0.1 Hierarchical Clustering (HC) 聚类算法

层次聚类（Hierarchical Clustering, HC）是一种用于分析数据集内嵌层次结构的聚类算法。HC 通过反复地将数据点进行分割或合并来形成一个树状的簇结构，称为树状图（Dendrogram）。HC 的两种主要方法是自底向上（凝聚法）和自顶向下（分裂法）。

凝聚法（Agglomerative Method）：从每个数据点自身作为一个簇开始，逐步合并最近的簇，直到所有点都被合并到一个簇中。合并步骤通常基于以下几种距离度量：

- 最小距离法（Single Linkage）：两个簇之间的距离定义为它们之间最近点的距离：
$$d_{\text{single}}(C_i, C_j) = \min_{x \in C_i, y \in C_j} \text{dist}(x, y)$$
- 最大距离法（Complete Linkage）：两个簇之间的距离定义为它们之间最远点的距离：
$$d_{\text{complete}}(C_i, C_j) = \max_{x \in C_i, y \in C_j} \text{dist}(x, y)$$
- 平均距离法（Average Linkage）：两个簇之间的距离定义为它们之间所有点的平均距离：
$$d_{\text{average}}(C_i, C_j) = \frac{1}{|C_i| \cdot |C_j|} \sum_{x \in C_i} \sum_{y \in C_j} \text{dist}(x, y)$$
- 质心法（Centroid Method）：两个簇之间的距离定义为它们质心之间的距离：
$$d_{\text{centroid}}(C_i, C_j) = \text{dist}(\mu_i, \mu_j)$$

Ward 法：Ward 法通过最小化每次合并后簇内方差的增加来决定簇的合并顺序：

$$d_{\text{ward}}(C_i, C_j) = \frac{|C_k|}{|C_k| + |C_j|} \|\mu_k - \mu\|^2 + \frac{|C_j|}{|C_k| + |C_j|} \|\mu_j - \mu\|^2$$

层次聚类的结果通常以树状图的形式展示，树状图展示了数据点合并或分裂的过程。通过剪切树状图可以得到不同数量的簇。层次聚类的优点在于它不需要预先指定簇的数量，并且能够生成一个多层次的聚类结果。缺点在于算法的计算复杂度较高，特别是在处理大规模数据集时。

HC 算法的时间复杂度通常为 $O(n^2 \log n)$，其中 n 为样本数量。在某些情况下，复杂度可以达到 $O(n^3)$。