PART5 聚类

2020-05-25

河北师范大学 软件学院 基本内容：

1. 什么是聚类？什么是分类？二者的区别与联系。
2. (1)以连续实值特征向量描述的样本点之间的距离度量方式欧式距离、切比雪夫距离、曼哈顿距离、马氏距离

例：若多维特征空间特征分布的协方差矩阵为Σ，则该空间任意两点之间

𝒙𝒊, 𝒙𝒋的马氏距离为：

.

𝑑 𝒙𝒊, 𝒙𝒋𝒙𝒊 𝒙𝒋 Σ 𝒙𝒊 𝒙𝒋

对于 1 维特征空间，特征分布标准差𝜎，则该空间任意两点之间𝒙𝒊, 𝒙𝒋的马氏

距离为：

𝑥 𝑥

𝑑 𝑥 , 𝑥

𝜎

(2)样本点与集合(例：聚类簇)之间距离

(3)集合与集合之间(例：两个聚类簇之间)的距离最小距离、最远距离、平均距离

1. 动态聚类。掌握 K-Means Clustering 算法(目标函数？哪些因素影响聚类性能？实现步骤？)
2. 系统聚类。以聚合式系统聚类为例，掌握系统聚类(实现步骤？哪些因素影响聚类性能？)。
3. 密度聚类。以 DBSCAN 算法为例，理解密度聚类实现的基本流程，掌握有关概念，哪些因素会影响聚类的效果。

练习:

1. 给定数据集𝑫 𝒙𝒊, 𝑖 1, … , 𝑚 ，其中𝑥 ∈ 𝑅 .若采用 K-均值聚类 算法将该

数据集𝑫划分为𝐊簇𝑪𝟏, ⋯ , 𝑪𝑲 ，请完成如下工作：

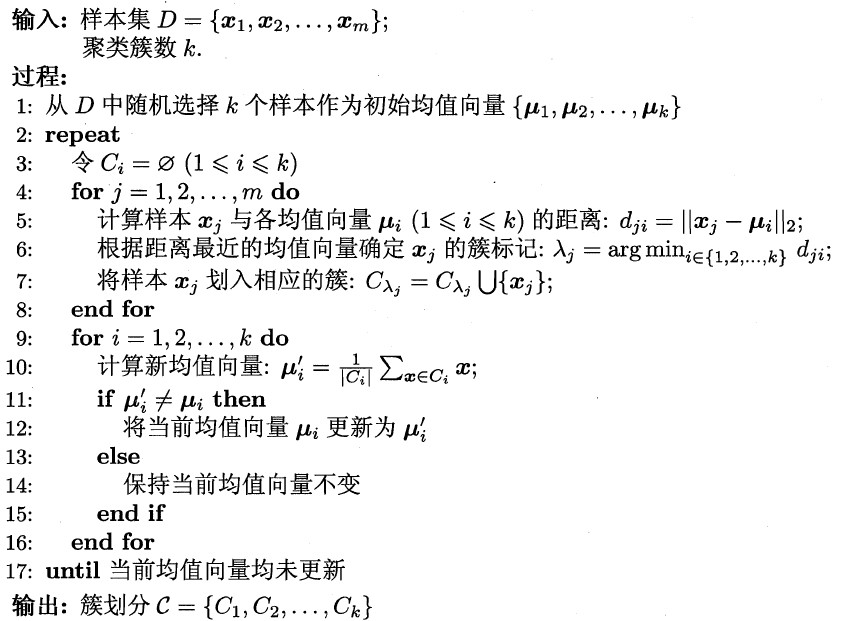
1. 写出 K-Means 算法所对应的准则函数，并给出必要的参数说明；
2. 对 K-Means 算法的实现过程进行描述；
3. 指出影响数据集𝑫划分结果的可能因素.

解：

(1)

|  |
| --- |
| **聚类准则 总的簇内误差平方和--**" "最小将样本集 划分成 个簇:***D k D***  ***C***1 ***Ck***  **总的簇内误差平方和** ***E***  1,..., ***k ,C***1,,***Ck***  ***x*** ***i*** 2  ***i*** 1 ***x Ci***  其中 ***Ci*** 第***i***个簇 聚类 ,  ***i*** 1,...,***k***  ***Ni***  ***Ci*** 第***i***个簇的样本数目  ***i*** 第***i***个聚类簇中心，***i***  1  ***x***  ***Ci x C*** ***i*** |

(2)



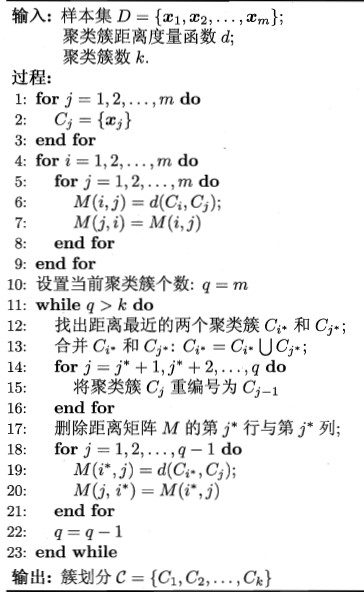
(3) K值的大小、样本集是否进行规范化预处理、聚类中心的初始化方式

1. (1)若要采用合并式层次聚类将样本集***D x*** *i*,i 1,,*m*划分为k个聚类簇，其中***x****i* *R p* . 请对该聚类算法的实现流程予以描述.

(2)上述聚类过程中，需要进行不同聚类簇之间的距离计算，请分别采用最近距离、最远距离，估算任意两个聚类簇***C Cj*** , ***l*** 之间的距离.

解：

**(1)**



**(2)**

最近距离***d***min ***C Cj*** , ***l***  ***x C***min ***j dist x z*** , 

***z******Cl***

最远距离***d***max ***C Cj*** , ***l***  max***x C*** ***j dist x z*** , 

***z******Cl***

1. DBSCAN是一种基于密度的聚类算法，若要基于该算法,对观测点集

***D x*** *i*,i 1,,*N*进行聚类，需要提供两个全局参数,*MinPts*，两个参数的意义是什么？如何确定核心对象？什么是密度直达？什么是密度可达？什么是密

度相连？

解：

 1 两个**全局邻域参数**,***MinPts***

邻域最大半径**--**

***MinPts*--**给定样本的**邻域**内最小样本数.

其中：

**邻域**  对于 ***x j D***, 的 邻域为***x j***  ***N******x j*** ***xi*** ***D dist x x***|  ***i*** , ***j*** 

 2 **核心对象*****core object*** 

若***x j*** ***D***并且***N******x j*** ***MinPts***，则称 为一个***x j* 核心对象**.

 3 **密度直达*****directly density*** ***reacheable***

若***x j*** ***N******xi*** ,并且 为一个核心对象，则称 为由***xi xj xi*密度直达**.

 4 **密度可达*****density*** ***reacheable*** 对于***x xi*** , ***j***，若存在样本序列***p p***1, 2,..., ***pn***,其中***p***1**=*x pi*** , ***n* =*xj***,且 由***pi***1 ***pi***  密度直达，则称 由***x j xi*密度可达**.

(5)**密度相连*****density******connected***

对于***xi***与***x j***，若存在样本 使得***xk*** , ***xi***与***x j***均由***xk*密度可达**，则称***xi***与***x j*  密度相连**