EDU

CSDN学院　IT实战派

DBSCAN聚类法

讲师：刘顺祥

课程目标

1. 熟悉密度聚类中的几个概念
2. 理解密度聚类的几步过程
3. 密度聚类相比Kmeans聚类的优势
4. 掌握密度聚类的应用实战

密度聚类的几个概念

模型介绍

Kmeans聚类存在两个致命缺点，一是聚类效果容易受到异常样本点的影响；二是该算法无法准确地将非球形样本进行合理的聚类。

基于密度的聚类则可以解决非球形簇的问题，“密度”可以理解为样本点的紧密程度，如果在指定的半径领域内，实际样本量超过给定的最小样本量阈值，则认为是密度高的对象，就可以聚成一个簇。

密度聚类的几个概念

概念讲解

- **点的ε领域**：在某点p处，给定其半径e后，所得到的覆盖区域
- **核心对象**：对于给定的最少样本量$MinPts$而言，如果某点p的e领域内至少包含$MinPts$个样本点，则点p就为核心对象
- **直接密度可达**：假设点p为核心对象，且在点p的e领域内存在点q，则从点p出发到点q是直接密度可达的
- **密度可达**：假设存在一系列的对象链$p_1, p_2, \cdots, p_n$，如果p_i是关于半径e和最少样本点$MinPts$的直接密度可达p_{i+1} $(i = 1,2, \cdots, n)$，则p_1密度可达p_n

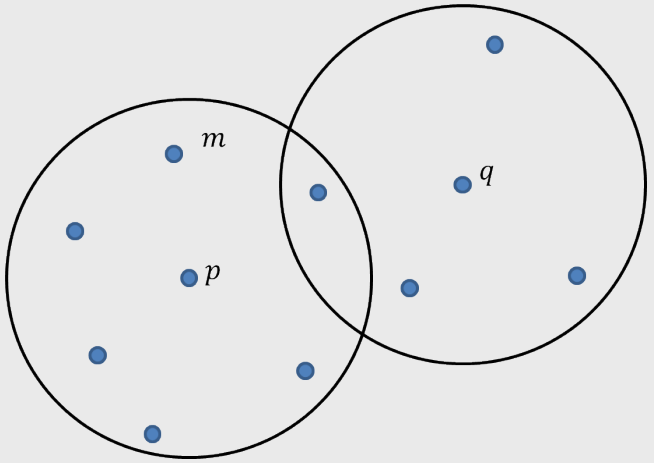
密度聚类的几个概念

概念讲解

- **密度相连：**假设点o为核心对象，从点o出发得到两个密度可达点p和点q，则称点p和点q是密度相连的
- **聚类的簇**：簇包含了最大的密度相连所构成的样本点
- **边界点**：假设点p为核心对象，在其领域内包含了点b，如果点b为非核心对象，则称其为点p的边界点。
- **异常点**：不属于任何簇的样本点

密度聚类的几个概念

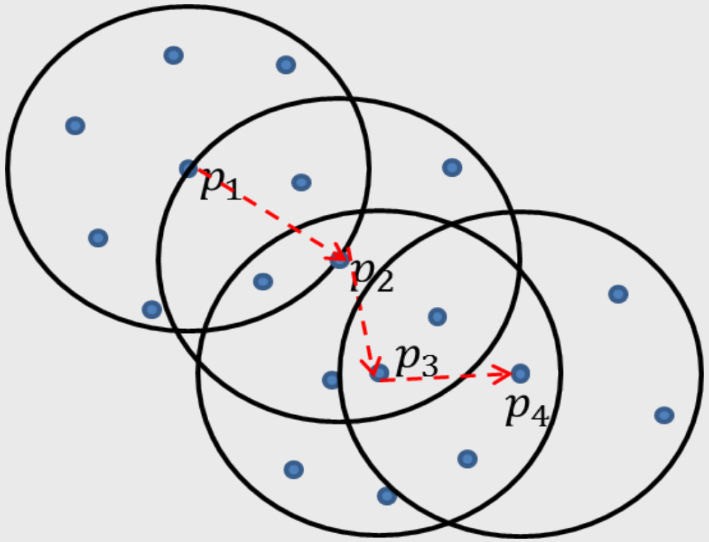
概念讲解



如图所示，如果ε为3、$MinPts$为7，则点p为核心对象（因为在其领域内至少包含了7个样本点）；点q为非核心对象；点m为点p的直接密度可达（因为它在点p的e领域内）。

