统计方法与机器学习

第五章：聚类方法

倪 葎

DaSE@ECNU
(lni@dase.ecnu.edu.cn)

目录

1. 聚类思想
2. 距离的定义
   - 点间距离
   - 类间距离
3. 聚类方法
   - 层次聚类
   - K 均值聚类
   - 混合高斯模型
   - DBSCAN
4. 聚类方法的评价
   - 外部聚类有效性
   - 内部聚类有效性
5. 参数选择
   - 轮廓法
   - CH 指数

聚类思想

聚类方法的核心思想

- **“以类识物”**是人类认识世界的一种重要方式。
- **原因**：人类自身无法认知大量复杂信息。
- **解决方案**：人类通常对个体的特征进行归纳，并将相似的个体归并为一类，以类的特征代替个体信息，以此达到信息的整体性认识。
- 聚类分析就是如何确定“类”的一种途径。

聚类思想

聚类分析的作用

- 作用一：识别从属特定总体的个体。
  - 例如，研究消费者行为从而将市场进行细分，对消费者进行精准广告投放或者商品推荐。

- 作用二：识别异常个体。
  - 例如，监测用户的上网行为从而判断其行为正常或异常，对政府、企业等重要数据库进行保护，并防止黑客攻击。

聚类思想

聚类分析的核心问题

- 聚类分析的核心问题包括以下几个部分：
  - 如何定义个体之间的相似性?
  - 如何确定类别的数目?
  - 如何选取个体的特征?
  - 如何评价聚类方法的结果?

基本概念

基本定义

- 在聚类问题中，我们主要研究的是无标签的数据集。
- 聚类分析是无监督学习中最为常用且重要的方法之一。
- 数据集可以写成矩阵的形式，如下：

变量（特征）

	1	2	$\cdots$	j	$\cdots$	p
1	x_{11}	x_{12}	$\cdots$	x_{1j}	$\cdots$	x_{1p}
2	x_{21}	x_{22}	$\cdots$	x_{2j}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
i	x_{i1}	x_{i2}	$\cdots$	x_{ij}	$\cdots$	x_{ip}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
n	x_{n1}	x_{n2}	$\cdots$	x_{nj}	$\cdots$	x_{np}

基本概念

两个角度

- 每一行表示一个样本，第 i 个样本 $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})'$ 可以看作 p 维空间中的一个点；
- 按**行**进行聚类，将相似的个体聚成一类，由此在数据集 X 中进行**集群发现**。
- 每一列表示一个特征，第 j 个特征 $\boldsymbol{x}_j^* = (x_{1j}, x_{2j}, \cdots, x_{nj})'$ 可以看作 n 维空间中的一个点；
- 按**列**进行聚类，将相似的变量聚成一类，可以对数据集 X 进行**降维**。

距离的意义

动机

- 在聚类分析中，**距离**的定义尤其重要。
- 原因：点或类的远近均依赖于距离的定义。
- 称两个点或类之间的**距离**小，它们的**相似性或亲密度**大；反之，它们的相似性或亲密度小。

点间距离

考虑两个样本 $\boldsymbol{x}_k = (x_{k1}, \cdots, x_{kp})'$ 和 $\boldsymbol{x}_l = (x_{l1}, \cdots, x_{lp})'$

连续变量的点间距离

- 欧氏距离

$$\|\boldsymbol{x}_k - \boldsymbol{x}_l\|_2 = \sqrt{\sum_{j=1}^{p}(x_{kj} - x_{lj})^2}$$

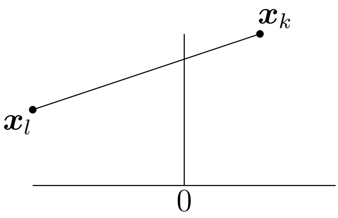
点间距离

考虑两个样本 $\boldsymbol{x}_k = (x_{k1}, \cdots, x_{kp})'$ 和 $\boldsymbol{x}_l = (x_{l1}, \cdots, x_{lp})'$

连续变量的点间距离

- 欧氏距离

$$\|\boldsymbol{x}_k - \boldsymbol{x}_l\|_2 = \sqrt{\sum_{j=1}^{p}(x_{kj} - x_{lj})^2}$$



点间距离

连续变量的点间距离

- 欧氏距离

$$\|\boldsymbol{x}_k - \boldsymbol{x}_l\|_2 = \sqrt{\sum_{j=1}^{p}(x_{kj} - x_{lj})^2}$$

- 平方欧氏距离

$$\|\boldsymbol{x}_k - \boldsymbol{x}_l\|_2^2 = \sum_{j=1}^{p}(x_{kj} - x_{lj})^2$$

点间距离

连续变量的点间距离

- 欧氏距离

$$\|\boldsymbol{x}_k - \boldsymbol{x}_l\|_2 = \sqrt{\sum_{j=1}^{p}(x_{kj} - x_{lj})^2}$$

- 平方欧氏距离

$$\|\boldsymbol{x}_k - \boldsymbol{x}_l\|_2^2 = \sum_{j=1}^{p}(x_{kj} - x_{lj})^2$$

连续变量的点间距离

- 欧氏距离

$$\|\boldsymbol{x}_k - \boldsymbol{x}_l\|_2 = \sqrt{\sum_{j=1}^{p} (x_{kj} - x_{lj})^2}$$

- 闵氏（Minkowski）距离

$$\left(\sum_{j=1}^{p} (x_{kj} - x_{lj})^q \right)^{\frac{1}{q}}$$

点间距离

连续变量的点间距离

- 欧氏距离

$$\|\boldsymbol{x}_k - \boldsymbol{x}_l\|_2 = \sqrt{\sum_{j=1}^{p}(x_{kj} - x_{lj})^2}$$

