统计方法与机器学习

第四章：聚类方法

倪 蒨

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

SCHOOL OF DATA
SCIENCE & ENGINEERING
数据科学与工程学院

目录

目录

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。

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：从行来看

- 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**。

基本概念

两个角度：从列来看

- 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**。

基本问题

聚类分析的本质

- 聚类模型或聚类算法本质上就是如何确定一个**划分**。
- 另外，有几个核心的问题：
  - 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?

目录

距离的定义

动机

- 在聚类分析中，如何定义**距离**是尤为重要的。
- 原因如下：
  - 我们将个体或者特征看作空间中的两个点，距离常用于度量两个点的远近程度。
  - 通常，我们认为两个点距离小，则它们相似性高；
  - 反之，它们相似性低。

点间距离

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

表: 常用的点间距离计算方法

常用距离	计算方法
欧式距离	$\|\boldsymbol{x}_k - \boldsymbol{x}_l\|_2 = \sqrt{\sum_{j=1}^{p}(x_{kj} - x_{lj})^2}$
欧式平方距离	$\|\boldsymbol{x}_k - \boldsymbol{x}_l\|_2^2 = \sum_{j=1}^{p}(x_{kj} - x_{lj})^2$
曼哈顿距离（绝对距离）	$\|\boldsymbol{x}_k - \boldsymbol{x}_l\|_1 = \sum_{j=1}^{p}\|x_{kj} - x_{lj}\|$
切比雪夫距离（最大距离）	$\|\boldsymbol{x}_k - \boldsymbol{x}_l\|_\infty = \max_j \|x_{kj} - x_{lj}\|$
闵氏距离	$\left(\sum_{j=1}^{p}(x_{kj} - x_{lj})^q\right)^{\frac{1}{q}}$
兰氏距离	$\sum_{j=1}^{p}\frac{\|x_{kj}-x_{lj}\|}{\|x_{kj}\|+\|x_{lj}\|}$
马氏距离（广义欧式距离）	$\sqrt{(\boldsymbol{x}_k - \boldsymbol{x}_l)'\Sigma^{-1}(\boldsymbol{x}_k - \boldsymbol{x}_l)}$

点间距离

基于相关系数的距离定义

- **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

基于相关系数的距离定义

- **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

基于相关系数的距离定义

- **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

基于相关系数的距离定义

- **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** $= 1 - \text{Spearman } \rho$

目录

层次聚类

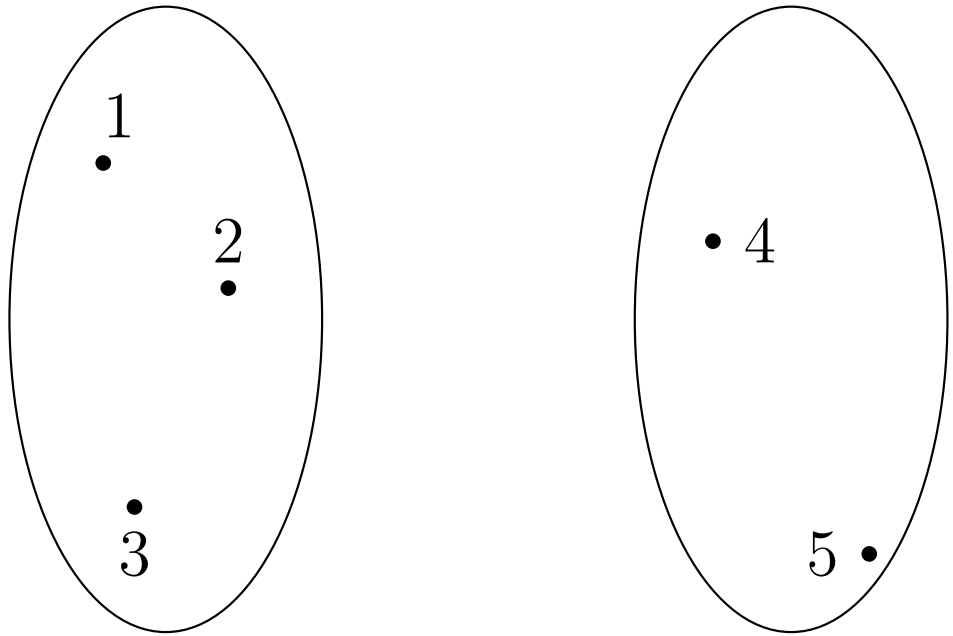
概述

- 层次聚类是一种易于解释的聚类方法；
- 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。
- 问题：如何定义**类间距离**？

