lec14.2 Agglomerative Clustring

凝聚层次聚类（Agglomerative Hierarchical Clustering）

凝聚层次聚类的主要思想

1. 逐步聚合：各个独立的对象逐步聚合成更高层次的簇
2. 方法的差异：不同方法的主要差异在于选择合并簇时所使用的目标函数

凝聚层次聚类的步骤：

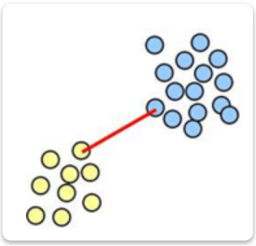
1. input 数据集 $\mathscr{D}$
2. init：将数据集中的每个对象放置在其各自的簇中，即每个对象自己成一个簇
3. repeat：
   - 找到 最接近 (需要指定集群之间的邻近度度量) 的簇对 i 和 j
     - 常见的度量方法包括最小距离、最大距离、平均距离和质心距离等
   - 将簇 i 和簇 j 合并成一个新簇
4. until 满足终止条件：终止条件可以是
   - 达到预定的簇数量
   - 簇间的最小距离大于某个阈值
5. return 当前聚类或层次结构

簇之间的接近度度量：单连接（single linkage）

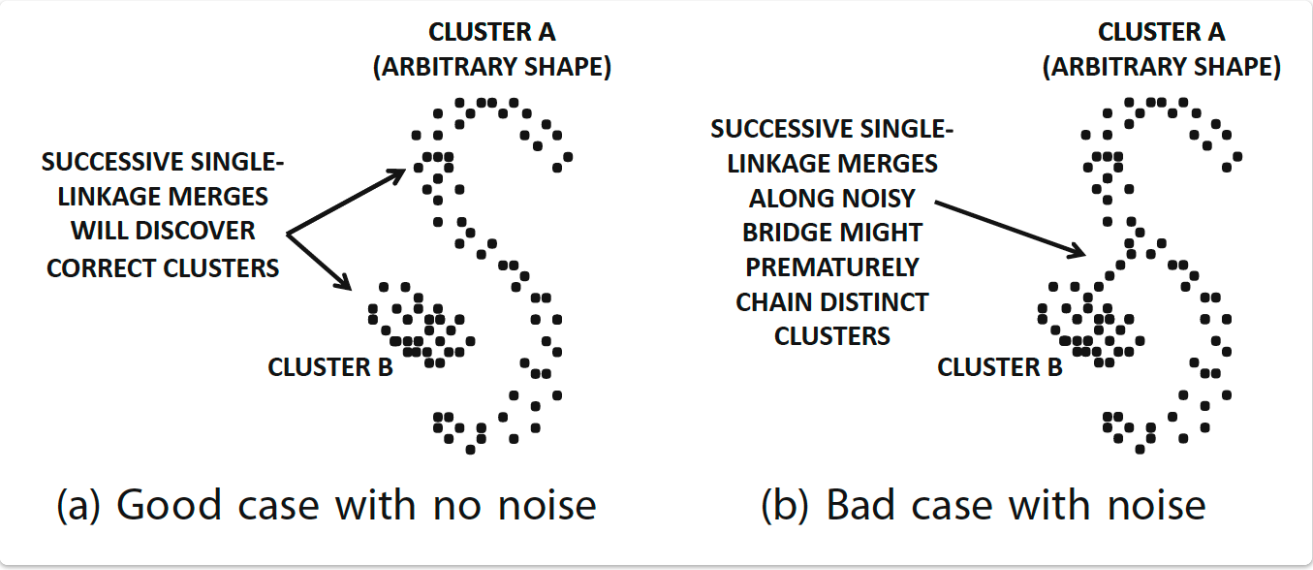
设 P、Q 是两个簇，假设我们有一个用于对象的距离函数 $d(\cdot,\cdot)$

单连接距离：
两个簇中所有元素之间距离最小的那个

$$\operatorname{dist}(P,Q)=\min_{X\in P,Y\in Q} d(X,Y)$$



有噪声的情况：



由于单连接方法是基于最小距离来合并簇的，这些噪声点可能会导致不同簇之间的错误连接，即所谓的“链状效应”（chaining effect）。在这种情况下，单连接方法可能会过早地将簇 A 和簇 B 连接在一起，导致错误的聚类结果（如图 b）

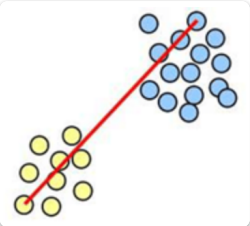
- 没有噪声的情况下，单连接方法能够很好地识别并聚类数据
- 噪声和离群点会显著影响单连接方法的效果。
- 噪声点会导致不同簇之间的错误连接，无法正确识别出独立的簇。

簇之间的接近度度量：完全连接（complete linkage）

完全链接距离：

两个簇中所有元素之间距离最大的那个

$$\text{dist}(P,Q) = \max_{X \in P, Y \in Q} d(X,Y)$$



簇之间的接近度度量：组平均连接（group-average linkage）

组平均连接距离：

两个簇中所有元素之间所有对象的平均距离

$$\operatorname{dist}(P,Q) = \frac{1}{p \cdot q} \sum_{X \in P, Y \in Q} d(X,Y)$$