

第14章 聚类方法

1 聚类的基本概念

相似度或距离

- n 个样本，每个样本由 m 个属性的特征向量组成，样本合集用矩阵 X 表示

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

矩阵的第 j 列表示第 j 个样本【列向量】， $j = 1, 2, \dots, n$ ；第 i 行表示第 i 个属性， $i = 1, 2, \dots, m$ 。矩阵元素 x_{ij} 表示第 j 个样本的第 i 个属性值， $i = 1, 2, \dots, m; j = 1, 2, \dots, n$

- 聚类的核心概念是相似度(similarity)或距离(distance)
- 相似度选择是聚类的根本问题

闵可夫斯基距离

- ▶ 闵可夫斯基距离越大相似度越小，距离越小相似度越大
- ▶ 定义14.1给定样本集合 X ， X 是 m 维实数向量空间 \mathbf{R}^m 中点的集合，其中 $x_i, x_j \in X, x_i = (x_{1i}, x_{2i}, \dots, x_{mi})^T, x_j = (x_{1j}, x_{2j}, \dots, x_{mj})^T$ ，样本 x_i 与样本 x_j 的闵可夫斯基距离(Minkowski distance)定义为

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

其中 $p \geq 1$

闵可夫斯基距离

➤ $p = 2$, 欧氏距离(Euclidean distance)

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

➤ $p = 1$, 曼哈顿距离(Manhattan distance)

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

➤ $p = \infty$, 切比雪夫距离(Chebyshev distance), 取各个坐标数值差的绝对值的最大值

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

马哈拉诺比斯距离

马哈拉诺比斯距离(Mahalanobis distance), 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。马哈拉诺比斯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

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

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

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。马氏距离是欧氏距离的推广。

马氏距离考虑了分布的各向异性。

相关系数(correlation coefficient)

相关系数的绝对值越接近于1，表示样本越相似；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

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

其中 $\bar{x}_i = \frac{1}{m} \sum_{k=1}^m x_{ki}$, $\bar{x}_j = \frac{1}{m} \sum_{k=1}^m x_{kj}$ ，不同属性的平均

定义了样本不同属性之间偏离均值的相关性

夹角余弦(cosine)

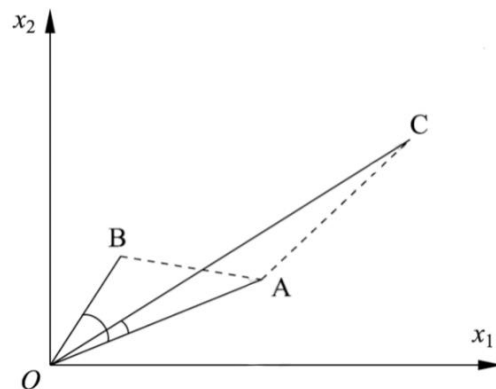
夹角余弦越接近于1，表示样本越相似；越接近于0，表示样本越不相似。

定义14.4 样本 x_i 与样本 x_j 之间的夹角余弦定义为

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

从距离的角度看，A和B比A和C更相似

从相关系数的角度看，A和C比A和B更相似。



类或簇

- ▶ 通过聚类得到的类或簇，本质是样本的子集
- ▶ 一个样本只能属于一个类，或类的交集为空集，那么该聚类方法称为硬聚类(hard clustering)方法
- ▶ 一个样本可以属于多个类，或类的交集不为空集，那么该聚类方法称为软聚类(soft clustering)方法

类或簇(cluster)

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

定义14.5 T 为给定的正数, 若 G 中任意两个样本 x_i, x_j , 有 $d_{ij} \leq T$, 则称 G 为一个类或族。

定义14.6 T 为给定的正数, 若对集合 G 的任意样本 x_i , 存在 G 中另一样本 x_j , 使得 $d_{ij} \leq T$, 则称 G 为一个类或族。

定义14.7 T 为给定的正数, 若对集合 G 中任意样本 x_i , G 中的另一样本 x_j 满足

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

其中 n_G 为 G 中样本的个数, 则称 G 为一个类或族。

定义14.8 T 和 V 为给定的正数, 如果 G 中任意样本 x_i, x_j 的距离 d_{ij} 满

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

则称 G 为一个类或族。

类的特征

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 ，又称为类的中心

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

其中 n_G 是类 G 的样本个数

2) 类的直径 (diameter) D_G 类的直径 D_G 是类中任意两个样本之间的最大距离

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

3) 类的样本散布矩阵 (scatter matrix) A_G

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

【整体偏离均值的程度】

与样本协方差矩阵 (covariance matrix)

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

类与类之间的距离

类 G_p 与类 G_q 之间的距离 $D(p, q)$,也称为连接(linkage)。

类 G_p 包含 n_p 个样本, G_q 包含 n_q 个样本, \bar{x}_p 和 \bar{x}_q 表示 G_p 和 G_q 的均值。

1)最短距离或单连接(single linkage)

类 G_p 的样本与 G_q 的样本之间的最短距离 $D_{pq} = \min\{d_{ij} \mid x_i \in G_p, x_j \in G_q\}$

