# 第14章 聚类方法

- 聚类是什么
  - 针对给定的样本，依据它们特征的**相似度**或**距离**，将其归并到若干个类或簇的数据分析问题
- 聚类的目的
  - 通过得到的类或簇来发现数据的特点或对数据进行处理，在数据挖掘、模式识别等领域有着广泛的作用
- 聚类属于**无监督学习**
  - 根据相似度或距离划分，初始时多少类并不知道
- 聚类算法：
  - 层次聚类（hierarchical clustering）
    - **聚合法**：**自下而上**，即开始时将**每个**样本各自分为**一个类**，之后将相距**最近**的两类**合并**，建立一个**新的类**，**重复**此操作直至满足条件，得到层次化的类别
    - **分裂法**：**自上而下**，即开始时将**所有**样本归为**一类**，之后将已有的类中距离相距**最远**的样本**分**到两个新的类，**重复**此操作直至满足条件，得到层次化的类别
  - k均值聚类（k-means clustering）：基于中心的聚类，通过迭代，将样本分到 $k$ 个类中，使得每个样本与其所属类的中心或均值最近，得到 $k$ 个平坦的、非层次化的类别，构成对空间的划分

## 14.1 聚类的基本概念

### 14.1.1 相似度或距离

- 聚类的对象是观测数据或样本集合。假设有 $n$ 个样本，每个样本有 $m$ 个属性的特征向量组成。样本集合表示为:

$$X = [x_{ij}]_{m\times n} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

  - 元素 $x_{ij}$ 表示第 $i$ 个样本第 $j$ 个属性，$i = 1, 2, \ldots, n, j = 1, 2, \ldots, m$
- 聚类的核心概念是相似度或距离，有多种相似度或距离的定义。因为相似度直接影响聚类的结果，所以其选择是聚类的根本问题

#### 闵可夫斯基距离

- 在聚类中，可以将样本集合想象成向量空间中的点，以该空间的距离表示样本之间的相似度

**定义14.1** 给定样本集合$X$，$X$是$m$维实数向量空间$R^m$中点的集合，其中$x_i, x_j \in X$，$x_i = \{x_{1i}, x_{2i}, \cdots, x_{mi}\}^T$，$x_j = \{x_{1j}, x_{2j}, \cdots, x_{mj}\}^T$，样本$x_i$与$x_j$之间的闵可夫斯基距离定义为：

$$d_{ij} = (\sum_{k=1}^{m} |x_{ki} - x_{kj}|^p)^{\frac{1}{p}}$$

这里$p \geq 1$。当$p = 2$时称为欧式距离，即

$$d_{ij} = (\sum_{k=1}^{m} |x_{ki} - x_{kj}|^2)^{\frac{1}{2}}$$

当$p = 1$时称为曼哈顿距离，即

$$d_{ij} = \sum_{i=1}^{m} |x_{ki} - x_{kj}|$$

当$p = \infty$时称为切比雪夫距离，取各个坐标差点最大值，即

$$d_{ij} = max_k |x_{ki} - x_{kj}|$$

## 马哈拉诺比斯距离（马氏距离）

- 考虑各个分量(特征)之间的相关性并与各个分量的尺度无关
- 马氏**距离越大相似度越小，距离越小相似度越大**

**定义14.2** 给定一个样本集合$X$，$X = (x_{ij})m \times n$，其协方差矩阵记作$S$。样本$x_i$与样本$x_j$之间的马哈拉诺比斯距离$d_{ij}$定义为

$$d_{ij} = [(x_i - x_j)^T S^{-1} (x_i - x_j)]^{\frac{1}{2}}$$

其中，

$$x_i = (x_{1i}, x_{2i}, \cdots, x_{mi})^T, \quad x_j = (x_{1j}, x_{2j}, \cdots, x_{mj})^T$$

当$S$为单位矩阵时，即样本数据的各个分量互相独立且各个分量的方差为1时，马氏距离就是欧氏距离，可以将马氏距离看作是欧氏距离的推广。

## 相关系数

- 相关系数的绝对值越接近于**1**，表示样本越相似
- 相关系数的绝对值越接近于**0**，表示样本越不相似

**定义14.3** 样本$x_i$与样本$x_j$之间的相关系数定义为

$$r_{ij} = \frac{\sum_{k=1}^{m} (x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j)}{[\sum_{k=1}^{m} (x_{ki} - \overline{x}_i)^2 \sum_{i=1}^{m} (x_{kj} - \overline{x}_j)^2]^{\frac{1}{2}}}$$

