1. 自己提出的问题的理解（罗列全部）：
2. 提出的问题1：当使用SSE时，局部最小如何理解？为什么对大量数据集来说局部最小是不可行的？

讨论后的理解：局部最小值是在选取某一种初始种子的情况下达到的最优解，而全局最小值则是问题的最优解，问题的最优解一般不易得到。

1. 提出的问题2：为何距离函数需要有平方？是为了保证SSE可以用增量来计算吗？

讨论后的理解：若不平方则计算根号将增加计算量，平方后利于计算。而平方后不能用于增量计算，因为SSE中含有mj，每次计算时应重新计算mj，故不能达到计算增量的目的。

1. 别人提出的问题的理解（选择几个问题罗列，并给出理解）：
2. 问题3：如果一个聚类任务的数据中，其属性既有离散值又有连续值怎么办？是否可以将连续属性按照数据集中的数据离散化分为几个区间，将区间视为离散的继续做？

自己的理解：对于一个属性是离散的还是连续的在求距离时是无所谓的，不需要做过多的处理。若对于范畴属性，则可用k-模算法解决问题。

1. 问题4：我想问第一个是，为什么解决空聚类的时候，选择离一个含有大量数据的聚类的聚类中心最远的数据点？

自己的理解：远离一个含有大量数据的聚类，最远的数据点更贴近一个潜在的聚类中心，否则需要花费更多的时间去迭代初结果。

1. 问题5：在4.2.2节中提到，公式(2)的求和运算是分开计算的，那么这个分开计算是如何实现的呢？  
   自己的理解：先把一个数据从硬盘中加载到内存，将这个数据归入某一聚类（归类时用之前的聚类中心），再利用这一个数据的信息一步一步计算新的聚类中心。
2. （必填）读书计划
3. 本周完成的内容章节：第四章
4. 下周计划：第五章

四、读书摘要及理解或伪代码的具体实现（读书摘要、伪代码的具体实现代码等可以写到这个部分）

读书摘要及理解：

# 无监督学习

俱来是一种发现数据中的内在结构的技术。

可以把关联规则的挖掘看成是一种无监督学习。

## 4.1 基本概念

有划分聚类和层次聚类

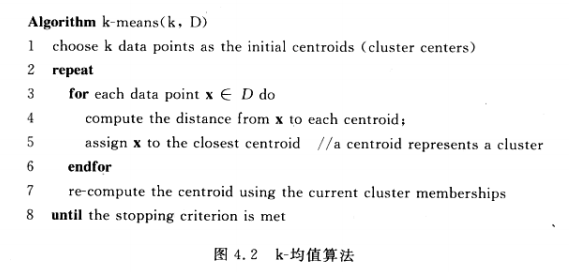
据类的目的就是通过使用某个聚类算法和某个距离函数来发现数据中内在的分组结构。

## 4.2 k-均值聚类

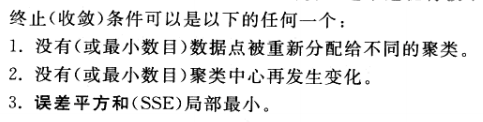
K-均值算法是最著名的划分聚类。

### 4.2.1 k-均值算法

算法：



终止条件：



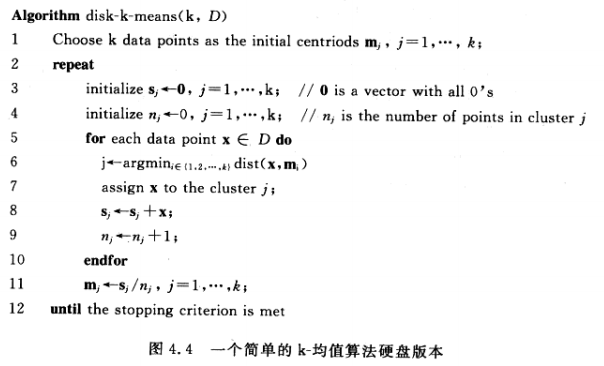
特殊问题：

可能出现没有任何数据分配给聚类中心成为**空聚类。**可以选择一个数据点替代这个聚类中心。

如果停止条件取决于SSE误差平方和，具有最大SSE的可以被用来寻找另外的聚类中心。

### 4.2.2 k-均值算法的硬板版本

应用：有时需要满足不把所有的数据集都加载到内存中，这在处理大规模数据时很有用。



（一个数据进入内存，增量地计算聚类中心，这一聚类的个数再加一）

若使用SSE，可以修改算法使SSE也成为增量式。

### 4.2.3 优势和劣势

优势：时间复杂度为O(tkn)（数据个数n，聚类个数k，循环次数t）

n远远大于k，t，所以认识此算法是线性的。

劣势：

1. 只能用于均值能被定义的数据集。难以应用到范畴数据，可用一种变体叫**k-模算法。**
2. 聚类的模是一个元组，由聚类中每个属性出现次数最多的值组成。距离为相匹配的属性数目。需事先指定聚类数目k
3. 对于异常值十分敏感。

处理异常值的方法：

1. 循环中监控异常值，决定是否去除。用阈值作为异常值的判定标准。
2. 随机采样。先用采样点进行预先聚类，再将其他数据点分配给这些聚类。
3. 把剩下的点分配个距离最近的点（最简单）
4. 利用聚好类的采样数据进行监督的学习，每个聚类被认为是一个类别，由分类器将剩下的点放入合适的聚类。
5. 利用采样数据产生的聚类作为种子进行**半监督的学习**（从少量有类别数据和大量的无类别数据中进行学习的学习模式）**。**
6. 对初始种子十分敏感。SSE作为终止条件时，只能达到局部最小值，对于大量数据时不可行的。

选择理想初始种子的方法：

1. 先计算全体数据的均值中心m，选择离m最远点x1作为第一个种子，距离x1最远点x2作为第二个种子，选择每个种子都距离以选择的种子距离和最大。（若有异常值则无效。可先采样，再利用采用数据选择初始种子）
2. 先采样，利用采样数据进行层次聚类，层次聚类的k个聚类中心作为种子。
3. 人工选择种子。
4. 不适用于发现形状不是超维椭圆体的聚类。

## 4.3 聚类的表示

### 4.3.1 聚类的一般表示方法

1. 利用聚类中心。可计算聚类的半径和标准差确定伸展度。用于高位球体，若被拉长了或者其他形状则不太合适。

2. 利用分类模型。

3. 利用最常见的值。对范畴属性时采用。（文本聚类）

### 4.3.2 任意形状的聚类

聚类结果优劣的判断适合具体应用相关的。

发现任意形状的聚类，可以采用近邻搜索算法（层次算法和基于密度的聚类方法），再空间和图像处理等中十分有用。