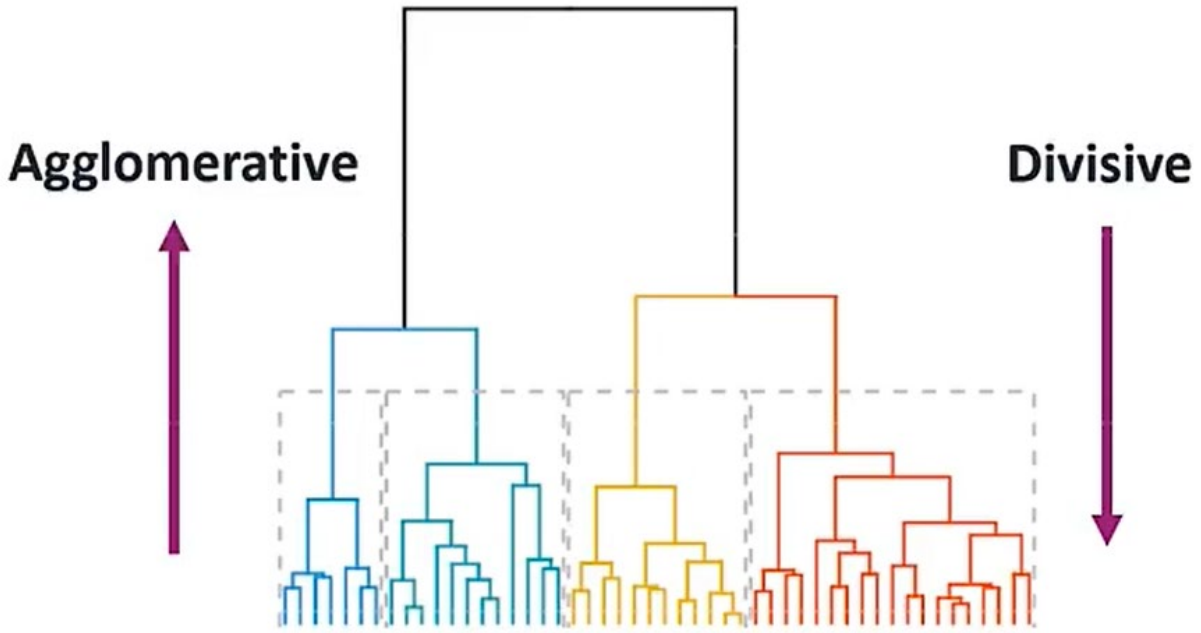




聚类
层次聚类方法

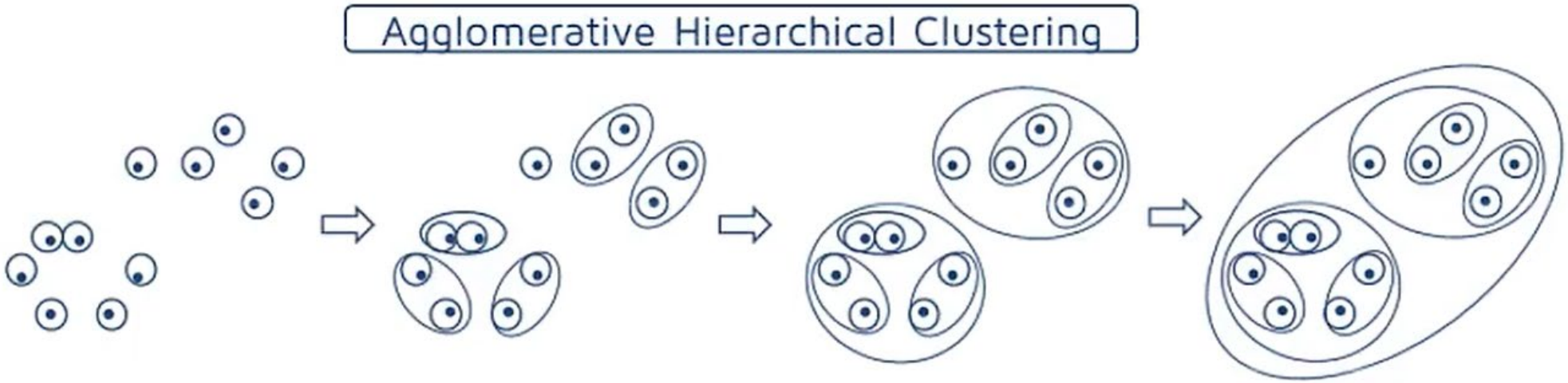
层次聚类（Hierarchical Clustering）

- 层次聚类通过计算不同类别数据点间的相似度创建一棵有层次的嵌套聚类树
  - 聚合（Agglomerative）或自底向上（Bottom-up）
  - 分裂（Divisive）或自顶向下（Top-down）



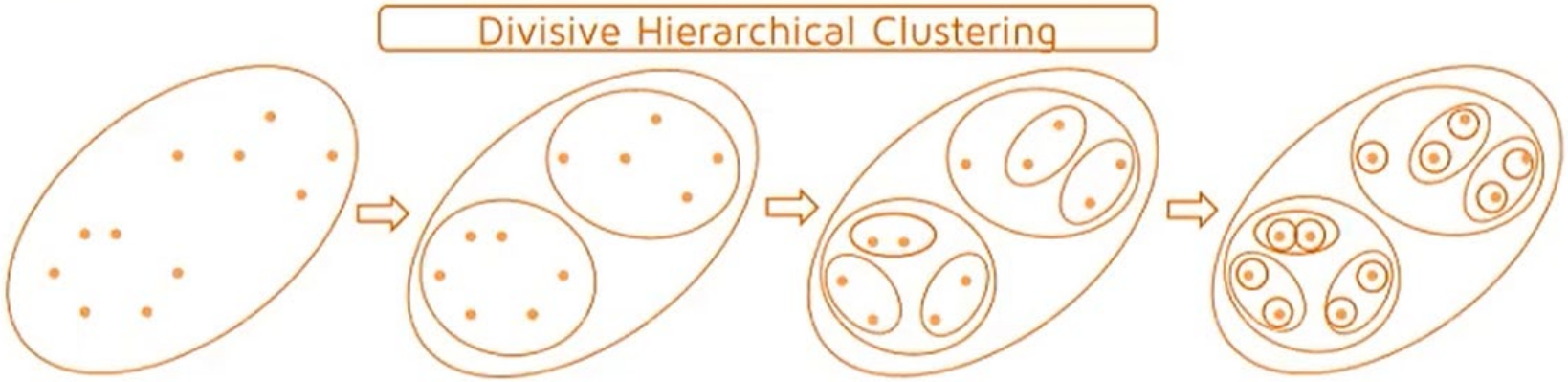
聚合（Agglomerative）

- 聚类过程
  - 将每个样本分到单独的类
  - 不断迭代以下过程，直至满足终止条件：
    - 计算类簇两两之间的距离，找到距离最小的两个类簇；
    - 合并这两个类簇为一个。



分裂（Divisive）

- 聚类过程
  - 将所有样本分到同一个类
  - 不断迭代以下过程，直至满足终止条件：
    - 在同一类簇c中，计算样本两两之间的距离，找出距离最远的两个样本a、b；
    - 将样本a、b分配到不同的类簇；
    - 计算原类簇c中剩余的其他样本点与a、b的距离，如果dis(a)<dis(b)则划入a所在的类簇，否则划入b所在的类簇。



层次聚类的优缺点

- 优点：
  - 原理简单，容易实现

- 缺点：
  - 合并点、分裂点的选择并不容易
  - 合并、分裂的操作不能撤销
  - 执行效率低$O(TN^2)$，T为迭代次数，N为样本点数