开课吧
kaikeba.com
慧科集团旗下企业

无监督学习-聚类1

- 1.无监督学习
- 2.Kmeans
- 3.DBSCAN

- 无监督：

  - 人们很容易就获得大量未标记的样本。
  - 获取有标签的数据成本较高。

  - 通过无标签的样本进行找数据规律的方法

- 分类和聚类？傻傻分不清楚

- 对于分类来说，在对数据集分类时，我们是知道这个数据集是有多少种类的，比如对一个学校的在校大学生进行性别分类，我们会下意识很清楚知道分为“男”“女”
- 而对于聚类来说，在对数据集操作时，我们是不知道该数据集包含多少类，我们要做的，是将数据集中相似的数据归纳在一起。

- 无监督学习目的

  - 从庞大的样本集合中选出一些具有代表性的加以标注，用于有监督学习
  - 在无类别信息情况下，寻找好的特征

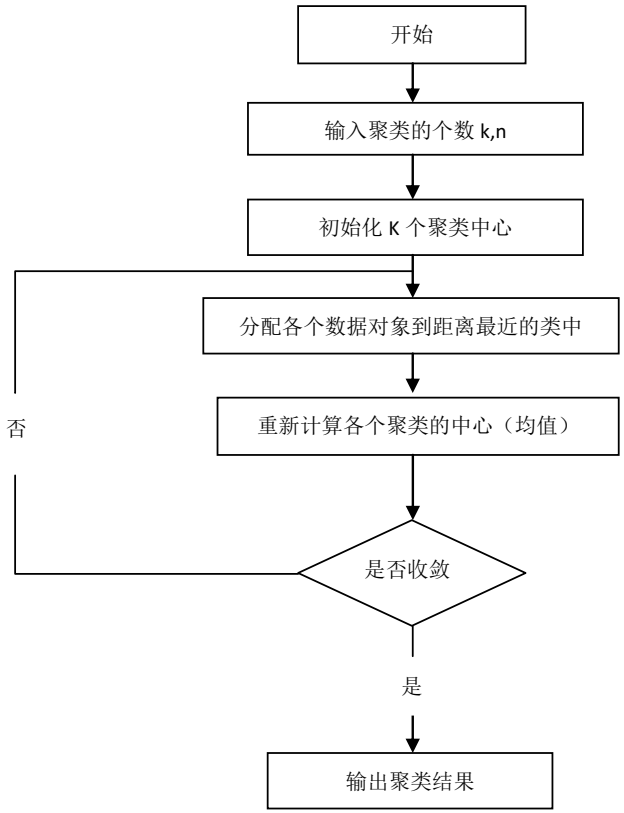
- 常用的聚类算法：
  - K-means 聚类
  - DBSCAN聚类

- K-means 思想

  - 通过迭代过程把数据集划分为不同的类别，使得评价聚类性能的准则函数达到最优，从而使生成的每个聚类内紧凑，类间独立。

- **算法描述**

1. 为中心向量$\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k$初始化$k$个种子
2. 分组:
    1. 将样本分配给距离其最近的中心向量
3. 确定中心:
    1. 用各个聚类的中心向量作为新的中心
4. 重复分组和确定中心的步骤，直至算法收敛



