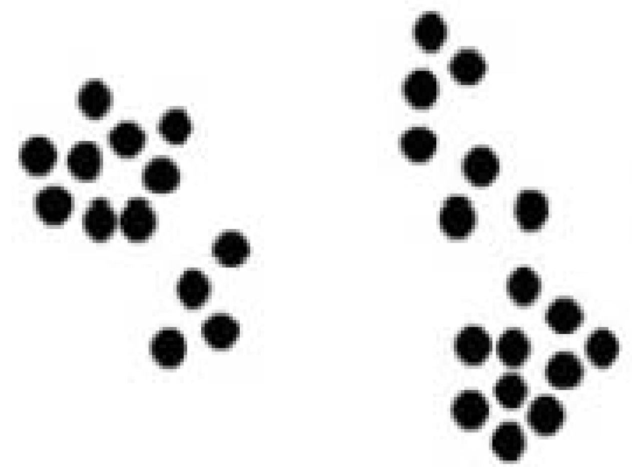
机器学习

4. 聚类-k均值

主要内容

- 什么是聚类
- 层次聚类方法
- k均值聚类

什么是聚类



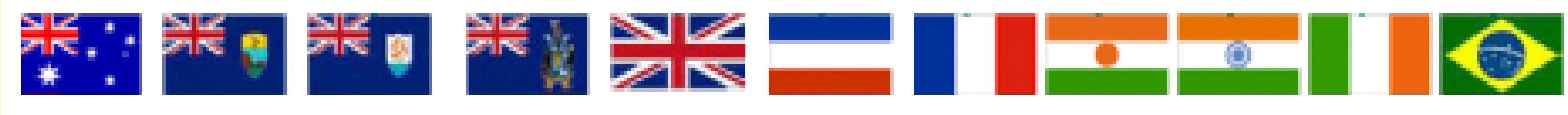
- 在以上点集中是否存在“类”
- 几个类?
- 每个类是什么?
- 怎样识别这些类?

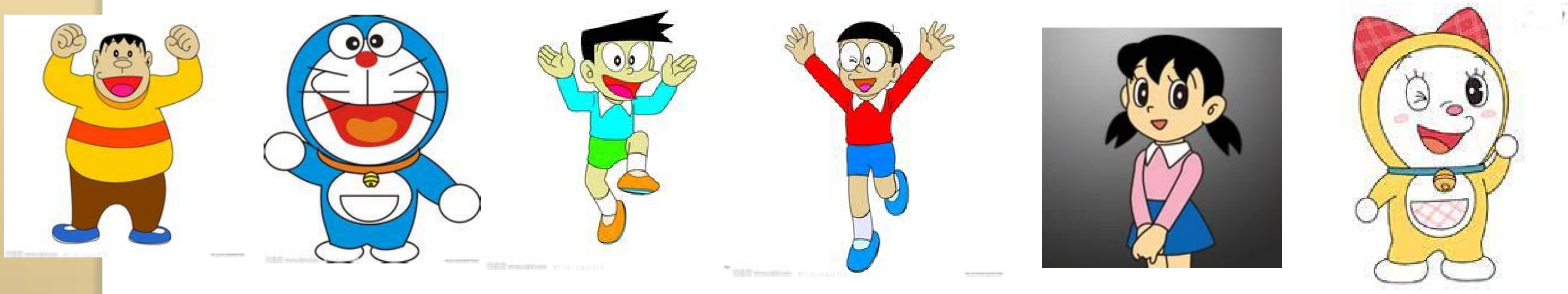
什么是聚类

- 聚类：将同类型的对象聚为不同类别的过程
  - 高类内相似性
  - 低类间相似性
  - 一种无监督学习的常见学习形式
- 无监督学习：
  - 从原始样本（无标注信息）中学习知识
- 一种对于科学、工程很多领域非常常见的学习目标
  - 基因分类
  - 用户甄别
  - 文本主题分类
  - 图片/视频目标分类
  - 。。。

什么是聚类

- 下面的例子怎样聚类?





什么是聚类

- 基本问题
  - 什么是一群目标数据的自然聚类?
  - 如何度量目标数据间的“关系”
  - 数据如何表达
  - 类数目如何度量?
  - 聚类算法
  - 算法是否收敛?

什么是聚类

- 聚类是主观的！

什么是聚类

- 聚类最重要的概念：
  - 相似度

- 相似度的定义是一个哲学问题
- 依赖于数据表达方式与算法导向
- 如何实际操作?

什么是聚类

- 距离！
  - $D(A,B) = D(B,A)$
  - $D(A,A) = 0$
  - $D(A,B) = 0$ IIf A= B
  - $D(A,B) \leq D(A,C) + D(B,C)$

直观意义？

什么是聚类

- 典型相似度度量（距离）
  - 两个p维向量：

$$x=(x_1,x_2,\cdots,x_p)$$

$$y=(y_1,y_2,\cdots,y_p)$$

  - Minkowski距离（Lp范数）

$$d(x,y)=\sqrt[r]{\sum_{i=1}^{p}|x_i-y_i|^r}$$

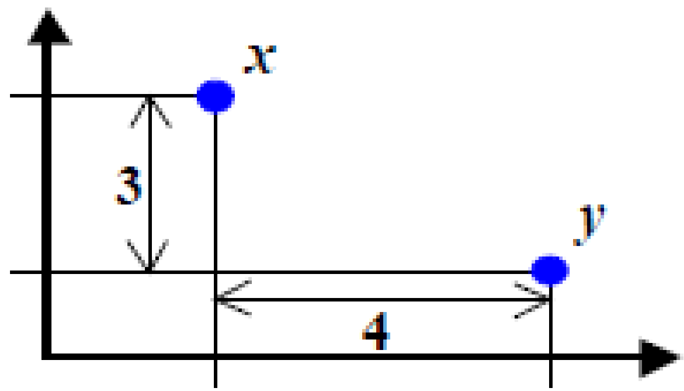
  - 最常见的Lp范数

$$d(x,y)=\sqrt[2]{\sum_{i=1}^{p}|x_i-y_i|^2}$$

$$d(x,y)=\sum_{i=1}^{p}|x_i-y_i|$$

$$d(x,y)=\max_{1\le i\le p}|x_i-y_i|$$

什么是聚类



- L2距离（欧氏距离）：
- L1距离：
- L无穷距离（最大距离）：

什么是聚类

- 海明距离（曼哈顿距离）：对全部特征为二值的向量对定义
  - 基因表达
  - 文本分类

	关键词1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
文本1	1	0	0	1	1	1	0	1	1	0	0	0	1	1	1
文本2	1	1	0	1	1	0	1	0	0	0	1	1	1	0	0