其中，

$$\overline{x}_i = \frac{1}{m} \sum_{k=1}^{m} x_{ki}, \quad \overline{x}_j = \frac{1}{m} \sum_{k=1}^{m} x_{kj}$$

## 夹角余弦

- 夹角余弦越接近于**1**，表示样本越相似
- 夹角余弦越接近于**0**，表示样本越不相似

**定义14.4** 样本$x_i$与样本$x_j$之间的余弦夹角定义为

$$s_{ij}=\frac{\sum_{k=1}^{m}x_{ki}x_{kj}}{[\sum_{k=1}^{m}x_{ki}^2\sum_{k=1}^{m}x_{kj}^2]^{\frac{1}{2}}}$$

## 14.1.2 类或簇

- 通过聚类得到的类或簇，本质是样本的子集
  - **硬聚类方法**：一个聚类方法假定一个样本**只能属于一个类**，或类的**交集为空集**
  - **软聚类方法**：一个聚类方法假定一个样本可以属于**多个类**，或类的交集**不为空集**

用$G$表示类或簇，用$x_i,x_j$表示类中的样本，用$n_G$表示$G$中样本的个数，用$d_{ij}$表示样本$x_i$与样本$x_j$之间的距离。

**定义14.5** 设 $T$ 为给定的正数，若集合$G$中任意两个样本$x_i$，$x_j$，有

$$d_{ij}\leqslant T$$

则称 $G$ 为一个类或簇。

**定义14.6** 设 $T$ 为给定的正数，若集合$G$的任意样本$x_i$，一定存在$G$中的另一个样本$x_j$，使得

$$d_{ij}\leqslant T$$

则称$G$为一个类或簇。

**定义14.7** 设 $T$ 为给定的正数，若集合$G$的任意样本$x_i$，$G$中的另一个样本$x_j$满足

$$\frac{1}{n_G-1}\sum_{x_j\in G}d_{ij}\leqslant T$$

其中 $n_G$ 为 $G$ 样本的个数，则称 $G$ 为一个类或簇。

**定义14.8** 设 $T$ 和 $V$ 为给定的两个正数，如果集合 $G$ 中的任意两个样本$x_i$，$x_j$的距离$d_{ij}$满足

$$\frac{1}{n_G(n_G-1)}\sum_{x_i\in G}\sum_{x_j\in G}d_{ij}\leqslant T$$

$$d_{ij}\leqslant V$$

则称 $G$ 为一个类或簇。

- **类的特征**可以通过不同角度来刻画，常用的特征有下面三种

(1) 类的均值$\bar{x}_G$，由称为类的中心

$$\bar{x}_G = \frac{1}{n_G}\sum_{i=1}^{nG} x_i$$

式中$n_G$是类$G$的样本个数。

(2) 类的直径$D_G$

类的直径$D_G$是类中任意两个样本之间的最大距离，即

$$D_G = max_{x_i,x_j \in G} d_{ij}$$

(3) 类的样本散布矩阵$A_G$与样本协方差矩阵$S_G$

类的散布矩阵$A_G$

$$A_G = \sum_{i=1}^{nG}(x_i - \bar{x}_G)(x_i - \bar{x}_G)^T$$

样本协方差矩阵$S_G$为

$$S_G = \frac{1}{m-1}A_G = \frac{1}{m-1}\sum_{i=1}^{nG}(x_i - \bar{x}_G)(x_i - \bar{x}_G)^T$$

其中$m$为样本的维数(样本属性的个数)。

### 14.1.3 类与类之间的距离

下面考虑类 $G_p$ 与类 $G_q$ 之间的距离 $D(p,q)$，也称为连接（linkage）。类与类之间的距离也有多种定义。

设类 $G_p$ 包含 $n_p$ 个样本，$G_q$ 包含 $n_q$ 个样本，分别用 $\bar{x}_p$ 和 $\bar{x}_q$ 表示 $G_p$ 和 $G_q$ 均值，即类的中心。