2)最长距离或完全连接(complete linkage)

类 G_p 的样本与 G_q 的样本之间的最长距离 $D_{pq} = \max\{d_{ij} \mid x_i \in G_p, x_j \in G_q\}$

3)中心距离

类 G_p 与类 G_q 的中心 \bar{x}_p 与 \bar{x}_q 之间的距离 $D_{pq} = d_{\bar{x}_p \bar{x}_q}$

4)平均距离

类 G_p 与类 G_q 任意两个样本之间距离的平均值 $D_{pq} = \frac{1}{n_p n_q} \sum_{x_i \in G_p} \sum_{x_j \in G_q} d_{ij}$

2 层次聚类

层次聚类

- ▶ 层次聚类假设类别之间存在层次结构，将样本聚到层次化的类中
 - ▶ 聚合(agglomerative)或自下而上(bottom-up)聚类
 - ▶ 分裂(divisive)或自上而下(top-down)聚类
- ▶ 每个样本只属于一个类，层次聚类属于硬聚类

层次聚类

➤ 聚合聚类

- 首先，每个样本都作为一个类
- 之后，将相距最近的两类合并，建立一个新的类
- 重复操作，直到满足停止条件

➤ 分裂聚类

- 首先，所有样本作为一个类
- 之后，将已有类中相距最远的样本，分裂成两个新的类
- 重复操作，直到满足停止条件

聚合聚类三要素

➤ 距离或相似度

- 闵可夫斯基距离
- 马哈拉诺比斯距离
- 相关系数
- 夹角余弦

➤ 合并规则

- 类间距离最小(类间距离可以是最短距离、最长距离、中心距离、平均距离)

➤ 停止条件

- 类的个数达到阈值(极端情况，类的个数是1)
- 类的直径超过阈值

聚合聚类算法

算法14.1(聚合聚类算法)

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，否则回到步3)。

可以看出聚合层次聚类算法的复杂度是 $O(n^3m)$,其中 m 是样本的维数， n 是样本个数。

例

例14.1 给定5个样本的集合，样本之间的欧氏距离由如下矩阵 D 表示：

$$D = [d_{ij}]_{5 \times 5} = \begin{bmatrix} 0 & 7 & 2 & 9 & 3 \\ 7 & 0 & 5 & 4 & 6 \\ 2 & 5 & 0 & 8 & 1 \\ 9 & 4 & 8 & 0 & 5 \\ 3 & 6 & 1 & 5 & 0 \end{bmatrix}$$

其中 d_{ij} 表示第 i 个样本与第 j 个样本之间的欧氏距离。 D 为对称矩阵。应用聚合层次聚类法对这5个样本进行聚类。

解 1) 首先用5个样本构建5个类， $G_i = \{x_i\}, i = 1, 2, \dots, 5$ ，这样，样本之间的距离也就变成类之间的距离，所以5个类之间的距离矩阵亦为 D 。

2) 由矩阵 D 可以看出， $D_{35} = D_{53} = 1$ 为最小，所以把 G_3 和 G_5 合并为一个新类，记作 $G_6 = \{x_3, x_5\}$ 。

3) 计算 G_6 与 G_1, G_2, G_4 之间的最短距离，有 $D_{61} = 2$ ， $D_{62} = 5$ ， $D_{64} = 5$

又注意到其余两类之间的距离是 $D_{12} = 7$ ， $D_{14} = 9$ ， $D_{24} = 4$

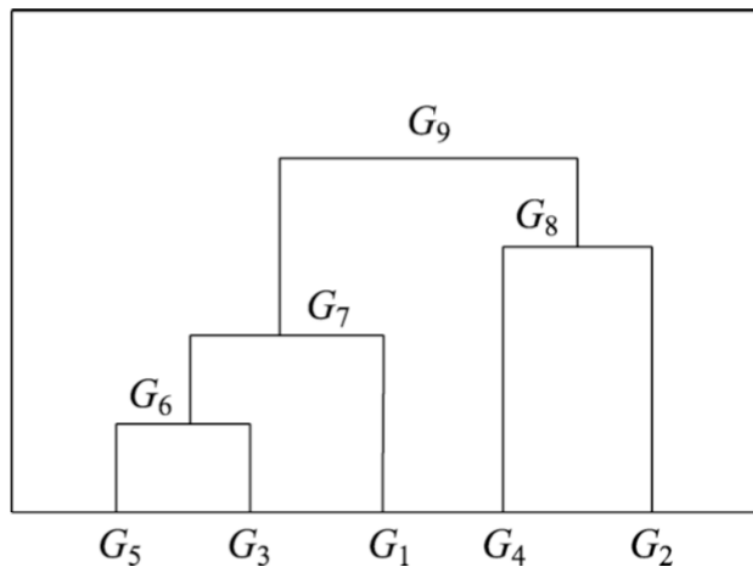
显然， $D_{61} = 2$ 最小，所以将 G_1 与 G_6 合并成一个新类，记作 $G_7 = \{x_1, x_3, x_5\}$ 。