例子 数据对象集合S见表1，作为一个聚类分析的二维样本，要求的簇的数量k=2。

O	x	y
1	0	2
2	0	0
3	1.5	0
4	5	0
5	5	2

(1)选择 $O_1(0,2)$, $O_2(0,0)$ 为初始的簇中心，即 $M_1 = O_1 = (0,2)$, $M_2 = O_2 = (0,0)$。

(2)对剩余的每个对象，根据其与各个簇中心的距离，将它赋给最近的簇。

对 O_3：

$$d(M_1, O_3) = \sqrt{(0-1.5)^2 + (2-0)^2} = 2.5$$

$$d(M_2, O_3) = \sqrt{(0-1.5)^2 + (0-0)^2} = 1.5$$

显然 $d(M_2, O_3) \le d(M_1, O_3)$，故将$O_3$分配给 C_2

O	x	y
1	0	2
2	0	0
3	1.5	0
4	5	0
5	5	2

- 对于 O_4：$d(M_1,O_4)=\sqrt{(0-5)^2+(2-0)^2}=\sqrt{29}$

  $d(M_2,O_4)=\sqrt{(0-5)^2+(0-0)^2}=5$
- 因为 $d(M_2,O_4)\le d(M_1,O_4)$ 所以将 O_4 分配给 C_2
- 对于 O_5：$d(M_1,O_5)=\sqrt{(0-5)^2+(2-2)^2}=5$

  $d(M_2,O_5)=\sqrt{(0-5)^2+(0-2)^2}=\sqrt{29}$
- 因为 $d(M_1,O_5)\le d(M_2,O_5)$ 所以将 O_5 分配给 C_1
- 更新，得到新簇 $C_1=\{O_1,O_5\}$ 和 $C_2=\{O_2,O_3,O_4\}$
- 计算平方误差准则，单个方差为

$E_1=\left[(0-0)^2+(2-2)^2\right]+\left[(0-5)^2+(2-2)^2\right]=25$ $M_1=O_1=(0,2)$

$E_2=27.25$ $M_2=O_2=(0,0)$

总体平均方差是：$E = E_1 + E_2 = 25 + 27.25 = 52.25$

O	x	y
1	0	2
2	0	0
3	1.5	0
4	5	0
5	5	2

（3）计算新的簇的中心。

$$M_1 = ((0+5)/2, (2+2)/2) = (2.5, 2)$$

$$M_2 = ((0+1.5+5)/3, (0+0+0)/3) = (2.17, 0)$$

重复（2）和（3），得到O_1分配给C_1；O_2分配给C_2，O_3分配给C_2，O_4分配给C_2，O_5分配给C_1。更新，得到新簇 $C_1 = \{O_1, O_5\}$ 和 $C_2 = \{O_2, O_3, O_4\}$ 。 中心为 $M_1 = (2.5, 2)$ ， $M_2 = (2.17, 0)$ 。

单个方差分别为

$$E_1 = \left[(0-2.5)^2 + (2-2)^2\right] + \left[(2.5-5)^2 + (2-2)^2\right] = 12.5 \quad E_2 = 13.15$$

总体平均误差是：$E = E_1 + E_2 = 12.5 + 13.15 = 25.65$

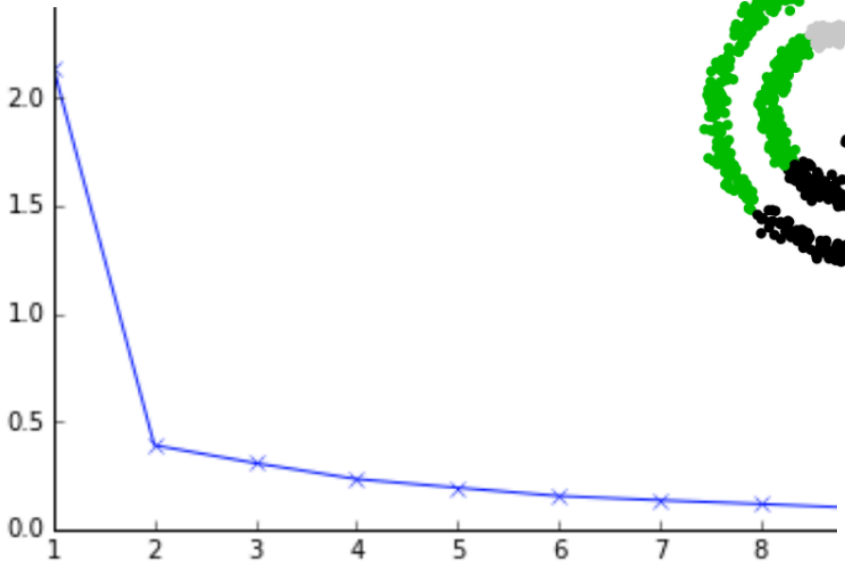
由上可以看出，第一次迭代后，总体平均误差值**52.25~25.65**，显著减小。由于在两次迭代中，簇中心不变，所以停止迭代过程，算法停止。