类间距离

基本想法

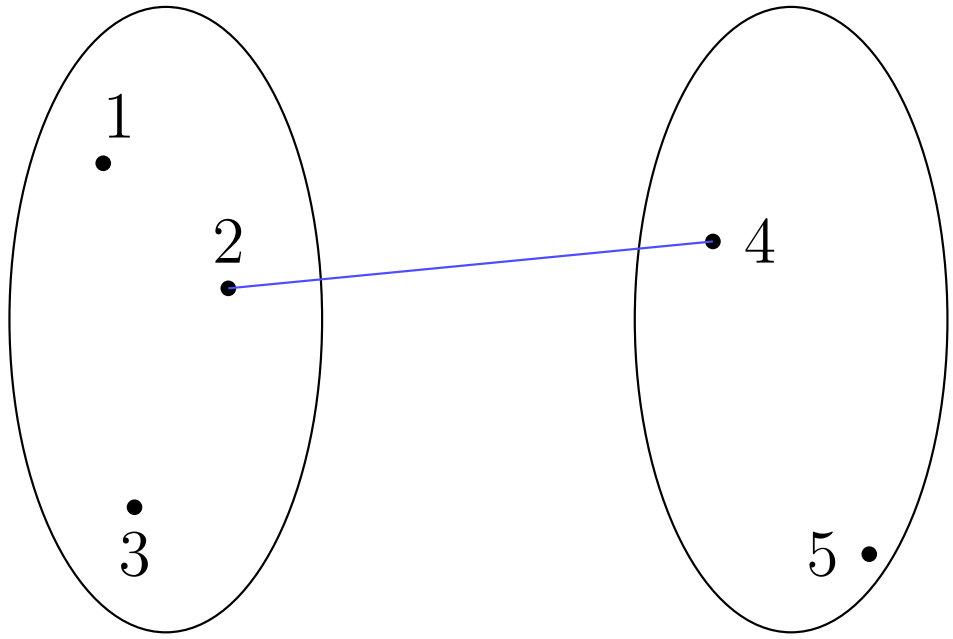
- 如果每个类都只是由一个点组成，那么两个类之间的距离就是这样两个点之间的距离。
- 如果每个类**包含不止一个点**，那么如何定义两个类之间距离就是我们核心关注的问题。
- 在层次聚类中，类间距离的定义方式也称**关联准则**。
- 例如，



类间距离

关联规则：简单连接（Simple Linkage）

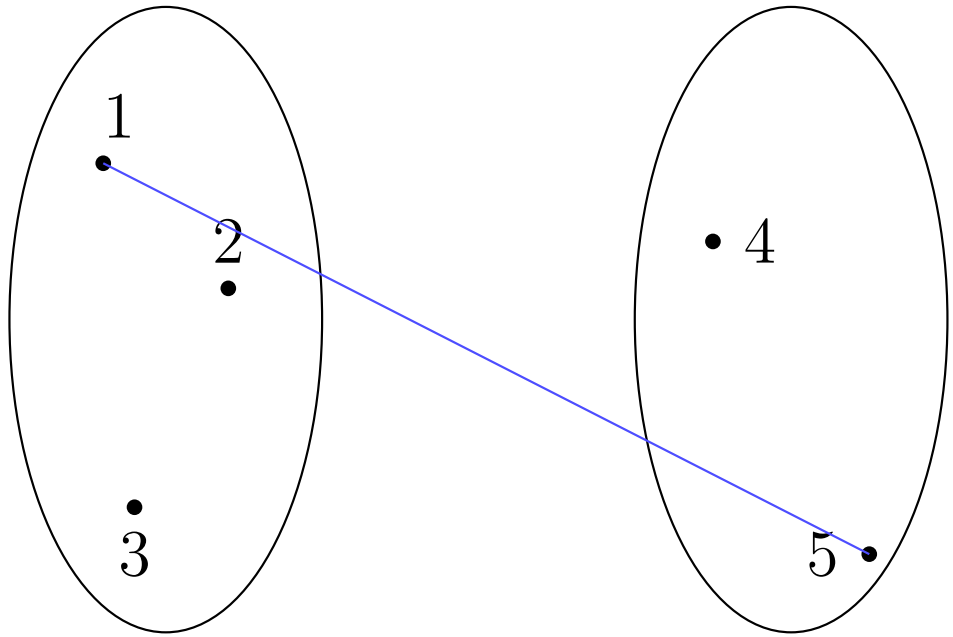
- 将两个类中**距离最短**的两个点之间的距离定义为类间距离。
- 在下图中，左侧类中选取样本点 2，右侧类中选取样本点 4，此时，类间距离为 d_{24}。



