§9 聚类

1. 聚类任务与性能度量

无监督：训练样本的标记信息是未知的。

有效性指标（validity index）

聚类性能度量 { 外部指标：将聚类结果与某个参考模型比较
　　　　　　　{ 内部指标：直接考察聚类结果而不利用任何参考模型

外部：Jaccard 系数、FM 指数、Rand 指数

内部：DB 指数、Dunn 指数

2. 距离

闵氏 $dist_{mk}(x_i, x_j) = \left(\sum_{u=1}^{n} |x_{iu} - x_{ju}|^p\right)^{\frac{1}{p}}$

p=2 欧式 $dist_{ed}(x_i, x_j) = ||x_i - x_j||_2 = \sqrt{\sum_{u=1}^{n} |x_{iu} - x_{ju}|^2}$

$dist_{man}$ 曼哈顿 $= ||x_i - x_j||_1 = \sum_{u=1}^{n} |x_{iu} - x_{ju}|$

3. 原型聚类

(1) k-means　$E = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||_2^2$　最小化距离

基本思想：对于给定的类别数目k，首先给定初始划分，通过迭代改变样本和簇的隶属关系，使每次处理后得到比上次好

步骤：选择初始化中心k个 $a_1, a_2 \cdots a_k$ → 对于每个样本 x_i，将其标记为距离类别中心 a_j 最近的类别 j　$\mathrm{argmin}\sqrt{\sum_{k=1}(x_{ik} - a_{jk})^2}$ → 更新 a_j 为均值 → 重复直到中止

中止：迭代次数、样本到中心距离平方和、簇中心点变化率

优点：说明简、较好伸缩性和高效率、高斯分布效果会较好

缺点：k未知、对初始值敏感、对离群点影响大、不适合发现非凸形状的簇

算法：k-medoids　kmeans++　kmeans||

(2) 学习向量量化 (LVQ)

利用样本之监督信息来聚类.

随机选择样本 → 计算样本 x_j 与 p_i 距离 $d_{ji}=\|x_j-p_i\|_2$ → 找出与 x_j 距离最近的原型向量 p_{i^*} → 更新向量 p' → 迭代中止

(3) GMM 常采用 EM 迭代

(4) 密度聚类

思想: 只要样本点密度大于某个阈值, 即将该样本加到最近之簇中

可用于任意形状, 对噪声 data 不敏感, 但计算量大

DBSCAN: 用一个点之ε邻域内的邻居点数衡量该点所在空间之密度

不需要设定 k

流程: 若点 x 的ε邻域包含多余 m 个对象, 则创建一个 x 作为核的 cluster

→ 寻找并合并核心对象直接密度可达对象 → 无新点可更新到簇之时候结束

优点: 无需给 k, 可任意形状 cluster, 寻找 outlier 不敏感, 只需 2 个参数.

几乎不依赖节点之遍历顺序

缺点: 依赖距离公式 (欧式)   不适合数据集中密度差异很大之情况

(5) 层次聚类

凝聚: AGNES. 自底向上. 两簇间距离为由距离最近之数据点之相似度确定. 合并过程反复至所有对象满足簇数目

分裂: DIANA 自顶向下. 按规则细分为越来越小之簇, 直至 cluster 数中止.

优缺点: 合并/分裂点选择不易. 不可撤消操作. 大数据不适合.