- 闵氏（Minkowski）距离

$$\left(\sum_{j=1}^{p}(x_{kj} - x_{lj})^q\right)^{\frac{1}{q}}$$

点间距离

连续变量的点间距离

- 曼哈顿（Manhattan）距离（绝对距离）

$$\|\boldsymbol{x}_k - \boldsymbol{x}_l\|_1 = \sum_{j=1}^{p} |x_{kj} - x_{lj}|$$

- 切比雪夫（Chebychev）距离（最大距离）

$$\|\boldsymbol{x}_k - \boldsymbol{x}_l\|_\infty = \max_j |x_{kj} - x_{lj}|$$

点间距离

连续变量的点间距离

- 曼哈顿（Manhattan）距离

$$\|\boldsymbol{x}_k - \boldsymbol{x}_l\|_1 = \sum_{j=1}^{p} |x_{kj} - x_{lj}|$$

- 兰氏（Canberra）距离

$$\sum_{j=1}^{p} \frac{|x_{kj} - x_{lj}|}{|x_{kj}| + |x_{lj}|}$$

  - 兰氏距离可看作加权的曼哈顿距离；
  - 兰氏距离对接近于零的值的变化非常敏感；
  - 兰氏距离对量纲不敏感。

点间距离

连续变量的点间距离

- 欧氏距离

$$\|\boldsymbol{x}_k - \boldsymbol{x}_l\|_2 = \sqrt{\sum_{j=1}^{p}(x_{kj} - x_{lj})^2}$$

- 马氏（Mahalanobis）距离，又称为广义欧式距离

$$\sqrt{(\boldsymbol{x}_k - \boldsymbol{x}_l)'\Sigma^{-1}(\boldsymbol{x}_k - \boldsymbol{x}_l)}$$

  - 当 $\Sigma = \boldsymbol{I}_p$，即 Σ 为单位阵时，马氏距离等于欧式距离；
  - 马氏距离可以看作标准化后的欧式距离。

点间距离

连续变量的点间距离

- **皮尔逊线性相关系数**定义为

$$\text{Pearson } r = \frac{\sum_{j=1}^{p}(x_{kj}-\bar{x}_k)(x_{lj}-\bar{x}_l)}{\sqrt{\sum_{j=1}^{p}(x_{kj}-\bar{x}_k)^2\sum_{j=1}^{p}(x_{lj}-\bar{x}_l)^2}}$$

  - 这里，$\bar{x}_k = \frac{1}{p}\sum_{j=1}^{p} x_{kj}$ 和 $\bar{x}_l = \frac{1}{p}\sum_{j=1}^{p} x_{lj}$；
  - 在概率论中，相关系数用于度量两个随机变量相关性，即

$$\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}},$$

  - 相关系数的取值范围为 $-1 \leq \text{Corr}(X,Y) \leq 1.$
- **皮尔逊线性相关距离** $= 1 - \text{Pearson } r$

点间距离

连续变量的点间距离

- **余弦相似度**定义为

$$\cos\theta = \frac{\sum_{j=1}^{p} x_{kj}x_{lj}}{\sqrt{\sum_{j=1}^{p} x_{kj}^2 \sum_{j=1}^{p} x_{lj}^2}}$$

  - 对于两个向量 $\boldsymbol{a}$ 和 $\boldsymbol{b}$，其夹角的余弦公式为

$$\cos\theta = \frac{\boldsymbol{a}'\boldsymbol{b}}{|\boldsymbol{a}||\boldsymbol{b}|}$$

  - 余弦相似度的取值范围为 $-1 \leq \cos\theta \leq 1$
- **余弦相关距离** $= 1 - \cos\theta$

点间距离

连续变量的点间距离

- **肯德尔秩相关系数**是基于观测值中两个特征同时增加或同时减少的个数从而计算的相关系数。
  - 协同对 (concordant pairs)：$(x_{kj}-x_{kj'})(x_{lj}-x_{lj'})>0$；
  - 不协同对 (discordant pairs)：$(x_{kj}-x_{kj'})(x_{lj}-x_{lj'})<0$
- 肯德尔秩相关系数定义为

$$\text{Kendall } \tau = \frac{n_c - n_d}{p(p-1)/2}$$

其中，n_c 表示协同对的个数，n_d 表示不协同对的个数。

- **肯德尔相关距离** $= 1 - \text{Kendall } \tau$

点间距离

连续变量的点间距离

- **斯皮尔曼秩相关系数**类似于皮尔逊相关系数，只不过将原始的数值 x_{kj} 用其秩 r_{kj} 来代替。
- 将 $\boldsymbol{x}_k$ 的各个分量 $x_{k1}, x_{k2}, \cdots, x_{kp}$ 按从小到大排序, 计算每一个分量所对应的秩，记为 $r_{k1}, r_{k2}, \cdots, r_{kp}$；

$$\text{Spearman } \rho = \frac{\sum_{j=1}^{p}(r_{kj} - \bar{r}_{kj})(r_{lj} - \bar{r}_{lj})}{\sqrt{\sum_{j=1}^{p}(r_{kj} - \bar{r}_{kj})^2 \sum_{j=1}^{p}(r_{lj} - \bar{r}_{lj})^2}}$$

- **斯皮尔曼相关距离** $= 1 - \text{Spearman } \rho$

点间距离

混合变量的点间距离

- 令 $s_j = s_j(x_{kj}, x_{lj})$ 为第 k 个观测值和第 l 个观测值在第 j 个特征或变量之间的相似性。
- 通过定义观测值在第 j 个特征或变量上的相似性 s_j 来定义点与点之间的距离 d_j，即 $s_j = 1 - d_j$。

混合变量的相似性

- 比如考虑性别为男性、产品的颜色是黑色等，这些特征 x_{kj} 和 x_{lj} 均为**定性变量**。
- 通常定义相似性为

$$s_j = s_j(x_{kj}, x_{lj}) = \begin{cases} 1 & \text{如果 } x_{kj} \text{ 和 } x_{lj} \text{ 相同} \\ 0 & \text{如果 } x_{kj} \text{ 和 } x_{lj} \text{ 不同} \end{cases}$$