类间距离

关联规则：复杂关联（Complete Linkage）

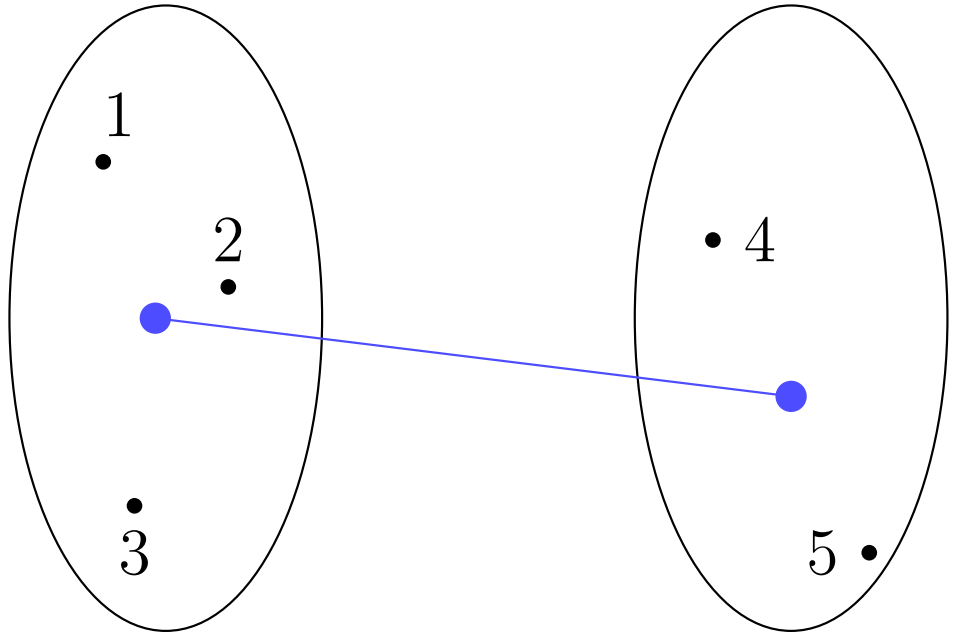
- 将两个类中**距离最长**的两个点之间的距离定义为类间距离。
- 在下图中，左侧类中选取样本点 1，右侧类中选取样本点 5，定义类间距离为 d_{15}。



类间距离

关联规则：质心连接（Centroid Linkage）

- 将两个类中所有点的**重心**的距离定义为类间距离。
- 在下图中，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。



类间距离

关联准则

- 除了上述常见的关联规则，还有其他关联规则。
- 这些关联规则有一个统一个公式——Lance-Williams **公式**，即

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

其中，

- $\alpha_k, \alpha_l, \beta, \gamma$ 为参数。
- $d(k, l)$ 表示类 k 与类 l 之间的距离。
- 类 k 与类 l 聚合成新的一个类，记为 $k \cup l$。

类间距离

关联准则

名称	α_k	α_k	β	γ
简单连接法 (single-linkage)	0.5	0.5	0	-0.5
复杂连接法 (complete-linkage)	0.5	0.5	0	0.5
平均连接法 (average linkage)	$\frac{n_k}{n_k+n_l}$	$\frac{n_l}{n_k+n_l}$	0	0
加权平均连接法 (McQuitty 法)	0.5	0.5	0	0
中位数连接法 (median linkage)	0.5	0.5	-0.25	0
质心连接法 (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 最小方差连接法 (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

