

模式识别

Pattern Recognition

第7讲 聚类分析

本章主要内容：

- ▣ 什么是聚类？
- ▣ 如何度量样本间的“相似性”。
- ▣ 如何构建聚类准则函数。
- ▣ 基本的聚类方法。

1 无监督学习与聚类

□有监督学习

- 确切地知道每一个训练样本所属的类别；
- 利用训练样本学习分类器，对未知类别样本分类。
- 关键问题：如何度量样本间的“相似性”。

□无监督学习

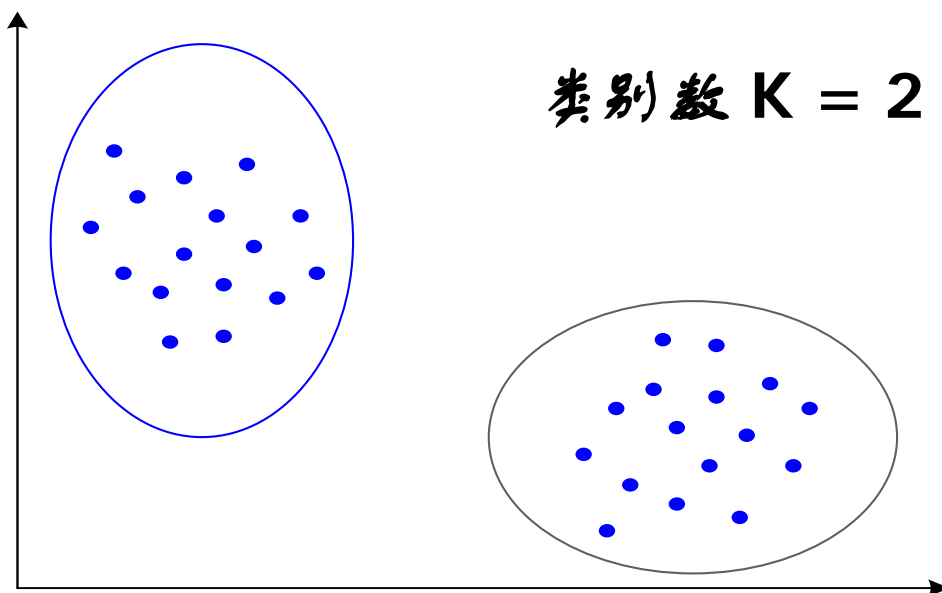
- 无监督样本集：已知训练样本集合中的样本，不能确切知道每个样本所属的类别
- 利用无监督样本集，希望能够学习出某种规律性的东西，构造相应的分类器。

?

1 无监督学习与聚类

□ 聚类 (Clustering)

- 根据某种原则将样本或者数据集合划分为若干个“有意义”的子集（聚类）
- 同一子集内的样本之间具有较大的“相似性”
- 不同子集样本之间具有较大的“差异性”。



1.1为什么要进行无监督学习？

□理论意义——科学归纳

- 知识都是实践过程中通过经验的积累、整理和对事物规律的发现所得到的
- 对各种纷繁复杂的事物进行分类，是认识世界的重要科学方法

□实践意义——揭示观测数据的内部结构和规律

- 对海量数据进行分析
 - 网络的普及和发展，获取大量训练样本变得容易，逐一对这些样本进行标注，非常耗时耗力。需要利用聚类对这些数据进行概括和解释
- 描述样本集合的有关结构信息，帮助监督学习更好设计分类器

1.2 聚类分析的应用

□ 信息检索

- 检索结果聚类

□ 商业应用：

- 用户的属性聚类，商品属性聚类

□ 图像分割

- 图像中所有像素点的聚类过程

□ 数据压缩

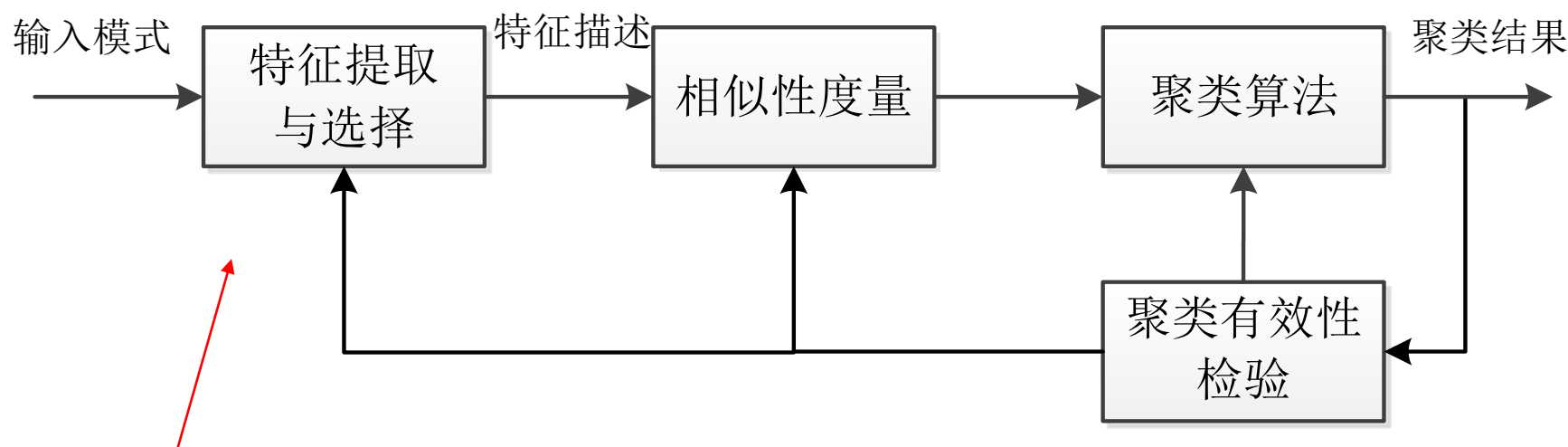
- 通过聚类找到相似数据的“中心”

□ 医学应用：

- 药物反应、医疗数据挖掘

1.3 聚类分析的过程

1) 聚类分析过程



同样的一组对象采用不同的特征聚类，结果可能是完全不同

1.3 聚类分析的过程

2) 聚类和分类有什么不同？

聚类（Clustering）	分类（Classification）
给出/构造相似性测度 （Similarity measure）	给出标签（Labels）
通过相似性推测标签	通过标签推测相似性
只讨论如何对当前样本集合中样本进行分类	对样本集之外的其它样本进行分类

发现知识

学习知识，解决问题

1.3 聚类分析的过程

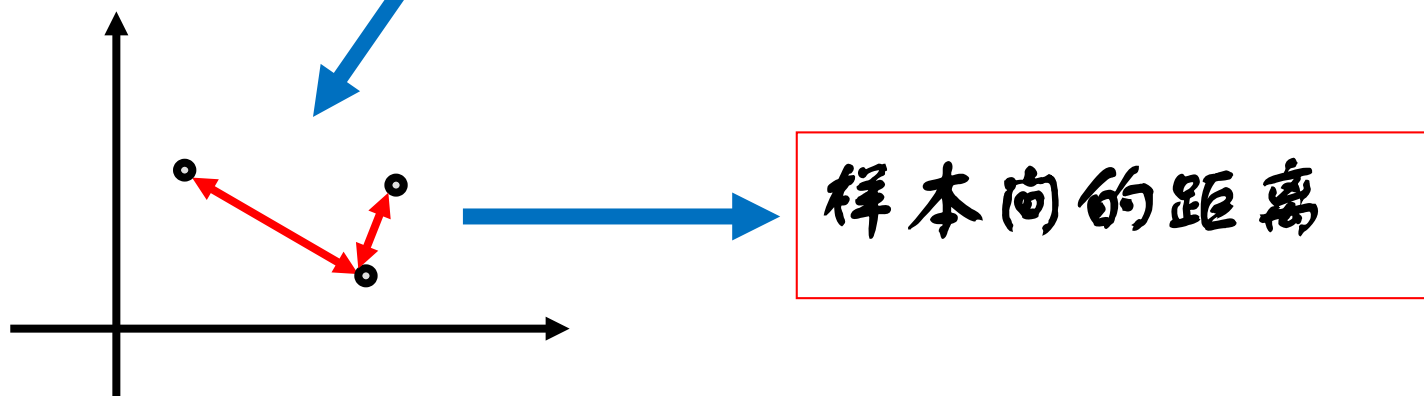
3) 聚类中的两个关键问题

➤ 怎样度量样本间的“相似性”？

如何定义相似性测度 (Similarity measure)

➤ 怎样才算好的聚类？

如何构建聚类的准则函数，用于评价聚类效果。



1.4 聚类问题的描述

聚类的数学描述

无监督样本集合 $D = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ 中包含 k 个聚类,

聚类数 k 可能是先验已知的, 也可能需要在聚类过程中确定,

k 个聚类 C_1, \dots, C_k 需要满足如下条件:

1. $C_i \neq \emptyset, \quad i = 1, \dots, k;$

2. $\bigcup_{i=1}^k C_i = D;$

3. $C_i \cap C_j = \emptyset, \quad i \neq j, \quad i, j = 1, \dots, k。$

每个聚类至少包含一个样本

任何一个样本属于且只属于一个聚类。

聚类结果是对集合 D 的一个划分

怎样才算一个好的划分?

