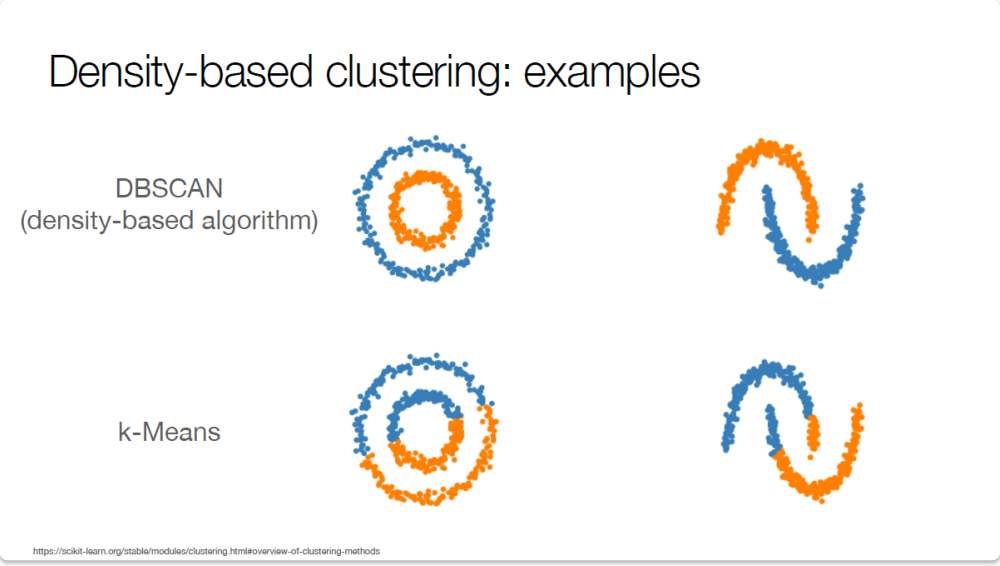
lec15 DBSCAN Clustering Algorithm

DBSCAN

基于密度的聚类方法（Density-Based Clustering）

1. 簇的定义：
   - 簇被定义为数据集中密度较高的区域
   - 这些高密度区域与数据集的其他部分相比，包含更多的对象
2. 稀疏区域中的对象：
   - 位于稀疏区域的对象通常被认为是噪声和边界点
   - 稀疏区域中的对象是需要用来分隔不同簇的对象



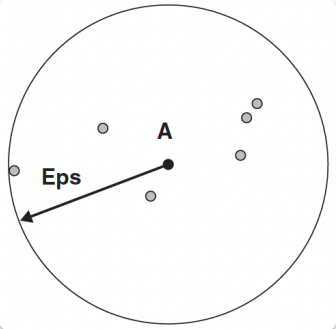
DBSCAN 主要思想

DBSCAN，全称为Density-Based Spatial Clustering of Applications with Noise，是一种基于密度的聚类算法，用于处理带有噪声的数据集

使用中心基于方法（Centre-Based Approach）来测量密度。这种方法的基本思想是通过指定一个邻域半径 ϵ 和一个最小点数 `MinPts` 来定义密度

