

聚类方法

聚类分析：无监督学习方法。

目标：给定样本，依据它们特征的相似或者距离，将其归并到若干个“类”或者“簇”的数据分析问题。这里的相似或者距离起着重要作用。

应用：客户细分、客户画像等。

聚类算法众多，本部分具体介绍层次聚类（hierarchical clustering）和 k -均值聚类（ k -means clustering）。

1 聚类的基本概念

1. 相似 (Similarity) & 距离 (Distance) 度量

假设有 N 个样本，每个样本的特征属性维度为 p 。设收集到的第 i 个样本为 $x_i = (x_{i1}, \dots, x_{ip})^\top$ ，所有样本构成的矩阵为 $X = (x_1, \dots, x_N)^\top \in \mathbb{R}^{N \times p}$ 。

聚类分析的核心概念是相似或者距离。相似或者距离的选择有多种方式，可以根据实际问题选择合适的相似或者距离度量。

(1) 闵可夫斯基距离 (Minkowski distance)

将样本集合看作向量空间中的点的集合，以该空间的距离代表样本之间的距离。常用的距离度量有闵可夫斯基距离。

定义：样本 x_i 与 样本 x_j 的闵可夫斯基距离定义为：

$$d_{ij} = \left(\sum_{k=1}^p |x_{ik} - x_{jk}|^p \right)^{\frac{1}{p}}. \quad (1.1)$$

这里 $p \geq 1$ 。其中 $p = 2$ 时称为欧氏距离（Euclidean distance）； $p = 1$ 时称为曼哈顿距离（Manhattan distance）； $p = \infty$ 时称为切比雪夫距离（Chebyshev distance），即 $d_{ij} = \max_k |x_{ik} - x_{jk}|$ 。

(2) 马氏距离 (Mahalanobis distance)

马氏距离考虑了各个分量之间的相关性并与各个分量的尺度无关。

定义: 给定样本集合 X , 其协方差矩阵记作 S . 样本 x_i 与 x_j 之间的马氏距离 d_{ij} 定义为:

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

注: 当 $S = I$ (各个特征不相关且方差为1) 时, 马氏距离就是欧式距离。

(3) 相关系数 (Correlation coefficient)

定义: 样本 x_i 与 x_j 之间的相关系数定义为:

$$r_{ij} = \frac{\sum_{k=1}^p (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{[\sum_{k=1}^p (x_{ik} - \bar{x}_i)^2 \sum_{k=1}^p (x_{jk} - \bar{x}_j)^2]^{1/2}}$$

其中

$$\bar{x}_i = \frac{1}{p} \sum_{k=1}^p x_{ik}, \quad \bar{x}_j = \frac{1}{p} \sum_{k=1}^p x_{jk}.$$

(4) 夹角余弦

定义: 样本 x_i 与 x_j 之间的夹角余弦定义为:

$$s_{ij} = \frac{\sum_{k=1}^p x_{ik}x_{jk}}{[\sum_{k=1}^p x_{ik}^2 \sum_{k=1}^p x_{jk}^2]^{1/2}}$$

注: 距离与相似度之间可以相互转化, 例如:

距离($d(x, y)$) \rightarrow 相似度($s(x, y)$): $s(x, y) = \frac{1}{1+d(x, y)}$

相似度 \rightarrow 距离: $d(x, y) = \sqrt{2(1 - s(x, y))}$.

2. 类或簇

(1) 类或簇的定义

用 G 表示类或簇 (cluster), 用 x_i 与 x_j 表示类中的样本, 用 n_G 代表 G 中的样本个数, d_{ij} 表示样本 x_i 与 x_j 之间的距离。类或簇有多种定义, 常见定义如下:

定义1: 设 T 为给定的正数, 若集合 G 中任意两个样本 x_i, x_j , 有

$$d_{ij} \leq T.$$

则称 G 为一个类或簇。

定义2: 设 T 为给定的正数, 若集合 G 中任意样本 x_i , 存在 G 中的另外一个样本 x_j , 使得

$$d_{ij} \leq T.$$

则称 G 为一个类或簇。

定义3: 设 T 为给定的正数, 若对集合 G 中任意样本 x_i 成立

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

则称 G 为一个类或簇。

定义4: 设 T 和 V 为给定的两个正数, 若对集合 G 中任意两个样本 x_i, x_j 成立

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

则称 G 为一个类或簇。

以上四个定义中, 定义1较常用。

(2) 常用特征

(a) 类的均值 (类中心):

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

(b) 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，即

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

(c) 类的样本散布矩阵 (scatter matrix) A_G 与样本协方差矩阵 (covariance matrix) S_G

类的样本散布矩阵 A_G :

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

样本的协方差矩阵 S_G :

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

3. 类与类之间的距离

类与类之间的距离 (也称为类连接, linkage) 定义有多种。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 表示两个类的类中心。

(a) 最短距离或单连接 (single linkage)

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

(b) 最长距离或完全连接 (complete linkage)