算法实例：自下而上

- 假设有 4 个点，距离矩阵为

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

- 找到距离最近的两个类：A 和 B，把他们聚成一类；

层次聚类

算法实例：自下而上

- 采用**简单连接法**来重新计算距离矩阵为

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

- 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**简单连接法**来重新计算距离矩阵为

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

- 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

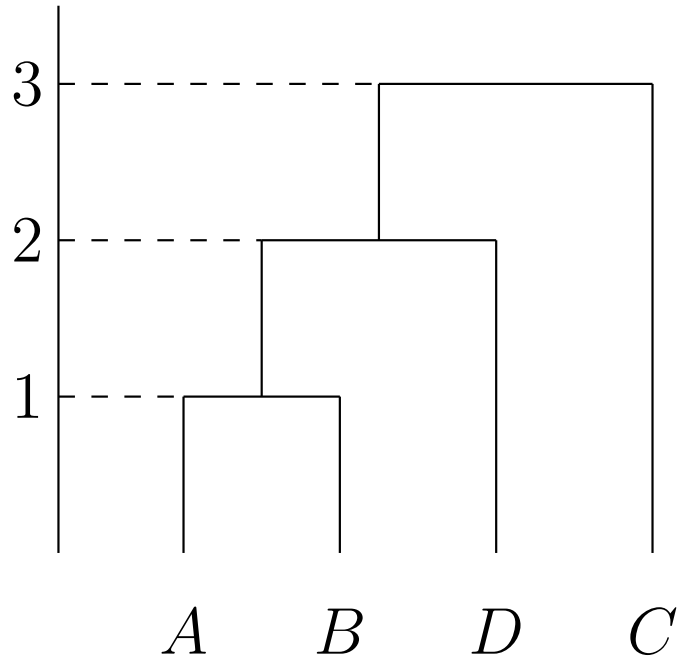


图: 聚合算法的层次聚类结果

K 均值聚类

概述

- K 均值（K-means）聚类的优点在于计算速度快，也称为快速聚类。
- 聚类数目在 K 均值聚类算法中是一个超参数，但需要提前确定；
- 假定聚类数目为 $K(K < n)$；
- 给定 n 个样本集 $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}'$；
- 我们希望，将 n 个样本划分到 K 个不同的类中。也就是说，找到数据集 $\boldsymbol{X}$ 的一种划分 $\mathcal{C} = \{\boldsymbol{C}_1, \boldsymbol{C}_2, \cdots, \boldsymbol{C}_K\}$，即

$$\boldsymbol{C}_k \cap \boldsymbol{C}_l = \emptyset, \quad \cup_{k=1}^{K} \boldsymbol{C}_k = \boldsymbol{X}.$$

K 均值聚类

概述

- K 均值聚类的目标：找到一个**最优**划分 $\mathcal{C}^*$——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

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

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

$$L(\mathcal{C}) = \sum_{k=1}^{K} \sum_{i \in C_k} \|\boldsymbol{x}_i - \boldsymbol{m}_k\|_2^2$$

其中，

- $\boldsymbol{m}_k$ 表示第 k 类中所有样本的均值；
- 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

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

- 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：假定来自一个类的样本均服从同一个正态分布。
- 对于第 k 个正态分布 $N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), k = 1, 2, \cdots, K$，
  - $\boldsymbol{\mu}_k$ 表示均值向量；
  - $\boldsymbol{\Sigma}_k$ 表示协方差矩阵；
- 如果样本 $\boldsymbol{x}_i$ 来自于第 k 类，那么 $\boldsymbol{x}_i$ 的密度函数为

$$f(\boldsymbol{x}_i) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}_k|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right\}, i = 1, 2, \cdots, n.$$

- K 表示聚类数目，与 K 均值聚类算法类似，是一个超参数。
- n 表示样本量。

混合高斯模型

由来

- 如果我们知道第 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

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

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

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

- 上式也是未知参数

$$
\boldsymbol{\theta} = (\pi_1, \cdots, \pi_K, \boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\Sigma}_K)
$$

  的似然函数。

- 理论上，基于（完全）似然函数，我们可以估计参数 $\boldsymbol{\theta}$。

混合高斯模型

由来

- 不幸的是，事实上仅仅能够观测到样本 $\{\boldsymbol{x}_i\}_{i=1}^n$；
- 无法观测到 $\boldsymbol{\delta}_i$，即无法得知真实每个样本的类别标签；
- 因此，我们无法直接估计未知参数 $\boldsymbol{\theta}$。
- 样本 $\boldsymbol{x}_i$ 的（边际）密度函数为

$$
\begin{aligned}
f(\boldsymbol{x}_i) &= \sum_{k=1}^{K} \pi_k (2\pi)^{-p/2} |\boldsymbol{\Sigma}_k|^{-1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right\} \\
&= \sum_{k=1}^{K} \pi_k \cdot f(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)
\end{aligned}
$$