点间距离

混合变量的相似性

- 比如考虑年龄、产品的价格等，这些特征 x_{kj} 和 x_{lj} 均为**定量变量**。
- 通常定义相似性为

$$s_j = s_j(x_{kj}, x_{lj}) = 1 - \frac{|x_{kj} - x_{lj}|}{R_j}$$

其中，R_j 表示第 j 个特征的极差，即 $R_j = \max_i x_{ij} - \min_i x_{ij}$。

点间距离

混合变量的相似性

- 比如考虑文化程度、空气质量指数级别等，这些特征 x_{kj} 和 x_{lj} 均为**定序变量**。
- 将 $x_{1j}, x_{2j}, \cdots, x_{nj}$ 从小到大进行排序，并分别计算其秩，记为 $r_{1j}, r_{2j}, \cdots, r_{nj}$
- 通常定义相似性为

$$s_j = s_j(x_{kj}, x_{lj}) = 1 - \frac{|r_{kj} - r_{lj}|}{\max_k r_{kj} - \min_k r_{kj}}$$

点间距离

相似度的定义

- 两个观测 $\boldsymbol{x}_k$ 和 $\boldsymbol{x}_l$ 之间的相似度定义为

$$s(\boldsymbol{x}_k, \boldsymbol{x}_l) = \frac{\sum_{j=1}^{p} s_j(x_{kj}, x_{lj})\delta(x_{kj}, x_{lj})w_j}{\sum_{j=1}^{p} \delta(x_{kj}, x_{lj})w_j}$$

其中，

- $\delta(x_{kj}, x_{lj})$ 表示观测值中是否存在缺失观测，即

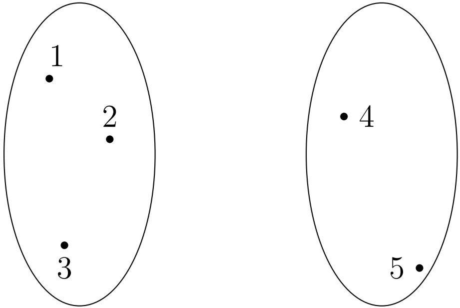
$$\delta(x_{kj}, x_{lj}) = \begin{cases} 0, & \text{如果 } x_{kj} \text{ 或 } x_{lj} \text{ 存在缺失观测} \\ 1, & \text{其他} \end{cases}$$

- w_j 表示权重，一般取值为 1，但是如果事先知道第 j 个特征尤其重要，可以增加相应的权重。

类间距离

定义

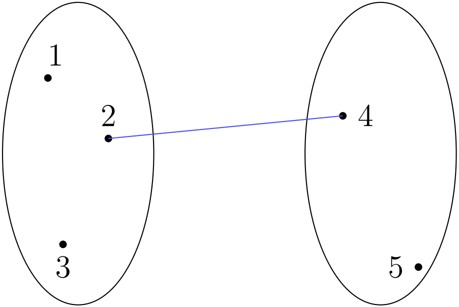
- 由一个点组成的类是最基本的类，如果每一类都由一个点组成，那么**点间的距离就是类间距离**。
- 如果某一个类包含不止一个点，就要定义**类间距离**，也称**关联准则**。
- 例如，



类间距离

关联规则

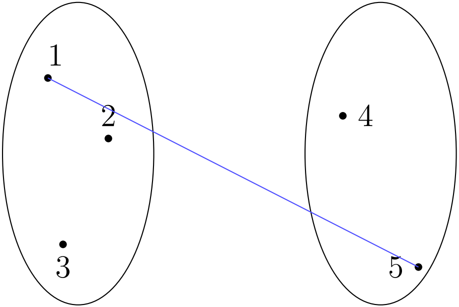
- 将两个类中**距离最短**的两个点之间的距离定义为类间距离，称为 simple linkage。
- 左侧类中选取样本点 2，右侧类中选取样本点 4，定义类间距离为 d_{24}。如



类间距离

关联规则

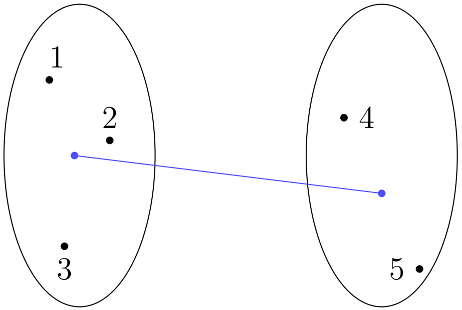
- 将两个类中**距离最长**的两个点之间的距离定义为类间距离，称为 complete linkage。
- 左侧类中选取样本点 1，右侧类中选取样本点 5，定义类间距离为 d_{15}。如



类间距离

关联规则

- 将两个类中所有点的重心的距离定义为类间距离，称为 centroid linkage。
- 左侧类中计算 3 个样本点的重心，右侧类中计算 2 个样本点的重心，将这两个重心之间的距离定义类间距离。如



类间距离

关联准则

- 为了统一符号，我们将点也看作类，因此把点（类）k 与 l 之间的距离用 $d(k,l)$ 表示。如果点（或者类）k 与 l 聚合成一个类，记为 $k\cup l$，对于任何其他的一个点（或者类）i，那么，类 $(k\cup l)$ 与点或类 i 之间的距离记为 $d(k\cup l,i)$。

- Lance-Williams 公式为

$$d(k\cup l,i)=\alpha_k d(k,i)+\alpha_l d(l,i)+\beta d(k,l)+\gamma|d(k,i)-d(l,i)|$$

其中，$\alpha_k,\alpha_l,\beta,\gamma$ 为参数。

类间距离

关联准则

名称	α_k	α_k	β	γ
最短距离法 (single-linkage)	0.5	0.5	0	-0.5
最长距离法 (complete-linkage)	0.5	0.5	0	0.5
类平均法 UPGMA (average linkage)	$\frac{n_k}{n_k+n_l}$	$\frac{n_l}{n_k+n_l}$	0	0
加权类平均法 WPGMA (McQuitty 法)	0.5	0.5	0	0
中位数法 WPGMC (median linkage)	0.5	0.5	-0.25	0
中心法 UPGMC (centroid linkage)	$\frac{n_k}{n_k+n_l}$	$\frac{n_l}{n_k+n_l}$	$-\frac{n_k n_l}{(n_k+n_l)^2}$	0
Ward 法 (minimum variance)	$\frac{n_k+n_i}{n_k+n_l+n_i}$	$\frac{n_l+n_i}{n_k+n_l+n_i}$	$-\frac{n_i}{n_k+n_l+n_i}$	0