——聚类准则

2 简单聚类方法

1) 顺序聚类——1967年Hall发表于《Nature》

- 初始化：第一个样本 \mathbf{x}_1 作为第一个聚类，

$$C_1 = \{\mathbf{x}_1\}, l = 1;$$

- 顺序输入每个训练样本 \mathbf{x}_i :

- 计算 \mathbf{x}_i 距离最近的类别 C_k :

$$d(\mathbf{x}_i, C_k) = \min_{1 \leq j \leq l} d(\mathbf{x}_i, C_j);$$

- 如果: $d(\mathbf{x}_i, C_k) > \theta$ 并且 $l < M$,

$$\text{则 } l = l + 1, \quad C_l = \{\mathbf{x}_i\};$$

- 否则: $C_k = C_k \cup \{\mathbf{x}_i\}$

- 输出: 聚类 $\{C_1, \dots, C_l\}$, 聚类数 l

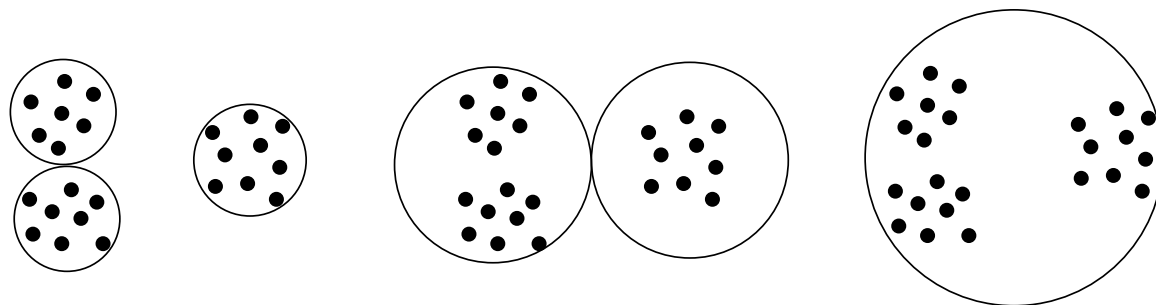
1) 顺序聚类

□ 优点：计算简单。只需计算不超过 $k \times n$ 个样本间的距离（ k 为类别数）

□ 缺点：

➤ 对初始的第一个聚类中心的选择依赖性比较强。

➤ 聚类效果还要受到阈值的影响。



□ 实际问题中，需要对不同的初始聚类中心和不同的阈值进行试探，直到得到一个满意的聚类结果为止。

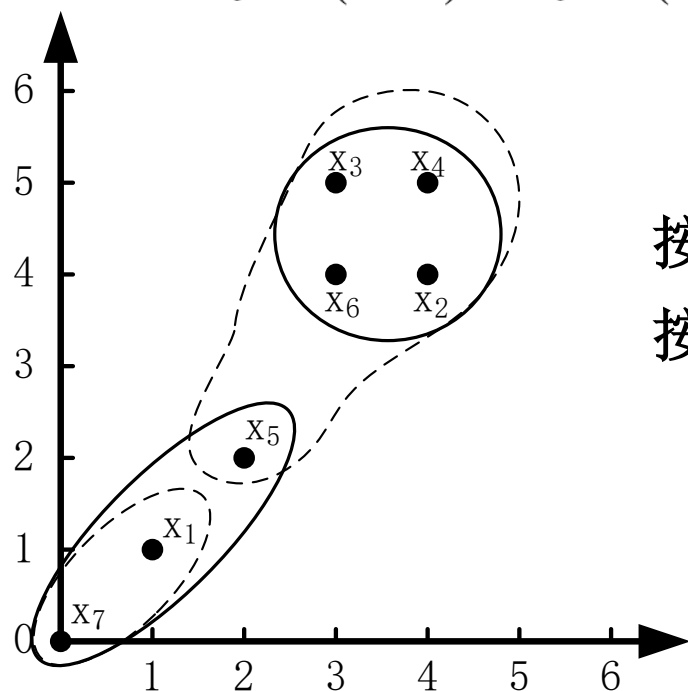
例：由 7 个样本组成的样本集合，顺序聚类

阈值 $\theta = 2$ ，采用欧氏距离度量样本之间的相似度

中心距离度量样本与聚类之间的相似度。

$$\mathbf{x}_1 = (1, 1)^t, \mathbf{x}_2 = (4, 5)^t, \mathbf{x}_3 = (3, 5)^t, \mathbf{x}_4 = (4, 4)^t$$

$$\mathbf{x}_5 = (2, 2)^t, \mathbf{x}_6 = (3, 4)^t, \mathbf{x}_7 = (0, 0)^t$$



按照 $\mathbf{x}_1 \rightarrow \mathbf{x}_7$ 的顺序输入样本得到实线效果
按照 $\mathbf{x}_7 \rightarrow \mathbf{x}_1$ 的顺序输入样本得到虚线效果

2 简单聚类方法

2) 最大最小距离算法

▣基本思想：在样本集中以最大距离原则选取新的聚类中心，以最小距离原则进行模式归类。

▣步骤：

- 1) 根据最小最大距离产生聚类中心
- 2) 根据聚类中心对样本分类

■ 确定聚类数量和聚类中心:

- 初始化: \mathbf{x}_1 作为第一个聚类中心, $\mathbf{m}_1 = \mathbf{x}_1$;
- 确定第二个聚类中心: 寻找距离 \mathbf{m}_1 最远样本:

$$i_{\max} = \arg \max_{1 \leq i \leq n} d(\mathbf{x}_i, \mathbf{m}_1), \quad \mathbf{m}_2 = \mathbf{x}_{i_{\max}}, \quad l = 2;$$

- 循环, 直到没有新的聚类中心产生为止:
 - 计算样本 \mathbf{x}_i 到当前 l 个中心的距离, 寻找最小:

$$d_i = \min_{1 \leq k \leq l} d(\mathbf{x}_i, \mathbf{m}_k)$$

- 寻找所有样本最小距离中的最大距离:

$$d_{\max} = \max_{1 \leq i \leq n} d_i, \quad i_{\max} = \arg \max_{1 \leq i \leq n} d_i$$

- 如果 $d_{\max} > \theta \|\mathbf{m}_1 - \mathbf{m}_2\|$, 则产生新的聚类中心,

$$\mathbf{m}_{l+1} = \mathbf{x}_{i_{\max}}, \quad l = l + 1;$$

■ 分类训练样本：

□ 初始化各个聚类： $C_k = \phi$ ， $1 \leq k \leq l$

□ 顺序输入每个训练样本 \mathbf{x}_i ：

● 计算 \mathbf{x}_i 距离最近的聚类： $k = \arg \min_{1 \leq t \leq l} d(\mathbf{x}_i, \mathbf{m}_t)$

