Clustering

Mathematical Modeling

Prof. Dr. Jingzhi Li

Department of Mathematics,
Southern University of Science and Technology

2025 Spring

1. Clustering
2. K-means

什么是聚类

聚类或聚类分析（Clustering）是一种在机器学习和数据分析中使用的无监督学习方法，目标是将数据集中的样本分为几个类（称为簇），使得每一类内部样本的特征都尽可能相近。

聚类的应用

- 探索性数据分析
- 在半监督学习中，用作有监督学习之前的预处理步骤
- 异常检测
- 图像分割

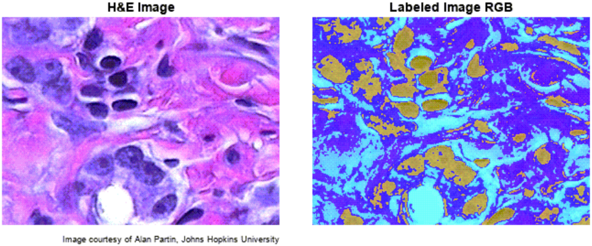

图 1: 左图：用苏木精-伊红染色的组织的原始图像。右图：MATLAB 将图像分成三个簇，从而将组织分割为三个类。

常用的聚类类型

- 划分聚类
  - K-Means 聚类：是一种最常见的划分聚类算法。它将数据划分为 K 个簇，通过不断迭代更新簇中心，使得每个数据点到其所属簇中心的距离之和最小。
  - K-Medoids 聚类：与 K-Means 类似，但它选择簇中的实际数据点作为簇中心（medoids），而不是计算平均值。
- 层次聚类
  - 凝聚层次聚类：从每个数据点作为一个单独的簇开始，逐步合并最相似的簇，直到达到预设的簇数量或满足其他终止条件。
  - 分裂层次聚类：从所有数据点作为一个簇开始，逐步分裂簇，直到每个数据点都成为一个单独的簇。
- 密度聚类
  - DBSCAN（Density-Based Spatial Clustering of Applications with Noise）：基于数据点的密度进行聚类，能够发现任意形状的簇，并识别噪声点。

聚类的工作原理

1. 数据准备
2. 定义相似性度量
3. 选择正确的聚类算法
4. 评估和细化聚类

原理

其目标是将数据分成 K 个相互独立、不重叠且方差相等的簇，并最小化一种称为惯性（inertia）或簇内平方和（within-cluster sum-of-squares, WCSS）的准则。

$$\sum_{i=0}^{n} \min_{\mu_j \in C} \left(\|\mathbf{x}_i - \mu_j\|^2 \right)$$

算法

1. 选择簇的数量，即 K 值。
2. 打乱数据集并随机选择 K 个数据点作为簇的中心（质心）来初始化中心点。
3. 将每个数据点分配到距离最近的质心所属的簇中。
4. 通过计算分配到每个簇的所有数据点的均值来更新质心位置。
5. 重复步骤 3 和 4，直到达到设定的迭代次数，或者质心在连续迭代之间的变化趋于稳定。

An Easy Example

假设有以下四个点，每个点是二维坐标：

$$X = (2, 10), (2, 5, (8, 4), (5, 8)$$

我们设定 K 值（簇的个数）$K = 2$, 并随机构造初始簇中心：

$$C_1 = (2, 10), C_2 = (5, 8)$$

第一轮迭代：
我们使用欧几里得距离公式计算数据点到簇中心的距离：

$$d(A, B) = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$$

数据点	到 $C_1=(2,10)$ 的距离	到 $C_2=(5,8)$ 的距离	分配的簇
(2,10)	$\sqrt{(2-2)^2+(10-10)^2}=0$	$\sqrt{(2-5)^2+(10-8)^2}=\sqrt{9+4}=\sqrt{13}\approx 3.61$	C₁
(2,5)	$\sqrt{(2-2)^2+(5-10)^2}=\sqrt{0+25}=5$	$\sqrt{(2-5)^2+(5-8)^2}=\sqrt{9+9}=\sqrt{18}\approx 4.24$	C_2
(8,4)	$\sqrt{(8-2)^2+(4-10)^2}=\sqrt{36+36}=\sqrt{72}\approx 8.49$	$\sqrt{(8-5)^2+(4-8)^2}=\sqrt{9+16}=\sqrt{25}=5$	C_2
(5,8)	$\sqrt{(5-2)^2+(8-10)^2}=\sqrt{9+4}=\sqrt{13}\approx 3.61$	$\sqrt{(5-5)^2+(8-8)^2}=0$	C_2

新的簇分配如下：

$$C_1:(2,10),\ C_2:(2,5),(8,4),(5,8)$$

重新计算新的簇中心：

$$C_{\text{new}}=\frac{1}{N}\sum_{i=1}^{N}X_i$$

新的簇中心：

$$C_{1,\text{new}} = (2, 10),\ C_{2,\text{new}} = (\frac{2+8+5}{3}, \frac{5+4+8}{3}) = (5, 5.67)$$

第二轮迭代：

数据点	到 $C_1 = (2, 10)$ 的距离	到 $C_2 = (5, 5.67)$ 的距离	分配的簇
(2,10)	0	$\sqrt{(2-5)^2 + (10-5.67)^2} = \sqrt{9 + 18.22} = \sqrt{27.22} \approx 5.22$	C_1
(2,5)	$\sqrt{(2-2)^2 + (5-10)^2} = 5$	$\sqrt{(2-5)^2 + (5-5.67)^2} = \sqrt{9 + 0.4489} = \sqrt{9.45} \approx 3.07$	C_2
(8,4)	$\sqrt{(8-2)^2 + (4-10)^2} = 8.49$	$\sqrt{(8-5)^2 + (4-5.67)^2} = \sqrt{9 + 2.78} = \sqrt{11.78} \approx 3.43$	C_2
(5,8)	$\sqrt{(5-2)^2 + (8-10)^2} = 3.61$	$\sqrt{(5-5)^2 + (8-5.67)^2} = \sqrt{0 + 5.38} = \sqrt{5.38} \approx 2.32$	C_2

新的分配仍然是

$$C_1 : (2, 10),\ C_2 : (2, 5), (8, 4), (5, 8)$$

由于分配没有发生变化，算法收敛，K-Means 终止。

通过 MATLAB 实现 Kmeans

1. 使用内置的 kmeans 函数。
2. 使用数据聚类实时编辑器任务以交互方式执行 k 均值聚类和层次聚类。

K-means 算法的特点

1. 对 K 值的选取敏感：需要预先选取 K 值，而实际的数据的最佳 K 值难以事先确定。
   1. Elbow Method(手肘法)
   2. Silhouette Analysis(轮廓系数法)
2. 对初始值敏感：不同的初始质心可能会导致不同的聚类结果，因此 K-Means 可能会陷入局部最优解，影响稳定性。
   1. K-means++
3. 不适用于各向异性数据或方差不等的数据。

评估聚类质量-轮廓系数

轮廓系数（Silhouette Coefficient）是一种用于评估聚类质量的指标。它用于衡量数据点在其所属簇中的紧密程度（同簇内相似性）以及不同簇之间的分离程度（簇间分离度）。
轮廓系数的取值范围为 $[-1, 1]$，数值越大，说明聚类效果越好。
计算公式: 对于数据集中的每个样本点 i，轮廓系数 $s(i)$ 为

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

其中 $a(i)$ 为簇内平均距离，$b(i)$ 为簇间最小平均距离。
在 MATLAB 中，可以使用 silhouette() 函数计算轮廓系数。