层次聚类

概述

- 易于解释的一种聚类方法是层次聚类；
- 层次聚类一般有两种不同的形式：
  - **自下而上**：每个样本各自分到一个类中，之后将类间距离最近的两类关联，并建立一个新的类，反复此过程直到所有的样本聚合至一个类中；
  - **自上而下**：将所有样本归到一个类中，之后将在类中相距最远的样本记为两个新的类，基于这两个类，将未进行聚类的点逐一比较其与两个新的类的距离，这样所有样本划分成了两类，在每一个类中重复此过程直到每个样本点各自分到一个类中；

算法实例：自下而上

- 假设有 4 个点，距离矩阵为

$$
\begin{array}{cccc}
A & B & C & D \\
\end{array}
$$
$$
\begin{pmatrix}
0 & 1 & 3 & 2 \\
1 & 0 & 5 & 6 \\
3 & 5 & 0 & 4 \\
2 & 6 & 4 & 0
\end{pmatrix}
$$

- 找到距离最近的两个类：A 和 B，把他们聚成一类；

算法实例：自下而上

- 采用simple linkage，重新计算距离矩阵为

$$\begin{array}{ccc} A,B & C & D \end{array}$$
$$\begin{pmatrix} 0 & 3 & 2 \\ 3 & 0 & 4 \\ 2 & 4 & 0 \end{pmatrix}$$

- 找出距离最近的两个类：A, B 和 D，把他们聚成一类；

层次聚类

算法实例：自下而上

- 采用simple linkage，重新计算距离矩阵为

$$\begin{array}{cc} A,B,D & C \\ \multicolumn{2}{c}{\begin{pmatrix} 0 & 3 \\ 3 & 0 \end{pmatrix}} \end{array}$$

- 最后可以将 A,B,C,D 聚成一类。

层次聚类

算法实例：自下而上

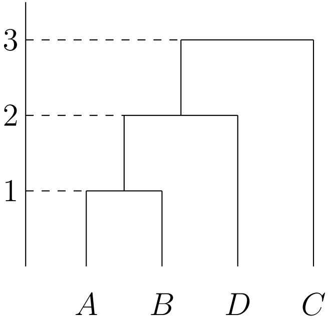


图: 聚合算法的层次聚类结果

K 均值聚类

概述

- K 均值聚类是最简单的无监督学习方法之一，且计算速度快，也称为快速聚类，
- 相较于层次聚类，K 均值聚类事先确定聚类数目；
- 这里假定聚类数目为 $K(K < n)$；
- 给定 n 个样本集 $X = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)'$；
- K 均值聚类的目标是将 n 个样本划分到 K 个不同的类中，这 K 个类 $C_1, C_2, \cdots, C_K$ 形成了样本集 X 的划分，即

$$C_k \cap C_l = \emptyset, \cup_{k=1}^{K} C_k = X$$

- 划分 $\mathcal{C} = \{C_1, C_2, \cdots, C_K\}$ 可以对应一个聚类结果；

K 均值聚类

概述

- 目标：我们希望能够找到一个最优划分 $\mathcal{C}^*$，使得类内距离足够小而类间距离足够大；
- 在 K 均值聚类方法中，通常采用平方欧式距离来表示点与点之间的距离，即

$$d_{kl} = d(k,l) = \|\boldsymbol{x}_k - \boldsymbol{x}_l\|_2^2 = \sum_{j=1}^{p}(x_{kj} - x_{lj})^2$$

K 均值聚类

概述

- 由此，我们可以定义一个合理的损失函数，即

$$W(\mathcal{C}) = \sum_{k=1}^{K} \sum_{i \in C_k} \|\boldsymbol{x}_i - \boldsymbol{m}_k\|^2$$

  其中，
  - $\boldsymbol{m}_k$ 表示第 k 类的均值或中心；
  - 这里 n_k 是第 k 类中样本的个数；
- 而 K 均值聚类实际上就是解决一个最优化问题

$$\mathcal{C}^* = \arg\min_{\mathcal{C}} W(\mathcal{C})$$

- 这是一个 NP-hard 的问题，可采用迭代法求解。

K 均值聚类

概述

- 通常采用迭代法来求解 K 均值聚类的问题，每次迭代包括两个步骤：
    - 确定 K 个类的中心 $\boldsymbol{m}_k$，将样本逐一分配到其最近的中心所对应的类中，得到一个聚类结果；
    - 更新每个类的样本均值，作为类的更新后的中心；重复此过程，直到收敛为止。

K 均值聚类

说明

- 收敛条件，通常可以设置为：聚类结果不变；
- 复杂度是 $O(pnK)$，其中 p 表示特征个数，n 表示样本个数，K 是聚类数目；
- 如果各个类的数据集非凸，K 均值聚类算法难以收敛；

混合高斯模型

概述

- 核心：分布的假定；
- 假定第 i 个样本 $\boldsymbol{x}_i$ 来自于第 k 类正态分布 $N_p(\boldsymbol{\mu}_k, \Sigma_k)$，
    - $\boldsymbol{\mu}_k$ 表示均值向量；
    - Σ_k 表示协方差矩阵；
- $\boldsymbol{x}_i$ 的密度函数为

$$f(\boldsymbol{x}_i) = (2\pi)^{-p/2}|\Sigma_k|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)'\Sigma_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right\}$$

- K 表示聚类数目，可作为一个超参数；
- n 表示样本量；

混合高斯模型

由来

- 如果我们能够确定第 i 个样本是来自于第 k 个高斯分布总体时，那么我们可以构造变量

$$\delta_{ik} = \begin{cases} 1, & \text{当第 } i \text{ 个样本 } \boldsymbol{x}_i \text{ 属于第 } k \text{ 个总体；} \\ 0, & \text{当第 } i \text{ 个样本 } \boldsymbol{x}_i \text{ 不属于第 } k \text{ 个总体.} \end{cases}$$

