**聚类的好坏不存在绝对标准（幼儿园小朋友的拼图可以按大小/形状/颜色等方式划分）**

簇=类: 数据对象的集合

在**同一个类**中，数据对象是**相似**的

**不同类**之间的数据对象是**不相似**的

类不是事先给定的，而是**后来根据数据的相似性/距离来划分（无监督的算法）**

聚类算法：根据给定的**相似性评价标准**，将一个数据集合分组成几个聚类（簇）

每个样本表征为**特征向量**（**点**）

**特征**对聚类的影响很大（可以按事物的某一种或结合多种特性进行划分），所以需要**选取和设计特征**

**点与点之间的距离**可作为相似性度量依据

距离度量对聚类有很大影响（可以粗聚类也可以细聚类），所以需要**选取度量函数**

聚类分析是根据不同样本之间的差异，根据距离函数的规律（大小）进行模式分类（聚类）的。

一个好的聚类算法

类（簇）**内部高相似性**

类（簇）**之间低相似性**

聚类分析的有效性

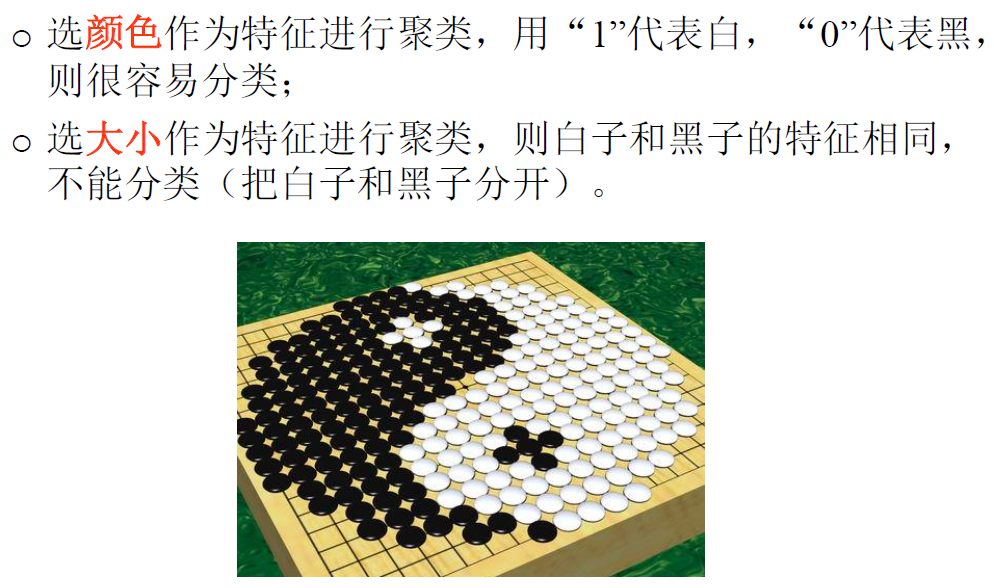
聚类分析方法是否有效与**向量点的分布形式**有很大关系。

若向量点的分布是**一群一群**的，同一群样本密集（距离很近），不同群样本距离很远，则很容易聚类；

若样本集的向量分布聚成一团，不同群的样本**混在一起**，则很难分类；

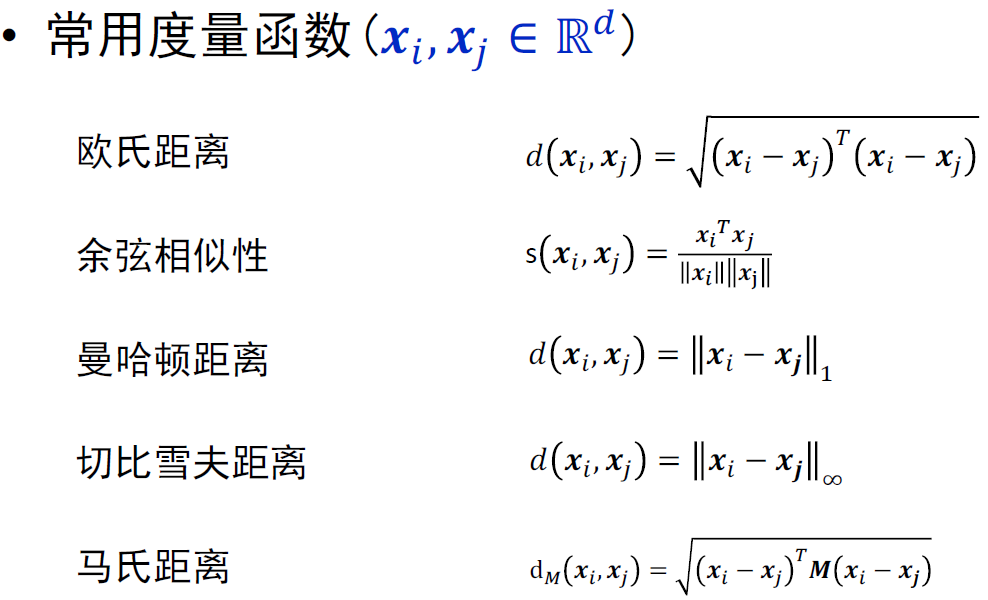
对具体数据做聚类分析的关键是**选取合适的特征**。特征选取得好，容易区分，选取得不好，很难区分。