- Kmeans 优缺点：

- 优点：
  - 当结果簇是密集，效果较好。

- 缺点：
  - 必须事先给出k，对初值敏感，对于不同的初始值，可能会导致不同结果
  - 非簇类聚类结果不好

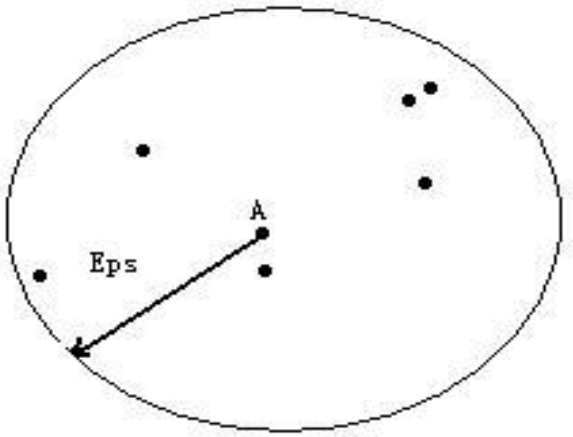
- 如何选择K？肘部法则



- DBSCAN
  - DBSCAN（Density-Based Spatial Clustering of Applications with Noise，具有噪声的基于密度的聚类方法）是一种基于密度的空间聚类算法。该算法将具有足够密度的区域划分为簇，并在具有噪声的空间数据库中发现任意形状的簇，它将簇定义为密度相连的点的最大集合。

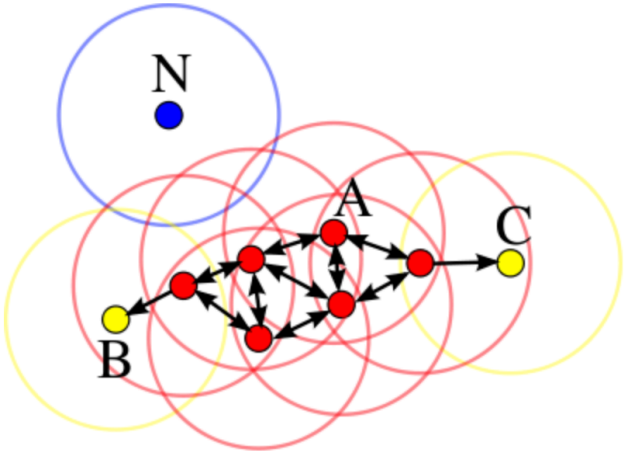
传统的密度定义：基于中心的方法

- 传统基于中心的密度定义为：
  - 数据集中特定点的密度通过该点Eps半径之内的点计数(包括本身)来估计。
  - 显然，密度依赖于半径。



DBSCAN

- 基于密度定义，我们将点分为：
  - 核心点(core point) :在半径Eps内含有超过MinPts数目的点，则该点为核心点
  - 边界点(border point):在半径Eps内点的数量小于MinPts，但是在核心点的邻居
  - 噪音点(noise point):任何不是核心点或边界点的点.