- 于是，$\boldsymbol{\delta}_i = (\delta_{i1}, \delta_{i2}, \cdots, \delta_{iK})'$ 满足
  - 独立同分布的随机向量；
  - 服从多维分布 $M(1, \pi_1, \pi_2, \cdots, \pi_K)$
  - $\pi_k = P(\delta_{ik} = 1)$ 且满足

$$0 < \pi_k < 1, \quad \sum_{i=1}^{K} \pi_k = 1$$

混合高斯模型

由来

- $\boldsymbol{\delta}_i = (\delta_{i1}, \delta_{i2}, \cdots, \delta_{iK})'$ 的密度函数为

$$f(\boldsymbol{\delta}_i) = \prod_{k=1}^{K} (\pi_k)^{\delta_{ik}}, i = 1, 2, \cdots, n$$

- 给定 $\boldsymbol{\delta}_i$ 后，$\boldsymbol{x}_i$ 的密度函数为

$$f(\boldsymbol{x}_i|\boldsymbol{\delta}_i) = \prod_{k=1}^{K} \left((2\pi)^{-p/2} |\Sigma_k|^{-1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)' \Sigma_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right\} \right)^{\delta_{ik}}$$

混合高斯模型

由来

- 样本 $\{\boldsymbol{x}_i, \boldsymbol{\delta}_i\}, i = 1, 2, \cdots, n$ 的联合密度函数为

$$\begin{aligned} & \prod_{i=1}^{n} f(\boldsymbol{x}_i, \boldsymbol{\delta}_i) \\ = & \prod_{i=1}^{n} f(\boldsymbol{\delta}_i) \cdot f(\boldsymbol{x}_i | \boldsymbol{\delta}_i) \\ = & \prod_{i=1}^{n} \left(\pi_k (2\pi)^{-p/2} |\Sigma_k|^{-1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)' \Sigma_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right\} \right)^{\delta_{ik}} \end{aligned}$$

- 上式也是未知参数 $\boldsymbol{\theta} = (\pi_1, \cdots, \pi_K, \boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K, \Sigma_1, \cdots, \Sigma_K)$ 的似然函数。
- 理论上，基于似然函数，我们可以估计参数 $\boldsymbol{\theta}$。

混合高斯模型

由来

- 实际中，我们仅仅能够观测到样本 $\{\boldsymbol{x}_i\}, i = 1, 2, \cdots, n$；
- 而无法观测到 $\boldsymbol{\delta}_i$；
- 我们无法直接估计未知参数 $\boldsymbol{\theta}$；
- 样本 $\boldsymbol{x}_i$ 的密度函数为

$$f(\boldsymbol{x}_i) = \sum_{k=1}^{K} \pi_k (2\pi)^{-p/2} |\Sigma_k|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)' \Sigma_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right\}$$

- 这个密度函数是由 K 个正态分布的密度函数加权组合而成的，常被称为**高斯混合模型**。

混合高斯模型

估计方法：EM 算法

- 变量 δ_{ik} 无法观测到，于是将其作为潜变量；
- 采用 EM **算法**；
- 未知参数 $\boldsymbol{\theta}$ 的对数似然函数为

$$\begin{aligned} l(\boldsymbol{\theta}) &= \ln L(\boldsymbol{\theta}) \\ &\propto -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\delta_{ik}\left((\boldsymbol{x}_i-\boldsymbol{\mu}_k)'\Sigma_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)+\ln|\Sigma_k|\right)+\sum_{i=1}^{n}\sum_{k=1}^{K}\delta_{ik}\ln(\pi_k) \\ &= Q_0(\boldsymbol{\theta}) \end{aligned}$$

混合高斯模型

估计方法：EM 算法

- **E 步：** 将潜变量 δ_{ik} 的期望 π^*_{ik} 代入 $Q_0(\boldsymbol{\theta})$，即

$$\begin{aligned} Q(\boldsymbol{\theta}) &= -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\pi^*_{ik}\Big((\boldsymbol{x}_i-\boldsymbol{\mu}_k)'\Sigma_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)+\ln|\Sigma_k|\Big)+\sum_{i=1}^{n}\sum_{k=1}^{K}\pi^*_{ik}\ln(\pi_k) \\ &=: Q_1(\boldsymbol{\theta})+Q_2(\boldsymbol{\theta}) \end{aligned}$$

- δ_{ik} 的期望为

$$\pi^*_{ik} = E(\delta_{ik}|\boldsymbol{x}_i) = P(\delta_{ik}=1|\boldsymbol{x}_i) = \frac{\pi_k\phi(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\Sigma_k)}{\sum_{k=1}^{K}\pi_k\phi(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\Sigma_k)}$$

其中，

$$\phi(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\Sigma_k) = (2\pi)^{-p/2}|\Sigma_k|^{-1/2}\exp\left\{-(\boldsymbol{x}_i-\boldsymbol{\mu}_k)'\Sigma_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)/2\right\}$$

混合高斯模型

估计方法：EM 算法

- **M 步：** 求 $Q(\boldsymbol{\theta})$ 的最大值而确定未知参数的估计。
- 我们发现，
  - $Q_1(\boldsymbol{\theta})$ 仅与未知参数 $\{\boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^K$ 有关；
  - $Q_2(\boldsymbol{\theta})$ 仅与未知参数 $\{\pi_k\}_{k=1}^K$；
- 于是，我们可以分别确定最大值点。

