KMeans理论与实践

# 基本介绍

## 概述

K-均值聚类是一种动态聚类的方法。其主要适用于分类问题。该算法给出一组对象（记录），聚类或分类的目标是把这些对象分割成组或集群，使得这些对象相比于组间，在组内更趋于相似。K-均值聚类是一种无监督学习的方法，因为不需要事先标记的数据。K-均值算法在实践中容易实施和运行，速度相对较快，算法内容也非常容易修改。

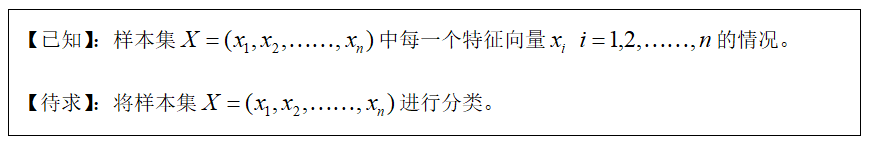
# KMeans详解

## 算法原理

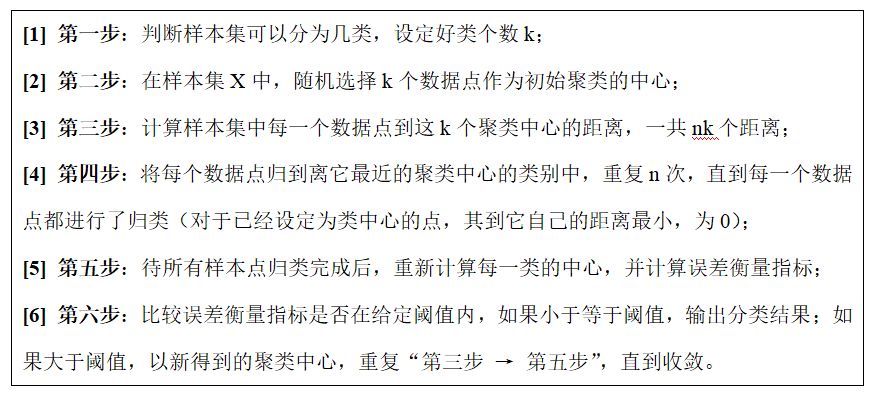
1. 随机选取k个中心点
2. 遍历所有数据，将每个数据划分到最近的中心点中
3. 计算每个聚类的平均值，并作为新的中心点
4. 重复2-3，直到这k个中线点不再变化（收敛了），或执行了足够多次的迭代

