0.1 K-Means 聚类算法

K-Means 是一种基于距离的聚类算法，通过迭代将数据集划分为 K 个簇。算法首先随机选择 K 个初始质心，然后将每个数据点分配到最近的质心，接着更新质心为簇内数据点的均值。这个过程重复进行，直到质心位置稳定或达到最大迭代次数为止。

K-Means 的目标是最小化簇内数据点到其质心的总距离平方和，具体表达为：

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

其中，x_i 表示数据点，μ_k 为第 k 个簇的质心，C_k 是簇 k 中的所有数据点。通过最小化 J 值，算法不断优化簇的划分。

K-Means 算法的优点在于简单易实现，且在处理大规模数据集时效率较高。然而，它需要预先指定簇的数量 K，且对初始质心选择较为敏感，不同的初始选择可能导致不同的聚类结果。此外，K-Means 更适用于凸形簇，对噪声和异常值的处理能力较弱。

K-Means 的平均时间复杂度为 $O(K \cdot n \cdot T)$，其中 K 为簇数，n 为样本数量，T 为迭代次数。最坏情况下，复杂度为 $O(n^{(K+2)/p})$，其中 p 为特征数量。处理高维数据时，算法的计算开销可能会增加，尤其是在初始质心选择不当的情况下。