混合高斯模型

求导

假定 X 是一个正定对称矩阵。

- 非线性的形式：

$$\frac{\partial \ln \det(X)}{\partial X} = X^{-1}$$

- 关于逆矩阵的求导：

$$\frac{\partial \mathrm{tr}(AX^{-1}B)}{\partial X} = -(X^{-1}BAX^{-1})'$$

混合高斯模型

估计方法：EM 算法

- $Q_1(\boldsymbol{\theta})$ 分别对 $\boldsymbol{\mu}_k$ 和 Σ_k 求导，并使得导函数为零。

$$\begin{cases} \dfrac{\partial Q_1(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}_k} = -\dfrac{1}{2}\displaystyle\sum_{i=1}^{n} \pi_{ik}^{*} \Sigma_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k) = 0 \\ \dfrac{\partial Q_1(\boldsymbol{\theta})}{\partial \Sigma_k} = \displaystyle\sum_{i=1}^{n} \pi_{ik}^{*} \left(\Sigma_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)'\Sigma_k^{-1} + \Sigma_k^{-1}\right) = 0 \end{cases}$$

- 由此解得

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{n} \pi_{ik}^{*} \boldsymbol{x}_i}{\sum_{i=1}^{n} \pi_{ik}^{*}} \tag{1}$$

$$\Sigma_k = \frac{\sum_{i=1}^{n} \pi_{ik}^{*} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)'}{\sum_{i=1}^{n} \pi_{ik}^{*}}. \tag{2}$$

混合高斯模型

估计方法：EM 算法

- 在求 $Q_2(\boldsymbol{\theta})$ 的最大值时，注意这里是对 π_k 有限制条件的，即

$$\sum_{k=1}^{K} \pi_k = 1, \quad 0 < \pi_k < 1, k = 1, 2, \cdots, K$$

- 采用拉格朗日乘子法，令

$$Q_2^*(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \pi_{ik}^* \ln(\pi_k) - \lambda(\sum_{k=1}^{K} \pi_k - 1)$$

混合高斯模型

估计方法：EM 算法

- 对 $Q_2^*(\boldsymbol{\theta})$ 关于 π_k 求导，并使得导函数为零，即

$$\frac{\partial Q_2^*(\boldsymbol{\theta})}{\partial \pi_k} = \sum_{i=1}^{n} \frac{\pi_{ik}^*}{\pi_k} - \lambda = 0$$

- 由此解得

$$\pi_k = \frac{1}{n} \sum_{i=1}^{n} \pi_{ik}^*.$$

概述

- DBSCAN (Density-Based Spatial Clustering and Application with Noise) 是一种典型的基于密度的聚类方法；
- DBSCAN 最早由 Ester 等人于 1996 年所提出的；
- 主要思想为：如果要判断两个样本属于同一类别，那么在这两个样本的附近，能够找到属于同一类别的样本。
- 数据集 $X = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)'$；
- 每个样本 $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})'$ 可以看作 R^p 空间中的一个点；
- 假定第 k 个点 $\boldsymbol{x}_k$ 和第 l 个点 $\boldsymbol{x}_l$ 之间的距离为 $d(k, l)$。

概述

- DBSCAN (Density-Based Spatial Clustering and Application with Noise) 是一种典型的基于密度的聚类方法；
- DBSCAN 最早由 Ester 等人于 1996 年所提出的；
- 主要思想为：如果要判断两个样本属于同一类别，那么在这两个样本的附近，能够找到属于同一类别的样本。
- 数据集 $X = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)'$；
- 每个样本 $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})'$ 可以看作 R^p 空间中的一个点；
- 假定第 k 个点 $\boldsymbol{x}_k$ 和第 l 个点 $\boldsymbol{x}_l$ 之间的距离为 $d(k, l)$。

基本概念

给定邻域半径 $\epsilon > 0$ 和可到达的最少样本个数 MinPts，

- 如果 $N_\epsilon(k) = \{\boldsymbol{x}_l \in X | d(k,l) \leq \epsilon\}$，那么，称点 $\boldsymbol{x}_k$ 的 **$\boldsymbol{\epsilon}$ 邻域**；
- 如果 $|N_\epsilon(k)| \geq \mathrm{MinPts}$, 那么，称点 $\boldsymbol{x}_k$ 为核心点。

DBSCAN

基本概念

给定邻域半径 $\epsilon > 0$ 和可到达的最少样本个数 MinPts,

- 如果点 $\boldsymbol{x}_l$ 满足
$$\boldsymbol{x}_l \in N_\epsilon(k) \quad 且 \quad |N_\epsilon(k)| \geq \text{MinPts},$$
那么，称点 $\boldsymbol{x}_l$ 可以从点 $\boldsymbol{x}_k$ **直接密度可达或密度直达**；
- 如果存在一列点 $\boldsymbol{x}_{k_0} = \boldsymbol{x}_k, \boldsymbol{x}_{k_1}, \boldsymbol{x}_{k_2}, \cdots, \boldsymbol{x}_{k_n} = \boldsymbol{x}_l \in X$, 使得点 $\boldsymbol{x}_{i+1}$ 可以从点 $\boldsymbol{x}_i$ 直接密度可达，那么，称点 $\boldsymbol{x}_l$ 可以从点 $\boldsymbol{x}_k$ **密度可达**；
- 如果存在一个点 $\boldsymbol{x}_i$ 使得点 $\boldsymbol{x}_k$ 和 $\boldsymbol{x}_l$ 均从点 $\boldsymbol{x}_i$ 密度可达，那么，称点 $\boldsymbol{x}_k$ 和点 $\boldsymbol{x}_l$ 密度连接。

说明

- 在**直接密度可达**和**密度可达**这两个定义中，点 $\boldsymbol{x}_k$ 均是核心点，而点 $\boldsymbol{x}_l$ 不一定是核心点。
- 因此，**（直接）密度可达**不是一种对称关系，即 $\boldsymbol{x}_l$ 可以从 $\boldsymbol{x}_k$（直接）密度可达，但是，$\boldsymbol{x}_k$ 不一定可以从 $\boldsymbol{x}_l$（直接）密度可达。
- 然而，在**密度连接**的定义中并未要求点 $\boldsymbol{x}_k$ 和 $\boldsymbol{x}_l$ 是核心对象，因此，**密度连接**是一种对称关系，即 $\boldsymbol{x}_l$ 可以从 $\boldsymbol{x}_k$ 密度连接，且 $\boldsymbol{x}_k$ 一定可以从 $\boldsymbol{x}_l$ 密度连接。