密度聚类的几个概念

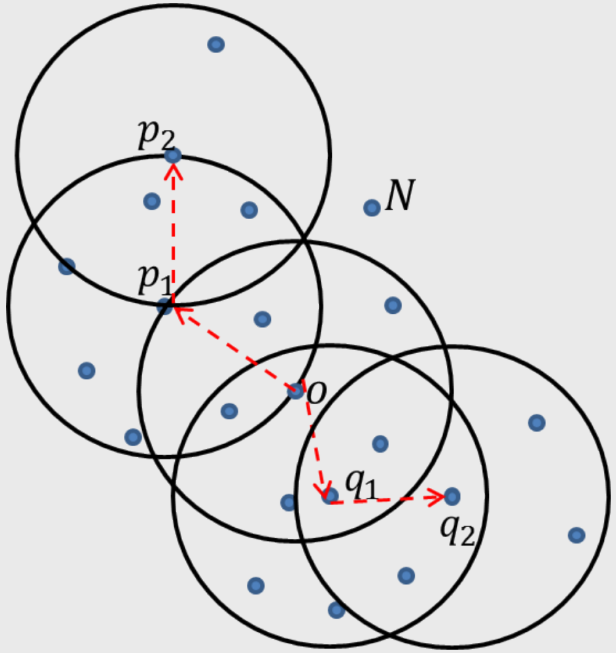
概念讲解



如图所示，如果ε为3、$MinPts$为7，则点p_1、p_2和p_3为核心对象，点p_4为非核心对象。点p_1直接密度可达点p_2、点p_2直接密度可达点p_3、点p_3直接密度可达点p_4，所以点p_1密度可达点p_4。点p_4为核心点p_3的边界点。

密度聚类的几个概念

概念讲解



如图16-3所示，如果ε为3、$MinPts$为7，则点o、p_1和q_1为核心对象，点p_2和q_2为非核心对象。由于点o密度可达点p_2，并且点o密度可达点q_2，则称点p_2和点q_2是密度相连的，如果点p_2和点q_2是最大的密度相连，则图中的所有样本点构成一个簇；由于点N不属于图中呈现的簇，故将其判断为异常点。

密度聚类的具体步骤

步骤讲解

（1）为密度聚类算法设置一个合理的半径ε以及ε领域内所包含的最少样本量$MinPts$。

（2）从数据集中随机挑选一个样本点p，检验其在ε领域内是否包含指定的最少样本量，如果包含就将其定性为核心对象，并构成一个簇C；否则，重新挑选一个样本点。

（3）对于核心对象p所覆盖的其他样本点q，如果点q对应的ε领域内仍然包含最少样本量$MinPts$，就将其覆盖的样本点统统归于簇C。

（4）重复步骤（3），将最大的密度相连所包含的样本点聚为一类，形成一个大簇。

（5）完成步骤（4）后，重新回到步骤（2），并重复步骤（3）和（4），直到没有新的样本点可以生成新簇时算法结束。

函数介绍

cluster.DBSCAN(eps=0.5, min_samples=5, metric= ‘euclidean’ , p=None)

