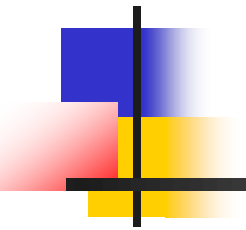


# K-means聚类

杜逆索

# K-means聚类算法

## 1.简介

K-means聚类算法就是基于距离的聚类算法
所谓的基于距离的聚类算法是指采用距离作为相似性度量的评价指标。

# K-means聚类算法

## 2. K-means聚类算法原理

K-means聚类算法的基本思想：

一、指定需要划分的簇的个数k值；

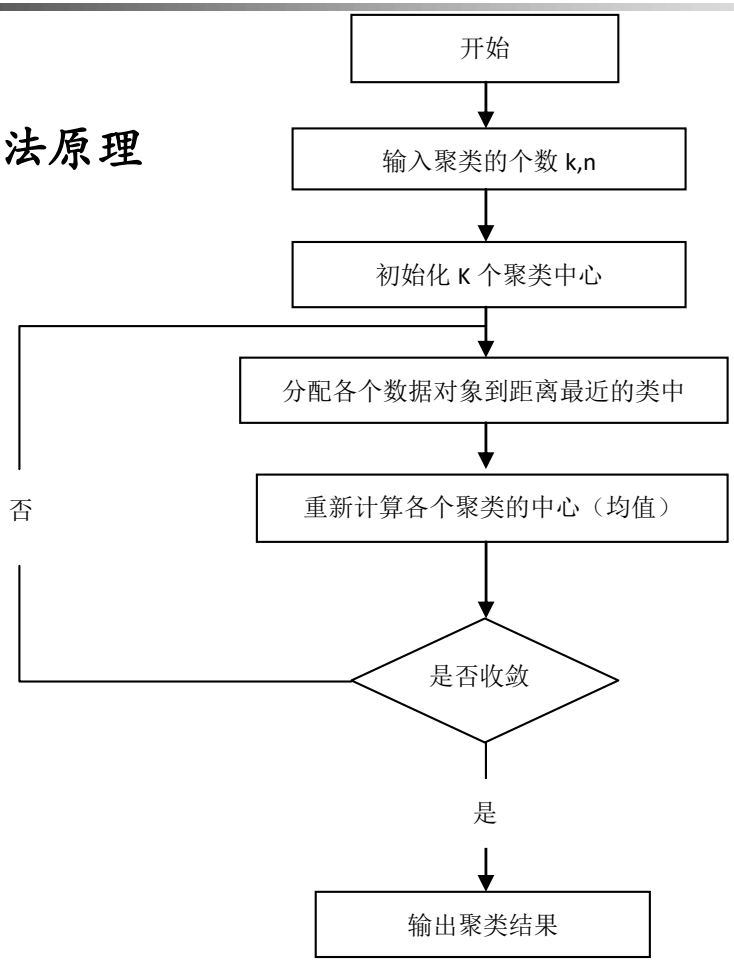
二、随机地选择k个初始数据对象点作为初始的聚类中心；

三、计算其余的各个数据对象到这k个初始聚类中心的距离，把数据对象划归到距离它最近的那个中心所处在的簇类中；

四、调整新类并且重新计算出新类的中心。

# K-means聚类算法

## 2. K-means聚类算法原理




K-means算法的工作流程

# K-means聚类算法

## 2. K-means聚类算法原理

K-Means算法的工作框架：

（1）给出 n 个数据样本,令 $I=1$,随机选择 K 个初始聚类中心 $Z_j(I)$, $j=1,2,3,\cdots,K$；

（2）求解每个数据样本与初始聚类中心的距离 $D\left(x_i,Z_j(I)\right),i=1,2,3,\cdots,n$

$j=1,2,3,\cdots,K$，若满足 $D\left(x_i,Z_j(I)\right)=\min\left\{D\left(x_i,Z_j(I)\right),i=1,2,\cdots,n\right\}$，那么 $x_i\in w_k$；

（3）令 $I=I+1$,计算新聚类中心 $Z_j(2)=\frac{1}{n}\sum_{i=1}^{n_j}x_i^{(j)}$， $j=1,2,\cdots,K$ 以及误差平方和准则函数 $J_c$ 的值： $J_c(2)=\sum_{j=1}^{K}\sum_{k=1}^{n_j}\left\|x_k^{(j)}-Z_j(2)\right\|^2$；

（4）判断:如果 $\left|J_c(I+1)-J_c(I)\right|<\xi$，那么表示算法结束,反之, $I=I+1$，重新返回第（2）步执行。

# K-means聚类算法

## 2. K-means聚类算法原理

K-Means算法的特点就是调整一个数据样本后就修改一次聚类中心以及聚类准则函数的值，当n个数据样本完全被调整完后表示一次迭代完成，这样就会得到新的簇和聚类中心的值。

K-Means聚类算法其本质是一个最优化求解的问题。

K-Means算法对聚类中心采取的是迭代更新的方法。

# K-means聚类算法

## 3 K-means聚类算法特点及应用

### 3.1 K-means聚类算法特点

优点：

（1）算法简单、快速。
（2）对处理大数据集，该算法是相对可伸缩的和高效率的。
（3）算法尝试找出使平方误差函数值最小的k个划分。

缺点：

（1）K-means聚类算法只有在簇的平均值被定义的情况下才能使用。
（2）要求用户必须事先给出要生成的簇的数目k。
（3）对初值敏感。
（4）不适合于发现非凸面形状的簇，或者大小差别很大的簇。
（5）对于“噪声”和孤立点数据敏感。

# K-means聚类算法

## 3 K-means聚类算法特点及应用

### 3.2 K-means聚类算法应用

（1）K-means 算法在散货船代货运系统中的应用

（2）K-Means 算法在客户细分中的应用

# K-means聚类算法

**4 小结**

本章详细地介绍了K-means算法的基本概念、基本原理,并介绍了该算法的特点和存在的缺陷,最后介绍了K-means算法的应用,从中可以看出K-means算法的应用非常广泛。