例子

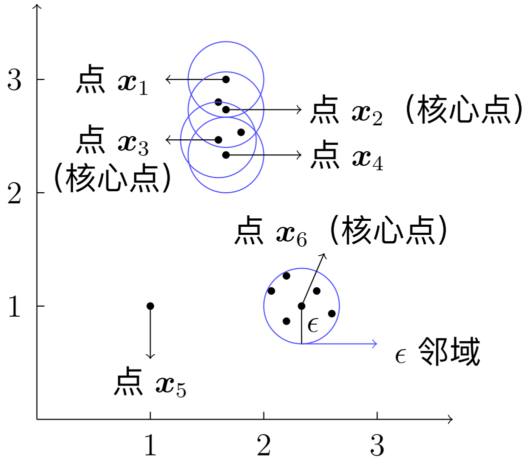


图: 二维平面 13 个点的示意图

例子

- 事先设定 $\epsilon = 0.5$ 且 $\mathrm{MinPts} = 5$；
- 不难发现：
  - 点 $\boldsymbol{x}_2$、点 $\boldsymbol{x}_3$ 和点 $\boldsymbol{x}_6$ 均是核心点，这是因为这些点的 ϵ 领域中点的个数大于等于 MinPts；
  - 点 $\boldsymbol{x}_1$ 可以从 $\boldsymbol{x}_2$ 直接密度可达，这是因为 $\boldsymbol{x}_1$ 在 $\boldsymbol{x}_2$ 的 ϵ 邻域内；
  - 点 $\boldsymbol{x}_4$ 可以由点 $\boldsymbol{x}_2$ 密度可达，这是因为点 $\boldsymbol{x}_3$ 也可以从点 $\boldsymbol{x}_2$ 直接密度可达，且点 $\boldsymbol{x}_4$ 可以从点 $\boldsymbol{x}_3$ 直接密度可达，而且点 $\boldsymbol{x}_2$ 和 $\boldsymbol{x}_3$ 均是核心点；
  - 点 $\boldsymbol{x}_4$ 可以由点 $\boldsymbol{x}_1$ 密度连接，因为我们可以找到点 $\boldsymbol{x}_2$ 使得 $\boldsymbol{x}_1$ 和 $\boldsymbol{x}_4$ 均可以从点 $\boldsymbol{x}_2$ 密度可达。

基本概念

- 称 X 的一个非空子集 C 关于 ϵ 和 MinPts **一个类**，如果集合 C 满足
  - 最大性（Maximality）：对于任意两个点 $\boldsymbol{x}_k$ 和 $\boldsymbol{x}_l$，如果点 $\boldsymbol{x}_k \in C$ 且点 $\boldsymbol{x}_l$ 可以从点 $\boldsymbol{x}_k$ 密度可达，那么 $\boldsymbol{x}_l \in C$；
  - 连接性（Connectivity）：对于任意两个点 $\boldsymbol{x}_k, \boldsymbol{x}_l \in C$，点 $\boldsymbol{x}_l$ 可以从点 $\boldsymbol{x}_k$ 密度连接。
- 假定 $C_1, C_2, \cdots, C_K$ 均是数据集 X 中关于 ϵ 和 MinPts 的类。如果点 $\boldsymbol{x}_i \in X$ 但 $\boldsymbol{x}_i \notin C_k, k = 1, 2, \cdots, K$，那么称点 $\boldsymbol{x}_i$ 为**噪声**。
- 说明：如果 $\boldsymbol{x}_i$ 是一个核心点，不难证明集合 $\mathcal{C} = \{\boldsymbol{x}_l \in X : \boldsymbol{x}_l$ 可以从 $\boldsymbol{x}_k$ 密度可达$\}$ 满足最大性和连接性。

DBSCAN

基本概念

Algorithm 5 DBSCAN 聚类算法

Require: 样本集 $X = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)'$;
邻域半径 ϵ;
可到达的最少样本个数 MinPts;

Ensure: DBSCAN 聚类的结果

```
初始化：类别序号 k = 0;
标记每个样本点为未被访问过 v_i = 0, i = 1, 2, ⋯, n;
for 从数据集 X 中随机选取一个未访问过的样本点 x_i do
    if 样本点 x_l 是未访问过的，即 v_i = 0 then
        标记点 x_i 被访问过，即 v_i = 1;
        计算 x_i 的 ϵ 邻域，即 N_ϵ(i) = {x_l ∈ X : dist(i, l) ≤ ϵ};
        if 样本点 x_i 不是一个核心点，即 |N_ϵ(i)| < MinPts then
            样本点 x_i 是一个噪声;
        else
            k = k + 1;
            令 G(i) 表示所有从样本点 x_i 密度可达的样本点，即 G(i) = {x_l ∈ X : 点 x_l 可以从点 x_i 密度可达};
            for 样本点 x_l ∈ G(i) do
                if 样本点 x_l 是未访问过的，即 v_l = 0 then
                    标记样本点 x_l 为被访问过，即 v_l = 1;
                    将样本点 x_k 归入类 C_k 中;
```

聚类方法的评价

概述

- 聚类有效性是评价聚类结果的方式；
- 聚类有效性的度量方法：
  - 外部聚类有效性；
  - 内部聚类有效性；
- 区别：是否使用外部的信息用来评价聚类的有效性。

外部聚类有效性

概述

对于 n 个测试样本 $\boldsymbol{x}_i (i = 1, 2, \cdots, n)$，

- 假定分类结果为 $\mathcal{C} = \{C_1, C_2, \cdots, C_K\}$ 并满足

$$C_k \cap C_l = \emptyset, \quad \cup_{k=1}^{K} C_i = X$$

- K 为聚类数目；
- 假设“真实的”标签划分 $\mathcal{P} = \{P_1, P_2, \cdots, P_{K'}\}$ 并满足

$$P_k \cap P_l = \emptyset, \quad \cup_{k=1}^{K'} P_i = X$$

- K' 为真实分类数目； 外部信息：真实类情况

外部聚类有效性

概述

- 可能性矩阵（Contingency Matix）定义为

	P_1	P_2	$\cdots$	$P_{K'}$	求和
C_1	n_{11}	n_{12}	$\cdots$	$n_{1K'}$	$n_{1\cdot}$
C_2	n_{21}	n_{22}	$\cdots$	$n_{2K'}$	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
C_K	n_{K1}	n_{K2}	$\cdots$	$n_{KK'}$	$n_{K\cdot}$
求和	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot K'}$	n