（1）最短距离或单连接（single linkage）

定义类 $G_p$ 的样本与 $G_q$ 的样本之间的最短距离为两类之间的距离

$$D_{pq} = \min\{d_{ij} | x_i \in G_p, x_j \in G_q\}$$

（2）最长距离或完全连接（complete linkage）

定义类 $G_p$ 的样本与 $G_q$ 的样本之间的最长距离为两类之间的距离

$$D_{pq} = \max\{d_{ij} | x_i \in G_p, x_j \in G_q\}$$

（3）中心距离

定义类 $G_p$ 与类 $G_q$ 的中心 $\bar{x}_p$ 与 $\bar{x}_q$ 之间的距离为两类之间的距离

$$D_{pq} = d_{\bar{x}_p \bar{x}_q}$$

（4）平均距离

定义类 $G_p$ 与类 $G_q$ 任意两个样本之间距离的平均值为两类之间的距离

$$D_{pq} = \frac{1}{n_p n_q} \sum_{x_i \in G_p} \sum_{x_j \in G_q} d_{ij}$$

## 14.2 层次聚类

**聚合聚类算法**

- 聚合聚类开始将每个样本各自分为**一个类**，之后将**距离最近**的两个类**合并**，建立一个**新类**，重复此操作直至满足停止条件，得到层次化的类别
- 具体步骤：
  - 输入：$n$ 个样本组成的样本集合及样本之间的距离
  - 输出：对样本集合的一个层次化聚类
    1. 计算 $n$ 个样本两两之间的欧氏距离 $d_{ij}$ ，记作矩阵 $D = [d_{ij}]_{n \times n}$
    2. 构造 $n$ 个类，每个类只包含一个样本
    3. 合并类间距离最小的两个类，其中**最短距离为类间距离**，构建一个新的类
    4. 计算新类与当前各类的距离，若类的个数为1，终止计算，否则回到(3)
  - 聚合层次聚类算法的复杂度为 $O(n^3 m)$ ，其中 $m$ 是样本的维数，$n$ 是样本个数

**分裂聚类算法**

- 分裂聚类算法开始将所有样本分为**一个类**，之后将已有类中**距离最远**的样本**分**到两个新类，重复此操作直至满足停止条件，得到层次化的类别
- 具体步骤：
  - 输入：$n$ 个样本组成的样本集合及样本之间的距离

- 输出：对样本集合的一个层次化聚类

1. 计算 $n$ 个样本两两之间的欧氏距离 $d_{ij}$ ，记作矩阵 $D=[d_{ij}]_{n\times n}$
2. 将样本集中的所有的样本归为一个类
3. 在同一个类 $c$ 中计算两两样本之间的距离，找出距离最远的两个样本 $a$ 和 $b$ ，之后将样本 $a$ 和 $b$ 分配到不同的类 $c1$ 和 $c2$ 中
4. 计算原类 $c$ 中剩余的其他样本点和 $a$ 和 $b$ 的距离，若是 $distance(a)<distance(b)$ ，则将样本点归到 $c1$ 中，否则归到 $c2$ 中
5. 重复步骤4直至达到聚类的数目或者达到设定的条件

# 14.3 $k$ 均值聚类

- $k$ 均值聚类将样本集合划分为 $k$ 个子集，构成 $k$ 个类，将 $n$ 个样本分到 $k$ 个类中，每个样本到其**所属类的中心的距离最小**
- k均值聚类属于**硬聚类**，每个样本属于一个类

## 14.3.1 模型

- 给定 $n$ 个样本的集合 $X=\{x_1,x_2,\dots,x_n\}$，每个样本由一个**特征向量**表示，特征向量的**维数**为 $m$ 。$k$ 均值聚类的**目标**是将 $n$ 个样本分到 $k$ 个不同的类或簇中，这里假设 $k<n$ 。$k$ 个类 $G_1,G_2,\dots,G_k$ 形成对样本集合 $X$ 的划分，其中 $G_i\bigcap G_j=\varnothing$ ， $\bigcup_{i=1}^{k}G_i=X$ 。用 $C$ 表示划分，**一个划分对应着一个聚类结果**
- 划分 $C$ 是一个**多对一**的函数。事实上，如果把每个样本用一个整数 $i\in\{1,2,\dots,n\}$ 表示，每个类也用一个整数 $l\in\{1,2,\dots,k\}$ 表示，那么划分或者聚类可以用函数 $l=C(i)$ 表示，其中 $i\in\{1,2,\dots,n\},\quad l\in\{1,2,\dots,n\}$ 。所以 $k$ 均值聚类的模型是一个**从样本到类的函数**