**理解：特征的选取会影响向量点的分布形式**，如下：

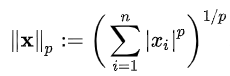


距离度量

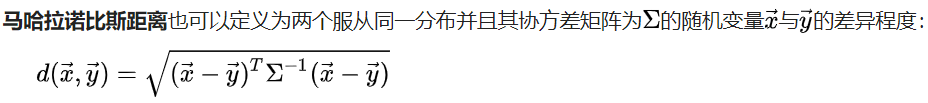
目的：度量**同一类样本间的类似性**和**不属于同一类样本间的差异性**。



注释：

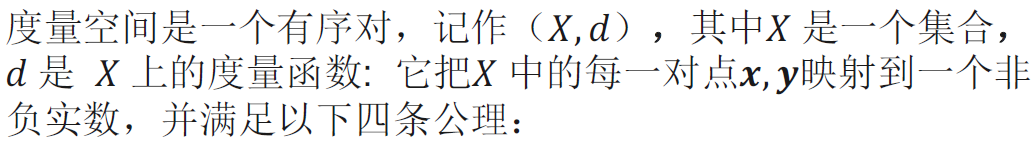
范数的计算：;

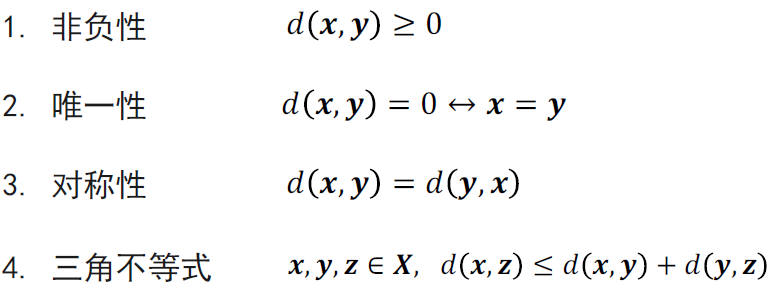
马氏距离：



度量函数和度量空间

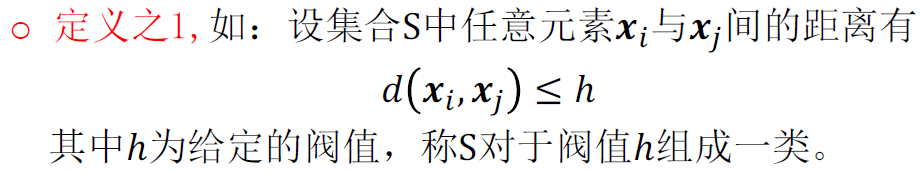
定义：





不满足的以上任意一条公理的度量只能算作**伪度量函数**，如KL散度

类的定义



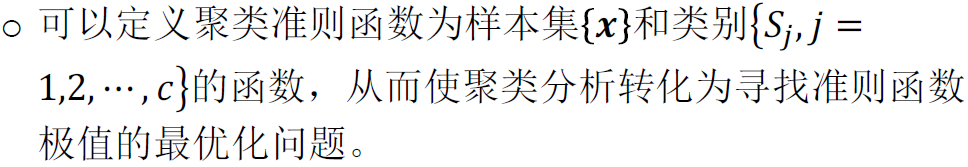
聚类准则

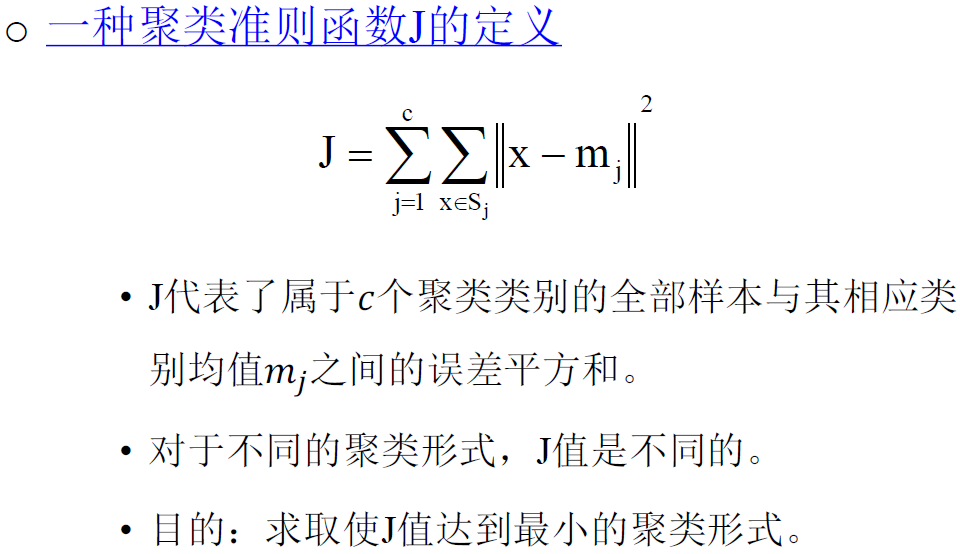
除了距离度量，还需要一种**基于数值的聚类准则**（比如阈值），才能将相似的样本分在同一类，相异的样本分在不同的类

