lec12.1 k-menas

k-Means

基于代表点的聚类算法（Representative-based algorithms）

设 k 为簇的数量，$\mathscr{D} = (X_1, \ldots, X_n)$ 为数据集

我们的目标是选择 k 个代表点 $Y_1, \ldots, Y_k$，使得以下目标函数最小化

$$\sum_{i=1}^{n} \left[\min_j d\left(X_i, Y_j\right) \right]$$

即，将对象与其最近代表点的距离之和最小化

而获得具体的算法需要

- 指定如何选择代表点。例如，可以使用k-means算法中的质心（centroid）作为代表点，或使用k-medoids算法中的实际数据点作为代表点
- 距离函数 $d(\cdot, \cdot)$ 的选择。例如，可以使用欧氏距离、曼哈顿距离等

一般情况下，代表点不一定是数据集中的对象，它们可以是数据集之外的点，如质心

通用的 k 个代表点方法（General k-representatives approach）

1. 初始化：我们选择 k 个初始的代表点
2. 迭代优化（Iteratively refine）
    - 分配步骤（assign step）：
        - 将每个对象分配给其最近的代表点，使用距离函数 $d(\cdot, \cdot)$
        - 将相应的簇记为 $C_1, \ldots, C_k$
    - 优化步骤（optimise step）：为每个簇 C_j 确定最佳的代表点 Y_j，使其局部目标函数最小化：

    $$\sum_{X_i \in C_j} d(X_j, Y_j)$$

    这一步旨在最小化每个簇内对象到代表点的总距离

k-Means 算法

- 代表点不一定需要从数据集中选择
- 使用平方欧氏距离（squared Euclidean distance）

目标函数：

$$\min_{Y_1, \ldots, Y_k} \sum_{i=1}^{k} \sum_{X \in C_i} \|X - Y_i\|^2$$

其中，C_i 表示与代表点 Y_i 最近的对象的集合

该目标函数称为簇内平方和 (WCSS) 目标，而我们的目标是最小化数据对象与其聚类代表点之间的总平方欧氏距离

算法介绍

假设 $C_1,...,C_k$ 已经确定，找到能使得下式最小化的 $Y_1,...,Y_k$

$$f_{C_1,...,C_k}(Y_1,...,Y_k)=\sum_{i=1}^{k}\sum_{X\in C_i}\|X-Y_i\|^2$$

这之后，我们需要找到极值

$$\frac{\partial f_{C_1,...,C_k}(Y_1,...,Y_k)}{\partial Y}=-\sum_{X\in C_i}2(X-Y_i)=0$$

解得

$$Y_i=\frac{1}{|C_i|}\sum_{X\in C_i}X$$

算法步骤：

1. *初始化：*
   从数据集中随机选择 k 个代表点 $Y_1,...,Y_k$
2. *分配：*
   将数据集中所有对象分配给最近的代表点，形成簇 $C_1,...,C_k$
3. *优化：*
   计算新的代表点 $Y_1,...,Y_k$，每个代表点为当前簇的质心
   $$Y_i=\frac{1}{|C_i|}\sum_{X\in C_i}X$$

这之后，迭代重复分配和优化阶段，直到收敛（即没有对象在簇之间移动或达到用户指定的最大迭代次数）

算法问题

1. **结果依赖于初始随机选择**：不同的初始代表点选择可能导致不同的聚类结果
2. **可能陷入局部最小值**：
   - 可能无法找到全局最优解
   - 可以通过多次不同的初始化重复聚类过程，并选择最优的最终聚类结果来改善这一问题
3. **离群点对均值的影响较大**：离群点会显著影响聚类中心（均值）和聚类结果
4. **聚类中心（均值）不是聚类中的实际实例**：均值点可能不存在于数据集中
5. **算法使用的欧氏距离不适合分类特征**：欧氏距离适用于连续数值特征，但对分类特征不适用

例子 k-Means 用于图像分割

1. 目标：将图像分割成同质的视觉部分
2. 每个像素是 (R,G,B) 空间中的一个点
3. 忽略不同像素的接近度：只考虑像素的颜色值，而不考虑像素之间的空间距离。

4. k 个像素簇由 k 种颜色表示