- 注意到，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

由来

- 不幸的是，事实上仅仅能够观测到样本 $\{\boldsymbol{x}_i\}_{i=1}^n$；
- 无法观测到 $\boldsymbol{\delta}_i$，即无法得知真实每个样本的类别标签；
- 因此，我们无法直接估计未知参数 $\boldsymbol{\theta}$。
- 样本 $\boldsymbol{x}_i$ 的（边际）密度函数为

$$
\begin{aligned}
f(\boldsymbol{x}_i) &= \sum_{k=1}^{K} \pi_k (2\pi)^{-p/2} |\boldsymbol{\Sigma}_k|^{-1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right\} \\
&= \sum_{k=1}^{K} \pi_k \cdot f(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)
\end{aligned}
$$

- 注意到，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

- EM 算法的核心在于如何构造潜变量?
- 将变量 $\boldsymbol{\delta}_i = (\delta_{i1}, \cdots, \delta_{iK})'$ 作为潜变量。
- 已知 $\{(\boldsymbol{x}_i, \boldsymbol{\delta}_i)\}_{i=1}^n$，未知参数 $\boldsymbol{\theta}$ 的对数似然函数为

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

混合高斯模型

估计方法：EM 算法

- EM 算法的核心在于如何构造潜变量?
- 将变量 $\boldsymbol{\delta}_i = (\delta_{i1}, \cdots, \delta_{iK})'$ 作为潜变量。
- 已知 $\{(\boldsymbol{x}_i, \boldsymbol{\delta}_i)\}_{i=1}^n$，未知参数 $\boldsymbol{\theta}$ 的对数似然函数为

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

混合高斯模型

估计方法：EM 算法

- E **步：**对 $l(\boldsymbol{\theta})$ 求期望。具体来说，将潜变量 δ_{ik} 的期望 π_{ik}^* 代入 $l(\boldsymbol{\theta})$，即

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

- δ_{ik} 的期望为

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

其中，

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

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 $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$ 有关；
  - $Q_2(\boldsymbol{\theta})$ 仅与未知参数 $\{\pi_k\}_{k=1}^K$ 有关；
- 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 $\boldsymbol{\Sigma}_k$ 求导，并使得导函数为零。

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

- 由此解得

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

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

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

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

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

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

- 而且

$$\sum_{k=1}^{K} \pi_k = 1.$$

- 由此解得

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

混合高斯模型

总结

- 高斯混合模型中的 EM 算法本质上也是迭代算法，每次迭代包括两个步骤：
  - 计算个体的类别概率期望，即

$$\pi_{ik}^* = \frac{\pi_k \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k=1}^K \pi_k \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}.$$

  - 更新参数 $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k\}_{k=1}^K$，即

$$\begin{aligned}
\boldsymbol{\mu}_k &= \frac{\sum_{i=1}^n \pi_{ik}^* \boldsymbol{x}_i}{\sum_{i=1}^n \pi_{ik}^*}, \\
\boldsymbol{\Sigma}_k &= \frac{\sum_{i=1}^n \pi_{ik}^* (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)'}{\sum_{i=1}^n \pi_{ik}^*}, \\
\pi_k &= n^{-1} \sum^n \pi_{ik}^*.
\end{aligned}$$

DBSCAN

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 $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}', \boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})'$；
- 每个样本 $\boldsymbol{x}_i$ 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 $d(k, l)$。

DBSCAN

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 $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}', \boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})'$；
- 每个样本 $\boldsymbol{x}_i$ 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 $d(k, l)$。

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

$$N_\epsilon(k) = \{\boldsymbol{x}_l \in \boldsymbol{X} | d(k,l) \leq \epsilon\},$$

那么，称点 $\boldsymbol{x}_k$ 的**ϵ 邻域**；

- 如果

$$|N_\epsilon(k)| \geq \text{MinPts},$$

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

$$\boldsymbol{x}_l \in N_\epsilon(k) \quad \text{且} \quad |N_\epsilon(k)| \geq \text{MinPts},$$

  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（$\epsilon = 0.5$，$\text{MinPts} = 5$）