## 算法原理分步骤描述

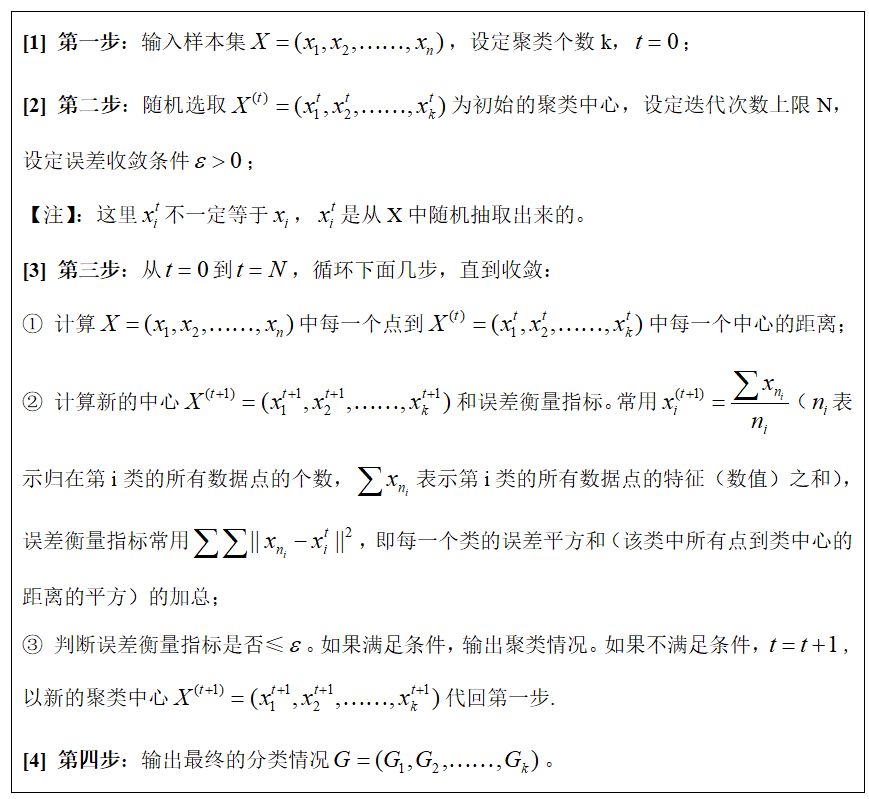
### 问题描述



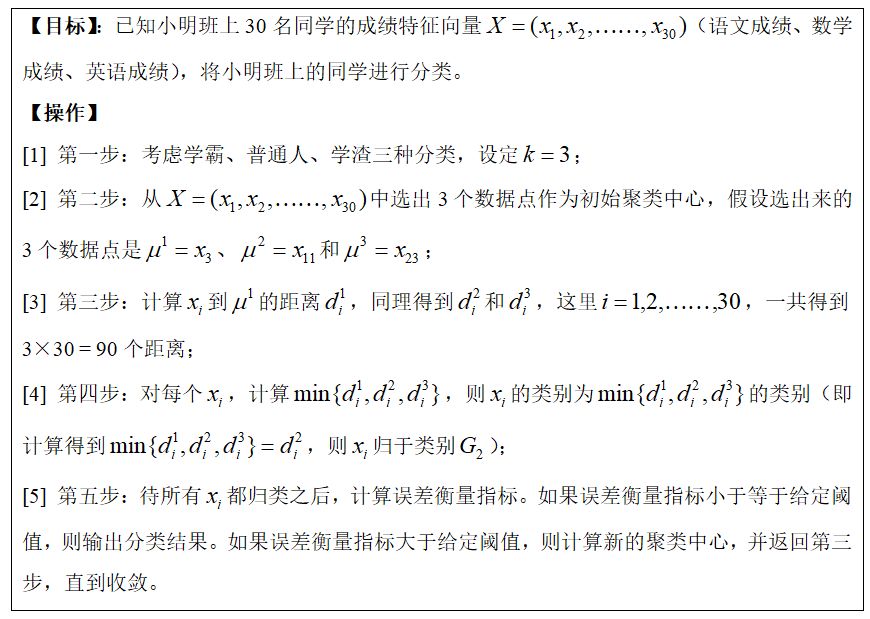
### 算法步骤（文字描述版）



### 算法步骤（数学描述版）



### 举例



## 优点

[1] 算法简单、迅速；

[2] 对于处理大数据集，该算法是相对可伸缩和高效的，因为它的复杂度大约是O (nkt)。其中n是所有对象的数目，k是分类的数目，t是迭代的次数，该算法经常以局部最优结束；

[3] 当类是密集、球状或者团状，且类与类之间区别明显时，该算法聚类效果很好。

## 缺点

[1] 该算法只有在类的平均值被定义的情况下才能使用，不适用于某些分类属性的数据；

[2] 对初值比较敏感，对于不同的初始值可能会导致不同的聚类结果；

[3] 不适合于发现非凸面形状的类，或者大小差别很大的类；

[4] 对于“噪声”和孤立点数据敏感，少量的该类数据能够对平均值产生极大影响。

## 注意事项

[1] 算法中的K值需要认真选取；

[2] 要慎重选取初始的聚类中心，如果选择不当可能很容易陷入局部最优；

[3] 样本要随机选取，可以提高算法的收敛速度。

# 实践

参考github：<https://github.com/jpegbert/MachineLearning/tree/master/kmeans>

参考：

<https://mp.weixin.qq.com/s/N-nk-1jcoGRTFEyygOrCkg>

<https://mp.weixin.qq.com/s/LtnpogzlO7fyUJzET2l8wQ>