● 分类 \mathbf{x}_i ： $C_k = C_k \cup \{\mathbf{x}_i\}$

■ 输出：聚类 $\{C_1, \dots, C_l\}$ ，聚类数 l 。

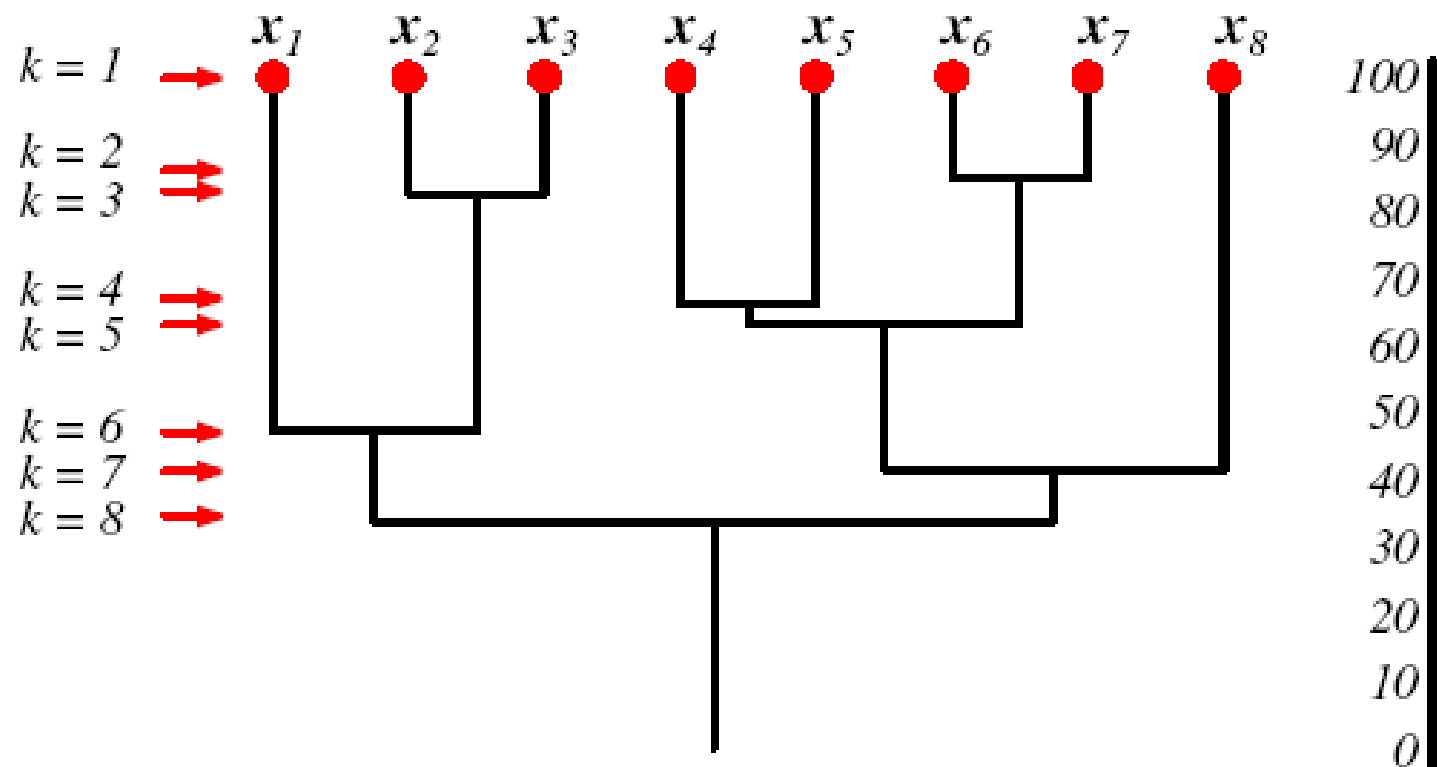
最大最小距离算法的结果只与第一个聚类中心 \mathbf{m}_1

以及阈值比例系数 θ 有关

计算距离次数： $k(k-1)/2 \times n + k \times n$

3 谱系聚类——基本思路

- 首先每一个样本自成一类，然后按照距离准则逐步合并，类别数由多到少，达到合适的类别数为止。
- 不仅是要产生出样本的不同聚类，而且要生成一个完整的样本层次分类谱系
- 已知：N个待识模式 $\{x_1, x_2, \dots, x_N\}$ ，类别数c。



- 第一步 建立N个初始类别，每个样本一个类别，计算距离矩阵 $D=(D_{ij})$ ；
- 第二步 寻找D中的最小元素，合并相应的两个类别，建立新的分类，重新计算距离矩阵D；
- 重复第二步，直到类别数为c为止。

类与类之间相似性度量

- 最短距离: $D_{ij} = \min \left(d \left(\mathbf{x}_l^{(i)}, \mathbf{x}_k^{(j)} \right) \right)$

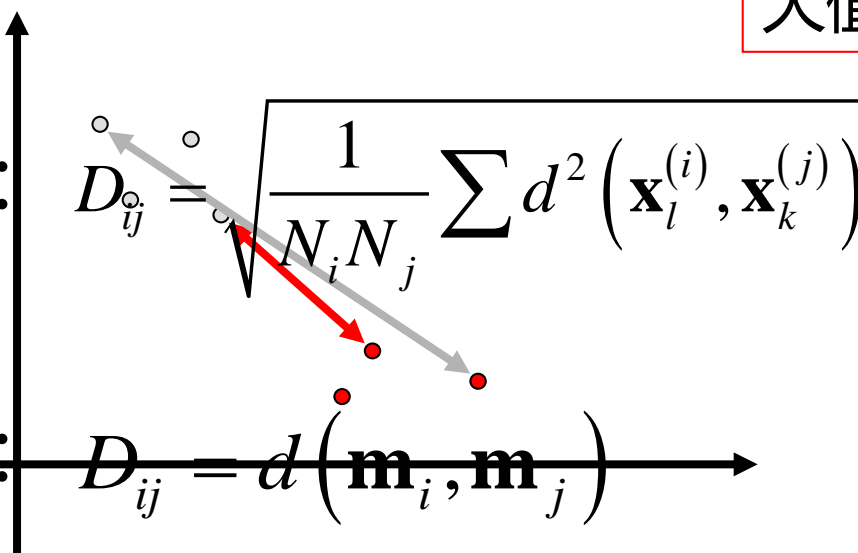
为第*i*类中所有样本与第*j*类中所有样本之间距离的最小值

- 最长距离: $D_{ij} = \max \left(d \left(\mathbf{x}_l^{(i)}, \mathbf{x}_k^{(j)} \right) \right)$

为第*i*类中所有样本与第*j*类中所有样本之间距离的最大值

- 平均距离: $D_{ij} = \sqrt{\frac{1}{N_i N_j} \sum d^2 \left(\mathbf{x}_l^{(i)}, \mathbf{x}_k^{(j)} \right)}$

- 平均样本: $D_{ij} = d \left(\mathbf{m}_i, \mathbf{m}_j \right)$



谱系聚类的特点

- ▣ 层次聚类不用初始化聚类中心，因此聚类结果不受初始聚类中心的影响；
- ▣ 需要定义类别之间的相似性度量；
- ▣ 当样本数较多时，算法的计算量大。
- ▣ 聚类结果是对平方误差准则函数的贪心优化结果。

3谱系聚类——实现

▣距离计算总次数

$$\sum_{k=1}^n \binom{n-k+1}{2} = \sum_{k=1}^n \frac{(n-k+1)(n-k)}{2} = \frac{n^3}{6} - \frac{n}{6}$$

k 轮合并之前需要计算 $n-k+1$ 个聚类之间的距离

▣计算量缩减:

- 每一轮只是将两个聚类的样本合并，生成一个新的聚类，只需要重新计算与新生成聚类有关的距离；
- 新生成聚类与原有聚类之间的距离可以由被合并的两个聚类与其它聚类之间的距离进行推算。

4 K-均值聚类

□历史

- 最早想法由Hugo Steinhaus于1957年提出
- Stuart Lloyd
 - 1957年在Bell实验室给出标准K-均值算法,
 - 1982年发表于IEEE Transactions on Information Theory。
- “K-Means”名称出现在1967年

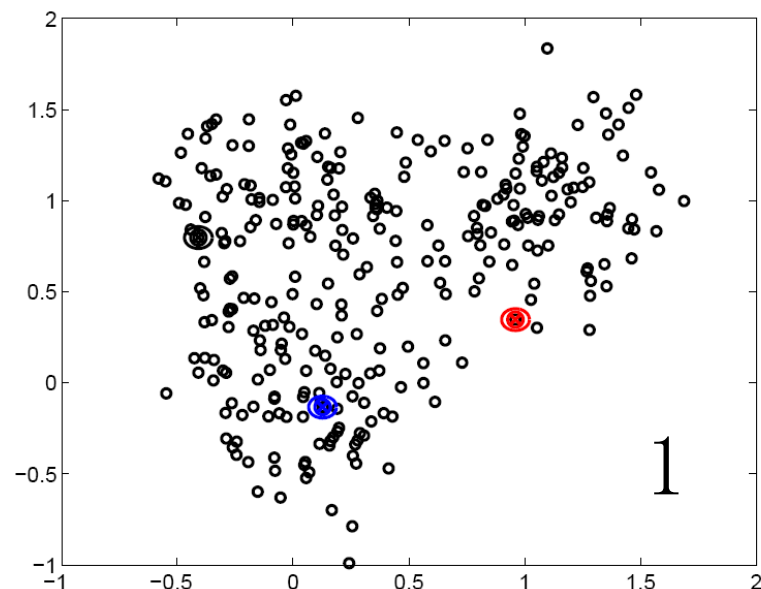
□优点:

- 算法实现简单, 计算复杂度和存储复杂度低
- 对很多简单的聚类问题可以得到令人满意的结果

最著名和最常用的样本聚类算法之一!

4.1 K-均值基本算法

- 第一步：任选K个初始聚类中心；
- 第二步：将每一个待分类样本分到K个类别中。
- 第三步：计算各簇的聚类中心；
- 第四步：检验新的聚类中心与旧的聚类中心是否相等，相等则算法结束；否则转第二步；



例: 现有6个样本 $\mathbf{x}_1 = (1, 1)^t$, $\mathbf{x}_2 = (2, 3)^t$, $\mathbf{x}_3 = (3, 2)^t$, $\mathbf{x}_4 = (9, 8)^t$, $\mathbf{x}_5 = (10, 13)^t$, $\mathbf{x}_6 = (11, 9)^t$, 距离度量采用街市距离, 请用 K 均值算法聚为2类别, 初始聚类中心设为: $\mathbf{m}_1 = (1, 1)^t$, $\mathbf{m}_2 = (2, 3)^t$.

- 1) 计算每个样本到两个聚类中心的距离
- 2) 根据距离聚类中心的远近, 聚为两类
- 3) 重新计算每个聚类的中心

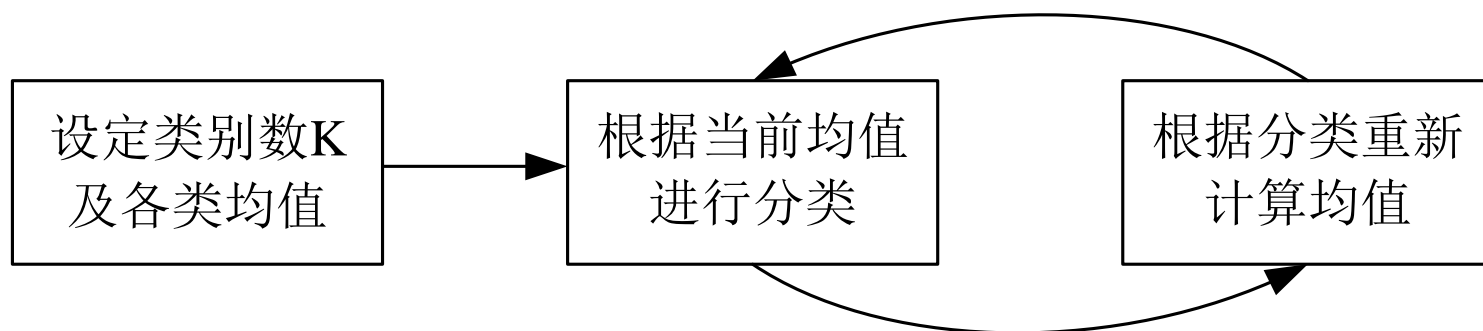
循环1-3步骤, 直至两次聚类结果相同, 聚类中心没有变化

4.1 K-均值基本算法

目标：将n个样本依据最小化类内距离准则分到 K个类别

$$\min_{C_1, \dots, C_K} J_W(C_1, \dots, C_K) = \frac{1}{n} \sum_{j=1}^K \sum_{\mathbf{x} \in C_j} \|\mathbf{x} - \mathbf{m}_j\|^2, \quad \mathbf{m}_j = \frac{1}{n_j} \sum_{\mathbf{x} \in C_j} \mathbf{x}$$