**聚类准则函数**：一个能对聚类过程或聚类结果的优劣进行评估的准则函数。

聚类准则函数选择得好，**聚类质量就会高**。

聚类准则应是反映类别间**相似性**或**分离性**的函数。

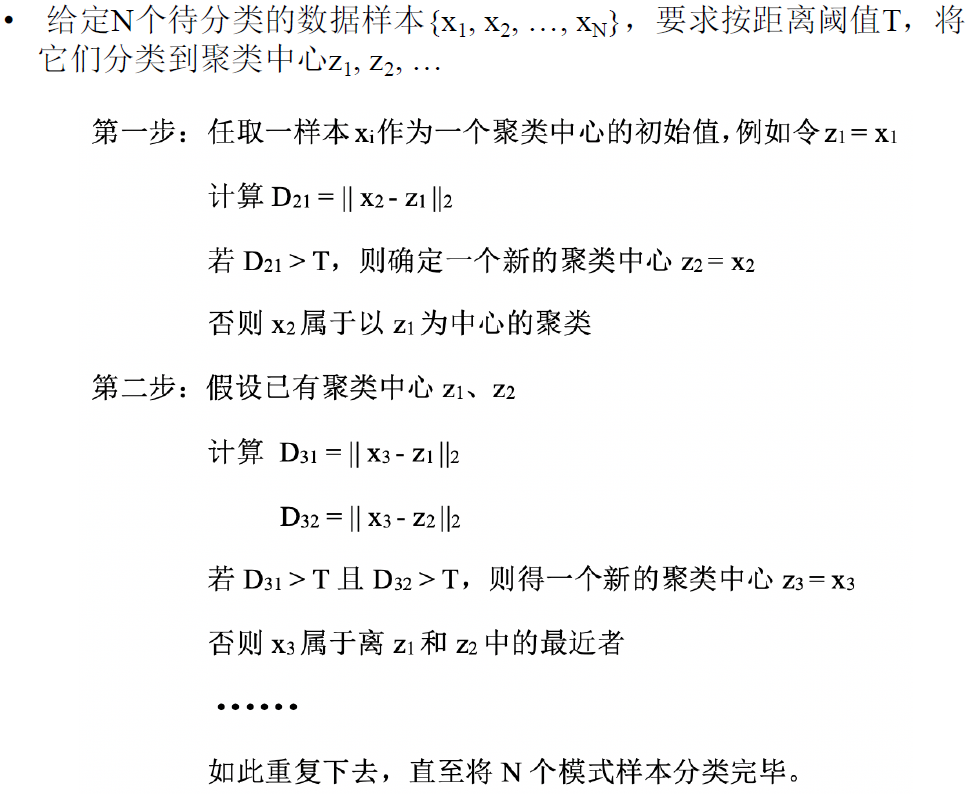


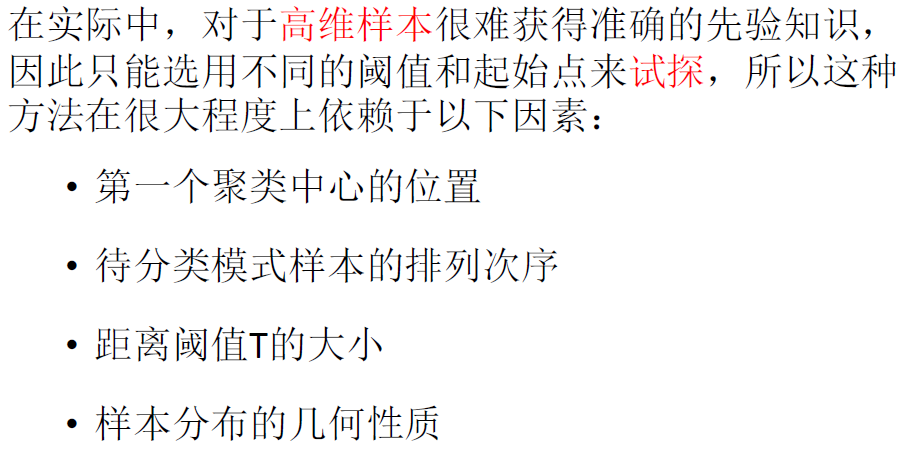


