聚类

概述

- “无监督学习”任务中研究最多，应用最为广泛的是“聚类”，其他常见的任务还有密度估计，异常检测等
- “无监督学习”目标是通过对无标记训练样本的学习来解释数据的内在性质及规律，为进一步的数据分析提供基础
- 聚类试图将数据集中样本划分为通常不相交的子集，每个子集成为一个“簇”，对应为潜在的类别
- 聚类过程仅能自动形成簇结构，其概念语义许人为把握命名

聚类性能度量

- 目标：聚类结果的“簇内相似度”高，且“簇间相似度”低

外部指标

- 定义：将聚类结果与某个“参考模型”进行比较的性能度量
- 对数据集 $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m\}$，假定通过聚类给出的簇划分为 $\mathcal{C}=\{C_1,C_2,\ldots,C_k\}$，参考模型给出的簇划分为 $\mathcal{C}^*=\{C_1^*,C_2^*,\ldots,C_s^*\}$，相应地，令 $\boldsymbol{\lambda}$ 与 $\boldsymbol{\lambda}^*$ 分别表示与 $\mathcal{C}$ 和 $\mathcal{C}^*$ 对应的簇标记向量．我们将样本两两配对考虑，定义

$$a=|SS|,\quad SS=\{(\boldsymbol{x}_i,\boldsymbol{x}_j)\mid \lambda_i=\lambda_j,\lambda_i^*=\lambda_j^*,i<j\},\tag{9.1}$$

$$b=|SD|,\quad SD=\{(\boldsymbol{x}_i,\boldsymbol{x}_j)\mid \lambda_i=\lambda_j,\lambda_i^*\neq\lambda_j^*,i<j\},\tag{9.2}$$

$$c=|DS|,\quad DS=\{(\boldsymbol{x}_i,\boldsymbol{x}_j)\mid \lambda_i\neq\lambda_j,\lambda_i^*=\lambda_j^*,i<j\},\tag{9.3}$$

$$d=|DD|,\quad DD=\{(\boldsymbol{x}_i,\boldsymbol{x}_j)\mid \lambda_i\neq\lambda_j,\lambda_i^*\neq\lambda_j^*,i<j\},\tag{9.4}$$

- 用于计算外部指标的值
- Jaccard系数：$\mathrm{JC}=\dfrac{a}{a+b+c}$
- FM指数：$\mathrm{FMI}=\sqrt{\dfrac{a}{a+b}\cdot\dfrac{a}{a+c}}$
- Rand指数：$\mathrm{RI}=\dfrac{2(a+d)}{m(m-1)}$
- 上述性能度量的结果值均在[0,1]，值越大越好

内部指标

- 定义：直接考察聚类结果而不利用任何参考模型的性能度量
- 考虑聚类结果的簇划分 $\mathcal{C}=\{C_1,C_2,\ldots,C_k\}$，定义

$$\mathrm{avg}(C)=\frac{2}{|C|(|C|-1)}\sum_{1\leqslant i<j\leqslant|C|}\mathrm{dist}(\boldsymbol{x}_i,\boldsymbol{x}_j),$$

$$\mathrm{diam}(C)=\max_{1\leqslant i<j\leqslant|C|}\mathrm{dist}(\boldsymbol{x}_i,\boldsymbol{x}_j),$$

$$d_{\min}(C_i,C_j)=\min_{\boldsymbol{x}_i\in C_i,\boldsymbol{x}_j\in C_j}\mathrm{dist}(\boldsymbol{x}_i,\boldsymbol{x}_j),$$

$$d_{\mathrm{cen}}(C_i,C_j)=\mathrm{dist}(\boldsymbol{\mu}_i,\boldsymbol{\mu}_j),$$

- 用于计算内部指标的值
- DB指数：$\mathrm{DBI}=\dfrac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\dfrac{\mathrm{avg}(C_i)+\mathrm{avg}(C_j)}{d_{\mathrm{cen}}(\boldsymbol{\mu}_i,\boldsymbol{\mu}_j)}\right)$
- Dunn指数：$\mathrm{DI}=\min_{1\leqslant i\leqslant k}\left\{\min_{j\neq i}\left(\dfrac{d_{\min}(C_i,C_j)}{\max_{1\leqslant l\leqslant k}\mathrm{diam}(C_l)}\right)\right\}$
- DBI的值越小越好，DI的值越大越好