$$\text{海明距离} = \#01 + \#10 = 9$$

什么是聚类

- 皮尔斯相关系数

$$s(x,y)=\frac{\sum_{i=1}^{p}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{p}(x_i-\bar{x})^2\times\sum_{i=1}^{p}(y_i-\bar{y})^2}}$$

$$\text{where } \bar{x}=\tfrac{1}{p}\sum_{i=1}^{p}x_i \text{ and } \bar{y}=\tfrac{1}{p}\sum_{i=1}^{p}y_i.$$

$$|s(x,y)|\le 1$$

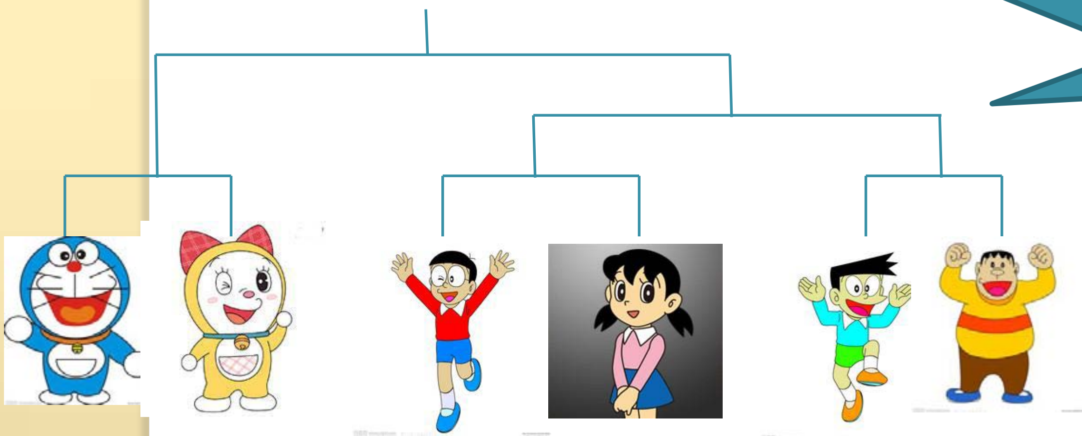
- 余弦距离

$$s(x,y)=\frac{\vec{x}\cdot\vec{y}}{|\vec{x}|\cdot|\vec{y}|}$$

什么是聚类

- 两类聚类问题

□ 层次化方法



Hierarchical Algorithms

□ 分部方法

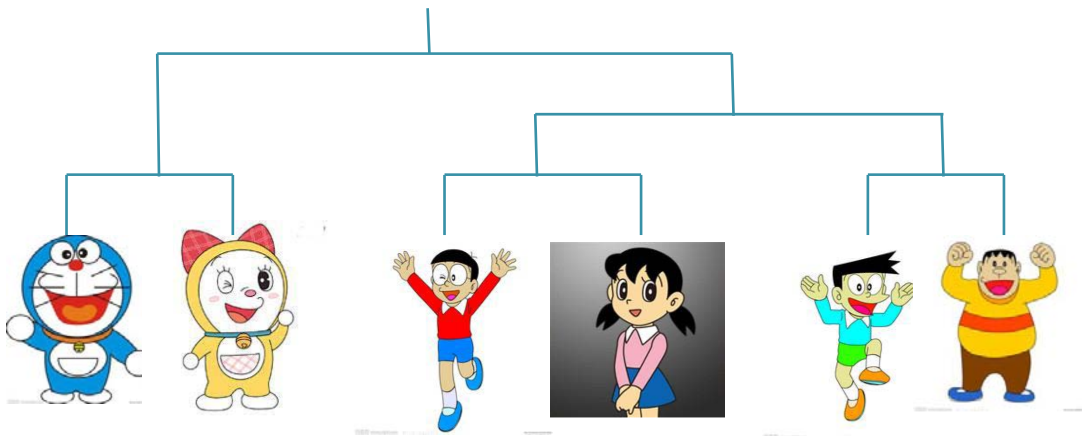
Partitional Algorithms

主要内容

- 什么是聚类
- 层次聚类方法
- k均值聚类

层次聚类方法

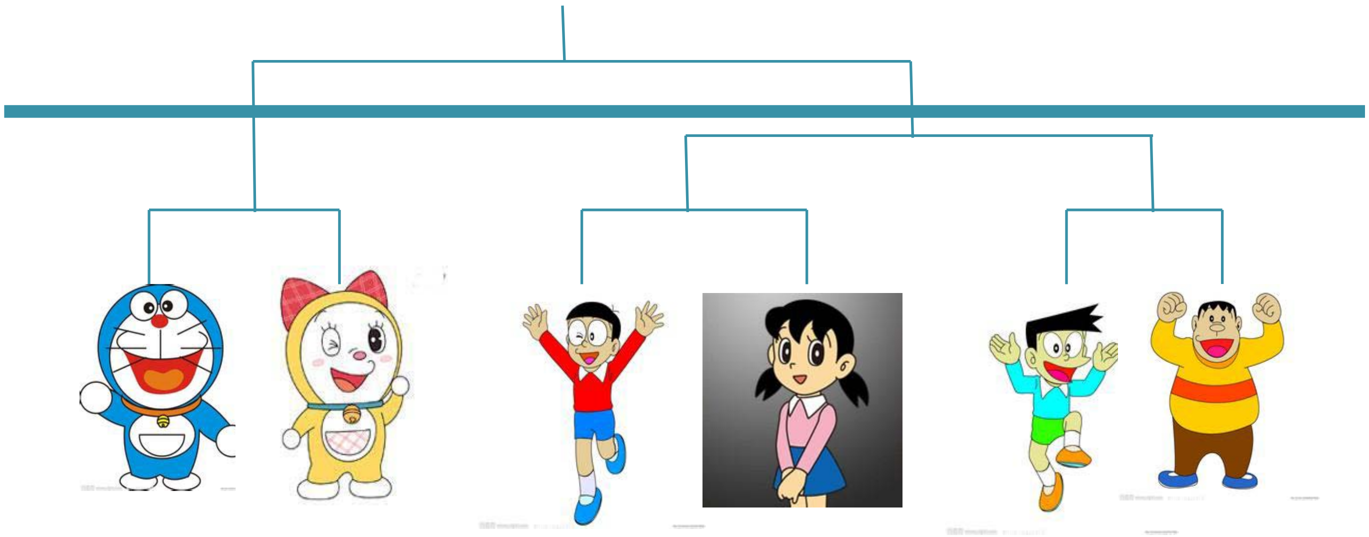