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

(c) 中心距离

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

(d) 平均距离

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

2 层次聚类

层次聚类有聚合聚类、分裂聚类两种方法. 本部分只介绍聚合聚类。

聚合聚类的基本步骤如下：

- (a) 开始时每个样本各成一类;
- (b) 将类别最近的两个类合并;
- (c) 重复以上操作直到满足停止条件（例如：所有样本被归为一类）

因此，聚合聚类需要确定以下三个要素：

- (1) 距离或相似度（闵可夫斯基距离，马氏距离，...）；
- (2) 合并规则（类间距离）；
- (3) 停止条件

聚合聚类算法如下：

输入： n 个样本组成的样本集合及样本之间的距离；

输出： 层次聚类结果

- (1) 计算 n 个样本之间的距离矩阵 $D = (d_{ij})_{n \times n}$
- (2) 构造 n 个类，每个类包含一个样本；
- (3) 合并类间距离最小的两个类，成为一个新类；
- (4) 计算新类与当前各类的距离。若类的个数为1，则终止计算。否则重复(3).

作业： [例14.1]

3 k 均值聚类

目标： 将 n 个样本分到 k 个不同的 类或簇。 k 个类 G_1, \dots, G_k 形成对样本集合 X 的划分： $G_i \cap G_j = \emptyset, \cup_{i=1}^k G_i = X$. 用 C 表示划分，一个划分对应着一个聚类结果。 划

分是一个多对一的函数。如果把每个样本用一个整数 $i \in \{1, \dots, n\}$ 表示，那么划分或者聚类可以用函数 $l = C(i)$ ($l \in \{1, 2, \dots, k\}$)表示。聚类模型就是一个从样本到类的函数。

3.1 策略

策略：通过损失函数的最小化选取最优的划分或者函数 C 。

采用欧式距离的平方作为样本间的距离 $d(x_i, x_j) = \|x_i - x_j\|^2$ 。定义样本与所属类中心的距离的总和为损失函数，即

$$W(C) = \sum_{l=1}^k \sum_{C(i)=l} \|x_i - \bar{x}_l\|^2 \quad (3.1)$$

其中 $\bar{x}_l = \sum_{C(i)=l} x_i$ 。 $W(C)$ 越小，代表同类别中样本的相似程度越高。

k 均值聚类就是求解最优化问题：

$$C^* = \arg \min_C \sum_{l=1}^k \sum_{C(i)=l} \|x_i - \bar{x}_l\|^2. \quad (3.2)$$

求解以上最优化问题是NP困难问题。在现实中往往采用迭代算法求解。

3.2 算法

k 均值聚类算法是一个迭代过程，包括两个基本步骤：

- (1) 给定 k 个类的中心，将样本逐个指派到与其最近的中心的类中，得到聚类结果；
- (2) 更新类中心：每个类样本的均值

注： (1) 首先，给定类中心 $\{m_1, m_2, \dots, m_K\}$ ，寻找划分 C 使得损失函数最小：

$$\min_C \sum_{l=1}^k \sum_{C(i)=l} \|x_i - m_l\|^2 \quad (3.3)$$

因此，应该将每个样本分到类中心距离它最近的类中。

(2) 给定划分，求类中心使得 $\{m_1, m_2, \dots, m_K\}$ ，使得损失函数最小。此时可求得：

$$m_l = \frac{1}{n_l} \sum_{C(i)=l} x_i \quad (3.4)$$

其中， n_l 代表类 G_l 中样本的个数。

输入： n 个样本的集合 X ，类别的个数 k ;

输出： 聚类结果 C

步骤：

(1) 初始化。令 $t = 0$ 。随机选择 k 个样本作为初始聚类中心： $m^{(0)} = (m_1^{(0)}, \dots, m_k^{(0)})$ 。

(2) 对样本进行聚类。给定类中心 $m^{(t)} = (m_1^{(t)}, \dots, m_k^{(t)})$ ，对每个样本计算样本到类中心的距离，将每个样本分配到距离最近的类别中，得到聚类结果 $C^{(t)}$ 。

(3) 更新类中心。计算各个类别样本的均值，得到更新的类中心。

(4) 如果迭代符合停止条件，则返回聚类结果；否则令 $t = t + 1$ ，重复步骤(2)–(3)。

[练习] 给定样本集合：

$$\begin{pmatrix} 0 & 2 \\ 0 & 0 \\ 1 & 0 \\ 5 & 0 \\ 5 & 2 \end{pmatrix}, \quad (3.5)$$

试用 k 均值算法将以上5个样本分到2个类别中（设定迭代类别中心初始值为前两个样本点取值）。

注：

(1) 收敛性：算法不能保证达到全局最优，最终迭代结果依赖于初值的选取；

(2) 初始点的选取：可以通过层次聚类，设定 k 个类别，选取离类中心点最近的点作为聚类的初始点；

(3) 类别数 k 的选取： k 均值聚类的类别数 k 需要预先指定。可以在不同的 k 值下检验聚类的质量。 例如，用类的平均直径衡量聚类的效果。 k 值变大时，类的直径变小，当 k 值超过某个值以后，类的平均直径会比较稳定。

(4) K 均值聚类不适合发现数据分布形状非凸的情况