

第九章 K-means聚类

一、聚类

1. 基本步骤

数据预处理——>选择相似度度量标准——>聚类——>分析

2. 什么是聚类

无监督学习——用于处理无类标号数据

根据数据的相似度和特有的模式划分数据

聚类的目的是为了将数据点划分为若干组，每组内数据点相互可比，组间数据点具有一定差异。它本质上是根据事物的相似性和差异度来划分的。

3. 聚类的应用

近年来聚类被广泛被各种行业应用

- 生物
- 医学图像
- 市场调查
- 推荐系统

二、基础思想

1. K-means的基本思想

K-means聚类会通过数据点到聚类重心的聚类来将数据点划分到K个聚类

将无标号数据划分到聚类中，在同一个聚类中数据点之间彼此相似，在不同聚类之间彼此存在差异

在聚类的时候使用的唯一信息就是样本之间的相似度

目标是将样本划分为k个部分

2. 衡量距离的方法

- 只能用欧几里得距离 $d(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$
- 针对非线性需要核化距离

三、K-means算法步骤

针对输入的n个样本和划分数量k

- 定义聚类重心：聚类重心的取值是这个聚类中数据点属性的均值
- 初始化聚类中心：使用的方法包含随机初始化，以及任意选择k个样本作为聚类重心
- 迭代：将每个样本分配给聚类最近的聚类中心，然后根据聚类样本点的属性更新聚类中心
- 收敛标准：聚类的中心点不再发生变化，或者说达到了最大循环步数

四、选择K值

1. 衡量模型性能的两个指标

- **肘部法**

核心思想：随着K的增大，聚类会变得更加精细，但是下降到一定程度后收益变少。

肘部法计算的是每个样本点到它所属的聚类中心的距离平方和

计算公式：

$$SSE = \sum_{j=1}^K \sum_{x_i \in C_j} \|x_i - \mu_j\|^2$$

其中 C_j 代表所属的簇， x_i 代表簇内的样本点， μ_j 代表簇的中心点

根据改变K计算SSE可以找到明显的拐点，这个拐点叫做肘部点，此时的K对应模型性能最好

- **轮廓系数**

轮廓系数可以用于衡量聚类结果的好坏，不仅可以选择K，也可以来查看聚类质量
计算公式：

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

其中

平均簇内距离： $a(i)$ 表示样本*i*与所在簇的其他样本的平均距离

平均最近簇距离： $b(i)$ 表示样本*i*与最近的其他簇的样本的平均距离

$$b(i) = \min(d(i, C_j)), \text{ where } x_i \in C_j$$

这里的 $d(i, C_j)$ 代表样本*i*与另外一个簇 C_j 中所有样本的平均距离

轮廓系数的取值范围是

$$-1 \leq s(i) \leq 1$$

轮廓系数为0时代表聚类是重叠的，只有一个类别

轮廓系数为1时代表聚类非常紧密并且完美分开

聚类系数为-1时代表聚类效果非常差，簇内彼此没有相似度

当聚类系数小于0的时候代表部分样本点可能被错误分类

五、总结

如果数据点的密度存在显著差异，K-means 算法可能会倾向于处理密度较高的簇，而忽略密度较低的簇，从而导致聚类结果不均衡。

优点	缺点
简单且适用于常规的不重叠集群	需要事先确定聚类中心的数量 K
迅速汇聚	数据中的噪声和异常值会降低聚类的准确性
适用于凸形物体	非凸形状的不良表现
	由于其具有凸面的形状 K , 具有不同密度的数据点会影响聚类的准确性,
	K-means算法对聚类中心的初始值非常敏感。 可能会陷入局部最优解状态 收敛速度太慢 可能需要进行多次初始化操作