- 我们可以计算

$$p_{ij} = \frac{n_{ij}}{n}, \quad p_i = \frac{n_{i\cdot}}{n}, \quad p_j = \frac{n_{\cdot j}}{n}$$

外部聚类有效性

常用指标

- 对 K 均值聚类算法而言，熵和纯度是两种最常用的外部度量。
- 熵 (Entropy, E)

$$E = -\sum_i p_i \left(\sum_j \frac{p_{ij}}{p_i} \ln \frac{p_{ij}}{p_i} \right)$$

- 纯度（purity，P）

$$P = \sum_i p_i (\max_j \frac{p_{ij}}{p_i})$$

内部聚类有效性

概述

- 内部聚类有效性的两个准则:
- 紧密度（Compactness）：在同一类内不同个体之间紧密关联的度量；
  - 方差可以体现数据的紧密度；低方差表明紧密度**好**；
  - 很多紧密度的定义是依赖于距离的，如：最大或平均两两距离，基于中心的最大或平均距离，等。
- 区分度（Separation）：不同类间区别程度的度量；
  - 例如，两个类中心的距离，或从两个不同类中任各选取一个体的最短距离，通常作为区分度的度量；
  - 密度（density）也会用于度量区分度。

内部聚类有效性

常用指标

- 均方标准差

$$RMSSTD = \left(\frac{\sum_{k=1}^{K} \sum_{i \in C_k} \|\boldsymbol{x}_i - \boldsymbol{m}_k\|^2}{p \sum_{k=1}^{K} (n_k - 1)} \right)^{1/2}$$

- R 平方（RS）

$$RS = 1 - \frac{\sum_{k=1}^{K} \sum_{i \in C_k} \|\boldsymbol{x}_i - \boldsymbol{m}_k\|^2}{\sum_{i=1}^{n} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|^2}$$

轮廓法

概述

- 轮廓法（silhouette）也是一种直观的且用于验证聚类结果的方法；
- 基本思想：同类相似，异类不同；
- 在轮廓法中，需要定义一个重要的概念——**轮廓值**；
- **轮廓值**是衡量一个样本相对其他类中的样本而言与本类样本的相似度，取值范围**介于 -1 与 $+1$ 之间**。
- 一般认为，轮廓值较高表示该样本被很好地聚到其所属的类，而不和其他类相似。
- 如果大部分的样本具有较高的轮廓值，那么聚类的结果是恰当的；如果存在许多样本具有较低的轮廓值，甚至是负值，那么聚类的个数可能不合适。

轮廓法

具体方法

- 给定某个聚类结果 $\mathcal{C} = \{\mathcal{C}_k : 1 \le k \le K\}$；
- 对于第 i 个样本，假定其属于第 k 类，n_k 表示第 k 类中的样本量，令 $a(i)$ 表示**第 i 个样本与第 k 类的其他样本的平均距离**，即

$$a(i) = \frac{1}{n_k - 1} \sum_{j \in \mathcal{C}_k, j \neq i} \mathrm{dist}(i, j)$$

- 注意到 $a(i)$ 表示第 i 个样本与其同属一类的样本的平均不相似度，如果该类只有第 i 个样本本身，那么 $a(i)$ 为零。

轮廓法

具体方法

- 对于第 i 个样本和另一个样本量为 $n_{k'}$ 的类 $\mathcal{C}_{k'}$，令 $d(i, \mathcal{C}_{k'})$ 为第 i 个样本与第 k' 个类的所有样本的平均不相似度，即

$$d(i, \mathcal{C}_{k'}) = \frac{1}{n_{k'}} \sum_{j \in \mathcal{C}_{k'}} \mathrm{dist}(i, j).$$

- 令

$$b(i) = \min_{k' \neq k} d(i, \mathcal{C}_{k'})$$

表示**第 i 个样本和不属于同一类的最近距离**。

轮廓法

具体方法

- 对于第 i 个样本，轮廓值 $s(i)$ 定义为

$$s(i) = \frac{b(i) - a(i)}{\max(a(i) - b(i))}.$$

- 对于所有 n 个样本，可以计算平均轮廓值，用于度量聚类数目 K 是否合适。
- 由此，以**最大的平均轮廓值**所对应的 K 作为最优聚类数目，这种方法称为轮廓法。

CH 指数

概述

- CH 指数（Calinski–Harabasz 指数）是与方差分析中 F 检验统计量相似的。
- 假定我们对 n 个样本进行聚类。在给定一个聚类结果 $\mathcal{C} = \{\mathcal{C}_k : 1 \le k \le K\}$ 后，我们分别考虑类间的平方和 $B(K)$ 与类内的平方和 $W(K)$。
- CH 指数定义为

$$\mathrm{CH}(K) = \frac{B(K)/(K-1)}{W(K)/(n-K)}.$$

其中，

- $W(K) = \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k\|^2, \quad \bar{\boldsymbol{x}}_k = n_k^{-1} \sum_{i \in \mathcal{C}_k} \boldsymbol{x}_i$
- $B(K) = \sum_{k=1}^{K} n_k \|\bar{\boldsymbol{x}}_k - \bar{\bar{\boldsymbol{x}}}\|^2, \quad \bar{\bar{\boldsymbol{x}}} = K^{-1} \sum_{k=1}^{K} \bar{\boldsymbol{x}}_k$

CH 指数

概述

- 基于方差分解公式，在给定样本时，$B(K)$ 与 $W(K)$ 的和是一个定值。
- 如果 $B(K)$ 越大，则 $W(K)$ 越小，那么满足对聚类的基本思想：**类内差异小，类间差异大**。
- 可以通过**最大化 CH 指数**可到最优的聚类数目 K。