理解：只考虑**类内的相似性**，没有考虑**类间的分离性**

**基于试探**的聚类搜索算法：

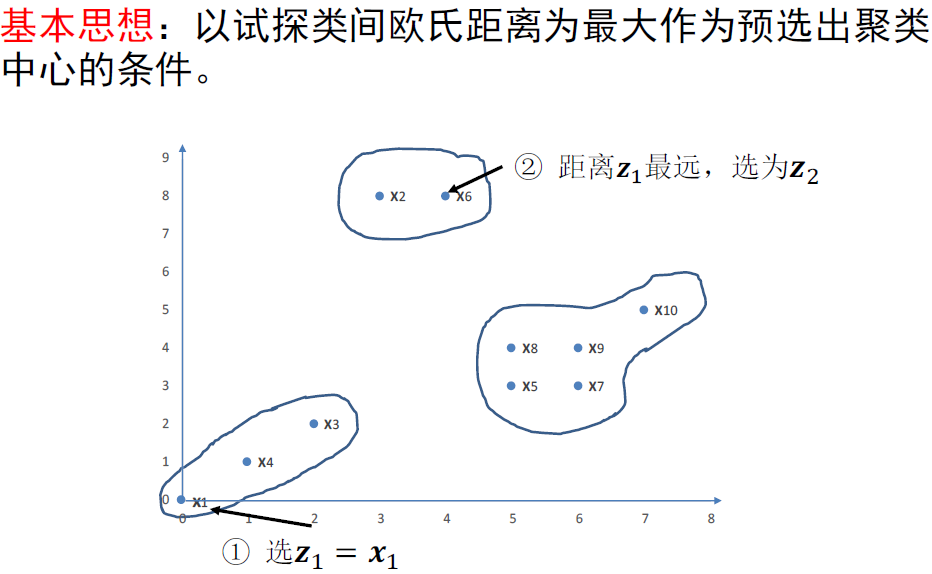
按**最近邻规则**的简单试探法

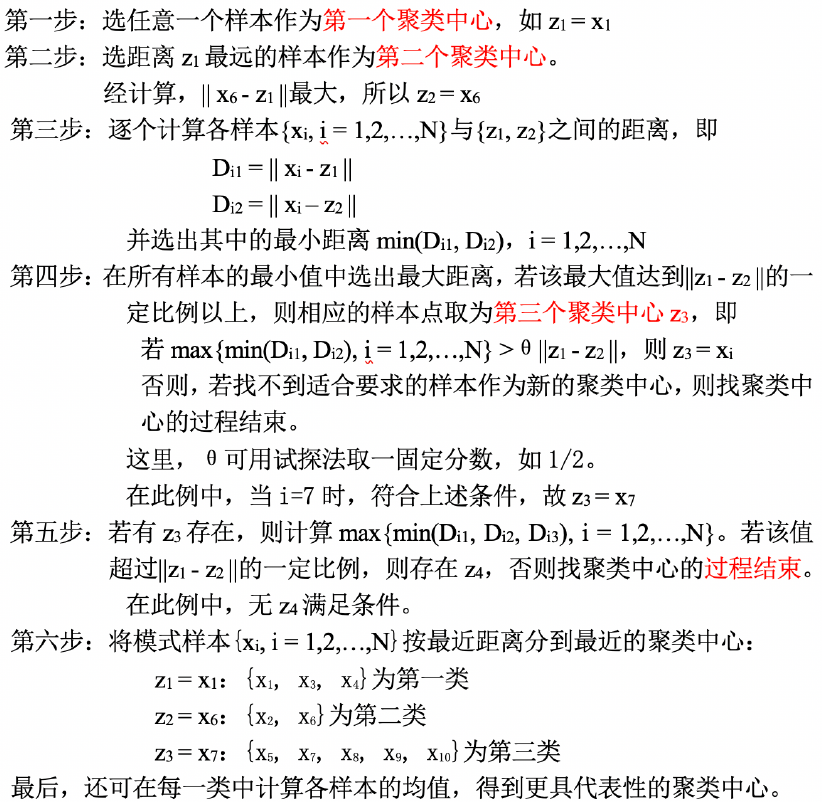




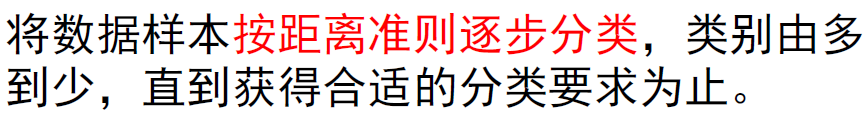
