Kmeans 聚类，层次聚类，基于密度的聚类
基于划分的方法：kmeans（计算），dbscan（不考计算）
基于层次的方法：层次聚类（计算）

一、分类与聚类区别
1、分类有类别标记信息，是有监督学习方法；聚类无标记信息，是无监督学习方法。
2、分类根据训练样本获得分类器，然后将每个数据样本归结到某个已知的类别，**进而可以预测未来数据类别的任务**；聚类是根据**信息相似度**原则进行聚类，通过聚类，人们能够**识别密集的和稀疏的区域**，进而发现**全局分布模式**以及**数据属性**的关系。
3、应用不同：分类可以应用于医疗诊断、信用卡的信用分级、图像模式识别；聚类可以应用于寻找具有类似行为的客户群，寻找具有相似特征的动物或植物群等。

二、Kmeans 优缺点
1、 优点：
（1）、原理简单，适用于常规不相交的簇和呈球形分布的数据。
（2）、易于实现，收敛速度较快。
（3）、具有有效性和可扩展性，其时间复杂度为 O(I×K×n×m)，I：收敛所需迭代次数; K：中心数; n：数据点数, m：类别数
2、 缺点：
（1）、需要提前确定 K 的值，一般来说，K 值的确定需要专业领域的知识。
（2）、可能会收敛到局部最优点，需要多次尝试不同的初始中心值才能获得最优点。
（3）、可能对噪声数据和异常值敏感。因为簇的中心是取平均，因此聚类簇很远的地方的噪声会导致簇的中心点偏移。
（4）、不适合非凸不规则形状的簇，普遍对球形分布样本聚类较好。

三、顺序领导者聚类
1、 优点：可以处理流数据，没有迭代，无需提前指定簇的个数 K。
2、缺点：簇聚类距离阈值选取困难。阈值太大，聚成的簇很少。阈值太小，聚成的簇很多
3、流程：

- 选择簇距离阈值
- 第一个点(称为领导者)代表一个簇
- 对于每个新数据点：
  - 计算新数据点与每个簇中心之间的距离。
  - 如果最小距离小于所选阈值，请将新数据点分配给相应的簇并重新计算簇中心。
  - 否则，创建一个以新数据点为中心的新簇。

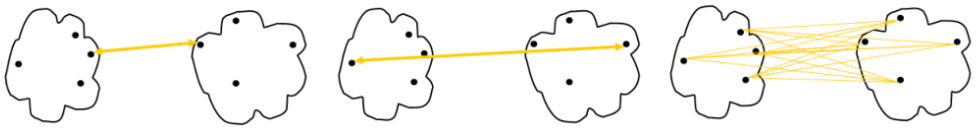
四、层次聚类簇之间临近性定义

◘簇之间的近邻性通常用特定的簇类型定义，主要有三种定义方式:

- **单链 (single link或MIN)**。MIN定义簇的邻近度为**不同簇的两个最近的点之间的邻近度**, 或者说不同的结点子集中两个节点之间的最短边。
- **全链 (complete link或MAX)**。MAX取**不同簇中两个最远点之间的邻近度**作为簇的邻近度, 或者说不同结点子集中两个结点之间的最长边。
- **组平均 (group average)**。定义簇邻近度为取自**不同簇的所有点对邻近度的平均值**(平均边长)。

◘两个点之间的邻近度度量是**距离(相异度)**，则MIN和MAX两个名字有提示作用，即**值越小表示点越接近(单链“小中取小”，全链“大中取小”)**。

◘两个点之间的邻近度度量是**相似度**，则**值越大表示点越接近(单链“大中取大”，全链“小中取大”)** 。



单链(MIN)　　全链(MIN)　　组平均

◘基于距离(相异度)的层次聚类

➢1. 单链层次聚类

- 步骤：① 找出所有点距离最小的两个点，第一个合并;
  ② 按照“小中取小”的原则依次合并剩余点，直至合并完所有点。

➢2. 全链层次聚类

- 步骤：① 找出所有点距离最小的两个点，第一个合并;
  ② 按照“大中取小”的原则依次合并剩余的点，直至所有点合并完成。

◘基于相似度矩阵的层次聚类

➢1. 单链层次聚类

- 步骤：① 找出所有点相似度最大的两个点，第一个合并;
  ② 按照“大中取大”的原则进行合并剩余的点，直至所有点合并完成为止。

➢2. 全链层次聚类

- 步骤：① 找出所有点相似度最大的两个点，第一个合并;
  ② 按照“小中取大”的原则进行合并剩余的点，直至所有点合并完成为止。

五、基于密度的聚类方法

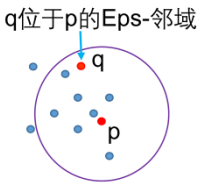
1、样本点分类

- **核心点** (Core point，稠密区域内部的点)。核心点的定义为: 如果该点的给定邻域内的点的个数超过给定的阈值MinPts (MinPts由用户指定), 则这些点为核心点。
- **边界点** (Border point， 稠密区域边缘上的点)。边界点不是核心点, 但它落在某个核心点的邻域内。边界点可能落在多个核心点的邻域内。
- **噪声**或背景点 (Noise point， 稀疏区域中的点)。噪声点是即非核心点也非边界点的任何点。