测量密度：
对于数据集中的特定点，通过计算该点指定半径（Eps）内的点的数量（包括该点本身）来估算密度

- 选择一个点 A
- 以 A 为中心，定义一个半径为 ϵ (Eps) 的邻域
- 计算在该邻域内的所有点的数量



图示中，点 A 的 ϵ-邻域内有 7 个点（包括 A 本身），因此 A 的密度为 7

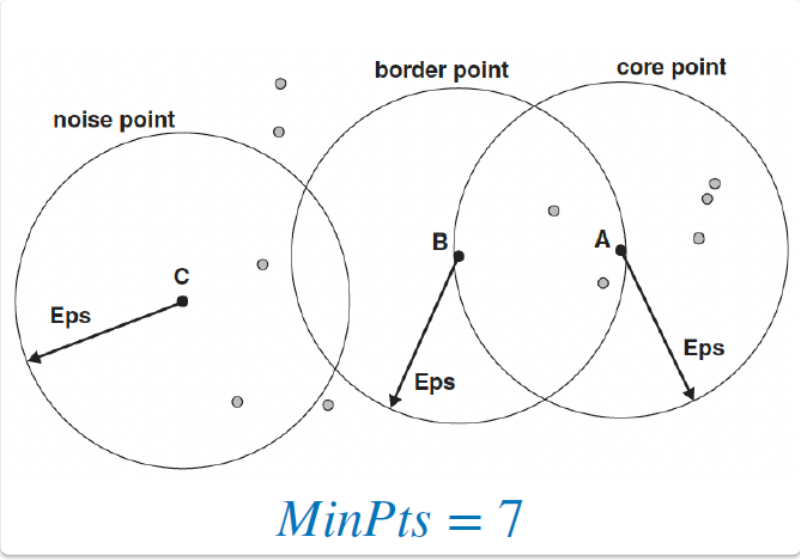
密度的测量高度依赖于参数 Eps 的选择：

1. ϵ 过大：
   - 如果 ϵ 太大，每个点的密度将接近数据集中点的总数 n
   - 这种情况下，所有点可能会被聚成一个簇，无法有效区分不同的簇
2. ϵ 过小：
   - 如果 ϵ 太小，每个点的密度将接近 1
   - 这种情况下，大多数点会被标记为噪声，难以形成有效的聚类

DBSCAN: 与密度相关的点的类型

DBSCAN 定义了三种类型的点：

1. 核心点（Core point）：
   - 定义：如果一个点的 ϵ-邻域内有至少 MinPts 个点（包括该点本身），则该点是核心点
   - 位置：核心点位于密集区域的内部
2. 边界点（Border point）：
   - 定义：边界点不是核心点，但属于某个核心点的 ϵ-邻域
   - 位置：边界点位于密集区域的边界
3. 噪声点或背景点（Noise or Background point）：
   - 定义：噪声点既不是核心点，也不是边界点
   - 位置：噪声点位于稀疏区域



DBSCAN 算法步骤

- 输入：
  - 数据集 $\mathscr{D}$
  - 参数：ϵ 和 MinPts
- 步骤：
  1. 标记所有点：
     - 将数据集中的所有点标记为核心点、边界点或噪声点
  2. 消除噪声点：
     - 从数据集中移除所有被标记为噪声点的点
  3. 添加边：
     - 在所有核心点之间添加边，如果它们之间的距离在 ϵ 范围内
  4. 形成簇：
     - 将所有连接的核心点分组，每组形成一个独立的簇
  5. 分配边界点：
     - 将每个边界点分配到其关联的核心点簇中
  6. 返回结果：
     - 返回最终的聚类结果和噪声点

选择 ϵ 和 MinPts 的方法：

基本方法：
调查点到其第 k 近邻点的距离的行为，我们称之为 $k-\text{dist}$

直观理解：

- 对于属于某个簇的点，如果 k 不大于簇的大小，那么 $k-\text{dist}$ 的值会很小
- 对于不属于任何簇的点，如噪声点，$k-\text{dist}$ 的值会相对较大

方法：

1. 计算：
   - 对于某个 k 值，计算所有数据点的 $k-\text{dist}$
2. 排序：
   - 按照递增顺序排列这些距离
3. 绘图：
   - 绘制排序后的距离值

有

- 我们希望在 $k-\text{dist}$ 的值中看到一个明显的变化，这对应一个合适的 ϵ 值
- 如果我们选择这个距离作为 ϵ 参数，并将 k 的值作为 MinPts 参数，那么对于那些 $k-\text{dist}$ 小于 ϵ 的点会被标记为核心点，而其他点会被标记为噪声或边界点

但是

- 如果 k 的值太小，即使是少数几个靠近的噪声点也会被错误地标记为簇
- 如果 k 的值太大，那么小簇（大小小于 k）可能会被标记为噪声