日期: /

聚类

一. 无监督学习.

1. 定义: 根据没有标记的样本，学习数据中的信息.

2. 聚类分析: 假设数据的特征允许将其划分为不同类别，但事先不知道数据是几个类 [无监督].

3. 类:

① 不同聚类方法得到不同的聚类结果

② 一个类是一组个体，这些个体离这个类的中心个体较近；不同类成员之间距离较远

4. 分类与聚类

- 分类:
  - 有类别标记信息, 因此是一种监督学习
  - 根据训练样本获得分类器，然后把每个数据归结到某个已知的类，进而也可以预测未来数据的归类。
  - 分类具有广泛的应用，例如医疗诊断、信用卡的信用分级、图像模式识别。
- 聚类:
  - 无类别标记, 因此是一种无监督学习
  - 无类别标记样本，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，因而发现全局的分布模式，以及数据属性之间的关系

二. 聚类 clustering.

{ 如何度量样本相似性.
  如何衡量某一种分组的好坏?

日期: /

1. 定义：聚集分析将数据对象划分子集的过程.

2. 目标：

- 同一簇样本尽可能彼此相似.
- 不同簇样本尽可能不同.

无标注，数据驱动

3. 相异性与相似性度量

① 相似性：样本相似程度.

- $0 \le s(x,y) \le 1$
- $s(x,x)=1$
- $s(x,y)=s(y,x)$.

② 相异性：多为距离. 常用欧式距离.

- $d(x,y) \ge 0 \quad (=\Leftrightarrow x=y)$.
- $d(x,x)=0$
- $d(x,y)=d(y,x)$
- $d(x,y) \le d(x,z)+d(z,y)$.

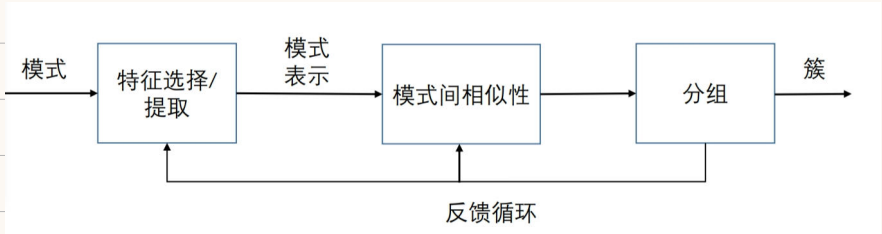
4. 聚类方法：

划分方法：
- K-Means.
- 顺序领导者方法. 流数据
- 基于密度的方法
- 基于模型的方法

层次方法

日期: /

① 基本流程



② k-means 聚类

<1> 聚类质量评价：

误差平方和：SSE

- 公式：

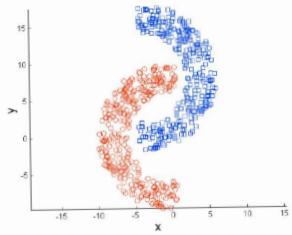
$$\mathrm{SSE}=\sum_{i=1}^{K}\sum_{x\in C_i}\|x-c_i\|_2^2,\quad c_i=\frac{1}{n_i}\sum_{x\in C_i}x$$

其中 K 是总的簇的个数。C_i 表示第 i 个簇，c_i 表示第 i 个簇的质心(均值)，x 表示第 i 个簇的任一样本，第 i 个簇 C_i 的样本数为 n_i。

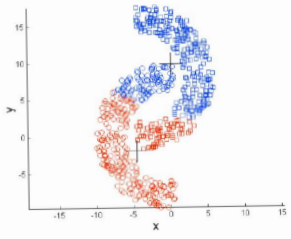
- 含义：计算每个样本到其类均值的距离平方，最后求所有类的和。
- SSE 取决于 K 个中心。
- SSE 刻画了簇内样本围绕簇均值的紧密程度，其值越小，簇内样本相似度越高。