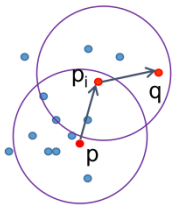
2、 概念

- **Eps领域**: 给定样本点p， 其半径为 Eps 内的区域称为该样本的 Eps 邻域。
- **核心点:** 如果给定点Eps 邻域内的样本数大于等于MinPts, 则该点为核心点。
- **直接密度可达**: 对于样本集合D, 如果样本点q 在p 的Eps邻域内, 并且p 为核心点, 那么点q 从点p 直接密度可达(又称密度直达)。
- **密度可达**: 对于样本集合D, 给定一串样本点p_1 , p_2 , … p_n , $p=p_1$, $q=p_n$, 假定样本p_i 从 p_{i-1} 直接密度可达, 那么点q 从点p 密度可达。

密度相连: 对于样本集合D 中的任意一点o, 如果存在点p 到点o 密度可达, 并且点q 到点o 密度可达, 那么点q 到点 p 密度相连。

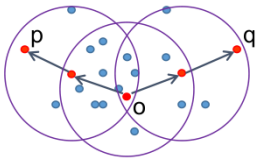


直接密度可达



密度可达

对于p和q，若存在样本序列$p_1,p_2,\ldots,p_n$使得p_i和 p_{i+1}密度直达，且$p=p_1$，$q=p_n$，则称点p和点q密度可达



密度相连

存在点o使得点p和点o密度可达，点q和点o密度可达，则称点p和点q密度相连

3、 流程

- 输入: 样本集合D， 聚类半径Eps, 密度阈值 MinPts
- 输出: 目标类簇集
- 方法: repeat
  - 1. 随机选取未被处理的点p， 判断输入点是否为核心点。
  - 2. 找出核心点的 Eps 领域中的所有密度可达点， 形成一个新的簇。
    –遍历数据集D, 直到所有输入点都判断完毕;
  - 3. 针对该核心点的 Eps 邻域内所有密度可达点找到最大密度相连的样本点集合, 产生最终的簇结果。
  - 4. 重复执行第2步和第3步， 直到数据集D中所有点都为“已处理”状态。

4、 优缺点

（1）优点

- 可以对任意形状的稠密数据集进行聚类，相对的，K-Means 之类的聚类算法一般只适用于凸数据集。
- 可以在聚类的同时发现异常点，对数据集中的异常点不敏感。
- 聚类结果没有偏倚，相对的，K-Means 之类的聚类算法初始值对聚类结果有很大影响。

（2）缺点

- 如果样本集的密度不均匀、聚类间距差相差很大时，聚类质量较差，这时用 DBSCAN 聚类一般不适合。
- 如果样本集较大时，聚类收敛时间较长。

- 调参相对于传统的 K-Means 之类的聚类算法稍复杂，主要需要对距离半径 Eps，邻域样本数阈值 MinPts 联合调参，不同的参数组合对最后的聚类效果有较大影响。

相异性 (距离) 与相似性度量

- 聚类就是发现数据中具有“相似性”(similarity) 的个体
- 选择合适的“相似性”度量是进行聚类的关键, 相似性度量函数 s(·, ·) 一般满足

$$
\begin{aligned}
&1.\ 0 \leq s(\mathbf{x}, \mathbf{y}) \leq 1 \\
&2.\ s(\mathbf{x}, \mathbf{x}) = 1 \\
&3.\ s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})
\end{aligned}
$$

- 也可以使用相异性(dissimilarity) 来度量数据之间的接近程度。下面我们以相异性为例. 相异性度量和相似性度量之间一般可以相互转换。
- 相异性度量多为某种“距离” 度量
- 样本点之间的相异性(距离) 函数 $d(\cdot, \cdot)$ 一般满足

$$
\begin{aligned}
&1.\ d(\mathbf{x}, \mathbf{y}) \geq 0, \text{等号成立当且仅当 } \mathbf{x} = \mathbf{y} \\
&2.\ d(\mathbf{x}, \mathbf{x}) = 0 \\
&3.\ d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x}) \\
&4.\ d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})
\end{aligned}
$$

- 相似度的度量并无统一的标准，实际中常用欧式距离，也可根据实际问题自己定义