结论：这种方法很受限

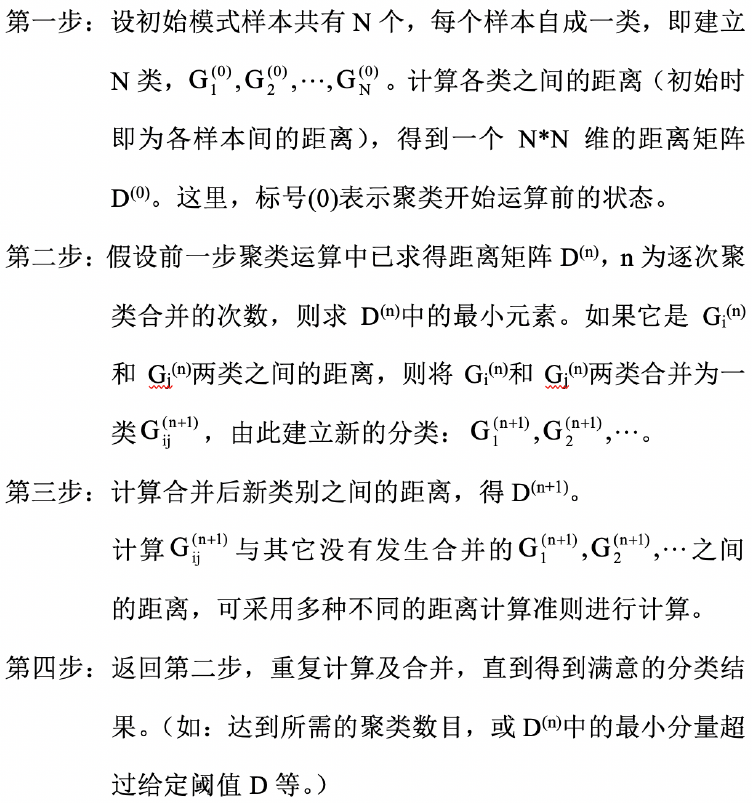
**最大最小距离**算法

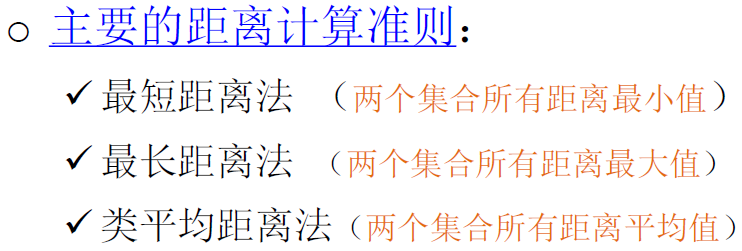




系统聚类法

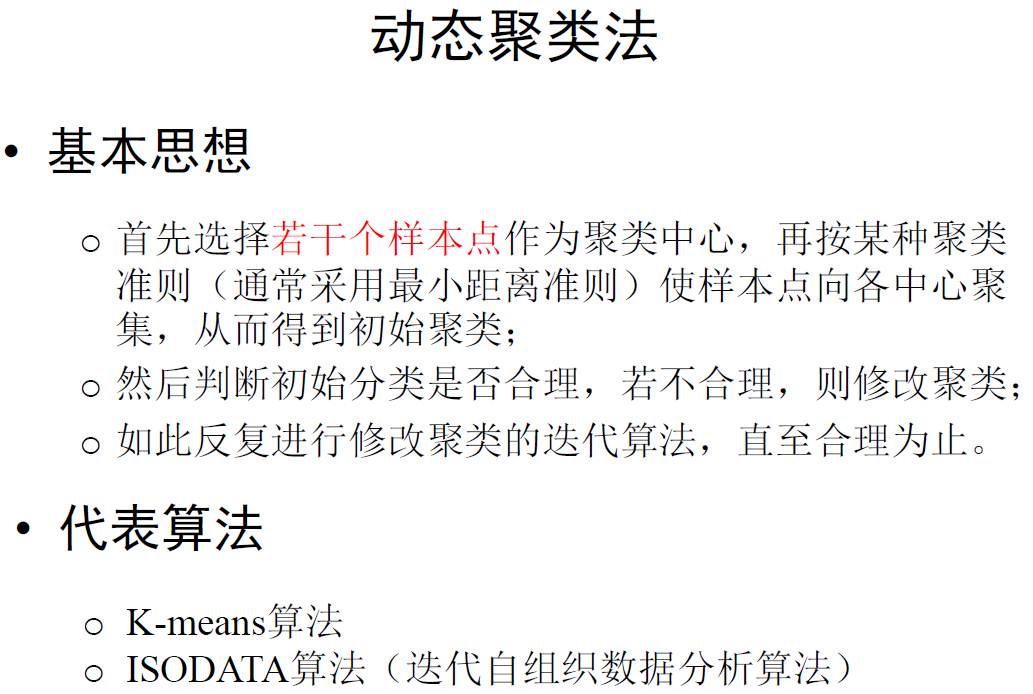