- 基本原理:将聚类过程分层次进行



- 与我们日常组织信息结构的方式非常类似
  - ✓ 图书馆书籍条目

层次聚类方法

- 有用性：可获得任意尺度，任意层次的聚类信息
  - 在需要的尺度切割聚类树



层次聚类方法

- 自底而上
  - 首先把每个目标数据视为一类
  - 不断将最近邻数据加入当前类
  - 最终形成一类

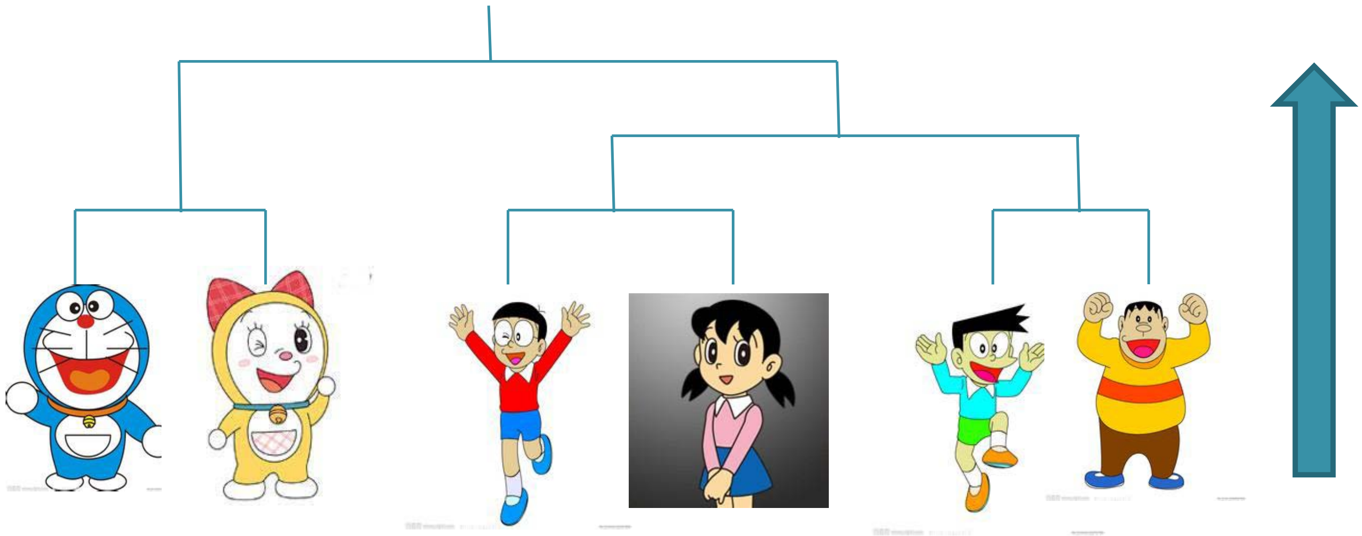
Bottom-Up
Agglomerative

- 自上而下
  - 首先把所有数据视为一类
  - 选择能将当前每类分离成两类的最佳分割
  - 直到所有数据分类一类