## 14.3.2 策略

- $k$ 均值聚类的策略是**通过损失函数最小化**选取最优的划分或函数 $C^*$
- 采用**欧氏距离平方**作为样本之间的距离

$$d(x_i,x_j)=\sum_{k=1}^{m}(x_{ki}-x_{kj})^2=\|x_i-x_j\|^2$$

- 定义**样本**与其**所属类**的**中心**之间的距离的总和为损失函数

$$W(C)=\sum_{l=1}^{k}\sum_{C(i)=l}\|x_i-\overline{x}_l\|^2$$

  - 式中 $\overline{x}_l=(\overline{x}_{1l},\overline{x}_{2l},\dots,\overline{x}_{ml})^T$ 是第 $l$ 个类的均值或中心
  - $n_l=\sum_{i=1}^{n}I(C(i)=l)$ ，$I(C(i)=l)$ 是指示函数，取值为1或0
  - 函数 $W(C)$ 也称为能量，表示相同类中的样本相似的程度

- $k$ 均值聚类就是**求解最优化问题**：

$$C^*=arg\min_C W(C)=arg\min_C\sum_{l=1}^{k}\sum_{C(i)=l}\|x_i-\overline{x}_l\|^2$$

  - 相似的样本被聚到同类时，损失函数值最小，这个目标函数的最优化能达到聚类的目的
- 该优化问题是 $n$ 个样本分到 $k$ 个类，所有可能**分法数量**：

$$S(n,k)=\frac{1}{k!}\sum_{l=1}^{k}(-1)^{k-l}\binom{k}{l}k^n$$

- 该数量是指数级的，采用迭代求解

## 14.3.3 算法

- **输入**：$n$ 个样本的集合 $X$
- **输出**：样本集合的聚类 $C^*$

1. **初始化**。令 $t=0$，**随机**选择 $k$ 个样本点作为**初始聚类中心** $m^{(0)}=(m_1^{(0)},\cdots,m_l^{(0)},\cdots,m_k^{(0)})$
2. **对样本进行聚类**。对固定的类中心 $m^{(t)}=(m_1^{(t)},\cdots,m_l^{(t)},\cdots,m_k^{(t)})$，其中 $m_l^{(t)}$ 为类 $G_l$ 的中心，计算每个样本到类中的距离，将每个样本**指派到**与其**最近**的中心的类中，构成聚类结果 $C^{(t)}$
3. **计算新的类中心**。对聚类结果 $C^{(t)}$，计算当前各个类中的样本的**均值**，作为**新的类**中心 $m^{(t+1)}=(m_1^{(t+1)},\cdots,m_l^{(t+1)},\cdots,m_k^{(t+1)})$
4. 如果**迭代收敛**或符合**停止条件**，输出 $C^*=C^{(t)}$。否则，令 $t=t+1$，返回**步骤2**

- $k$ 均值聚类算法的**复杂度**为 $O(mnk)$，其中 $m$ 是样本维数，$n$ 是样本个数，$k$ 是类别个数

## 14.3.4 算法特性

- **总体特点**
  - 基于**划分**的聚类算法
  - 类别数 $k$ **事先指定**
  - 以**欧氏距离平方**表示样本之间的距离，以**中心**或样本的**均值**表示类别
  - 以样本和其所属类的中心之间的距离的**总和**为最优化的目标函数
  - 得到的类别是**平坦的、非层次化**的
  - 算法是**迭代**算法，**不**能保证全局最优
- **收敛性**
  - **启发式**算法，无法保证全局最优
  - **初始中心点的选择会影响聚类结果**
  - 类中心随着训练移动，但是移动**不会太大**，因为在每一步中，样本分到与其最近的中心的类中
- **初始类的选择**
  - 选择不同的初始中心，会得到不同的聚类结果
  - 初始中心的先用**层次聚类**对样本进行聚类，得到 $k$ 个类是停止
- **类别数 $k$ 的选择**
  - 尝试用**不同**的 $k$ 值聚类
  - 一般而言，类别数变小时，平均直径会增加，类别数变大超过**某一个值**时，平均直径不变，即得到最优的 $k$ 值