

机器学习算法之 KMeans 聚类算法

作者: [Holyy](#)

原文链接: <https://ld246.com/article/1662962112346>

来源网站: [链滴](#)

许可协议: [署名-相同方式共享 4.0 国际 \(CC BY-SA 4.0\)](#)

一、K-Means 聚类算法原理

1. 算法思想

物以类聚，人以群分。

K-Means: 一种常见的无监督学习算法，名字叫做 K 均值算法。

是否为监督学习：只需要看输入的数据是否有标签。

K-Means 聚类算法是一种迭代求解的聚类分析算法。

算法思想是：

我们需要随机选择 K 个对象作为初始的聚类中心，然后计算每个对象和各个聚类中心之间的距离然后将每个对象分配给距离它最近的聚类中心。

聚类中心及分配给它们的对象就代表着一个聚类。每分配一个样本，聚类的中心会根据聚类中现的对象被重新计算。此过程将不断重复，直至满足设置的终止条件。

2. 算法步骤

ol>

li> 首先我们需要确定一个 k 值（随机），即我们希望数据经过聚类得到 k 不同的集合

li> 从给定的数据集中随机选择 K 个数据点作为质心

li> 对数据集中的每个点计算其与每一个质心的距离（比如欧式距离）；数据点离哪个质近，就划分到那个质心所属的集合

li> 第一轮将所有的数据归号集合后，一共有 K 个集合，然后重新计算每个集合的质心

li> 如果新计算出来的质心和原来的质心之间的距离小于某一个设置的阈值，则表示重新计算的质心位置变化不大，数据整体趋于稳定，或者说数据已经收敛。在这样的情况下，我们认为聚类效果已经到了期望的结果，算法可终止。

li> 反之，如果新质心和原来质心的距离变化很大，需要重复迭代 3-5 步骤直至位置变化不大，达到收敛状态。

3. 图解 K-Means

二、K-Means 聚类实战