直接优化存在困难，循环迭代优化：



问题：算法收敛吗？收敛到最优解吗？

□ 收敛到局部最优解

对平方误差准则函数的贪心搜索算法

4.2 K-均值算法的改进

□ 初始值的选择

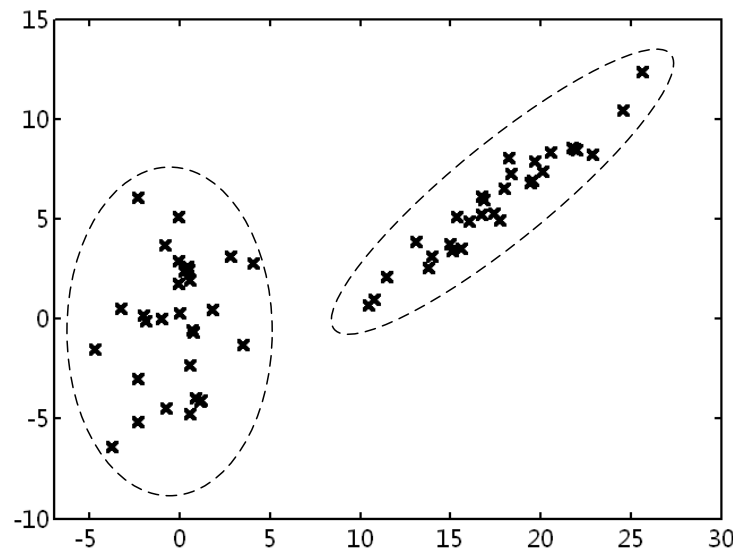
- 利用先验知识选择初值
- 找到相对距离远的样本作为中心

□ 聚类数的选择

- 试探的方式确定聚类数，然后进行聚类有效性检验

□ 距离函数的选择

- 欧氏距离：要求每个聚类的样本成“团型”分布
- 不成团分布——马氏距离



$$d(\mathbf{x}_i, C_j) = (\mathbf{x}_i - \mathbf{m}_j)^t \boldsymbol{\Sigma}_j^{-1} (\mathbf{x}_i - \mathbf{m}_j)$$

4.2 K-均值算法的改进

□ K-Medoids算法(对于序列、图等非矢量特征)

- 形式定义两个模式间相似程度
- 各聚类中，寻找与类中样本相似度之和最大者，代表新模式。

□ 模糊K-均值（类别间可以存在交叠）

- 用隶属度 u_{ij} 表示样本 \mathbf{x}_i 属于 C_j 类的程度

$$\sum_{j=1}^K u_{ij} = 1, \quad 0 \leq u_{ij} \leq 1$$

$b > 1$: 可调参数
控制不同聚类混合程度

- 聚类准则函数:

$$J_{WF}(\mathbf{m}_1, \dots, \mathbf{m}_K, u_{11}, \dots, u_{nK}) = \frac{1}{n} \sum_{j=1}^K \sum_{i=1}^n u_{ij}^b \|\mathbf{x}_i - \mathbf{m}_j\|^2$$

模糊K均值

- 当聚类的均值 $\mathbf{m}_1, \dots, \mathbf{m}_K$ 固定时，隶属度的最优解为：

$$u_{ij} = \frac{\left(1/\|\mathbf{x}_i - \mathbf{m}_j\|^2\right)^{1/(b-1)}}{\sum_{k=1}^K \left(1/\|\mathbf{x}_i - \mathbf{m}_k\|^2\right)^{1/(b-1)}}, \quad i = 1, \dots, n, \quad j = 1, \dots, K$$

- 当隶属度 u_{11}, \dots, u_{nK} 固定时，均值的最优解为：

$$\mathbf{m}_j = \frac{\sum_{i=1}^n u_{ij}^b \mathbf{x}_i}{\sum_{i=1}^n u_{ij}^b}, \quad j = 1, \dots, K$$

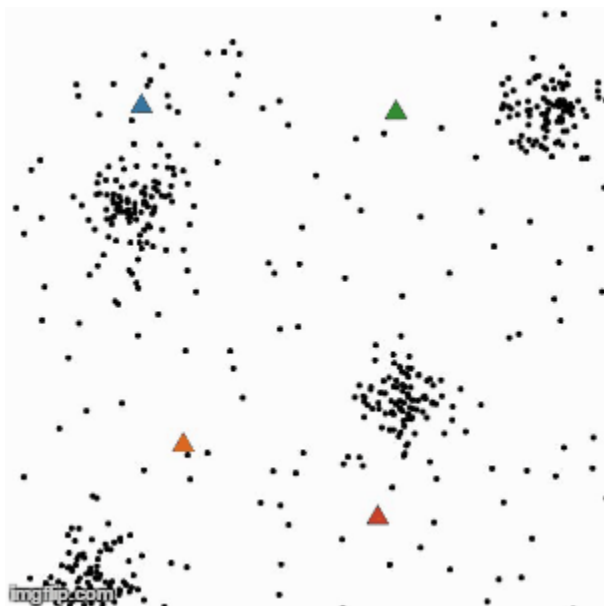
-
- 初始化：随机选择 K 个聚类均值 \mathbf{m}_j , $j = 1, \dots, K$;

- 循环，直到两次迭代的隶属度变化很小为止：

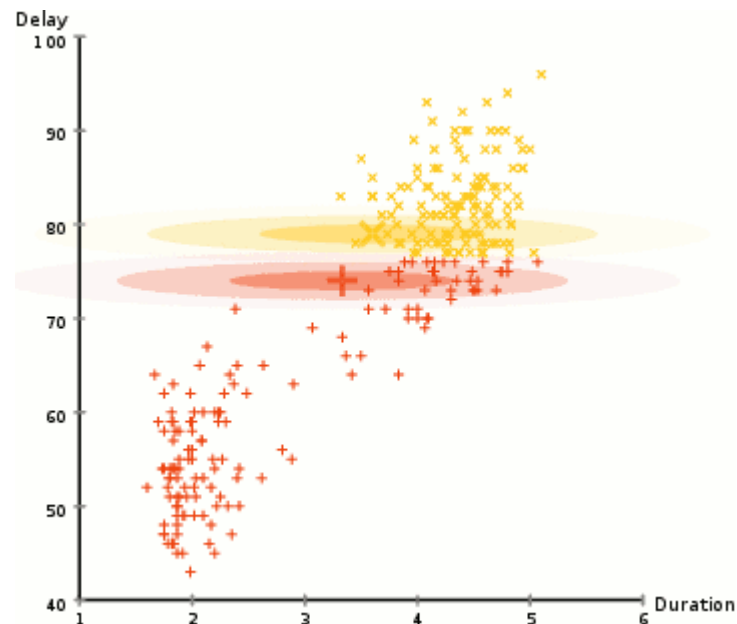
- 计算每个样本对于每个聚类的隶属度 u_{ij}

- 更新每个聚类的均值 \mathbf{m}_j ,

- 输出：样本集的隶属度 $\{u_{ij}\}_{i=1, \dots, n, j=1, \dots, K}$



K-均值



混合高斯

5.1 聚类结果的检验

1) Dunn指数

- 聚类间距：聚类之间最近的一对样本的距离

$$d(C_i, C_j) = \min_{\mathbf{x} \in C_i, \mathbf{y} \in C_j} d(\mathbf{x}, \mathbf{y})$$

- 聚类直径：类内距离最远的两个样本之间的距离

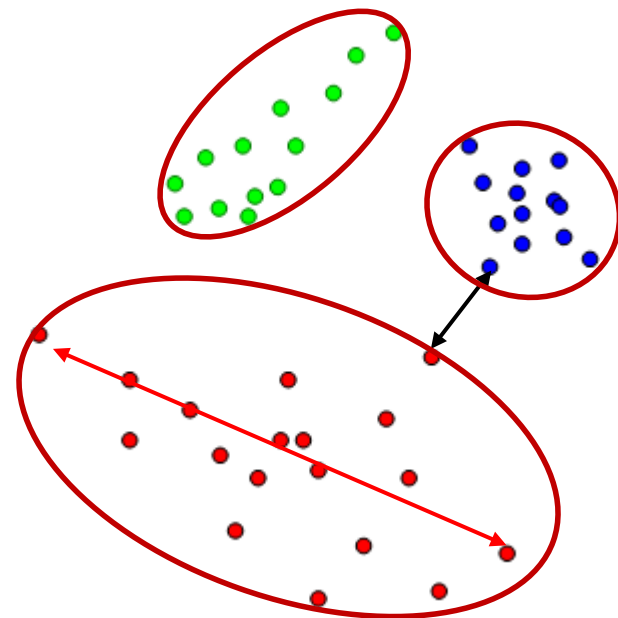
$$\text{diam}(C_i) = \max_{\mathbf{x}, \mathbf{y} \in C_i} d(\mathbf{x}, \mathbf{y})$$