<2> 适用场景：各类样本比较密集(球状分布)且样本数目差异不大的样本分布。

日期: /



聚成两类合乎人的直观



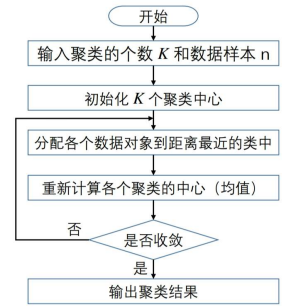
聚成两类误差平方和更小

<3> K-means 算法 流程

1. 输入训练数据和聚类数目K;
2. 执行下面二者之一:
   - 随机将数据分为K个类C_1, ……, C_K, 计算每个类的中心 c_i, $i = 1, \ldots, K$;
   - 指定K个类的中心 c_i, , $i = 1, \ldots, K$, 将所有数据点划分到离其最近的类中心所在的类
3. 计算每个数据点到其所属类的中心的平方距离

$$\mathrm{SSE} = \sum_{i=1}^{K} \sum_{x \in C_i} \|x - c_i\|_2^2, \quad c_i = \frac{1}{n_i} \sum_{x \in C_i} x$$

4. 重新将每个数据点划分到离其最近的类中心所在的类, 使得 SSE 减少. 完成后重新计算各类的中心 c_i, $i = 1, \ldots, K$;
5. 重复 3 和 4, 直到没有样本点需要调整（SSE不能再减少）;



<4> 优缺点...

优点

- 简单, 适用于常规不相交的簇。
- 收敛相对较快。
- 相对有效和可扩展。时间复杂度 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
- 假设数据是呈球形分布。实际任务中很少有这种情况。

日期: /

缺点

- 需要提前指定 K 的值。
  - 很难确定，领域知识可能会有所帮助。
- 可能会收敛到局部最优点。
  - 在实践中，需要尝试不同的初始中心点。
- 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
- 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

(5) 举例：

- 用k-均值算法将右表中的8个点聚为三个簇，假设第一次迭代选择序号1、序号4和序号7当作初始点，请给出第一次执行后的三个聚类中心以及最后的三个簇
- 参考答案：最后三个簇(1,4,8)、(3,5,6)、(2,7)

序号	属性1	属性2
1	2	10
2	2	5
3	8	4
4	5	8
5	7	5
6	6	4
7	1	2
8	4	9

③ 顺序领导者聚类

(1) 特点：有效的聚类算法；无迭代；单次传递数据；处理流数据

(2) 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。

日期: /

<3>缺点:

- **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

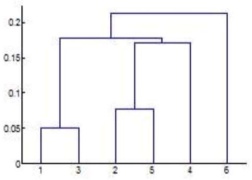
④ 层次聚类

<1> 2种基本方法

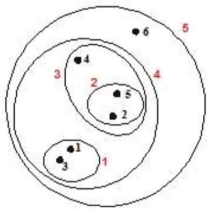
- **凝聚的（自下而上）**。从点作为个体簇开始，每一步合并两个最接近的簇。这需要定义簇的邻近性概念。
- **分裂的（自上而下）**。从包含所有点的某个簇开始，每一步分裂一个簇，直到仅剩下单点簇。在这种情况下，需要确定每一步分裂哪个簇，以及如何分裂。

<2> 2种表示方式:

- 树状图(dendrogram)。
- 嵌套簇图(nested cluster diagram)。



树状图



嵌套簇图

PS.

<3> 凝聚(自下而上)

- 自下而上的方法: 从个体点作为簇开始, 相继合并两个最接近的簇, 直到只剩下一个簇

将每个数据点看作簇.
计算邻近矩阵.
合并最近的簇.
重复,至只剩下一个簇.