MinPts=4
红色为核心点
黄色为边界点
蓝色为噪音点

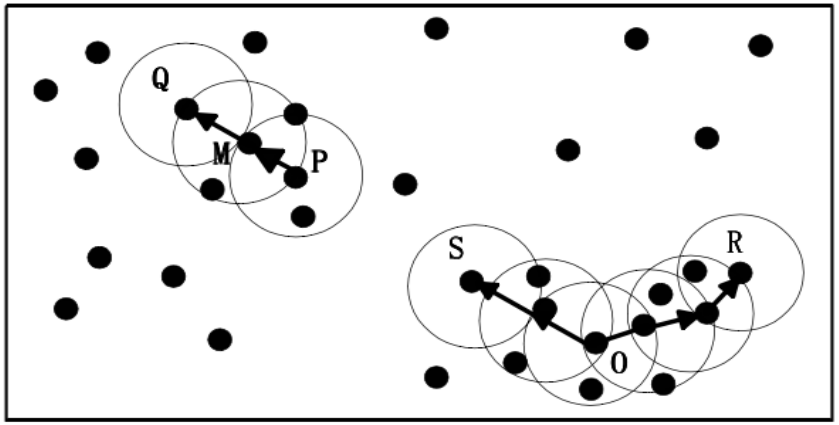
直接密度可达：给定一个对象集合D，如果p在q的Eps邻域内，而q是一个核心对象，则称对象p 从对象q出发时是直接密度可达的(directly density-reachable)。

密度可达：如果存在一个对象链 $p_1, p_2, \cdots, p_n, p_1 = q, p_n, = p$ 对于 $p_i \in D(1 \le i \le n)$，p_{i+1} 是从 p_i 关于Eps和MinPts直接密度可达的，则对象p是从对象q关于Eps和MinPts密度可达的(density-reachable)。

密度相连：如果存在对象O∈D，使对象p和q都是从O关于Eps和MinPts密度可达的，那么对象p到q是关于Eps和MinPts密度相连的(density-connected)。