Top-Down
Divisive

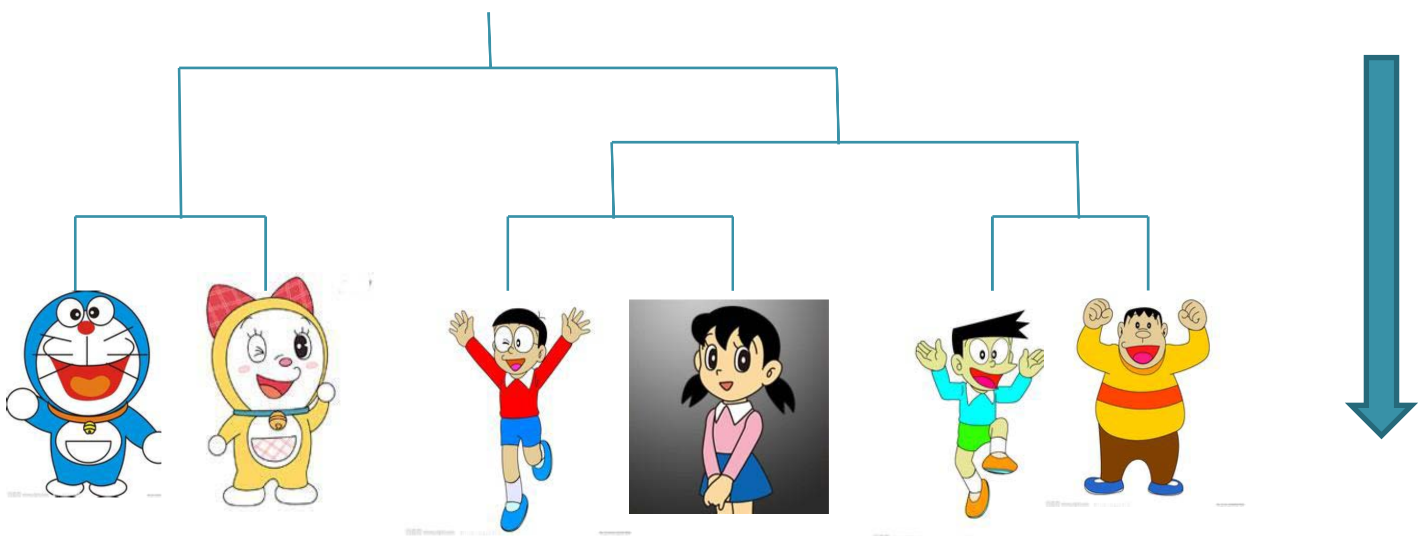
层次聚类方法

- 自底而上
  - 首先把每个目标数据视为一类
  - 不断将最近邻数据加入当前类
  - 最终形成一类



层次聚类方法

- 自上而下
  - 首先把所有数据视为一类
  - 选择能将当前每类分离成两类的最佳分割
  - 直到所有数据分类一类

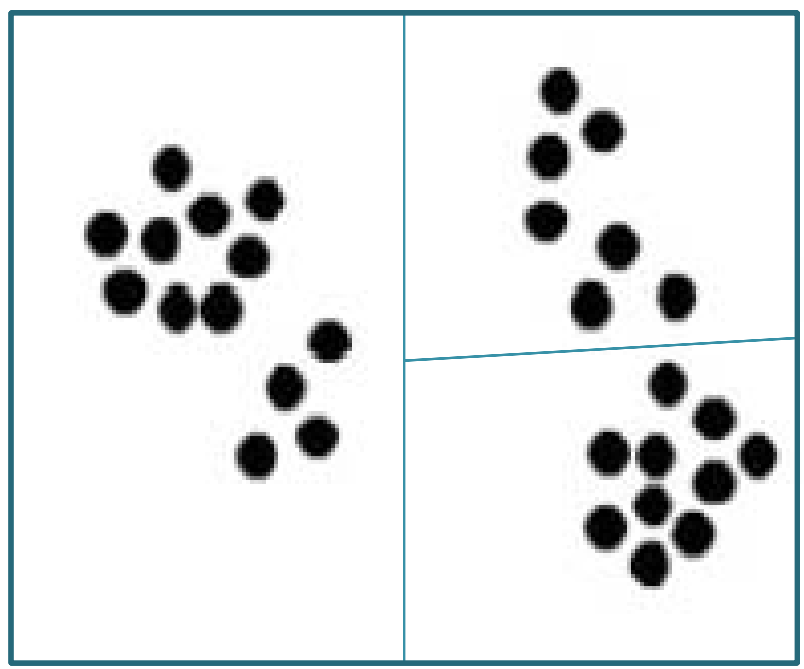


主要内容

- 什么是聚类
- 层次聚类方法
- k均值聚类

K-均值聚类

- 分部方法原理：
  - 将n个目标数据分割到预设的K个聚类中