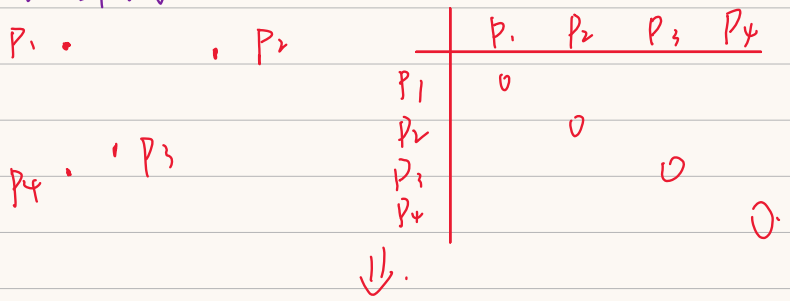
日期: /

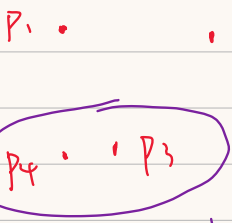
- 计算簇之间近邻性方法.

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)。

单链举例



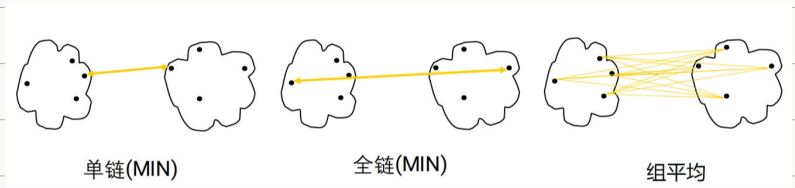


计算 P_1, P_2 距离

P_2, P_3 取 P_2 和 G_1 距离

P_1, P_4 取 P_1 和 G_1 距离

$G_1 = \{P_3, P_4\}$



日期: /

- 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")**。

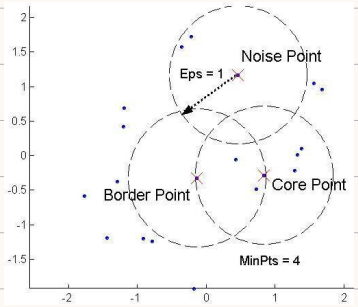
单链举例：以距离作为邻近度量. (P37.)

⑤ 基于密度的方法(DBSCAN)

<1> 核心：根据密度完成样本数据的聚类，通过不断生长足够高密度区域来进行聚类；可以从含有噪声空间数据库中发现任意形状聚类

<2> 基于中心方法将点分类

- **密度的基于中心的方法将点分类为**:
  - **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，其中MinPts是用户指定的，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。



Eps：邻域半径

MinPts：核心点邻域内点数目阈值

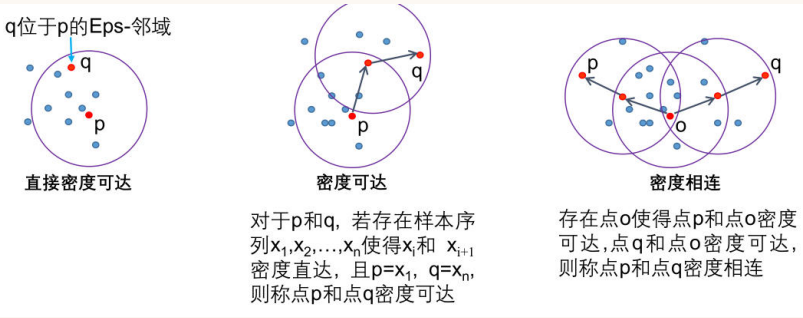
日期: /

<3> 概念

□ DBSCAN中的几个定义:

- **Eps领域**: 给定样本点p，其半径为 Eps 内的区域称为该对象的 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(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 , $p=p_1$, $q=p_n$, 假定对象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。



对于p和q，若存在样本序列$x_1,x_2,\ldots,x_n$使得x_i和 x_{i+1}密度直达，且$p=x_1$，$q=x_n$，则称点p和点q密度可达

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

<4> 过程

- 输入: 聚类半径Eps, 密度阈值 MinPts, 样本集合D
- 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。

日期: /

□DBSCAN的主要优点：

- 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。

□DBSCAN的主要缺点：

- 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，主要需要对距离阈值ϵ，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。