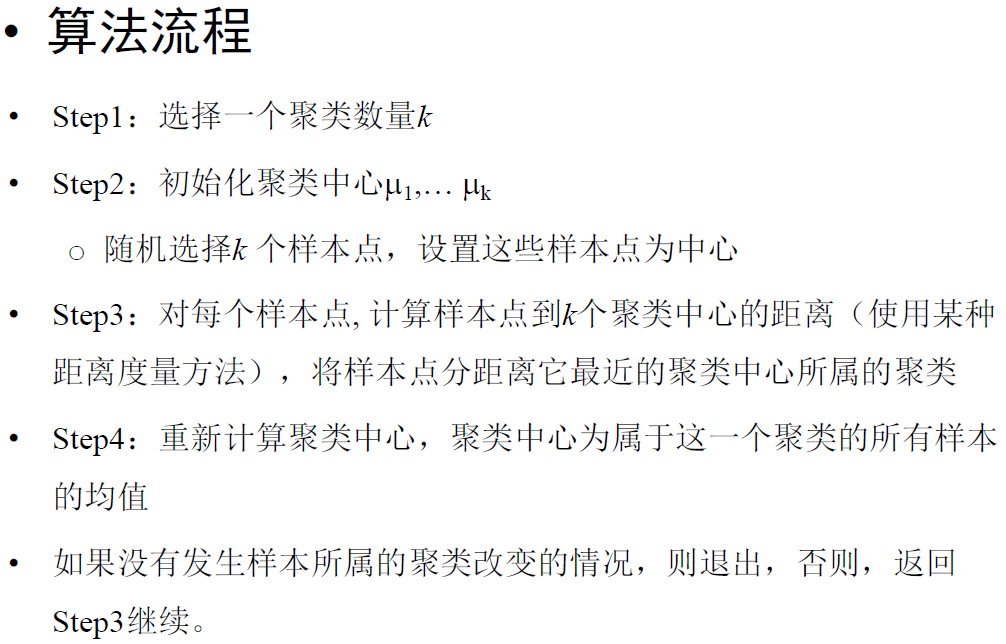


理解：**距离计算准则**是计算两个类之间而不是点之间的距离。

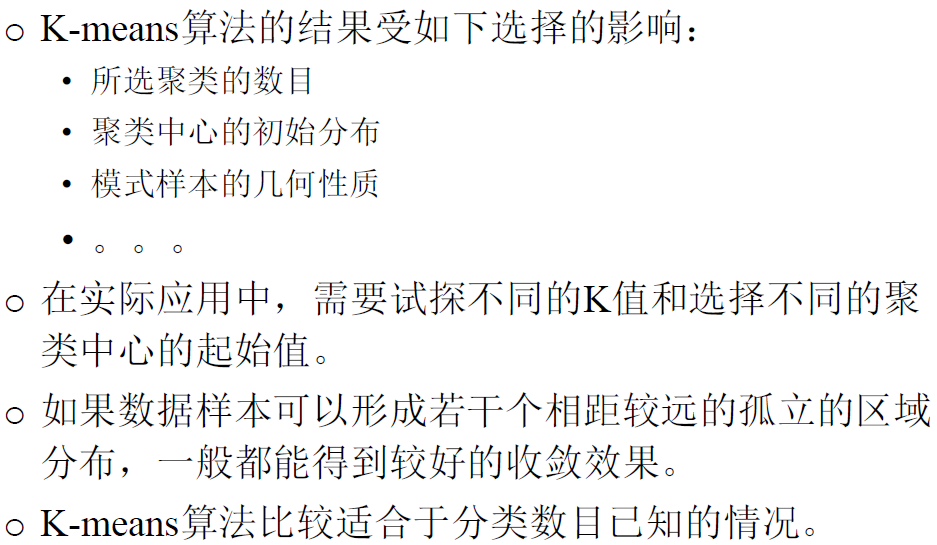
动态聚类法



**K-means算法**

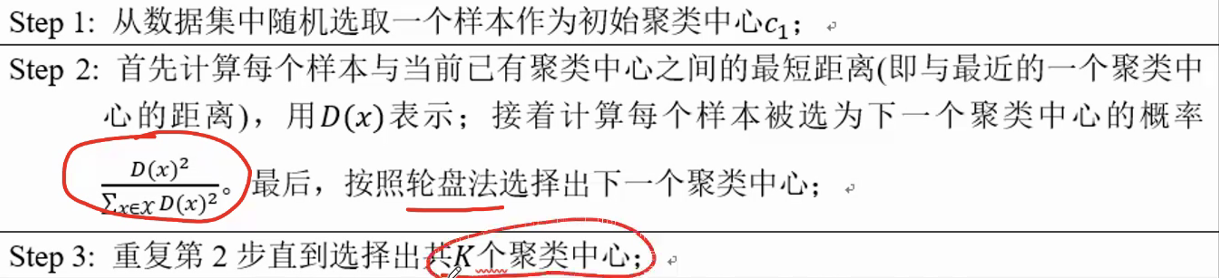


