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$\boldsymbol{x_i}, \boldsymbol{x_j}$的马氏距离为：

$$d_M(\boldsymbol{x_i}, \boldsymbol{x_j}) = \left[(\boldsymbol{x_i} - \boldsymbol{x_j})^T \Sigma^{-1} (\boldsymbol{x_i} - \boldsymbol{x_j})\right]^{0.5}$$

   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$\boldsymbol{x_i}, \boldsymbol{x_j}$的马氏距离为：

$$d_M(x_i, x_j) = \left|\frac{x_i - x_j}{\sigma}\right|$$

   (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

3. **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？)
4. **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？)。
5. **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$\boldsymbol{D} = \{\boldsymbol{x_i}, i = 1, \ldots, m\}$，其中$x_i \in R^d$.若采用 K-均值聚类 算法将该数据集$\boldsymbol{D}$划分为K簇$\{\boldsymbol{C_1}, \cdots, \boldsymbol{C_K}\}$，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$\boldsymbol{D}$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$\boldsymbol{D}$划分成$\boldsymbol{k}$个簇: $\boldsymbol{D}=\boldsymbol{C}_1\cup\cdots\cup\boldsymbol{C}_k$

总的簇内误差平方和 $\boldsymbol{E}\left(\mu_1,...,\mu_k,\boldsymbol{C}_1,\cdots,\boldsymbol{C}_k\right)=\sum_{i=1}^{k}\sum_{x\in\boldsymbol{C}_i}\left\|\boldsymbol{x}-\boldsymbol{\mu}_i\right\|^2$

其中 $\boldsymbol{C}_i$ --第$\boldsymbol{i}$个簇(聚类), $\boldsymbol{i}=1,...,\boldsymbol{k}$

$\boldsymbol{N}_i=\left|\boldsymbol{C}_i\right|$--第$\boldsymbol{i}$个簇的样本数目

$\boldsymbol{\mu}_i$--第$\boldsymbol{i}$个聚类簇中心， $\boldsymbol{\mu}_i=\frac{1}{\left|\boldsymbol{C}_i\right|}\sum_{x\in\boldsymbol{C}_i}\boldsymbol{x}$

(2)

输入: 样本集 $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m\}$;
聚类簇数 k.

过程:

从 D 中随机选择 k 个样本作为初始均值向量 $\{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\ldots,\boldsymbol{\mu}_k\}$
repeat
  令 $C_i=\varnothing\ (1\leqslant i\leqslant k)$
  for $j=1,2,\ldots,m$ **do**
    计算样本 $\boldsymbol{x}_j$ 与各均值向量 $\boldsymbol{\mu}_i\ (1\leqslant i\leqslant k)$ 的距离: $d_{ji}=||\boldsymbol{x}_j-\boldsymbol{\mu}_i||_2$;
    根据距离最近的均值向量确定 $\boldsymbol{x}_j$ 的簇标记: $\lambda_j=\arg\min_{i\in\{1,2,\ldots,k\}}d_{ji}$;
    将样本 $\boldsymbol{x}_j$ 划入相应的簇: $C_{\lambda_j}=C_{\lambda_j}\bigcup\{\boldsymbol{x}_j\}$;
  end for
  for $i=1,2,\ldots,k$ **do**
    计算新均值向量: $\boldsymbol{\mu}_i'=\frac{1}{|C_i|}\sum_{\boldsymbol{x}\in C_i}\boldsymbol{x}$;
    if $\boldsymbol{\mu}_i'\neq\boldsymbol{\mu}_i$ **then**
      将当前均值向量 $\boldsymbol{\mu}_i$ 更新为 $\boldsymbol{\mu}_i'$
    else
      保持当前均值向量不变
    end if
  end for
until 当前均值向量均未更新
输出: 簇划分 $\mathcal{C}=\{C_1,C_2,\ldots,C_k\}$

(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

2. (1)若要采用**合并式层次聚类**将样本集 $\boldsymbol{D}=\left\{\boldsymbol{x}_i,\mathrm{i}=1,\cdots,m\right\}$ 划分为k个聚类簇，其

中 $\boldsymbol{x}_i \in R^p$. 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 $\boldsymbol{C}_j, \boldsymbol{C}_l$ 之间的距离.

解：

(1)

输入： 样本集 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$;
聚类簇距离度量函数 d;
聚类簇数 k.

过程：

```
for j = 1,2,...,m do
    C_j = {x_j}
end for
for i = 1,2,...,m do
    for j = 1,2,...,m do
        M(i,j) = d(C_i, C_j);
        M(j,i) = M(i,j)
    end for
end for
设置当前聚类簇个数: q = m
while q > k do
    找出距离最近的两个聚类簇 C_i* 和 C_j*;
    合并 C_i* 和 C_j*: C_i* = C_i* ∪ C_j*;
    for j = j* + 1, j* + 2, ..., q do
        将聚类簇 C_j 重编号为 C_{j-1}
    end for
    删除距离矩阵 M 的第 j* 行与第 j* 列;
    for j = 1,2,...,q-1 do
        M(i*, j) = d(C_i*, C_j);
        M(j, i*) = M(i*, j)
    end for
    q = q - 1
end while
```

输出： 簇划分 $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$

(2)

最近距离 $\boldsymbol{d}_{\min}\left(\boldsymbol{C}_j, \boldsymbol{C}_l\right) = \min_{\substack{\boldsymbol{x}\in\boldsymbol{C}_j \\ \boldsymbol{z}\in\boldsymbol{C}_l}} \boldsymbol{dist}\left(\boldsymbol{x}, \boldsymbol{z}\right)$

最远距离 $\boldsymbol{d}_{\max}\left(\boldsymbol{C}_j, \boldsymbol{C}_l\right) = \max_{\substack{\boldsymbol{x}\in\boldsymbol{C}_j \\ \boldsymbol{z}\in\boldsymbol{C}_l}} \boldsymbol{dist}\left(\boldsymbol{x}, \boldsymbol{z}\right)$

3. 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 $\boldsymbol{D} = \left\{\boldsymbol{x}_i, \mathrm{i} = 1, \cdots, N\right\}$ 进行聚类，需要提供两个全局参数 $\left(\varepsilon, MinPts\right)$，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**$(\varepsilon, MinPts)$

ε--邻域最大半径

***MinPts*--**给定样本的$\varepsilon-$**邻域**内最小样本数.

其中：

$\varepsilon-$**邻域** 对于$\forall x_j \in D$， x_j的$\varepsilon-$邻域为$N_\varepsilon(x_j)=\{x_i \in D \mid dist(x_i, x_j) \le \varepsilon\}$

(2)**核心对象**$(core\ \ object)$

若$x_j \in D$并且$|N_\varepsilon(x_j)| \ge MinPts$，则称$x_j$为一个**核心对象**.

(3)**密度直达**$(directly\ \ density-reacheable)$

若$x_j \in N_\varepsilon(x_i)$,并且$x_i$为一个核心对象，则称$x_j$为由$x_i$**密度直达**.

(4)**密度可达**$(density-reacheable)$

对于x_i, x_j，若存在样本序列$p_1, p_2, ..., p_n$,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**.

(5)**密度相连**$(density-connected)$

对于x_i与x_j，若存在样本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**