K-均值聚类

- 游戏
  - 迭代进行以下步骤：
  - 每个人将自己归类于与自己最近的队伍核心
  - 将位于自己类的成员中心更新为新的队伍核心

K-means

- 人肉 K均值算法

K-均值聚类

- **算法基本步骤**

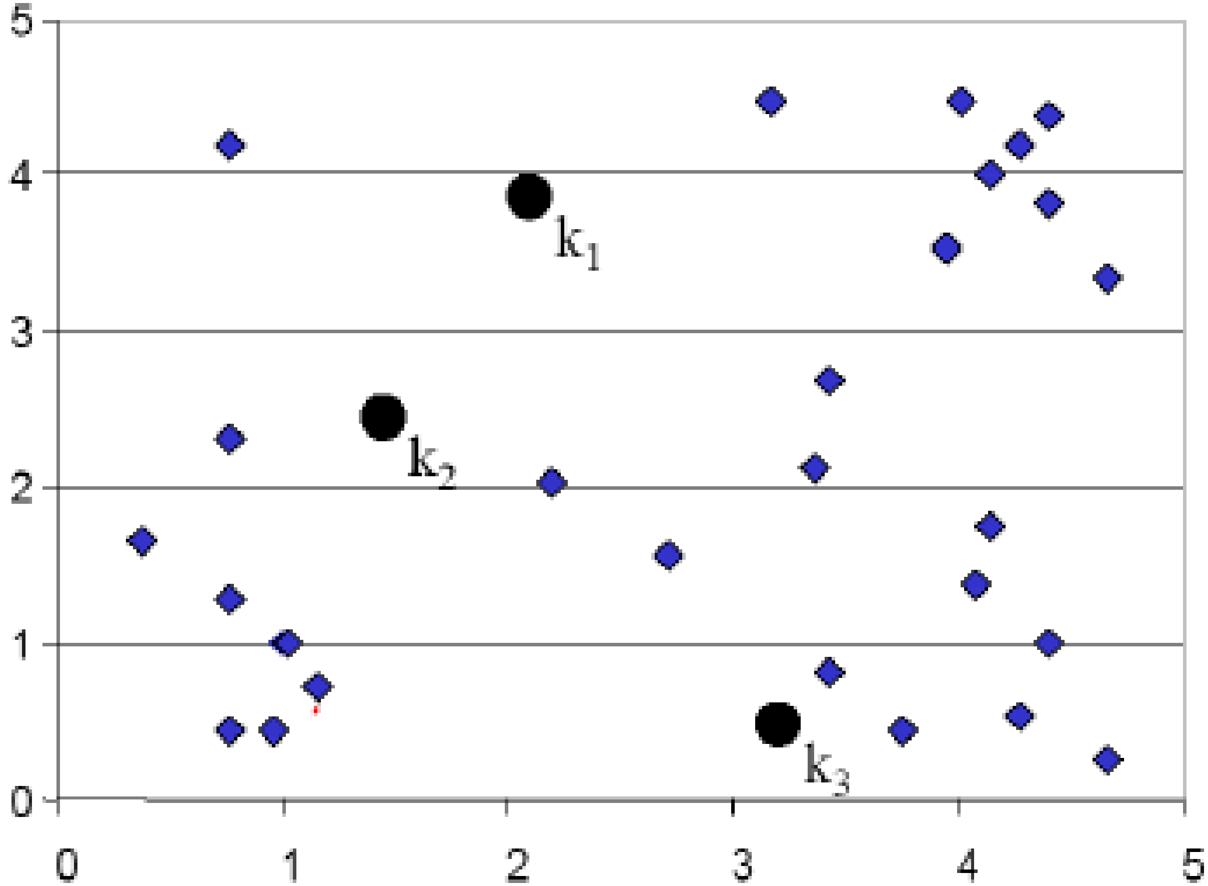
输入:数据,聚类个数 K

1. 初始化 K 个聚类中心
2. 开始如下迭代
   a) 对每一个样本进行归类，距离哪个聚类中心近，则将其归为哪一类；
   b) 重新估计 K 个聚类中心

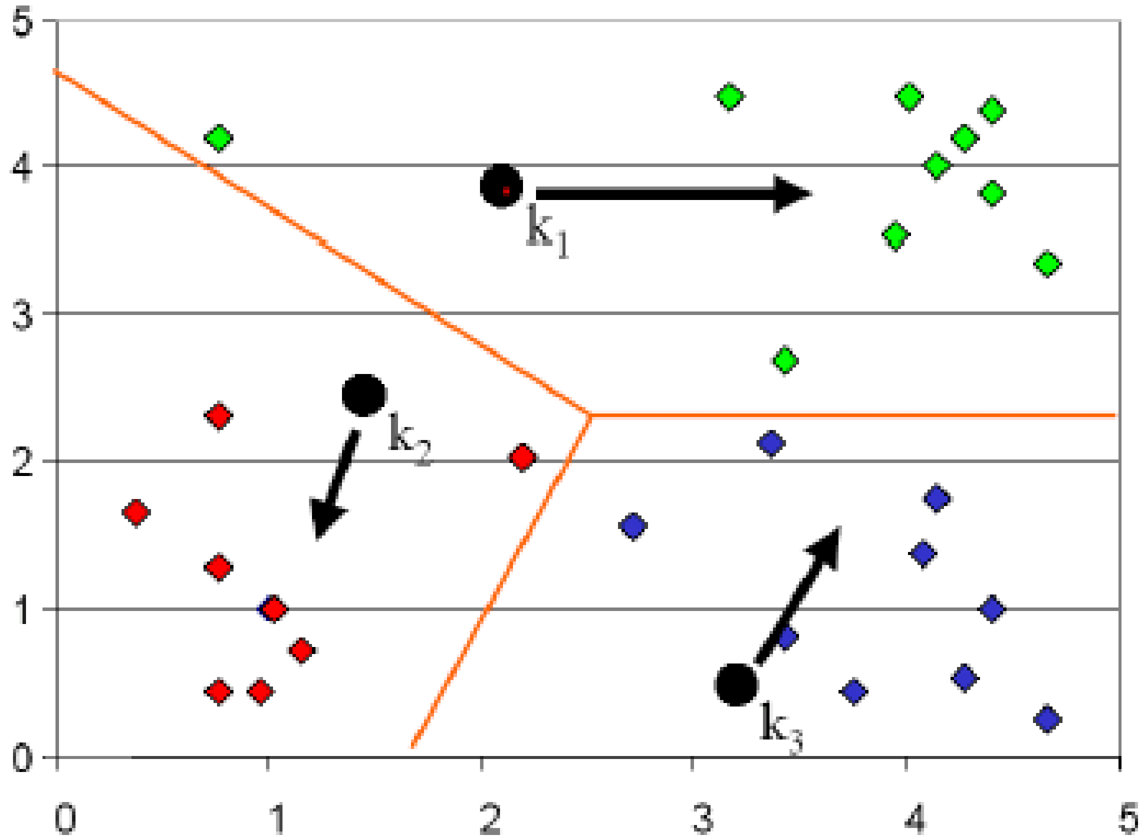
