lec13.1 k-medians

参考 lec12.1 基于代表点的聚类算法（Representative-based algorithms）、通用的 k 个代表点方法（General k-representatives approach） 与 k-means算法中的内容

k-medians 算法

Denotes：

- $Y_1,...,Y_k$：选择的 k 个代表点
- $C_1,...,C_k$：代表点对应的簇
- X：数据集

目标函数：

$$\min_{Y_1,...,Y_k}\sum_{i=1}^{k}\sum_{X\in C_i}\|X-Y_i\|_1$$

其中，C_i 包含了与 Y_i 在 L^1 距离上最近的对象

目标：
我们希望最小化数据对象与其簇代表 $Y_1,...,Y_k$ 之间的总 L^1 距离

固定簇：
假设簇 $C_1,...,C_k$ 已固定，找到代表点 $Y_1,...,$ Y_k 使得

$$f_{C_1,...,C_k}(Y_1,...,Y_k)=\sum_{i=1}^{k}\sum_{X\in C_i}\|X-Y_i\|_1$$

最小化

固定簇的新代表：
计算使得下式最小化的新代表 Y

$$\sum_{X\in C}\|X-Y\|_1$$

设 $X=(X^{(1)},...,X^{(d)})$，$Y=(Y^{(1)},...,Y^{(d)})$，每个坐标都是独立的，因此我们可以分别对每个坐标进行最小化
如果 $i=1,...,d$，则数值 $Y^{(i)}$ 使得

$$\sum_{X\in C}\left|X^{(i)}-Y^{(i)}\right|$$

最小化，则

$$Y=\left(Y^{(1)},...,Y^{(d)}\right)$$

最小化

$$\sum_{X\in C}\|X-Y\|_1$$

中位数:

给定 s 个数 $X_1^{(i)}, \ldots, X_s^{(i)}$，使得

$$Y^{(i)} = \text{median}\left(X_1^{(i)}, \ldots, X_s^{(i)}\right)$$

最小化

$$\sum_{t=1}^{s} \left| X_t^{(i)} - Y^{(i)} \right|$$

算法步骤:

1. **初始化阶段**:
    - 随机从数据集中选择 k 个簇代表 $Y_1, \ldots, Y_k$
2. **分配阶段**:
    - 将数据集中的所有对象分配给与其 L^1 最近的代表。得到簇 $C_1, \ldots, C_k$
3. **优化阶段**:
    - 计算新的代表 $Y_1, \ldots, Y_k$ 作为当前簇 $C_1, \ldots, C_k$ 的中位数