# KMEANS 算法

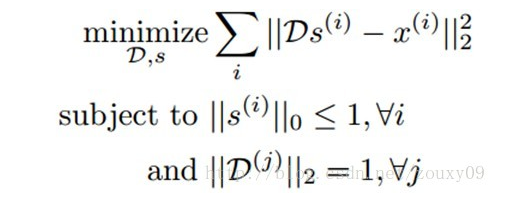
## 概述

无监督学习：没有分类标签，只有特征向量

【非监督学习】

从一组无标签数据中学习特征，然后用学习到的特征去提取有用的标签数据特征（一般存在很多hyper-parameter需要调整）

聚类算法通过最小化数据点/最近邻中心距离来寻找个各类的中心（矢量量化vector quantization）：构建一个字典D∈R，最小化重构误差。



寻找满足条件的字典D和每个样本对应的码矢量。

【聚类的预处理】

数据均值和对比度归一化

消除相关性

初始化聚类中心

监督学习：有分类标签，如朴素贝叶斯，SVM

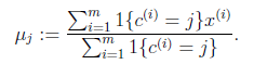
例子：三维空间样本的星星（聚类成星星团）

## 算法

1. 随机选取K个聚类质心点（cluster centroids聚类中心）分为
2. 重复下列过程到收敛
3. 对每个样例i，计算应该属于的类



1. 对每一类j，重新计算该类质心



K：事先给定的聚类数

c（i）：样例i与k类中最近的类，为k类中一个

uj:质心

随机选取k个宇宙中的点作为k星团质心，计算每个星星到质心的距离，然后选取距离最近的星团为c（j），

对每个星团，重新计算质心，重复迭代第一步第二步到质心不变。

如何保证收敛：

定义畸变函数（distortion function）



J表示每个样本点到质心的距离平方和，kmeans让j最小。

J为非凹函数，因此容易局部收敛，不能保证全局最优，需要假定一个极大似然估计P（X,Y）