- **Dunn指数**：“最小聚类间距”除以“最大聚类直径”

$$J_{Dunn}(C_1, \dots, C_K) = \frac{\min_{i, j=1, \dots, K, j \neq i} d(C_i, C_j)}{\max_{k=1, \dots, K} \text{diam}(C_k)}$$

仅考虑最近距离
能否进一步考虑样
本离散程度？

指数越大
聚类结果越好



5.1 聚类结果的检验

2) Davies-Bouldin指数

- 类间离散度：聚类均值之间的均方距离

$$d_{ij} = \|\mathbf{m}_i - \mathbf{m}_j\|$$

- 类内离散度：样本到聚类均值的均方距离度量

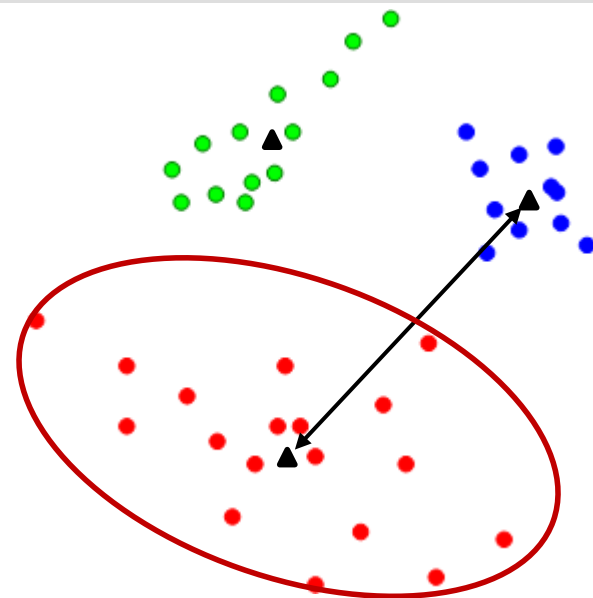
$$s_i = \sqrt{\frac{1}{n_i} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \mathbf{m}_i\|^2}$$

- 类间相似度：类内离散度之和/类间离散度

$$R_{ij} = (s_i + s_j) / d_{ij}$$

- **Davies-Bouldin指数**: 计算各类的最大相似度, 然后求均值

$$J_{DB}(C_1, \dots, C_K) = \frac{1}{K} \sum_{i=1}^K \left(\max_{j=1, \dots, K, j \neq i} R_{ij} \right)$$



如果不限制类别数量
类别越多，指数越大？

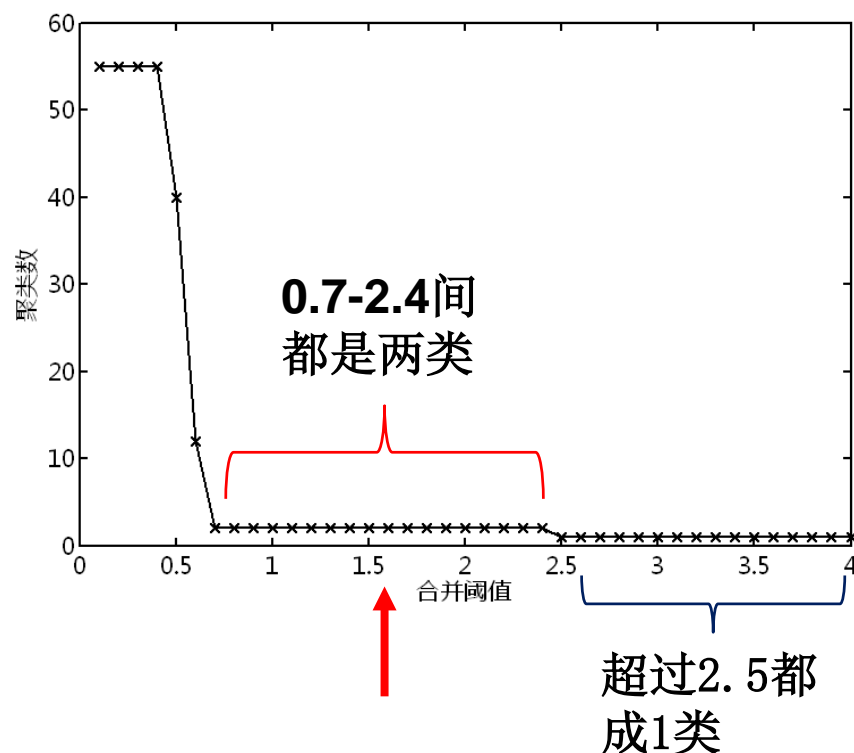
类别数量的选择十分
关键！

5.2 聚类数的间接选择

□ 谱系聚类参数选择

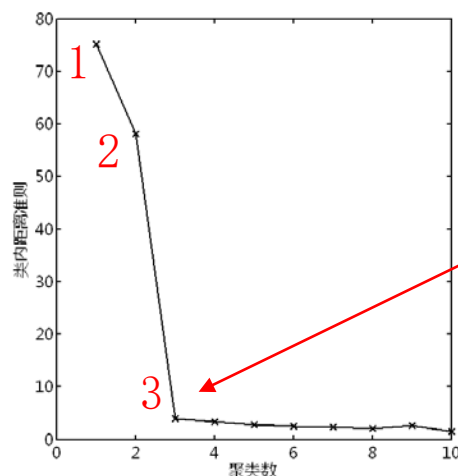
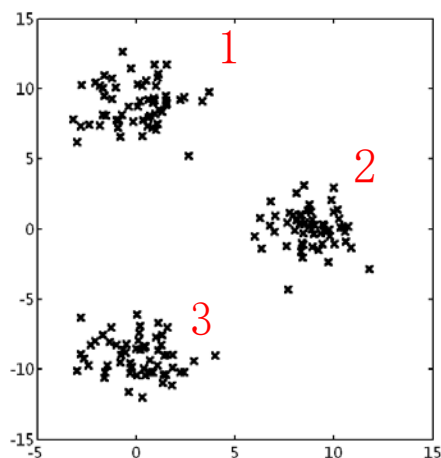
- 建立合并阈值与聚类数的对应关系
- 选择聚类数相同的最大参数区域
- 以区域的中点为最优参数

直接选择合并阈值很困难，通过实验的方法解决！

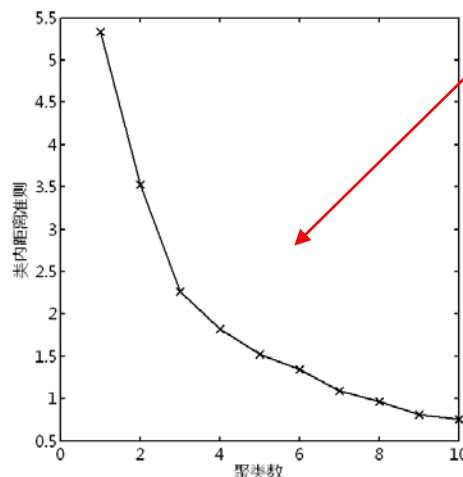
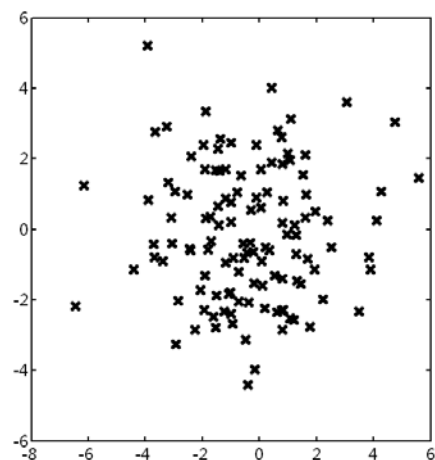


5.3 聚类数的直接选择

□ 尝试不同的聚类数，不同的聚类数对应不同的聚类评价



存在明显聚类，
对应明显“拐点”



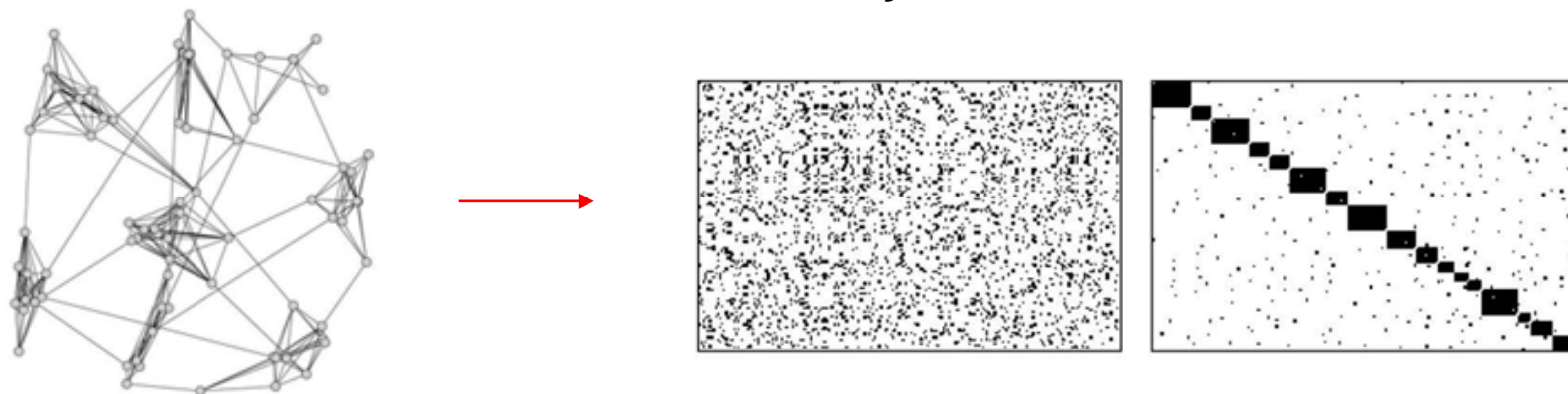
无明显聚类，
无明显“拐点”

k-均值聚类数的选择

- 在同一聚类数上，尝试不同初始条件，选择最优结果绘制曲线
- 寻找拐点，确定最佳聚类数

6 谱聚类简介

▣ **相似图**：样本集 $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ 可以表示为相似图 $G = (V, E)$ 。
图G也可以用邻接矩阵 $W = \{w_{ij}\}_{i,j=1,\dots,n}$ 描述。



