9.5 密度聚类

1. 密度聚类亦称"基于密度的聚类"，此类算法假设聚类结构能通过样本分布的紧密程度确定，通常情形下，密度聚类算法从样本密度的角度来考察样本之间的可连续性，并基于可连接样本不断扩展聚类簇以获得最终聚类结果。

2.
- ε-邻域：对 $x_j \in D$，其 ε-邻域包含样本集 D 中与 x_j 的距离不大于 ε 的样本
  $N_\varepsilon(x_j) = \{x_i \in D \mid dist(x_i, x_j) \le \varepsilon\}$；
- 核心对象：若 x_j 的 ε-邻域至少包含 MinPts 个样本，即 $|N_\varepsilon(x_j)| \ge MinPts$，则 x_j 是核心对象。
- 密度直达：若 x_j 位于 x_i 的 ε-邻域内，且 x_j 为核心对象，则 x_i 由 x_j 密度直达
- 密度可达：对 x_i 与 x_j，若存在样本序列 $p_1, p_2, \dots, p_n$，其中 $p_1 = x_i$，$p_n = x_j$ 且 p_{i+1} 由 p_i 密度直达，则称 x_j 由 x_i 密度可达。
- 密度相连：对 x_i 与 x_j，若存在 x_k 使得 x_i 与 x_j 均由 x_k 密度可达，则称 x_i 与 x_j 密度相连。

（若 x_j 为核心对象，则……）

3. "簇"定义为：由密度可达关系导出的最大的密度相连集合。形式化地说，给定邻域参数 $(\varepsilon, MinPts)$，簇 $C \subseteq D$ 是满足以下性质：

a. 连接性：$x_i \in C$，$x_j \in C \Rightarrow x_i$ 与 x_j 密度相连

b. 最大性：$x_i \in C$，x_j 由 x_i 密度可达，$\Rightarrow x_j \in C$.

DBSCAN 算法

输入：样本集 $D=\{x_1, x_2, \cdots, x_m\}$

邻域参数 $(\epsilon, MinPts)$.

过程：

初始化核心对象集合：$\Omega=\emptyset$

for $j=1,2,\cdots,m$ do

确定样本 x_j 的 ϵ-邻域 $N_\epsilon(x_j)$

若 $N_\epsilon(x_j) \geq MinPts$ then

将样本 x_j 加入核心对象集合：$\Omega=\Omega \cup \{x_j\}$

初始化聚类簇数：$k=0$

初始化未访问样本集合：$\Gamma=D$.

while $\Omega \neq \emptyset$ do

记录当前未访问样本集合：$\Gamma_{old}=\Gamma$

随机选取一个核心对象 $o \in \Omega$，初始化队列 $Q=<o>$;

$\Gamma=\Gamma \setminus \{o\}$.

while $Q \neq \emptyset$ do

取出队列 Q 中的首个样本 q;

若 $|N_\epsilon(q)| \geq MinPts$ then

令 $\Delta=N_\epsilon(q) \cap \Gamma$

将 Δ 中的样本加入队列 Q

$\Gamma=\Gamma \setminus \Delta$

end while

$k=k+1$，生成聚类簇 $C_k=\Gamma_{old} \setminus \Gamma$

$\Omega=\Omega \setminus C_k$

end while

输出：$C=\{C_1, C_2, \cdots C_k\}$