注意：聚类中心不落在样本点上

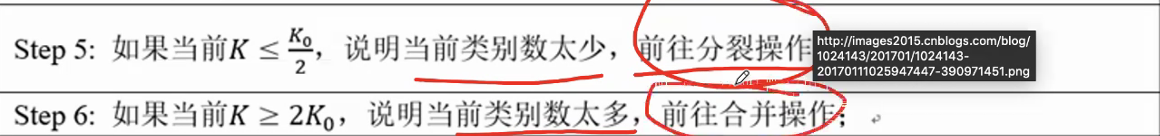


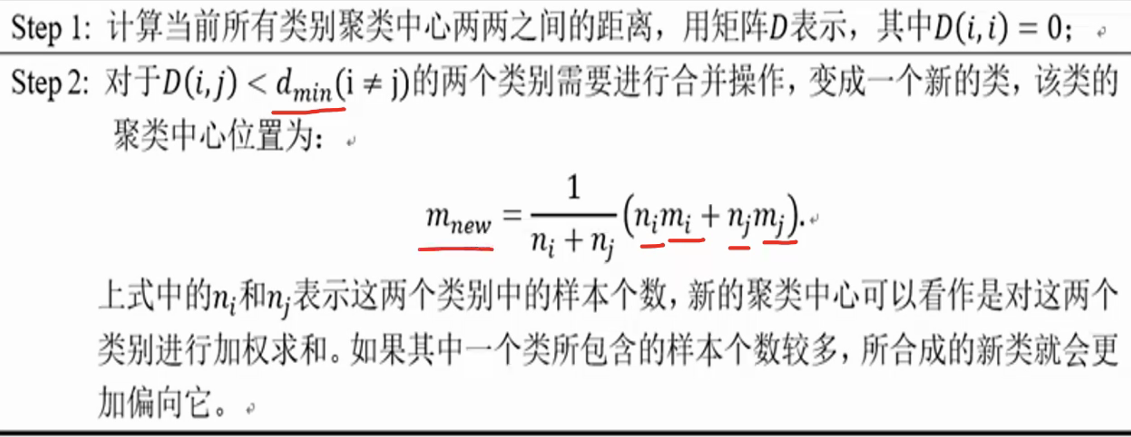
kmeans++

刚开始不是随机选取k个样本，而是

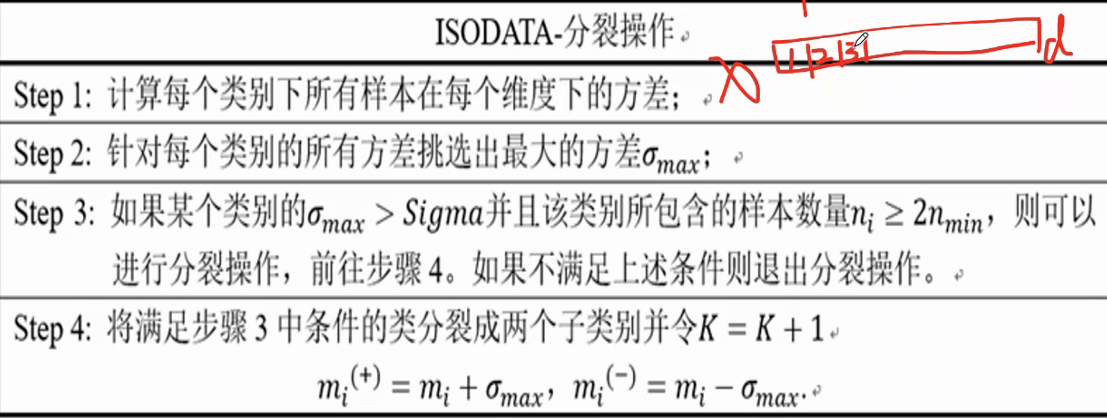


ISODATA的核心