▣ **最小割** mincut：将图的节点划分为 k 个子集 A_1, \dots, A_k ，使得子集之间的连接权重最小

$$\min_{A_1, \dots, A_k} \text{Cut}(A_1, \dots, A_k) = \frac{1}{2} \sum_{i=1}^k W(A_i, \bar{A}_i)$$

容易造成将单个样本划分为一个子集的现象，必须进一步考虑子集的“大小”

改进的图割: Normalized Cut

$$RatioCut(A_1, \dots, A_k) = \frac{1}{2} \sum_{i=1}^k \frac{W(A_i, \bar{A}_i)}{|A_i|} = \sum_{i=1}^k \frac{Cut(A_i, \bar{A}_i)}{|A_i|}$$

$|A|$ = A中节点的个数

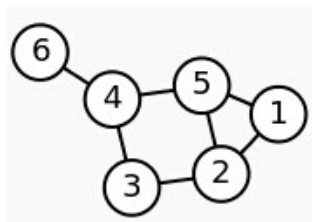
$$NCut(A_1, \dots, A_k) = \frac{1}{2} \sum_{i=1}^k \frac{W(A_i, \bar{A}_i)}{vol(A_i)} = \sum_{i=1}^k \frac{Cut(A_i, \bar{A}_i)}{vol(A_i)}$$

$$vol(A) = \sum_{i \in A} d_i \quad \text{其中 } d_i = \sum_{j=1}^n w_{ij} \text{ 为节点 } i \text{ 的连接权重和}$$

Normalized Cut的求解是一个NP完全问题，求近似解

normalized Cut 求解算法

□ 给定一个邻接矩阵为 W 的图，拉普拉斯矩阵定义为 $L=D-W$



$$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

W

$$\begin{pmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$D = \text{diag}(d_i)$

$$\begin{pmatrix} 2 & -1 & 0 & 0 & -1 & 0 \\ -1 & 3 & -1 & 0 & -1 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ -1 & -1 & 0 & -1 & 3 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix}$$

$L=D-W$

□ 利用拉普拉斯矩阵的特征值求解（非正则方法）：

1. 计算 L 前 k 个(最小)特征矢量 $\mathbf{u}_1, \dots, \mathbf{u}_k$;
2. 用 $\mathbf{u}_1, \dots, \mathbf{u}_k$ 作为列矢量构造矩阵 U ;
3. $\mathbf{y}_1, \dots, \mathbf{y}_n$ 为 U 的行矢量，用 K 均值算法将其聚成 k 个类别。

Laplacian矩阵的性质

1. 对任意矢量 f ，成立：

$$f'Lf = \frac{1}{2} \sum_{i,j=1}^n w_{ij}(f_i - f_j)^2.$$

证明

$$\begin{aligned} f'Lf &= f'Df - f'Wf = \sum_{i=1}^n d_i f_i^2 - \sum_{i,j=1}^n f_i f_j w_{ij} \\ &= \frac{1}{2} \left(\sum_{i=1}^n d_i f_i^2 - 2 \sum_{i,j=1}^n f_i f_j w_{ij} + \sum_{j=1}^n d_j f_j^2 \right) = \frac{1}{2} \sum_{i,j=1}^n w_{ij}(f_i - f_j)^2. \end{aligned}$$

Laplacian矩阵的性质

2. L 为对称的半正定矩阵

证明： D 和 W 为对称矩阵， L 为对称矩阵，实对称矩阵为半正定矩阵。

3. L 的最小特征值为0，对应的特征矢量为 $\mathbf{1}$

证明： $(D - W)\mathbf{1} = \mathbf{0}$ ，因此0为特征值， $\mathbf{1}$ 为特征矢量。半正定矩阵，所以0为最小特征值。

normalized Cut 求解思路

- 利用L矩阵性质，定义恰当的f，转化Cut优化函数

$$f'Lf = \frac{1}{2} \sum_{i,j=1}^n w_{ij} (f_i - f_j)^2.$$

- 舍去离散约束，进行近似谱求解。

详细参阅： [A_Tutorial_on_Spectral_Clustering](#)

RatioCut的近似谱求解: $k=2$

$$\text{RatioCut}(A_1, \dots, A_k) = \frac{1}{2} \sum_{i=1}^k \frac{W(A_i, \bar{A}_i)}{|A_i|} = \sum_{i=1}^k \frac{\text{Cut}(A_i, \bar{A}_i)}{|A_i|}$$

▣ **定义**: n 维矢量 \mathbf{f} (指示矢量) $f_i = \begin{cases} \sqrt{|\bar{A}|/|A|} & \text{if } v_i \in A \\ -\sqrt{|A|/|\bar{A}|} & \text{if } v_i \in \bar{A}. \end{cases}$

▣ \mathbf{f} 与RatioCut的关系:

$$\begin{aligned} \mathbf{f}'L\mathbf{f} &= \frac{1}{2} \sum_{i,j=1}^n w_{ij} (f_i - f_j)^2 \\ &= \frac{1}{2} \sum_{i \in A, j \in \bar{A}} w_{ij} \left(\sqrt{\frac{|\bar{A}|}{|A|}} + \sqrt{\frac{|A|}{|\bar{A}|}} \right)^2 + \frac{1}{2} \sum_{i \in \bar{A}, j \in A} w_{ij} \left(-\sqrt{\frac{|\bar{A}|}{|A|}} - \sqrt{\frac{|A|}{|\bar{A}|}} \right)^2 \\ &= \text{cut}(A, \bar{A}) \left(\frac{|\bar{A}|}{|A|} + \frac{|A|}{|\bar{A}|} + 2 \right) \\ &= \text{cut}(A, \bar{A}) \left(\frac{|A| + |\bar{A}|}{|A|} + \frac{|A| + |\bar{A}|}{|\bar{A}|} \right) \\ &= |V| \cdot \text{RatioCut}(A, \bar{A}). \end{aligned}$$

样本总数: $|V| = |A| + |\bar{A}|$

RatioCut的近似谱求解: $k=2$

□ \mathbf{f} 与矢量 $\mathbf{1}$ 正交:

$$\sum_{i=1}^n f_i = \sum_{i \in A} \sqrt{\frac{|\bar{A}|}{|A|}} - \sum_{i \in \bar{A}} \sqrt{\frac{|A|}{|\bar{A}|}} = |A| \sqrt{\frac{|\bar{A}|}{|A|}} - |\bar{A}| \sqrt{\frac{|A|}{|\bar{A}|}} = 0.$$

即:
$$\sum_{i=1}^n f_i = \mathbf{f}^t \mathbf{1} = 0$$

□ \mathbf{f} 的长度平方为 n :

$$\|\mathbf{f}\|^2 = \sum_{i=1}^n f_i^2 = |A| \frac{|\bar{A}|}{|A|} + |\bar{A}| \frac{|A|}{|\bar{A}|} = |\bar{A}| + |A| = n.$$

RatioCut的优化问题

严格的优化问题：

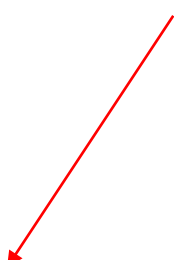
$$\min_{A \subset V} \mathbf{f}^t L \mathbf{f}$$

约束： $\mathbf{f}^t \mathbf{1} = 0$

$$\sum_{i=1}^n f_i^2 = n$$

$$f_i = \begin{cases} \sqrt{|\bar{A}|/|A|} & \text{if } v_i \in A \\ -\sqrt{|A|/|\bar{A}|} & \text{if } v_i \in \bar{A}. \end{cases}$$

离散性约束导致是一个NP问题！



近似的RatioCut的优化问题 (k=2)

□近似的优化问题：放松对 \mathbf{f} 中元素的离散性约束

$$\min_{\mathbf{f} \in R^n} \mathbf{f}' L \mathbf{f}$$

$$\text{约束: } \mathbf{f} \perp \mathbf{1}, \|\mathbf{f}\| = \sqrt{n}$$

□问题的解：对应L第2小特征值的特征矢量

1. 不考虑正交约束，问题变成Rayleigh商的优化，解是L的最小特征值对应的特征矢量；
2. 最小特征值对应特征矢量为 $\mathbf{1}$ ，不满足正交条件，第2小特征值对应特征矢量满足正交条件（L为实对称矩阵）；