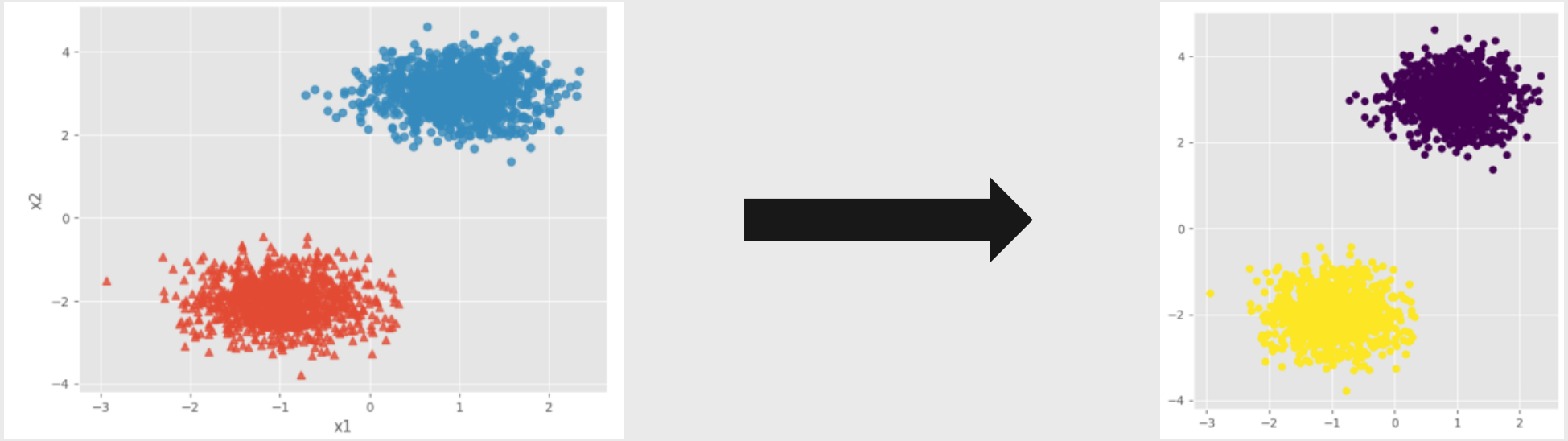
eps：用于设置密度聚类中的ε领域，即半径，默认为0.5

min_samples：用于设置ε领域内最少的样本量，默认为5

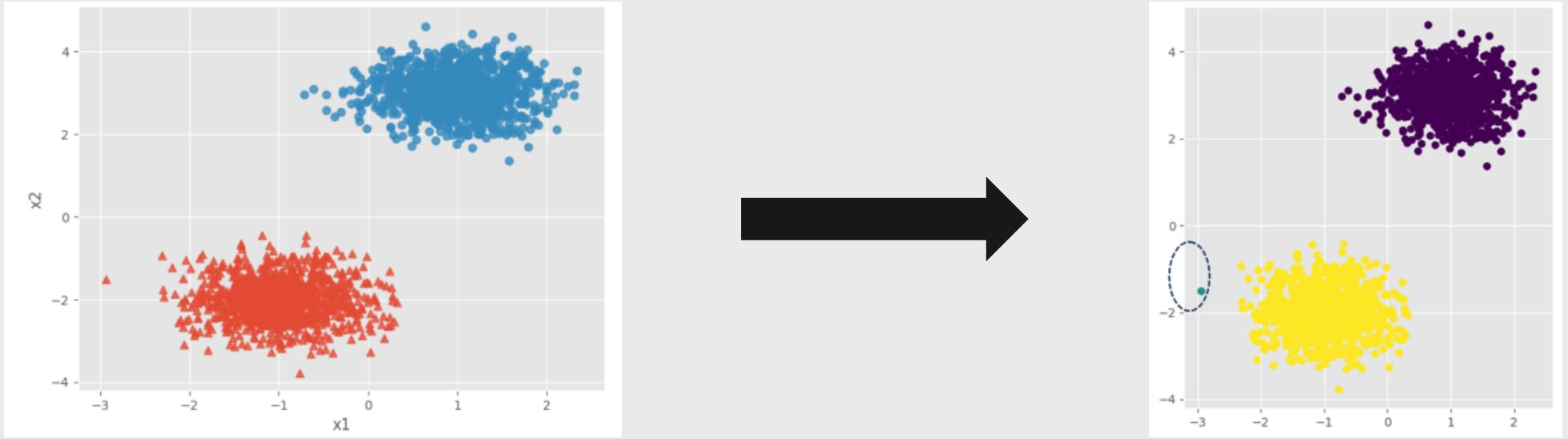
metric：用于指定计算点之间距离的方法，默认为欧氏距离

p：当参数metric为闵可夫斯基（'minkowski'）距离时，p=1，表示计算点之间的曼哈顿距离；p=2，表示计算点之间的欧氏距离；该参数的默认值为2

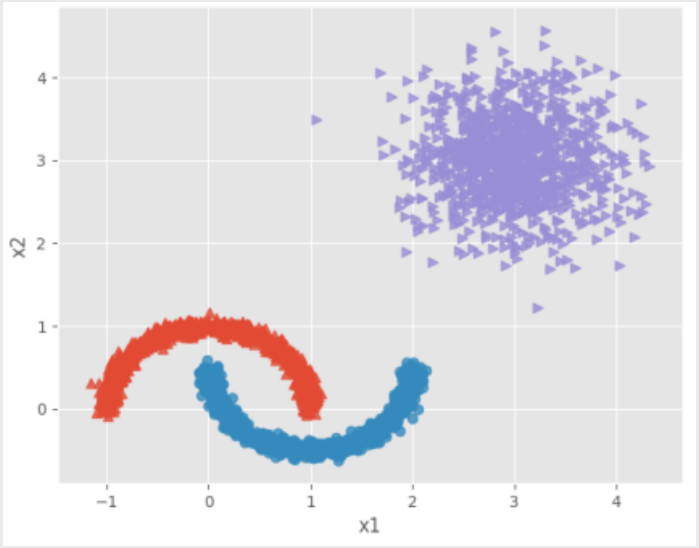
Kmeans聚类效果--球形簇的情况

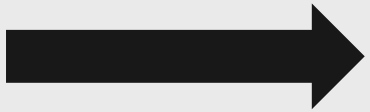