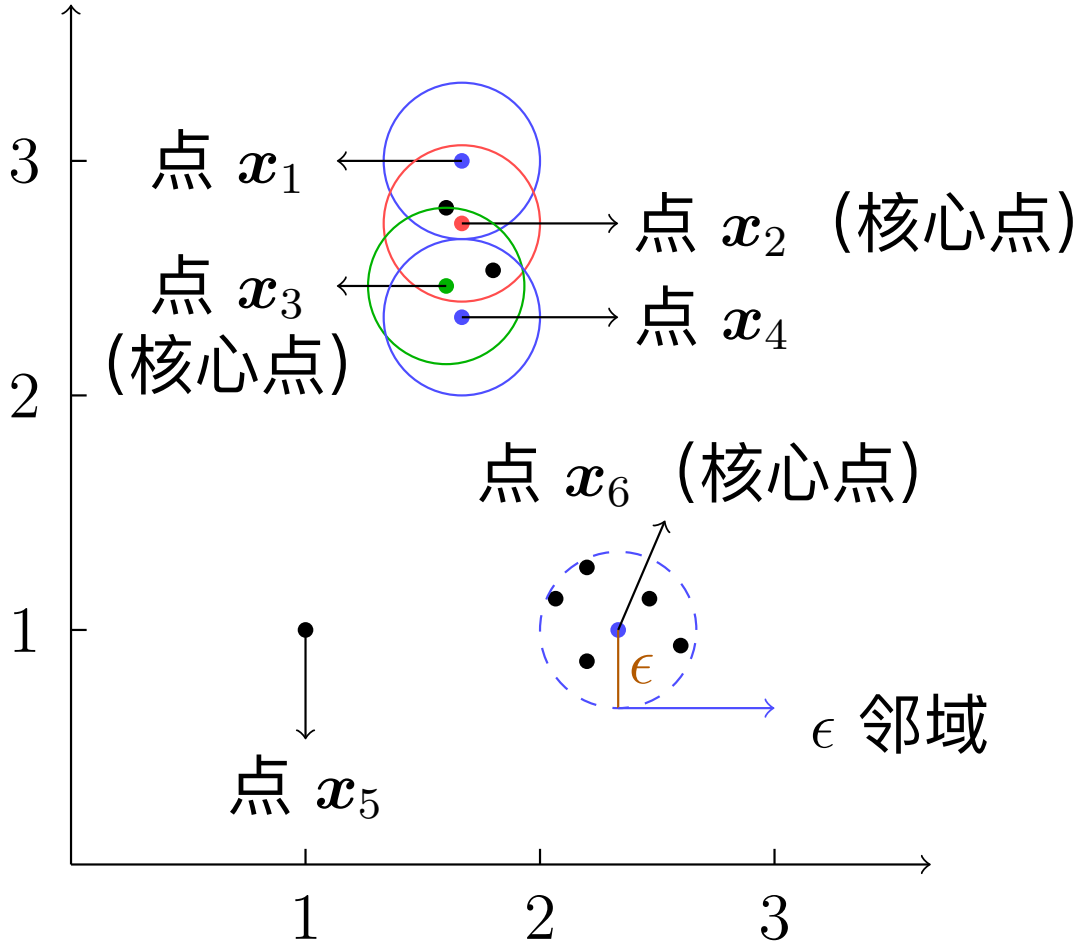


图: 二维平面 13 个点的示意图

例子（$\epsilon = 0.5$，$\mathrm{MinPts} = 5$）

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。

DBSCAN

基本概念

- 称 $\boldsymbol{X}$ 的一个非空子集 C 是关于 ϵ 和 MinPts 的一个**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 $C_1, \cdots, C_K$ 均是 $\boldsymbol{X}$ 中关于 ϵ 和 MinPts 的类。
- 如果点 $\boldsymbol{x}_i \in \boldsymbol{X}$ 但 $\boldsymbol{x}_i \notin C_k, k = 1, 2, \cdots, K$，那么称点 $\boldsymbol{x}_i$ 为**噪声**。
- 如果 $\boldsymbol{x}_i$ 是一个核心点，不难证明集合 $\mathcal{C} = \{\boldsymbol{x}_l \in \boldsymbol{X} : \boldsymbol{x}_l$ 可以从 $\boldsymbol{x}_i$ 密度可达$\}$ 满足最大性和连接性。

DBSCAN

Algorithm 5 DBSCAN 聚类算法

Require: 样本集 $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$;
邻域半径 ϵ;
可到达的最少样本个数 MinPts;
Ensure: DBSCAN 聚类的结果 $\mathcal{C} = \{\boldsymbol{C}_1, \boldsymbol{C}_2, \cdots, \boldsymbol{C}_K\}$