广义特征向量

\mathbf{A} 为 n 阶实对称矩阵, \mathbf{B} 为 n 阶实对称正定矩阵
满足 $\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x}$ 的数 λ 为 \mathbf{A} 相对于 \mathbf{B} 的特征值
与 λ 相对应的非零解 \mathbf{x} 称为属于 λ 的特征向量

广义瑞利商特性

非零向量 \mathbf{x}_0 是 $R(\mathbf{x})$ 的驻点的充要条件是

\mathbf{x}_0 为 $\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x}$ 的属于特征值 λ 的特征向量

$$R(\mathbf{x}) = \frac{\mathbf{x}^t \mathbf{A} \mathbf{x}}{\mathbf{x}^t \mathbf{B} \mathbf{x}}$$

$$\mathbf{x}^t \mathbf{B} \mathbf{x} R(\mathbf{x}) = \mathbf{x}^t \mathbf{A} \mathbf{x}$$

$$2\mathbf{B} \mathbf{x} R(\mathbf{x}) + \mathbf{x}^t \mathbf{B} \mathbf{x} \frac{dR}{d\mathbf{x}} = 2\mathbf{A} \mathbf{x}$$

$$\frac{dR}{d\mathbf{x}} = \frac{2(\mathbf{A} \mathbf{x} - R(\mathbf{x}) \mathbf{B} \mathbf{x})}{\mathbf{x}^t \mathbf{B} \mathbf{x}} = 0$$

$$\mathbf{A} \mathbf{x}_0 = R(\mathbf{x}_0) \mathbf{B} \mathbf{x}_0$$

\mathbf{x}_0 为 $\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x}$ 的属于特征值 $\lambda = R(\mathbf{x}_0)$ 的特征向量

RatioCut的近似谱求解: $k > 2$

□ **定义:** k 个指示矢量 $\mathbf{h}_j = (h_{1,j}, \dots, h_{k,j})^t$:

$$h_{i,j} = \begin{cases} 1/\sqrt{|A_j|}, & v_i \in A_j \\ 0, & \text{otherwise} \end{cases}$$

□ 与RatioCut的关系:

$$\begin{aligned} \mathbf{h}_k^t L \mathbf{h}_k &= \frac{1}{2} \sum_{i,j=1}^n w_{ij} (h_{i,k} - h_{j,k})^2 \\ &= \frac{1}{2} \sum_{\substack{i \in A_k \\ j \in A_k}} w_{ij} \left(\frac{1}{\sqrt{|A_k|}} \right)^2 + \frac{1}{2} \sum_{\substack{i \in \bar{A}_k \\ j \in A_k}} w_{ij} \left(-\frac{1}{\sqrt{|A_k|}} \right)^2 \\ &= \frac{1}{|A_k|} \text{Cut}(A_k, \bar{A}_k) \end{aligned}$$

$$\text{RatioCut}(A_1, \dots, A_k) = \sum_{j=1}^k \frac{1}{|A_j|} \text{Cut}(A_j, \bar{A}_j) = \sum_{j=1}^k \mathbf{h}_j^t L \mathbf{h}_j$$

RatioCut的优化问题

□严格的优化问题:

$$\min_{A_1, \dots, A_k} \sum_{j=1}^k \mathbf{h}_j^t L \mathbf{h}_j$$

约束: $\mathbf{h}_i^t \mathbf{h}_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$

$$h_{i,j} = \begin{cases} 1/\sqrt{|A_j|}, & v_i \in A_j \\ 0, & otherwise \end{cases}$$

□仍然是NP问题。

近似的RatioCut的优化问题

□近似的优化问题：放松对 \mathbf{h} 中元素的离散性约束

$$\min_{\mathbf{h}_1, \dots, \mathbf{h}_k} \sum_{j=1}^k \mathbf{h}_j^t L \mathbf{h}_j$$

$$\text{约束: } \mathbf{h}_i^t \mathbf{h}_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

□问题的解：最小 k 个特征值对应特征矢量。

NCut的近似谱求解: $k=2$

□**定义**: 指示矢量 \mathbf{f}

$$vol(A) = \sum_{i \in A} d_i \quad f_i = \begin{cases} \sqrt{\frac{vol(\bar{A})}{vol(A)}}, & v_i \in A \\ -\sqrt{\frac{vol(A)}{vol(\bar{A})}}, & v_i \in \bar{A} \end{cases}$$

1. $(D\mathbf{f})^t * \mathbf{1} = 0$

证明:

$$\begin{aligned} (D\mathbf{f})^t * \mathbf{1} &= \sum_{i=1}^n d_i f_i \\ &= \sum_{i \in A} d_i \sqrt{\frac{vol(\bar{A})}{vol(A)}} - \sum_{i \in \bar{A}} d_i \sqrt{\frac{vol(A)}{vol(\bar{A})}} \\ &= vol(A) * \sqrt{\frac{vol(\bar{A})}{vol(A)}} - vol(\bar{A}) * \sqrt{\frac{vol(A)}{vol(\bar{A})}}} \\ &= 0 \end{aligned}$$

NCut的近似谱求解： $k=2$

2. $\mathbf{f}^t D \mathbf{f} = \text{vol}(V)$

证明：

$$\begin{aligned}\mathbf{f}^t D \mathbf{f} &= \sum_{i=1}^n d_i f_i^2 \\ &= \sum_{i \in A} d_i \frac{\text{vol}(\bar{A})}{\text{vol}(A)} + \sum_{i \in \bar{A}} d_i \frac{\text{vol}(A)}{\text{vol}(\bar{A})} \\ &= \text{vol}(\bar{A}) + \text{vol}(A) \\ &= \text{vol}(V)\end{aligned}$$

NCut的近似谱求解: $k=2$

3. $\mathbf{f}^t L \mathbf{f} = \text{vol}(V) \text{NCut}(A, \bar{A})$

证明:

$$\begin{aligned}
 \mathbf{f}^t L \mathbf{f} &= \frac{1}{2} \sum_{i,j=1}^n w_{ij} (f_i - f_j)^2 \\
 &= \frac{1}{2} \sum_{\substack{i \in A \\ j \in \bar{A}}} w_{ij} \left(\sqrt{\frac{\text{vol}(\bar{A})}{\text{vol}(A)}} + \sqrt{\frac{\text{vol}(A)}{\text{vol}(\bar{A})}} \right)^2 + \frac{1}{2} \sum_{\substack{i \in \bar{A} \\ j \in A}} w_{ij} \left(-\sqrt{\frac{\text{vol}(\bar{A})}{\text{vol}(A)}} - \sqrt{\frac{\text{vol}(A)}{\text{vol}(\bar{A})}} \right)^2 \\
 &= \text{Cut}(A, \bar{A}) * \left(\frac{\text{vol}(\bar{A})}{\text{vol}(A)} + \frac{\text{vol}(A)}{\text{vol}(\bar{A})} + 2 \right) \\
 &= \text{Cut}(A, \bar{A}) * \left(\frac{\text{vol}(V)}{\text{vol}(A)} + \frac{\text{vol}(V)}{\text{vol}(\bar{A})} \right) \\
 &= \text{vol}(V) \text{NCut}(A, \bar{A})
 \end{aligned}$$

NCut的优化问题

□严格的优化问题:

$$\min_A \mathbf{f}^t L \mathbf{f}$$

$$\text{约束: } (D\mathbf{f})^t \mathbf{1} = 0$$

$$\mathbf{f}^t D \mathbf{f} = \text{vol}(V)$$

$$f_i = \begin{cases} \sqrt{\frac{\text{vol}(\bar{A})}{\text{vol}(A)}}, & v_i \in A \\ -\sqrt{\frac{\text{vol}(A)}{\text{vol}(\bar{A})}}, & v_i \in \bar{A} \end{cases}$$

NCut的近似优化问题

□近似的优化问题：放松 \mathbf{f} 的离散性约束

$$\min_{\mathbf{f}} \mathbf{f}^t L \mathbf{f}$$

$$\text{约束: } (D\mathbf{f})^t \mathbf{1} = 0$$

$$\mathbf{f}^t D \mathbf{f} = \text{vol}(V)$$

□问题的解：对应矩阵 $(D^{-1}L)$ 第2小特征值的特征矢量。

证明：

1. 不考虑正交性约束，是一个广义的Rayleigh商问题，解是 $(D^{-1}L)$ 最小特征值对应特征矢量；
2. 最小特征矢量不满足正交性，第2小特征矢量满足。

NCut的近似谱求解: $k > 2$

□**定义**: k 个指示矢量 $\mathbf{h}_1, \dots, \mathbf{h}_k$

$$h_{i,j} = \begin{cases} 1/\sqrt{\text{vol}(A_j)}, & v_i \in A_j \\ 0, & v_i \in \bar{A}_j \end{cases}$$

1. $\mathbf{h}_i^t D \mathbf{h}_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$

证明: $i \neq j$ 显然;

$$\mathbf{h}_i^t D \mathbf{h}_i = \sum_{t=1}^n d_t h_{t,i}^2 = \frac{1}{\text{vol}(A_i)} \sum_{t \in A_i} d_t = 1$$