例

4) 计算 G_7 与 G_2, G_4 之间的最短距离, $D_{72} = 5$, $D_{74} = 5$

又注意到 $D_{24} = 4$, 显然, 其中 $D_{24} = 4$ 最小, 所以将 G_2 与 G_4 合并成一类, 记作 $G_8 = \{x_2, x_4\}$ 。

5) 将 G_7 与 G_8 合并成一个新类, 记作 $G_9 = \{x_1, x_2, x_3, x_4, x_5\}$, 即将全部样本聚成1类, 聚类终止。



3 K均值聚类

k均值聚类

- ▶ k均值聚类是基于样本集合划分的聚类算法
 - ▶ 分成多少样本集合
- ▶ k均值聚类将样本集合分到k个类中，每个样本到其所属类的中心的距离最小
- ▶ 每个样本只能属于唯一的一个类，k均值聚类是硬聚类

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, 特征向量的维数是 m
- ▶ k 均值聚类的目标是将 n 个样本分到 k 个不同的类或族中。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 $G_i \cap G_j = \emptyset$, $\bigcup_{i=1}^k G_i = X$
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
- ▶ 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表示
 - ▶ $l = C(i)$, 其中 $i \in \{1, 2, \dots, n\}, l \in \{1, 2, \dots, k\}$

策略

k 均值聚类的策略：通过损失函数的最小化选取最优的划分或函数 C^* 。

首先，样本之间的距离 $d(x_i, x_j)$ ：欧氏距离平方(squared Euclidean distance)

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

然后，定义损失函数：样本与其所属类的中心之间的距离的总和

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

式中 $\bar{x}_l = (\bar{x}_{1l}, \bar{x}_{2l}, \dots, \bar{x}_{ml})^T$ 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。

策略

k 均值聚类就是求解最优化问题：

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

组合优化问题， n 个样本分到 k 类，所有可能分法的数目是：

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

k 均值聚类的最优解求解问题是NP困难问题。采用迭代的方法求解。

算法

k 均值聚类的算法是一个迭代的过程，每次迭代包括两个步骤：

首先选择 k 个类的中心，将样本逐个指派到与其最近的中心的类中，得到一个聚类结果；然后更新每个类的样本的均值，作为类的新的中心。重复以上步骤，直到收敛为止

1)对给定的中心值 (m_1, m_2, \dots, m_k) ,求划分 C ,使得目标函数极小化 $\min_C \sum_{l=1}^k \sum_{C(i)=l} \|x_i - m_l\|^2$

在类中心确定的情况下，将每个样本分到一个类中，使样本和其所属类的中心之间的距离总和最小。求解结果，将每个样本指派到与其最近的中心 m_l 的类 G_l 中

2)对给定的划分 C ,再求各个类的中心 (m_1, m_2, \dots, m_k) ,使得目标函数极小化：

$$\min_{m_1, \dots, m_k} \sum_{l=1}^k \sum_{C(i)=l} \|x_i - m_l\|^2$$

在划分确定的情况下，使样本和其所属类的中心之间的距离总和最小。求解结果，对于每个包含 n_l 个样本的类 G_l ,更新其均值 m_l : $m_l = \frac{1}{n_l} \sum_{C(i)=l} x_i$, $l = 1, \dots, k$

3)重复以上两个步骤，直到划分不再改变，得到聚类结果

算法14.2(k 均值聚类算法)

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 $m^{(0)} = (m_1^{(0)}, \dots, m_l^{(0)}, \dots, m_k^{(0)})$

2)对样本进行聚类。对固定的类中心 $m^{(t)} = (m_1^{(t)}, \dots, m_l^{(t)}, \dots, m_k^{(t)})$, 其中 $m_l^{(t)}$ 为类 G_l 的中心, 计算每个样本到类中心的距离, 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

$$m^{(t+1)} = (m_1^{(t+1)}, \dots, m_l^{(t+1)}, \dots, m_k^{(t+1)})$$

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(2)。

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

算法特性 - 总体特点

- 基于划分的聚类方法
- 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，不能保证得到全局最优

算法特性 - 收敛性

- ▶ k均值聚类属于启发式方法，不能保证收敛到全局最优，初始中心的选择会直接影响聚类结果
- ▶ 类中心在聚类的过程中会发生移动，但是往往不会移动太大，因为在每一步，样本被分到与其最近的中心的类中

算法特性 - 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
- ▶ 初始中心的选择，比如可以用层次聚类对样本进行聚类，得到k个类时停止。然后从每个类中选取一个与中心距离最近的点

算法特性 - 类别数k的选择

- k均值聚类中的类别数k值需要预先指定，而在实际应用中最优的k值是不知道的。
- 尝试用不同的k值聚类，检验得到聚类结果的质量，推测最优的k值
- 聚类结果的质量可以用类的平均直径来衡量
- 一般地，类别数变小时，平均直径会增加
- 类别数变大超过某个值以后，平均直径会不变，而这个值正是最优的k值。实验时，可以采用二分查找，快速找到最优的k值

