**四种聚类算法**

* **基于连通性的聚类或层次聚类**

\* 返回树状图数据，展示信息结构而不是集群分类，性能差；

* **基于质心的聚类**

\* 思想：聚类数为k，找到聚类中心并将对象分配给最近的聚类中心，以使与聚类的平方距离最小化；

\* 优点：

- k-means：①只能找到局部最优解；②需要使用不同的随机初始化进行多次运行；③需要提前指定k数；④倾向将对象分配给最近的质心，因此他们倾向于使用近似大小的集群，常会导致错误的切割集群边界。

\* 集群分配的可能数量与数据点的量成指数关系，k-means典型的方法涉及一个直观的迭代方法（expectation-maximization期望最大化）；

- 注：初始化的质心之间需要有足够远的距离（scikit-learn中init = ‘k-means++’）

\* 缺点：无法聚类哪些非凸的数据集（即聚类形状一般是球状）；

* **基于数据密度的聚类**

\* 思想：将数据集划分为聚类，计数器会输入r参数，即”邻居”距离。因此，如果目标点位于半径为r的圆/球内，则他属于该集群。

\* 优点：具有噪声的基于密度的聚类方法（DBSCAN），①对噪声不敏感；②能发现任意形状的簇；

\* DBSCAN缺点：①对两个参数的设置敏感（圆半径eps、阈值MinPts）；②使用固定参数聚类，如果密度不均匀，则很难使用该算法；③如果数据样本集越大，收敛时间越长，可以用KD树优化；④如果对象的位置太近，则无法估算r参数。