NCut的近似谱求解: $k > 2$

$$2. \text{NCut}(A_1, \dots, A_k) = \sum_{j=1}^k \mathbf{h}_j^t L \mathbf{h}_j$$

证明: $\mathbf{h}_i^t L \mathbf{h}_i = \frac{1}{2} \sum_{j,t=1}^n w_{jt} (h_{j,i} - h_{t,i})^2$

$$= \frac{1}{2} \sum_{\substack{j \in \underline{A}_i \\ t \in \bar{A}_i}} \frac{w_{jt}}{\text{vol}(A_i)} + \frac{1}{2} \sum_{\substack{j \in \bar{A}_i \\ t \in A_i}} \frac{w_{jt}}{\text{vol}(A_i)}$$

$$= \text{Cut}(A_i, \bar{A}_i) / \text{vol}(A_i)$$

$$\text{NCut}(A_1, \dots, A_k) = \sum_{i=1}^k \frac{\text{Cut}(A_i, \bar{A}_i)}{\text{vol}(A_i)} = \sum_{i=1}^k \mathbf{h}_i^t L \mathbf{h}_i$$

NCut的优化问题： $k > 2$

▣ 严格的优化问题：

$$\min_{\mathbf{h}_1, \dots, \mathbf{h}_k} \sum_{i=1}^k \mathbf{h}_i^t L \mathbf{h}_i$$

$$\text{约束: } \mathbf{h}_i^t D \mathbf{h}_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

$$h_{i,j} = \begin{cases} 1 / \sqrt{\text{vol}(A_j)}, & v_i \in A_j \\ 0, & v_i \in \bar{A}_j \end{cases}$$

NCut的近似优化问题： $k > 2$

▣近似的优化问题： 放松 \mathbf{h} 的离散性约束

$$\min_{\mathbf{h}_1, \dots, \mathbf{h}_k} \sum_{i=1}^k \mathbf{h}_i^t L \mathbf{h}_i$$

$$\text{约束: } \mathbf{h}_i^t D \mathbf{h}_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

▣问题的解： 为矩阵 $D^{-1}L$ 最小 k 个特征值对应的特征矢量。

谱聚类算法

➤非正则算法

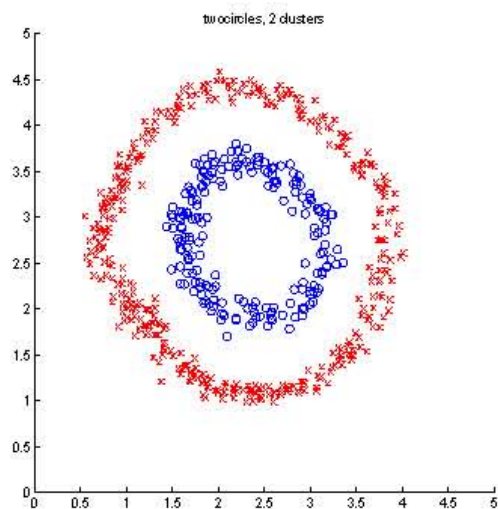
1. 计算相似图的邻接矩阵 W ;
2. 计算Laplacian矩阵 L ;
3. 计算 L 前 k 个(最小)特征矢量 u_1, \dots, u_k ;
4. 用 u_1, \dots, u_k 作为列矢量构造矩阵 U ;
5. y_1, \dots, y_n 为 U 的行矢量, 用 K 均值算法将其聚成 k 个类别。

谱聚类算法

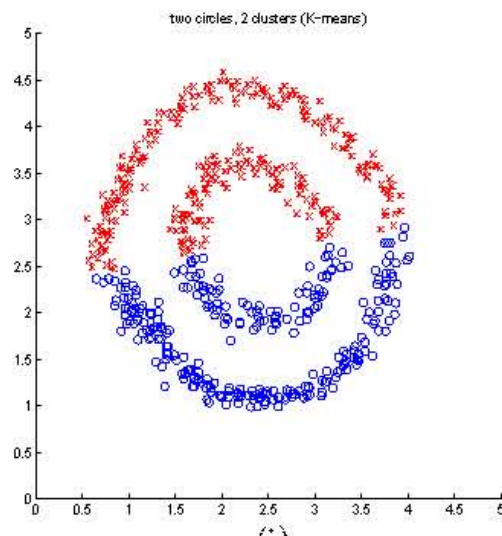
➤ 正则化算法

1. 计算相似图的邻接矩阵 W ;
2. 计算Laplacian矩阵 L ;
3. 求解广义特征值问题 $L\mathbf{u} = \lambda D\mathbf{u}$ 的前 k 个(最小)特征矢量 $\mathbf{u}_1, \dots, \mathbf{u}_k$;
4. 用 $\mathbf{u}_1, \dots, \mathbf{u}_k$ 作为列矢量构造矩阵 U ;
5. $\mathbf{y}_1, \dots, \mathbf{y}_n$ 为 U 的行矢量, 用K均值算法将其聚成 k 个类别。

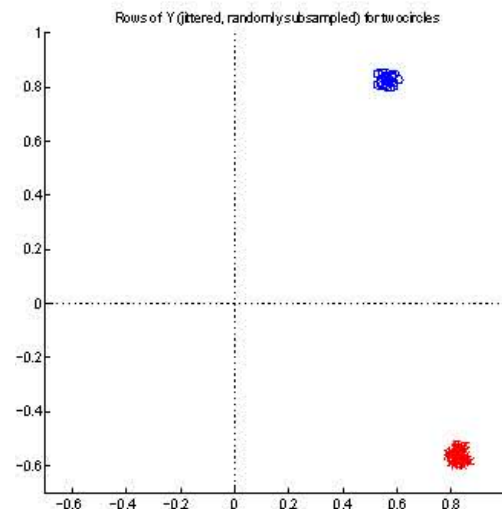
谱聚类示例



原样本分布



K均值聚类



特征值矩阵的行矢量

算法的实现

1. 相似图的构造

➤ 全连接:

$$w_{ij} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma^2}\right)$$

➤ ε -近邻:

$$w_{ij} = \begin{cases} 1, & \|\mathbf{x}_i - \mathbf{x}_j\| < \varepsilon \\ 0, & otherwise \end{cases}$$

➤ 互K-近邻:

$$w_{ij} = \begin{cases} \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma^2}\right), & \mathbf{x}_i \in KNN(\mathbf{x}_j) \wedge \mathbf{x}_j \in KNN(\mathbf{x}_i) \\ 0, & otherwise \end{cases}$$

otherwise

算法的实现

2. **特征矢量的计算**: 矩阵具有稀疏性, 存在快速算法 Lanczos method, 最小特征值0对应的特征矢量为1, 其它矢量与其正交;

3. **Laplacian矩阵的选择**:

➤非正规化: $L = D - W$

➤对称正规化: $L_{sym} = D^{-1/2}LD^{-1/2} = I - D^{-1/2}WD^{-1/2}$

➤随机游走正规化: $L_{rw} = D^{-1}L = I - D^{-1}W$

➤使用 时 L_{sym} , 特征矢量需要乘 $D^{-1/2}$ 。推荐使用 L_{rw} 。

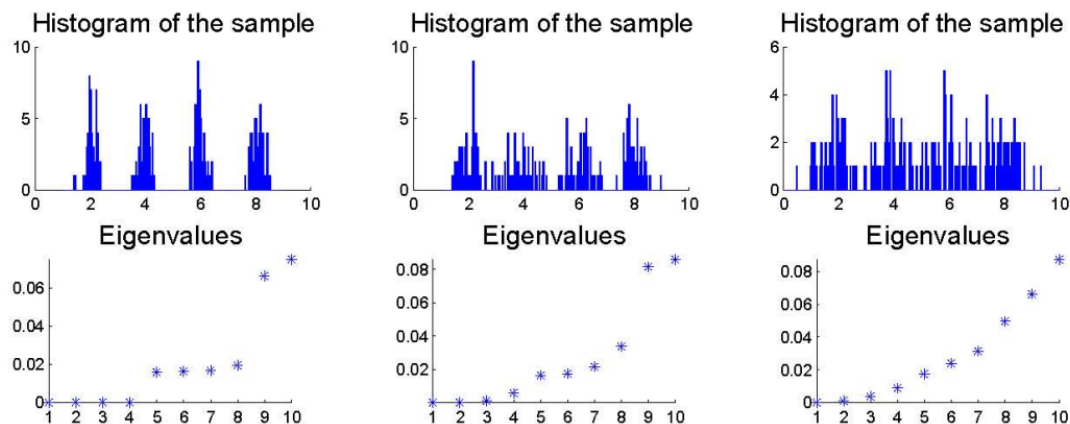
算法的实现

4. 聚类数的选择:

- 理想情况：样本按照类别的顺序排列，类别之间相似度为0，Laplacian矩阵前k个特征值为0

$$L = \begin{bmatrix} L_1 & & & \\ & L_2 & & \\ & & \ddots & \\ & & & L_k \end{bmatrix}$$

- 一般情况：根据特征值的分布确定聚类数



算法的实现

1) 构建表示对象集的相似度矩阵 W ;

$d = \text{pdist}(M);$

$W = \text{squareform}(d);$

2) 根据相似度矩阵 W 构建非正规拉普拉斯矩阵;

$N = \text{tril}(W, 0);$

$s = \text{sum}(N);$

$D = \text{diag}(s);$

$L = D - N;$

3) 计算拉普拉斯矩阵的前 k 个特征值与特征向量, 构建特征向量空间;

$[Q, A] = \text{eigs}(L, k, 'SR');$

4) 利用K-means对特征向量空间中的特征向量进行聚类。

$C = \text{kmeans}(Q, k);$