DBSCAN算法概念示例

- 如图所示，Eps用一个相应的半径表示，设MinPts=3，请分析Q,M,P,S,O,R这5个样本点之间的关系。

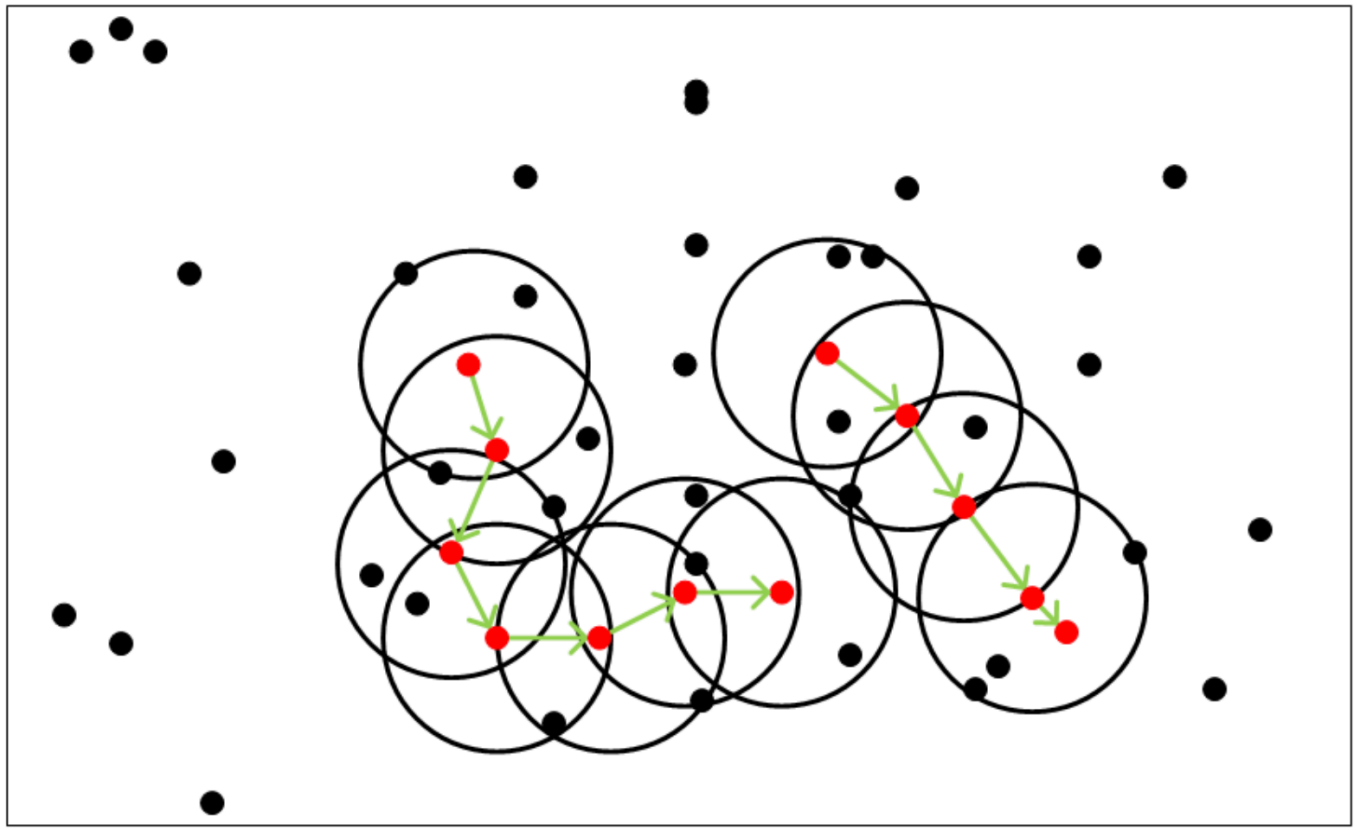


“直接密度可达”和“密度可达”概念示意描述

根据以上概念知道：由于有标记的各点M、P、O和R的Eps近邻均包含3个以上的点，因此它们都是核对象；M是从P“直接密度可达”；而Q则是从M“直接密度可达”；基于上述结果，Q是从P“密度可达”；但P从Q无法“密度可达”(非对称)。类似地，S和R从O是“密度可达”的；O、R和S均是“密度相连”的。

DBSCAN

核心点能够连通（密度可达），它们构成的以Eps长度为半径的圆形邻域相互连接或重叠，这些连通的核心点及其所处的邻域内的全部点构成一个簇。

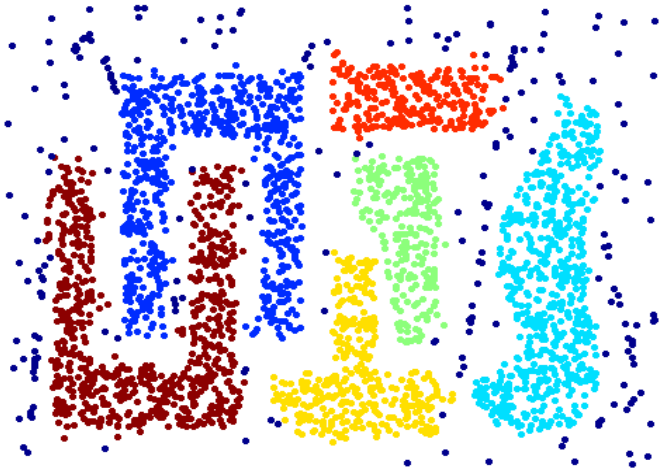


DBSCAN算法原理

- DBSCAN通过检查数据集中每点的Eps邻域来搜索簇，如果点p的Eps邻域包含的点多于MinPts个，则创建一个以p为核心对象的簇。
- 然后，DBSCAN迭代地聚集从这些核心对象直接密度可达的对象，这个过程可能涉及一些密度可达簇的合并。
- 当没有新的点添加到任何簇时，该过程结束.

DBSCAN运行效果好的时候

Original Points



Clusters

- 对噪音不敏感
- 可以处理不同形状和大小的数据

DBSCAN算法的优缺点

- 优点
  - 基于密度定义，相对抗噪音，能处理任意形状和大小的簇

- 缺点
  - 当簇的密度变化太大时，会有麻烦
  - 对于高维问题，密度定义是个比较麻烦的问题

开课吧
kaikeba.com
慧科集团旗下企业

- 谢谢大家