以上迭代当每个聚类数据不发生改变时终止。

K-均值聚类

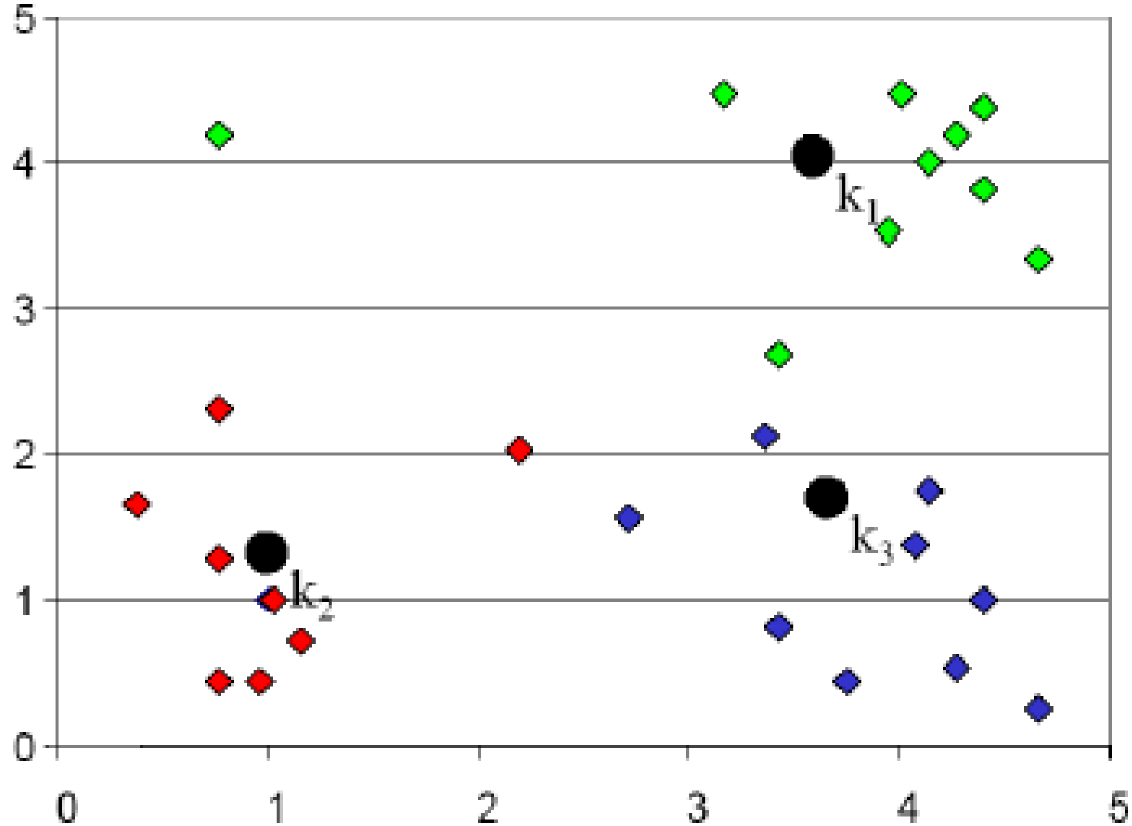
- 示例：



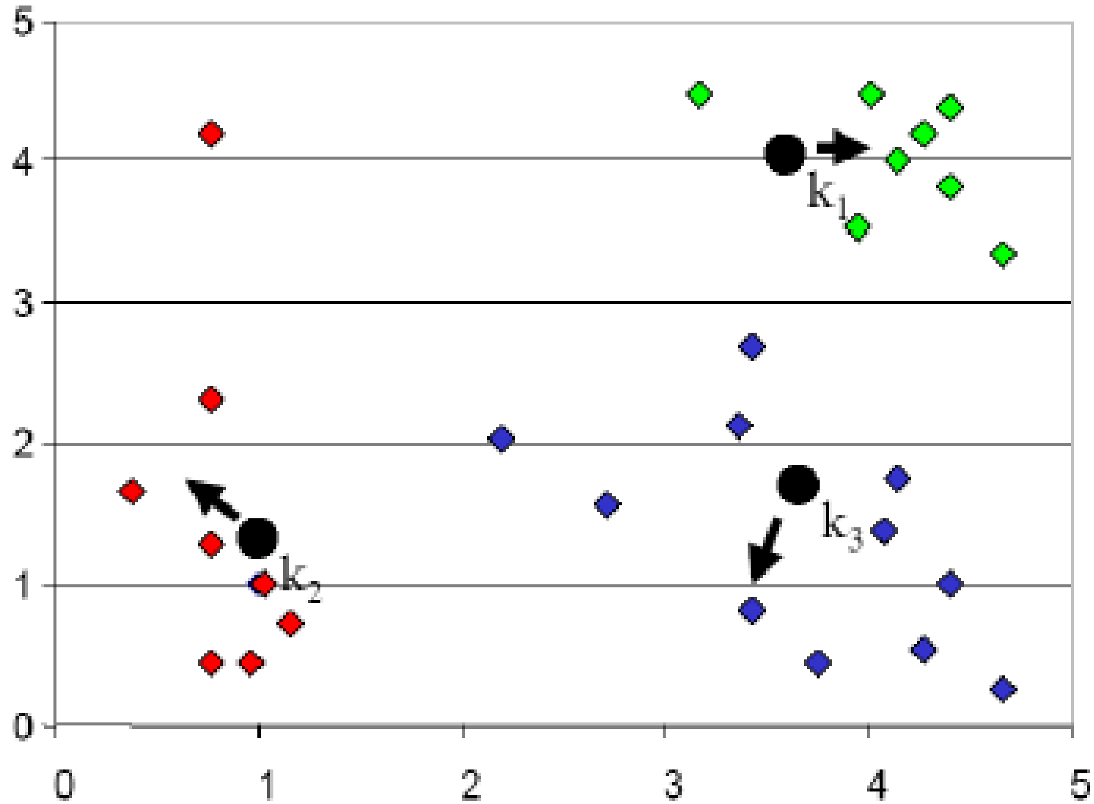
K-均值聚类



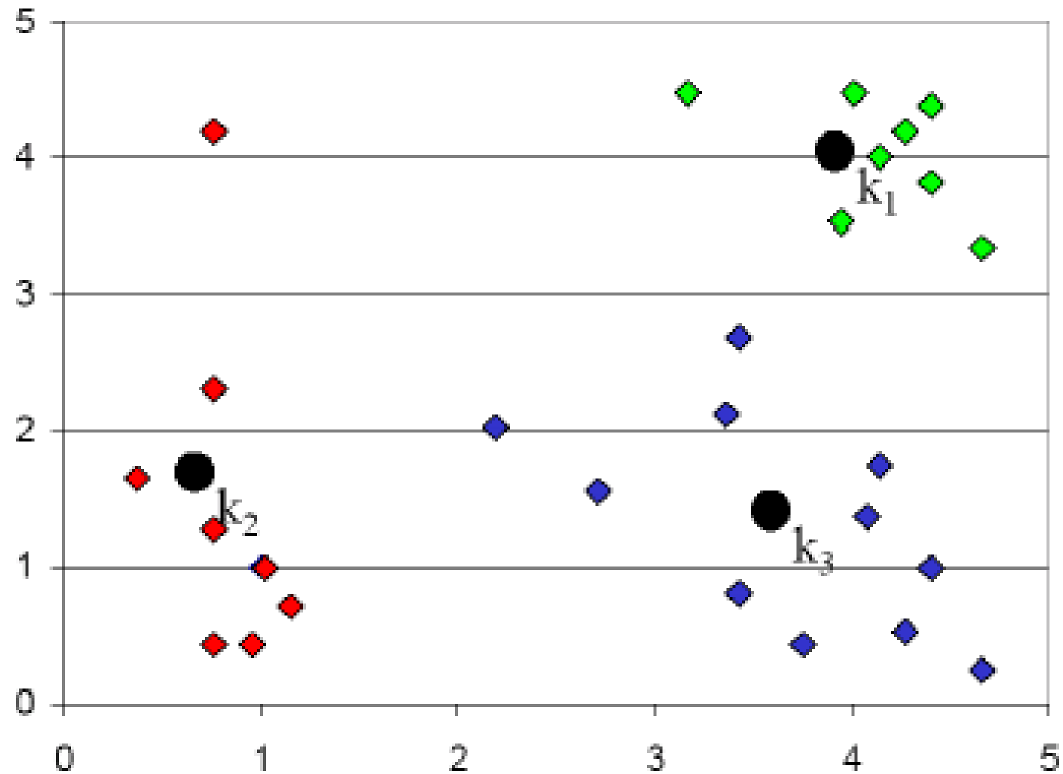
K-均值聚类



K-均值聚类



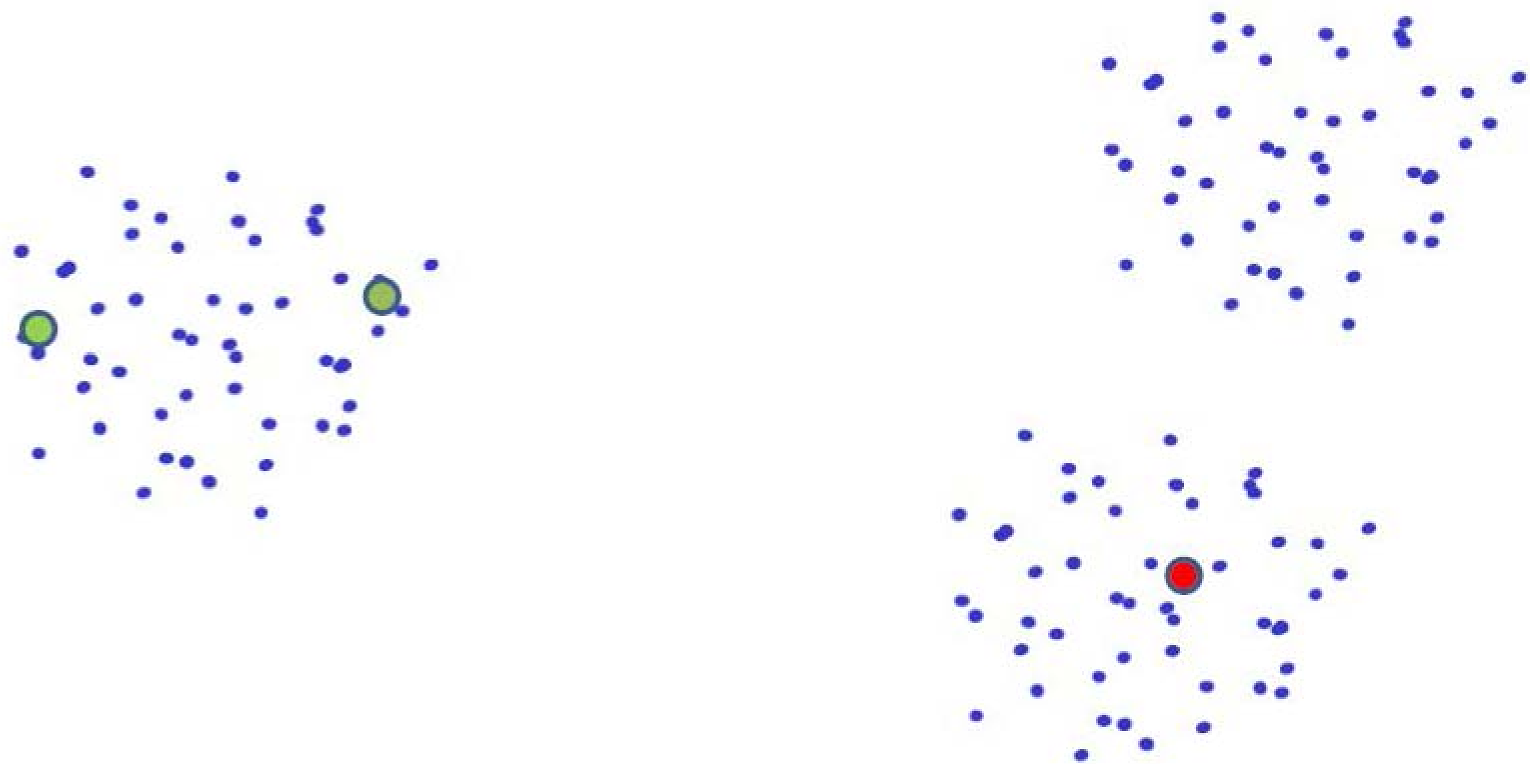
K-均值聚类



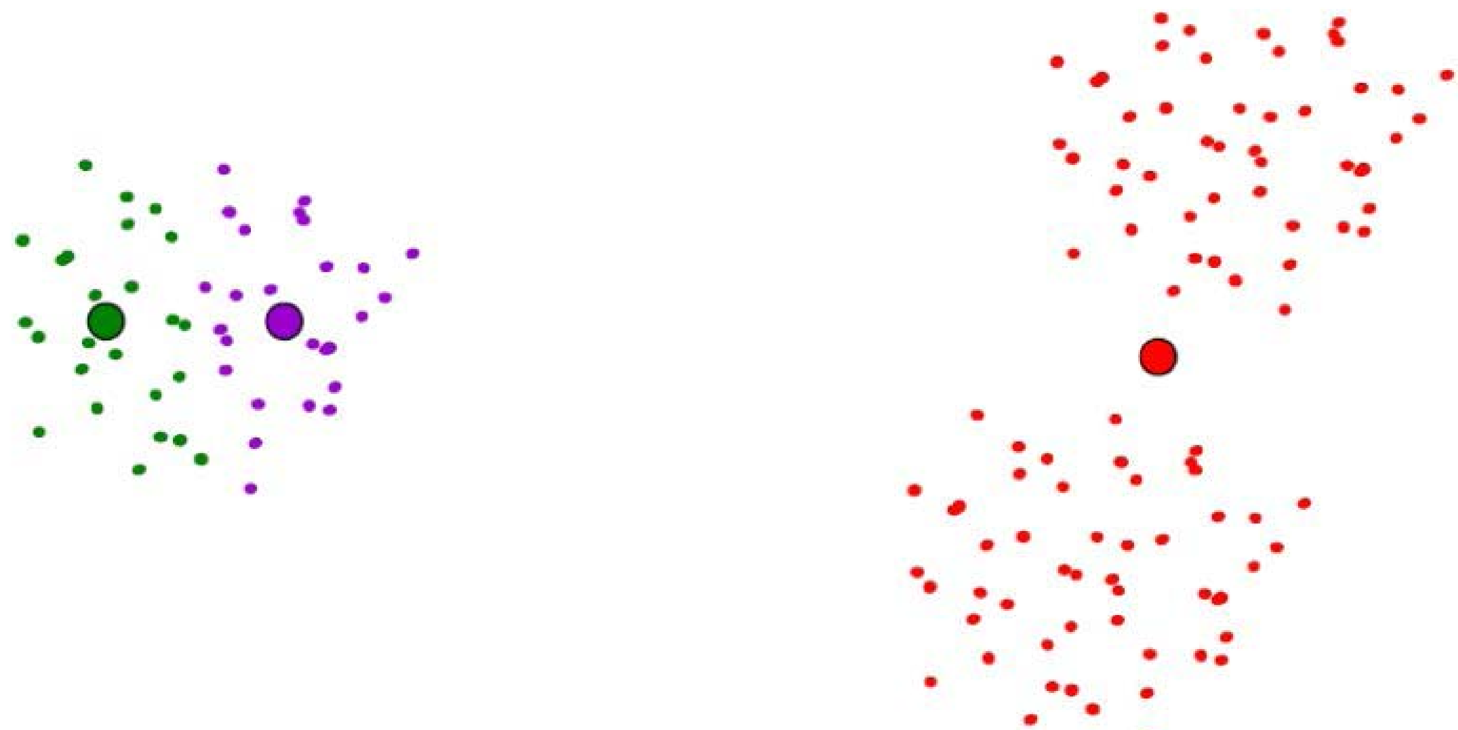
K-均值聚类

1. 是否一定收敛?
2. 是否一定收敛到一个合理值?

