1 K-means 聚类算法详解

1.1 算法基本思想

给定数据集 $X = \{x_1, x_2, \ldots, x_n\}$, 其中 $x_i \in \mathbb{R}^d$，目标是将数据划分为 K 个簇 $C_1, C_2, \ldots, C_K$。

1.2 数学原理

1.2.1 目标函数

最小化簇内平方误差和：

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

其中 μ_k 是第 k 个簇的中心：

$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i$$

1.2.2 距离度量

使用欧氏距离：

$$d(x_i, \mu_k) = \|x_i - \mu_k\| = \sqrt{\sum_{j=1}^{d} (x_{ij} - \mu_{kj})^2}$$

1.3 算法步骤

1.4 数值示例（1D数据）

数据点：$X = \{1, 2, 8, 9\}$，$K = 2$

1.4.1 迭代过程

- **初始化：** $\mu_1^{(0)} = 2, \mu_2^{(0)} = 9$
- **迭代1 - 分配：**

$$\text{簇1：}\{1, 2\}$$
$$\text{簇2：}\{8, 9\}$$

Algorithm 1 K-means 聚类算法

Require: 数据集 X, 簇数 K, 最大迭代次数 max_iter, 容差 tol

Ensure: 簇标签 $labels$, 簇中心 μ

初始化：随机选择 K 个初始中心 $\mu_1^{(0)}, \ldots, \mu_K^{(0)}$
for $t = 1$ to max_iter **do**
    分配步骤：
    for 每个样本 x_i **do**
        $labels[i] \leftarrow \arg\min_k \|x_i - \mu_k^{(t-1)}\|^2$
    end for
    更新步骤：
    for 每个簇 $k = 1$ to K **do**
        if 簇 k 非空 **then**
            $\mu_k^{(t)} \leftarrow \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i$
        else
            $\mu_k^{(t)} \leftarrow \text{handle\_empty\_cluster}()$
        end if
    end for
    收敛判断：
    if $\max_k \|\mu_k^{(t)} - \mu_k^{(t-1)}\| < tol$ 或标签未变化 **then**
        break
    end if
end for
return $labels$, $\mu^{(t)}$

- **迭代1 - 更新：**

$$\mu_1^{(1)} = \frac{1+2}{2} = 1.5$$
$$\mu_2^{(1)} = \frac{8+9}{2} = 8.5$$

- **迭代2 - 分配：** 标签不变，算法收敛

1.5 终止条件

$$\begin{aligned} \text{中心变化：} & \max_k \|\mu_k^{(t)} - \mu_k^{(t-1)}\| < tol \\ \text{标签不变：} & labels^{(t)} = labels^{(t-1)} \\ \text{最大迭代：} & t \geq max\_iter \\ \text{目标函数变化：} & J^{(t-1)} - J^{(t)} < \epsilon \end{aligned}$$

1.6 复杂度分析

- 每次迭代复杂度：$O(nKd)$
- 空间复杂度：$O(n(d+K))$

1.7 边界情况处理

1.7.1 空簇处理策略

- 重新初始化中心到随机数据点
- 选择距离当前中心最远的数据点
- 删除空簇，$K \leftarrow K-1$

1.7.2 初始化改进（K-means++）

1. 随机选择第一个中心 μ_1
2. 对于 $k = 2$ 到 K：

   $D(x)^2 = \min_{\mu \in M} \|x - \mu\|^2$

   $P(x) = \frac{D(x)^2}{\sum_{x \in X} D(x)^2}$

   按概率 $P(x)$ 选择下一个中心

1.8 算法特性

优点	缺点
简单易实现	需要预设 K 值
计算效率高	对异常值敏感
适合球形簇	对初始中心敏感
可扩展性强	不适合非凸形状簇

表 1: K-means 算法优缺点对比

1.9 目标函数单调性证明

$$\begin{aligned} J^{(t)} &= \sum_k \sum_{x_i \in C_k^{(t)}} \|x_i - \mu_k^{(t)}\|^2 \\ &\leq \sum_k \sum_{x_i \in C_k^{(t)}} \|x_i - \mu_k^{(t-1)}\|^2 \quad \text{(更新步骤最优性)} \\ &\leq \sum_k \sum_{x_i \in C_k^{(t-1)}} \|x_i - \mu_k^{(t-1)}\|^2 = J^{(t-1)} \end{aligned}$$