初始化核心点集合 $\boldsymbol{I} = \varnothing$;
for $i = 1, 2, \cdots, n$ **do**
    确定样本 $\boldsymbol{x}_i$ 的 ϵ 邻域 $N_\epsilon(i)$;
    if $|N_\epsilon(i)| \geq$ MinPts **then**
        更新核心点集合 $\boldsymbol{I} = \boldsymbol{I} \cup \{\boldsymbol{x}_i\}$;
初始化聚类数目 $k = 0$;
初始化未被访问的样本集合 $\boldsymbol{P} = \boldsymbol{X}$;
while $\boldsymbol{P} \neq \varnothing$ **do**
    保留当前未被访问的样本集合 $\tilde{\boldsymbol{P}} = \boldsymbol{P}$;
    随机选取一个核心点 $\boldsymbol{x}_j \in \boldsymbol{I}$;
    初始化队列 $\mathcal{Q} =< \boldsymbol{x}_j >$;
    将核心点集合中剔除 $\boldsymbol{x}_j$，即 $\boldsymbol{I} = \boldsymbol{I} - \{\boldsymbol{x}_j\}$;
    while $\mathcal{Q} \neq \varnothing$ **do**
        取出队列 $\mathcal{Q}$ 中的首个样本 $\boldsymbol{x}_q$;
        if $|N_\epsilon(q)| \geq$ MinPts **then**
            令 $\boldsymbol{R} = N_\epsilon(q) \cap \boldsymbol{P}$;
            将 $\boldsymbol{R}$ 的样本加入队列 $\mathcal{Q}$;
            $\boldsymbol{P} = \boldsymbol{P} - \boldsymbol{R}$;
    $k = k + 1$
    生成新的一个类 $\boldsymbol{C}_k = \tilde{\boldsymbol{P}} - \boldsymbol{P}$;
    更新核心点的集合 $\boldsymbol{I} = \boldsymbol{I} - \boldsymbol{C}_k$;

目录

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 $\mathcal{C}=\{\boldsymbol{C}_1,\boldsymbol{C}_2,\cdots,\boldsymbol{C}_K\}$ 并满足

$$\boldsymbol{C}_k\cap\boldsymbol{C}_l=\emptyset,\quad \cup_{k=1}^{K}\boldsymbol{C}_i=\boldsymbol{X}$$

- K 为聚类数目;
- 假设“真实的”标签划分 $\mathcal{Y}=\{\boldsymbol{Y}_1,\boldsymbol{Y}_2,\cdots,\boldsymbol{Y}_{K'}\}$ 并满足

$$\boldsymbol{Y}_k\cap\boldsymbol{Y}_l=\emptyset,\quad \cup_{k=1}^{K'}\boldsymbol{Y}_i=\boldsymbol{X}$$

- K' 为真实分类数目;

外部聚类有效性

概述

- 可能性矩阵（Contingency Matix）定义为

	$\boldsymbol{Y}_1$	$\boldsymbol{Y}_2$	$\cdots$	$\boldsymbol{Y}_{K'}$	求和
$\boldsymbol{C}_1$	n_{11}	n_{12}	$\cdots$	$n_{1K'}$	$n_{1.}$
$\boldsymbol{C}_2$	n_{21}	n_{22}	$\cdots$	$n_{2K'}$	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
$\boldsymbol{C}_K$	n_{K1}	n_{K2}	$\cdots$	$n_{KK'}$	$n_{K.}$
求和	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.K'}$	n

- 我们可以计算

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

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相似程度的度量；
  - 方差可以体现数据的紧密度；低方差表明数据差异小，则紧密度**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）：不同类间差异程度的度量；
  - 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

- 均方标准差为

$$\left(\frac{\sum_{k=1}^{K} \sum_{i \in C_k} \|\boldsymbol{x}_i - \boldsymbol{m}_k\|_2^2}{p \sum_{k=1}^{K} (n_k - 1)} \right)^{1/2}$$

- R^2 为

$$1 - \frac{\sum_{k=1}^{K} \sum_{i \in C_k} \|\boldsymbol{x}_i - \boldsymbol{m}_k\|_2^2}{\sum_{i=1}^{n} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|_2^2}$$