距离计算

- 距离度量需满足的条件：非负性，同一性，对称性，直递性
- 最常用的距离为“闵可夫斯基距离”，可用于测量有序属性：$\mathrm{dist}_{\mathrm{mk}}(\boldsymbol{x}_i,\boldsymbol{x}_j)=\left(\sum_{u=1}^{n}|x_{iu}-x_{ju}|^p\right)^{\frac{1}{p}}$
- 当p=2时，闵可夫斯基距离为欧氏距离
- 当p=1时，闵可夫斯基距离为曼哈顿距离
- 采用VDM可计算无序属性，将闵可夫斯基距离与VDM结合即可处理混合属性
- 注意：通常用于相似度度量的距离未必需要满足距离度量的所有基本性质，尤其是直递性。存在“非距离度量”，以及根据现实任务确定合适的距离计算式

原型聚类

- 基于原型的聚类，此类算法假设聚类结构能通过一组原型刻画

k均值算法

- “k均值”算法的优化目标式针对聚类所得簇划分C最小化其平方误差：$E=\sum_{i=1}^{k}\sum_{\boldsymbol{x}\in C_i}\|\boldsymbol{x}-\boldsymbol{\mu}_i\|_2^2$
- 实际中k均值采用贪心策略
- 工作机制
  1. 从D中随机选择k个样本作为初始均值向量
  2. 计算各样本与均值向量的距离，根据最近的均值向量确定样本的簇标记
  3. 根据各簇的样本重新计算其均值向量，更新均值向量
  4. 循环2-3步，直到均值向量不再更新

学习向量量化(LVQ)

- LVQ假设数据样本带有类别标记，学习过程利用样本的标记信息来监督辅助聚类
- 工作机制
  1. 设定原型向量及其个数，预设其类别标记（根据人为期望）
  2. 根据样本随机初始化原型向量（标记需和原型标记一一对应）
  3. 从样本集D随机选取样本，计算该样本与各原型之间的距离
  4. 找到最近的原型样本，对比该样本与原型样本标记，若相同则根据迭代公式拉近原型样本与该样本的距离，反之则使原型样本远离
  5. 循环3-4步，直到满足一定条件（例如以达到最大迭代轮数或更新幅度很小）

高斯混合聚类

- 高斯混合(Mixture-of-Gaussian)采用概率模型来表达聚类原型
- 设定样本服从高斯分布，采用EM算法迭代计算所需参数
- 工作机制
  1. 初始化高斯分布的模型参数
  2. 根据贝叶斯定理计算各混合成分生成的后验概率
  3. 利用计算生成的后验概率重新计算，新均值向量，新协方差矩阵，新混合系数
  4. 根据第3步得到的参数模型，继续循环2-3步，直到满足停止条件
  5. 根据各样本计算其所属不同簇的后验概率，将其划分到最大后验概率对应的簇里
  6. 输出簇划分

密度聚类

- 此类算法假设聚类结构能够通过样本分布的紧密程度确定：从考察密度的角度来考察样本之间的可连接性，并基于可连接性样本不断扩展聚类簇以获得最终的聚类效果。

代表 DBSCAN

- 基于一组邻域参数来刻画样本分布的紧密程度
- 工作机制
  1. 给定样本集与领域参数，找出所有的核心对象（在邻域内至少包含MinPts个样本的对象）
  2. 以任一核心对象为出发点，找出其由其密度可达的样本生成聚类簇
  3. 若第2步形成聚类簇的过程中包含了其他核心对象，则下次循环不再访问已经访问过的核心对象
  4. 重复循环2-3步，直到所有核心对象均被访问，余下的样本被视为噪声，不被分配到簇

层次聚类

- 层次聚类试图在不同层次对数据集进行划分，从而形成不同属性的聚类结构，可采用“自底向上”的聚合策略，也可采用“自顶向下”分拆策略
- AGNES：一种自底向上聚合策略的层次聚类算法
- 工作机制
  1. 将数据集中的每个样本看作一个初始聚类簇
  2. 在算法运行的每一步找出距离最近的两个聚类簇进行合并
  3. 不断重复第2步，直至达到预设的聚类簇个数
  - 该算法的关键是如何计算距离，集合间的距离计算常采用豪斯多夫距离（Hausdorff distance）