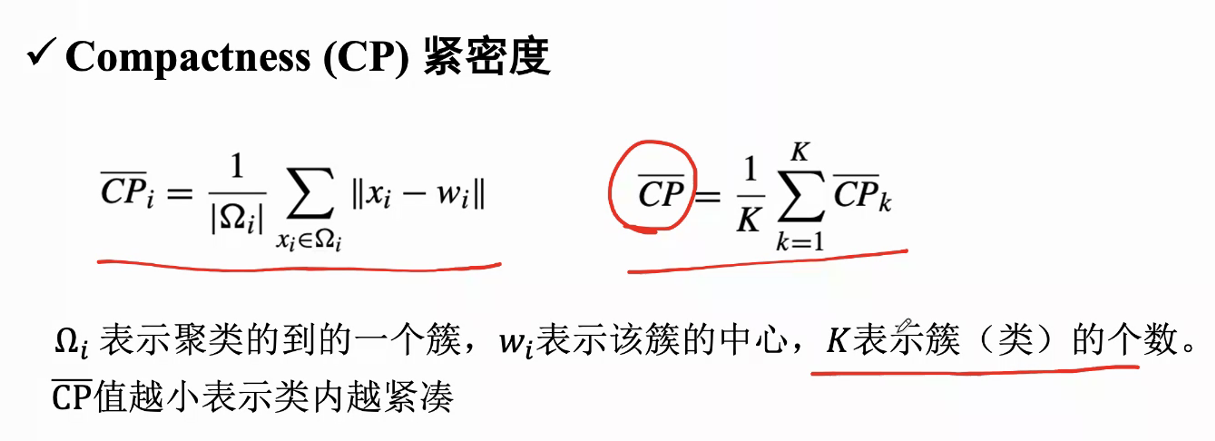




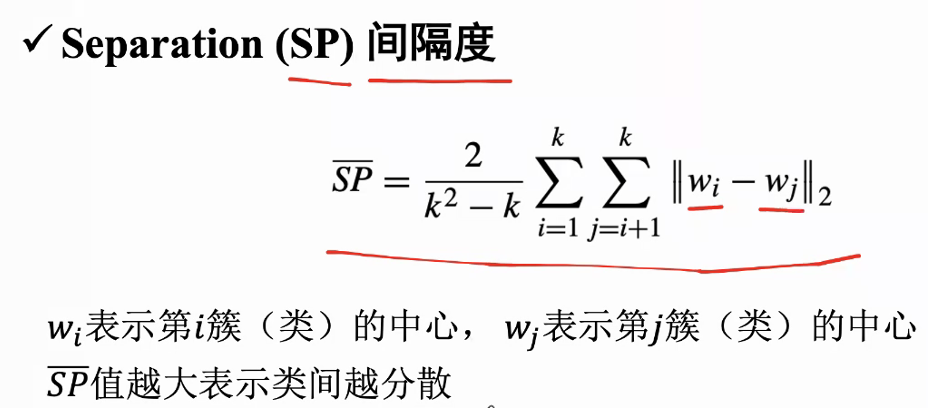
m是类的均值

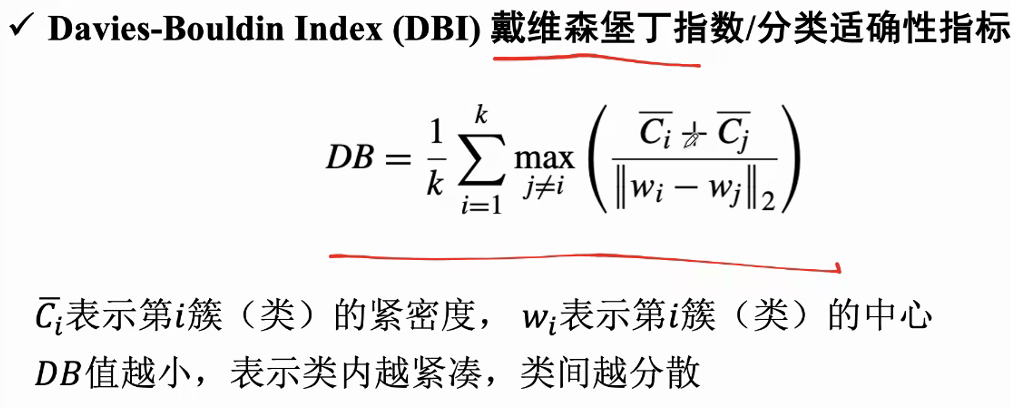






没有考虑类间





缺点：使用欧式距离

分子：任意两个簇之间的最小距离



