一、什么是聚类

1.1 聚类的定义

聚类(Clustering)是按照某个特定标准(如距离)把一个数据集分割成不同的类或簇，使得同一个簇内的数据对象的相似性尽可能大，同时不在同一个簇中的数据对象的差异性也尽可能地大。也即聚类后同一类的数据尽可能聚集到一起，不同类数据尽量分离。

1.2 聚类和分类的区别

- **聚类(Clustering)**：是指把相似的数据划分到一起，具体划分的时候并

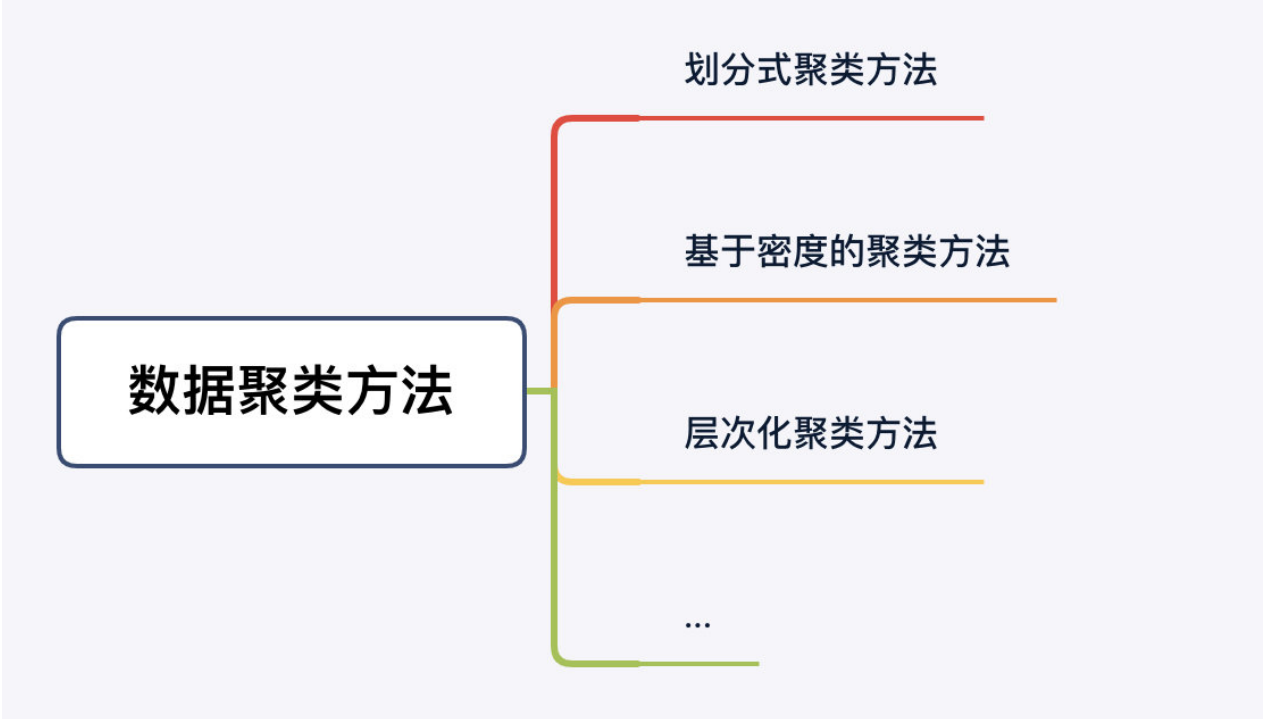
不关心这一类的标签，目标就是把相似的数据聚合到一起，聚类是一种**无监督学习(Unsupervised Learning)**方法。
- **分类(Classification)**：是需要标注数据是某种具体的类型，通过训练数据集获得一个分类器，再通过分类器去预测未知数据的过程，分类是一种**监督学习(Supervised Learning)**方法。

1.3 聚类的一般过程

1. 数据准备：特征标准化和降维
2. 特征选择：从最初的特征中选择最有效的特征，并将其存储在向量中
3. 特征提取：通过对选择的特征进行转换形成新的突出特征
4. 聚类：基于某种距离函数进行相似度度量，获取簇
5. 聚类结果评估：分析聚类结果

二、数据聚类方法

数据聚类方法主要可以分为**划分式聚类方法(Partition-based Methods)**、**基于密度的聚类方法(Density-based methods)**、**层次化聚类方法(Hierarchical Methods)**等。



2.1 划分式聚类方法

划分式聚类方法需要事先指定簇类的数目或者聚类中心，通过反复迭代，直至最后达到"簇内的点足够近，簇间的点足够远"的目标。经典的划分式聚类方法有**k-means**及其变体**k-means++**、**k-medians**、**kernel k-means**。