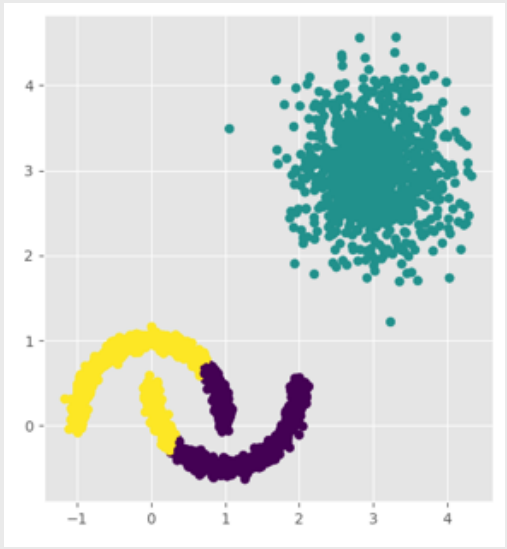
密度聚类效果--球形簇的情况



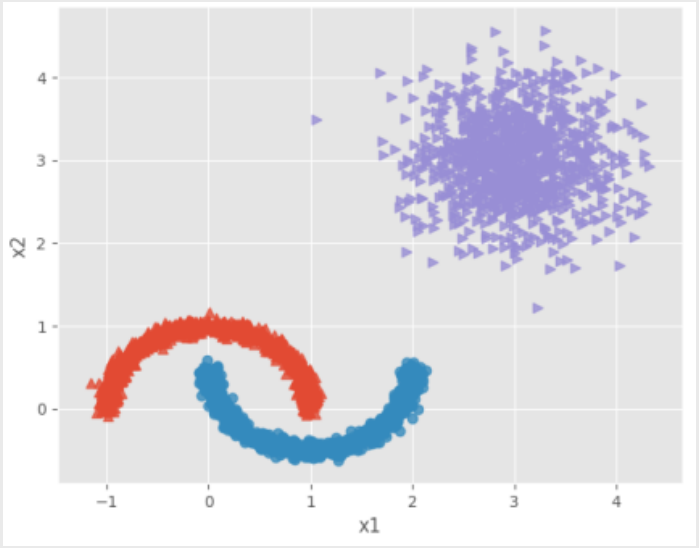
Kmeans聚类效果—非球形簇的情况

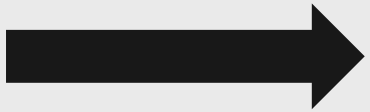


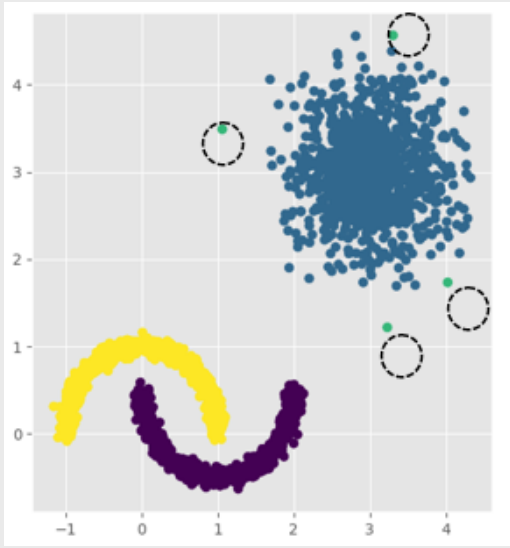




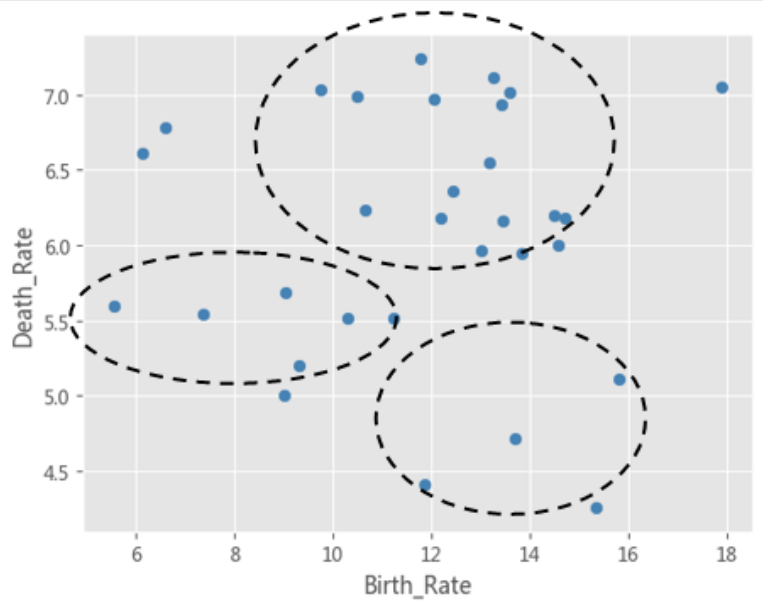
密度聚类效果—非球形簇的情况







各省份人口出生率与死亡率



如图所示，31个点分别代表了各省份人口的出生率和死亡率，通过肉眼就能够快速发现三个簇，即图中的虚线框，其他不在圈内的点可能就是异常点了。

EDU

CSDN学院　IT实战派