机器学习算法系列（11）：聚类（1）—简介

一、引言

聚类（Clustering）算法就是对大量未知标注的数据集，按照数据的内在相似性将数据集划分为多个类别，使类别内的数据相似度较大而类别间的数据相似度较小。聚类是一种无监督算法。给定一个有N个对象的数据集，构造数据的K个簇，$k \leq n$，同时满足，每个簇至少包含一个对象，每一个对象属于且仅属于一个簇，将满足上述条件的K个簇称作一个合理划分。它的主要思想是对于给定的类别数目K，首先给出初始划分，通过迭代改变样本和簇的隶属关系，使得每一次改进之后的划分方案都较前一次好。

聚类算法主要包括以下五类：

- 基于分层的聚类（hierarchical methods）

这种方法对给定的数据集进行逐层，直到某种条件满足为止。具体可分为合并型的“自下而上”和分裂型的“自下而上”两种方案。如在“自下而上”方案中，初始时每一个数据记录都组成一个单独的组，在接下来的迭代中，它把那些相互邻近的组合并成一个组，直到所有的记录组成一个分组或者某个条件满足为止。代表算法有：*BIRCH算法*（1996）、*CURE算法*、CHAMELEON算法等。

- 基于划分的聚类（partitioning methods）

给定一个有N个记录的数据集，分裂法将构造K个分组，每一个分组就代表一个聚类，K<N,而且这K个分组满足下列条件：（1）每一个分组至少包含一个数据记录；（2）每一个数据记录属于且仅属于一个分组（咋某些模糊聚类算法中可以放宽条件）。对于给定的K，算法首先给出一个初始的分组方法，以后通过反复迭代的方法改变分组，使得每一次改进之后的分组方案都较前一次好，而所谓好的标准是：同一分组中的记录越近越好，而不同分组中的记录越远越好。使用这个基本思想的算法有：*K-means算法*、*K-medoids算法*、*CLARANS算法*

- 基于密度的聚类（density-based methods）

基于密度的方法和其他方法的一个根本区别是：它不是基于各种各样的距离的，而是基于魔都的，这样就能克服基于距离的算法只能发现“类圆形”的聚类的缺点。这个方法的指导思想为：只要一个区域的点的密度大过某个阈值，就把它加到与之相近的聚类中去，代表算法有：*DBSCAN（Density-Based Spatial Clustering of Applic with Noise）算法（1996）*、*OPTICS（Ordering Points to Identify Clustering Structure）算法（1999）*、*DENCLUE算法*

（1998）、WaveCluster算法（1998，具有O（N）时间复杂性，但只适用于低维数据）

- 基于网格的聚类（grid-based methods）

这种方法首先将数据空间划分成为有限个单元（cell）的网络结构，所有的处理都是以单个的单元为对象的。这么处理的一个突出的优点就是处理速度很快，通常这是与目标数据库中记录的个数无关，它只与把数据空间分成多少个单元有关。代表算法有：*STING（Statistical Information Grid）、CLIQUE（Clustering In Quest）算法（1998）、WaveCluster算法*。其中STRING算法把数据空间层次地划分为单元格，依赖于存储在网格单元中的统计信息进行聚类；CLIQUE算法结合了密度和网格的方法。

- 基于模型的聚类（model-based methods）

基于模型的方法给每一个聚类假定一个模型，然后去寻找能够很好地满足这个模型的数据集。这样一个模型可能是数据点在空间中的密度分布函数或者其它。它的一个潜在的假定就是：目标数据集是由一系列的概率分布所决定的。通常有两种尝试方向：统计的方案和神经网络的方案。

二、相似度、距离计算方法

- 给定n维空间R^n中的两个向量$X = (x_1, x_2, \cdots, x_n)^T$和$y = (y_1, y_2, \cdots, y_n)^T$，$x, y$之间的距离可以反映两者的相似程度，一般采用$L_p$距离

$$dist\,(X, Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

，其中$p \geq 1$，也称为闵可夫斯基距离（Minkowski）距离。常用的p为$1, 2, +\infty$，此时相应的距离公式分别为

  - 1.当$p = 1$时，称为曼哈顿距离（Manhattan distance），改名字的由来起源于在纽约市去测量街道之间的距离就是由人不行的步数来确定的。

$$d\,(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

  - 当$p = 2$时，称为欧几里得距离（Euclidean distance）

$$d\,(x, y) = \left(\sum_{i=1}^{n} (x_i - y_i)^2 \right)^{\frac{1}{2}}$$

  - 当$p = +\infty$时，称为最大值距离（Maximum distance）

$$d(x, y) = \max_{1 \le i \le n} |x_i - y_i|$$

- 杰卡德相似系数（Jaccard）

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

- 余弦相似度（Cosine Similarity）

$$\cos(\theta) = \frac{x^T y}{|x| \cdot |y|}$$

- pearson相似系数

$$\rho_{XY} = \frac{cov(X, Y)}{\sigma_x \sigma_y} = \frac{E\left[(x - u_x)\left(y - u_y\right)\right]}{\sigma_x \sigma_y}$$

- 相对熵（K-L）距离

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = E_{p(x)} \log \frac{p(x)}{q(x)}$$

- Hellinger距离

$$D_a(p||q) = \frac{2}{1 - a^2}\left(1 - \int p(x)^{\frac{1+a}{2}} q(x)^{\frac{1-a}{2}} dx\right)$$

- 余弦相似度与pearson相似系数的比较

n维向量x和y的夹角记作θ，根据余弦定理，其余弦值为：

$$\cos(\theta) = \frac{x^T y}{|x| \cdot |y|} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \cdot \sqrt{\sum_{i=1}^n y_i^2}}$$

这两个向量的相关系数是：

$$\rho_{XY} = \frac{cov(X, Y)}{\sigma_x \sigma_y} = \frac{E\left[(x - u_x)\left(y - u_y\right)\right]}{\sigma_x \sigma_y}$$

$$= \frac{\sum_{i=1}^n (x_i - \mu_x)\left(y_i - \mu_y\right)}{\sqrt{\sum_{i=1}^n (x_i - \mu_x)^2} \sqrt{\sum_{i=1}^n \left(y_i - \mu_y\right)^2}}$$

相关系数即将x, y坐标向量各自平移到原点后的夹角余弦。这即揭示了为何文档间求距离使用夹角余弦，因为这个物理量表征了文档去均值化后的随机向量间的相关系数。