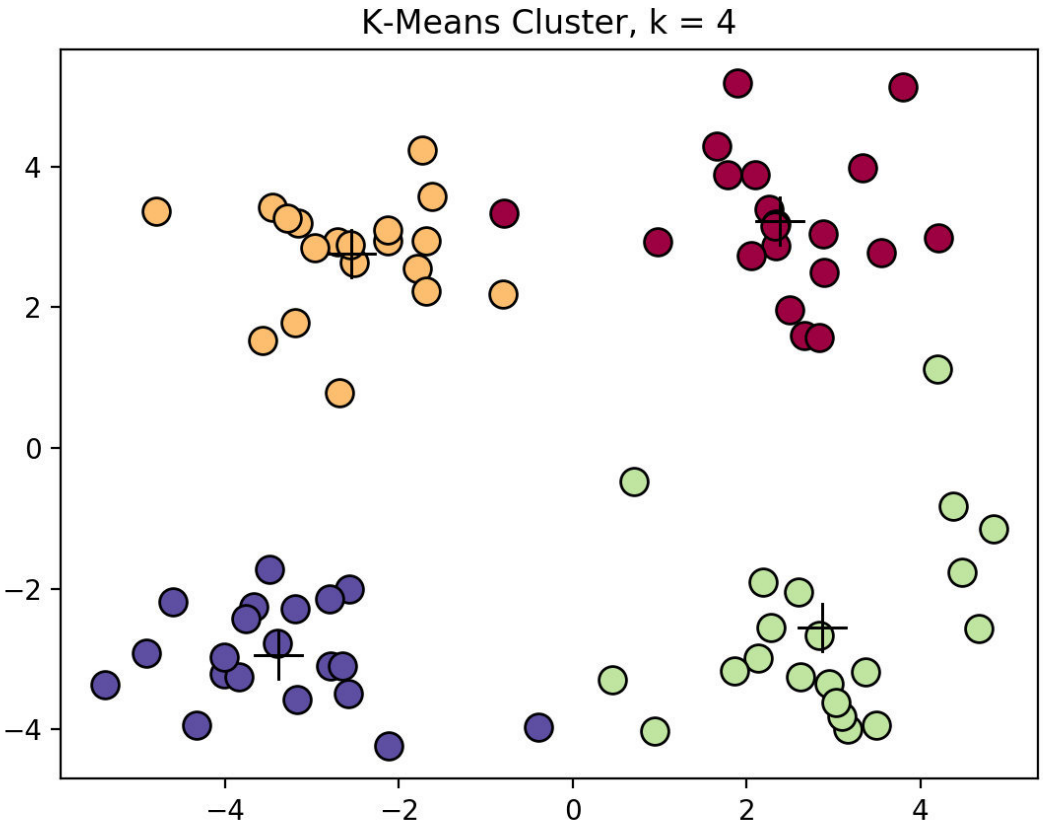
2.1.2 经典k-means算法

经典的k-means算法的流程如下：

> 1. 创建k个点作为初始质心(通常是随机选择)
> 2. 当任意一个点的簇分配结果发生改变时
>   1. 对数据集中的每个数据点
>   1. 对每个质心
>   1. 计算质心与数据点之间的距离
>   2. 将数据点分配到距其最近的簇
>   2. 对每个簇，计算簇中所有点的均值并将均值作为质心

经典k-means源代码:

测试数据效果



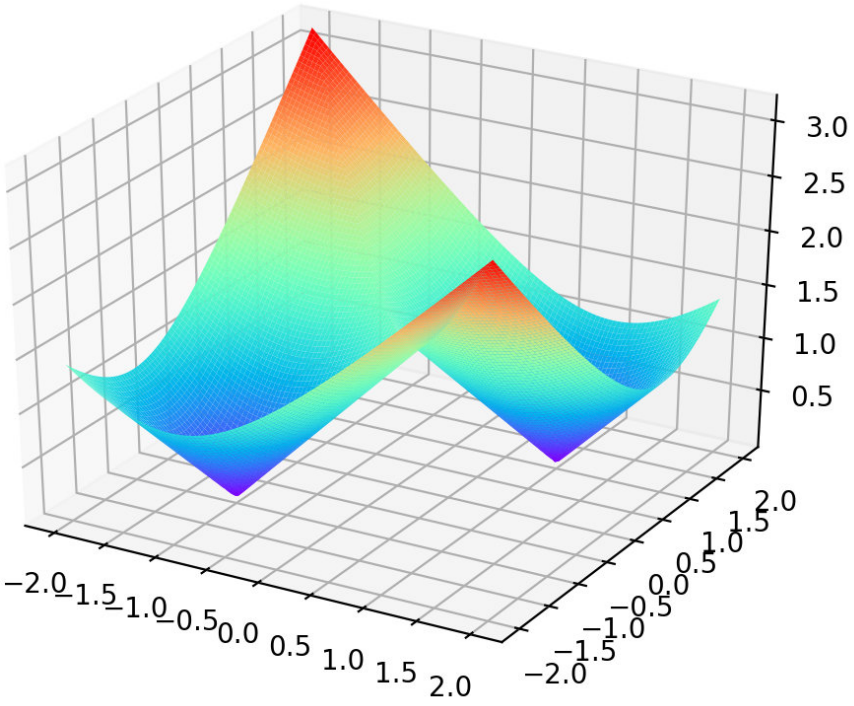
看起来很顺利，但事情并非如此，我们考虑k-means算法中最核心的部分，假设$x_i(i=1,2,\ldots,n)$是数据点，$\mu_j(j=1,2,\ldots,k)$是初始化的数据中心，那么我们的目标函数可以写成

$$\min \sum_{i=1}^{n} \min_{j=1,2,\ldots,k} ||x_i - \mu_j||^2$$

这个函数是非凸优化函数，会收敛于局部最优解，可以参考证明过程。举个🌰，$\mu_1 = [1,1], \mu_2 = [-1,-1]$，则

$$z = \min_{j=1,2} ||x_i - \mu_j||^2$$

该函数的曲线如下图所示



可以发现该函数有两个全局最优点，当时初始质心点取值不同的时候，最终的聚类效果也不一样，接下来我们看一个具体的实例。

2.2 基于密度的方法

2.3 层次化聚类方法

2.4 新方法

2.5 聚类方法比较

三、分布式聚类配置方法

四、小区融合项目应用

4.1 聚类方法的选取

4.2 聚类效果

五、参考文献

[1] https://www.zhihu.com/question/34554321

[2] T. Soni Madhulatha.AN OVERVIEW ON CLUSTERING METHODS