**Clustering**

**1. 聚类算法有哪些？**

1. 基于划分

给定一个有N个元组或者纪录的数据集，分裂法将构造K个分组，每一个分组就代表一个聚类，K<N。

特点：计算量大。很适合发现中小规模的数据库中小规模的数据库中的球状簇。

代表算法：K-MEANS算法、K-MEDOIDS算法、CLARANS算法

1. 基于层次

对给定的数据集进行层次似的分解，直到某种条件满足为止。具体又可分为“自底向上”和“自顶向下”两种方案。

特点：较小的计算开销。然而这种技术不能更正错误的决定。

算法：BIRCH算法、CURE算法、CHAMELEON算法

1. 基于密度

只要一个区域中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。

特点：能克服基于距离的算法只能发现“类圆形”的聚类的缺点。

算法：DBSCAN算法、OPTICS算法、DENCLUE算法

1. 基于网络

将数据空间划分成为有限个单元（cell）的网格结构,所有的处理都是以单个的单元为对象的。

特点：处理速度很快，通常这是与目标数据库中记录的个数无关的，只与把数据空间分为多少个单元有关。

算法：STING算法、CLIQUE算法、WAVE-CLUSTER算法

**2. 简单介绍下K-means算法？**

输入：样本数据集D，聚类簇数k

输出：簇的集合

过程：

1. 从样本中选取k个样本点作为初始的均值向量
2. 循环以下步骤直到满足停止条件
3. 令
4. 对所有样本点计算到k个均值向量之间的距离，取其中最短距离对应的均值向量的标记作为该点的簇标记，然后将该点加入相应的簇
5. 对每一个簇计算新的均值向量，如果相比之前的向量有变化，就更新，将其作为新的均值向量。

**3. K-means的复杂度是多少？**

时间复杂度：

空间复杂度：

其中，表示迭代次数，表示簇数，为样本个数

**4. . K-means的优缺点是什么？**

优点：

1. 有较好的可解释性；
2. 计算时间短，速度快

缺点：

1. 需要先确定k的个数；
2. 对噪声和离群点敏感；
3. 不能发现非凸形状的簇，或大小差别很大的簇；
4. 结果不一定是全局最优，只能保证局部最优；
5. 初始簇中心对结果有较大的影响。

**5. K-means的k如何确定？**

1. 手肘法：

手肘法的核心指标是SSE（平方误差和）

其中，是第个簇，是中的样本点，是的质心，SSE是所有样本的聚类误差，代表了聚类效果的好坏。

手肘法的核心思想是：随着聚类数k的增大，样本划分会更加精细，每个簇的聚合程度会逐渐提高，那么误差平方和SSE自然会逐渐变小。并且，当k小于真实聚类数时，由于k的增大会大幅增加每个簇的聚合程度，故SSE的下降幅度会很大，而当k到达真实聚类数时，再增加k所得到的聚合程度回报会迅速变小，所以SSE的下降幅度会骤减，然后随着k值的继续增大而趋于平缓，也就是说SSE和k的关系图是一个手肘的形状，而这个肘部对应的k值就是数据的真实聚类数。当然，这也是该方法被称为手肘法的原因。

1. 轮廓系数法：

的核心指标是轮廓系数：

其中，是与同簇的其他样本的平均距离，称为凝聚度，是与最近簇中所有样本的平均距离，称为分离度。而最近簇的定义是

其中是某个簇中的样本。事实上，简单点讲，就是用到某个簇所有样本平均距离作为衡量该点到该簇的距离后，选择离最近的一个簇作为最近簇。

求出所有样本的轮廓系数后再求平均值就得到了平均轮廓系数。平均轮廓系数的取值范围为[-1,1]，且簇内样本的距离越近，簇间样本距离越远，平均轮廓系数越大，聚类效果越好。那么，很自然地，平均轮廓系数最大的k便是最佳聚类数。

**6. K-means的初始簇中心如何选取？**

1. K-means++：

1、从输入的数据点集合中随机选择一个点作为第一个聚类中心

2、对于数据集中的每一个点x，计算它与最近聚类中心(指已选择的聚类中心)的距离D(x)

3、选择一个新的数据点作为新的聚类中心，选择的原则是：D(x)较大的点，被选取作为聚类中心的概率较大

4、重复2和3直到k个聚类中心被选出来

5、利用这k个初始的聚类中心来运行标准的k-means算法

1. 选用层次聚类或Canopy算法进行初始聚类，然后从k个类别中分别随机选取k个点来作为K-means的初始聚类中心点

**7. 如何衡量聚类算法的优劣？**

无监督的方法，即样本没有给定标签：

1. 紧密性CP：计算每一个类各点到聚类中心到平均距离，其数值越低意味着类内聚类距离越近。缺点是没有考虑类间距离。
2. 间隔性SP：计算各聚类中心两两之间的平均距离，其数值越高意味着类间聚类距离越远。缺点是没有考虑类内效果。
3. 戴维森堡丁指数DB：计算任意两类别的类内距离平均距离之和除以两聚类中心的最大值，其数值越小意味着类内距离越小，同时类间距离越大。缺点是因使用欧式距离，所以对于环状分布，聚类评测效果很差。
4. 邓恩指数DVI：计算任意两个簇元素的最短距离（类间）除以任意簇中的最大距离（类内），其数值越大意味着类间距离越大，同时类间距离越小。缺点是对离散点的聚类测评很高，对环状分布测评效果很差。

有监督的方法，需要基准数据，基准是一种理想的聚类，通常由专家构建，用一定的度量评判聚类结果与基准数据的符合程度。对数据集，假定通过聚类得到的簇划分为，参考模型给出的簇划分为，相应地，令与分别表示与和对应的簇标记向量。现将样本两两配对考虑，定义

1. Jaccard系数JC
2. FM指数FMI
3. Rand指数RI

上述性能度量的结果值均在区间，值越大越好。