K-均值聚类



K-均值聚类



K-均值聚类

- 效果与初值选择有关
- 如何选择初值?

K-均值聚类

- K均值聚类算法有无目标函数(表现度量)?

问题描述：

给定 n 个观察数据$(x_1, x_2, \dots, x_n)$，学习目标为将其归入 K 个类中：$C = \{C_1, C_2, \dots, C_K\}$，对应类具有类指示数据$\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_K)$,从而使得以下目标函数（类内数据最小二乘误差）最小：

$$\underset{C,\boldsymbol{\mu}}{\operatorname{argmin}} \sum_{i=1}^{K} \sum_{x_j \in C_i} \left\| x_j - \mu_i \right\|_2^2.$$

K-均值聚类

问题描述：

给定 n 个观察数据$(x_1, x_2, \dots, x_n)$，学习目标为将其归入 K 个类中：$C = \{C_1, C_2, \dots, C_K\}$，对应类具有类指示数据$\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_K)$,从而使得以下目标函数（类内数据最小二乘误差）最小：

$$\underset{C,\boldsymbol{\mu}}{\operatorname{argmin}} \sum_{i=1}^{K} \sum_{x_j \in C_i} \left\| x_j - \mu_i \right\|_2^2 .$$

- 问题求解复杂度：
  - 组合优化问题
  - NP-hard!

K-均值聚类

- 解决之道：
  - 迭代优化/搜索/更新算法

Althernative Optimization/Search

K-均值聚类

$$\underset{S,\boldsymbol{\mu}}{\operatorname{argmin}} \sum_{i=1}^{K} \sum_{x_j \in C_i} \left\| x_j - \mu_i \right\|_2^2.$$

- 初始化 k 个类中心 $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_K)$
- 迭代进行以下优化
  - 更新聚类：固定$\boldsymbol{\mu}$，优化C
  - 更新类中心：固定C，优化$\boldsymbol{\mu}$

K-均值聚类

$$\underset{S,\boldsymbol{\mu}}{\operatorname{argmin}} \sum_{i=1}^{K} \sum_{x_j \in C_i} \left\| x_j - \mu_i \right\|_2^2.$$

- 更新聚类：固定$\boldsymbol{\mu}$，优化C
- 等价于对每个x_j单独计算以下优化问题

$$\underset{x_j \in C_i}{\operatorname{argmin}} \left\| x_j - \mu_i \right\|_2^2 \quad \text{s.t. } i = 1,2,\ldots,K$$

➢ 解是?

K-均值聚类

$$\underset{S,\boldsymbol{\mu}}{\operatorname{argmin}} \sum_{i=1}^{K} \sum_{x_j \in C_i} \left\| x_j - \mu_i \right\|_2^2.$$

- 更新类中心：固定C，优化$\boldsymbol{\mu}$
- 等价于对每个类中心μ_i，计算以下优化问题：

$$\underset{\mu_i}{\operatorname{argmin}} \sum_{x_j \in C_i} \left\| x_j - \mu_i \right\|_2^2.$$

➢ 解是?

K-均值聚类

- K-均值算法基本步骤

输入:数据,聚类个数 K

1. 初始化 K 个聚类中心
2. 开始如下迭代
   a) 对每一个样本进行归类，距离哪个聚类中心近，则将其归为哪一类；
   b) 重新估计 K 个聚类中心

以上迭代当每个聚类数据不发生改变时终止。

- 完全对应于求解 $\operatorname{argmin}_{S,\mu} \sum_{i=1}^{K} \sum_{x_j \in C_i} \|x_j - \mu_i\|_2^2.$ 的迭代优化方法！
- 目标函数单调递减，一定收敛！

K-均值聚类

- 初值问题
  - 为何依赖于初值?
    - 目标函数非凸，因此不同初值会收敛于不同局部极优点
  - 怎样选择相对较好初值
    - 利用启发式方法（将初值尽量散开设置）
    - 尝试多个初值
    - 使用其它方法运行结果作为初值

K-均值聚类

- 什么是一个好的聚类
  - 内部标准：
    - 类内相似度高
    - 类间相似度低
    - 使用相似度度量的合理性
  - 外部标准
    - 能否挖掘一些隐藏在数据中的有用模式
    - 能否恢复真实类别

K-均值聚类

- k-medoid方法

$$\underset{S,\boldsymbol{\mu}}{\operatorname{argmin}}\sum_{i=1}^{K}\sum_{x_j\in C_i}\left\|x_j-\mu_i\right\|_2^2.$$

$$\underset{C,\boldsymbol{\mu}}{\operatorname{argmin}}\sum_{i=1}^{K}\sum_{x_j\in C_i}\left\|x_j-\mu_i\right\|_1$$

- 仔细思考与K-均值求解方法的异同

K-均值聚类

- 缺点?
  - 归类方式为hard，而非soft
  - 确定性模型而非生成模型
  - 如何预测?

要求

1. 聚类的基本概念与思想
2. 层次聚类的基本思想
3. K均值聚类算法、模型与优化原理

阅读:

[1] Pattern Recognition and Machine Learning, Christopher , M. Bishop, Springer, 2006. 9. Mixture Models and EM

编程作业1:

1. 实现K-means方法
2. 在一组自己设计的人工数据上尝试K-means计算效果,并从不同角度分析算法特点
3. 在UCI数据集或其他实际数据集上尝试K-means方法