第五章 非监督学习

1.K均值聚类

K均值算法的优缺点是什么？如何对其进行调优？

缺点：

（1）受初值和离群点的影响，每次的结果不稳定。

（2）结果通常不是全局最优解而是局部最优解。

（3）无法很好地解决数据簇分布差别比较大的情况（比如一类样本是另一类样本数量的100倍）。

（4）不太适用于离散分类等情况。

优点：

（1）对于大数据集，K均值聚类算法相对是可伸缩和高效的。

（2）虽然以局部最优结束，但一般情况下达到的局部最优已经可以满足聚类的需求。

调优：

（1）数据归一化和离群点处理。

（2）合理选择K值。

（3）采用核函数。传统的欧式距离度量方式，使得K均值算法本质上假设了各个数据簇的数据具有一样的先验概率，并呈现球形或者高维球形分布。面对非凸的数据分布时，可能需要引入核函数来优化。

针对K均值算法的缺点，有哪些改进的模型？

主要缺点：

（1）需要人工预先确定K值。

（2）K均值只能收敛到局部最优，效果受到初始值影响很大。

（3）易受到噪点的影响。

（4）样本只能被划分到单一的类中。

K-Means++

主要解决K已确定，初始点选择的问题：假设已经选取了n个初始聚类中心，则在选取第n+1个聚类中心时，距离当前n个聚类中心越远的点会有更高的概率被选为第n+1个聚类中心。在选取第1个聚类中心时同样通过随机的方法。

ISODATA算法

当K值的大小不确定时，可以使用ISODATA算法、
当属于某个类别的样本数过少时，把该类别取出；当属于某个类别的样本数过多、分散程度较大时，把该类别划分为两个子类别。等于在原来K-Means算法上增加了分裂和合并操作。需要考虑三个参数：
（1）预期的聚类中心数目K。该值为用户指定，该算法的聚类中心数目变动范围也由其决定。最终输出的聚类中心数目常见范围是K的一半到两倍K。
（2）每个类所要求的的最少样本数目N。分裂后导致某个子类别所包含样本数目小于该值，则不进行分裂。
（3）最大方差Sigma，用于空中某个类别中样本的分散程度。
（4）两个聚类中心之间所允许最小距离D。

2.高斯混合模型

高斯混合模型假设每个簇的数据都是符合高斯分布的，当前数据呈现的分布就是各个簇的高斯分布叠加在一起的结果。
高斯混合模型的核心是，假设数据可以看作从多个高斯分布中生成出来的。在该假设下，每个单独的分模型都是标准高斯分布，其中均值μ_i和方差∑_i是待估计的参数。此外，每个分模型都还有一个参数π_i，可以理解为权重或生成数据的概率。高斯混合模型是一个生成式模型。
高斯混合模型的简单例子参考书上P.104.

高斯混合模型的参数一般通过最大似然估计来求解，具体使用EM算法最大化目标函数。

高斯混合模型和K-Means相同点：
（1）都可用于聚类。
（2）都需要指定K值。
（3）都采用EM算法求解。
（4）只能收敛到局部最优。

高斯模型相比K-Means的优点：
（1）可以给出一个样本属于某个类的概率。
（2）可以用于概率密度的估计。
（3）可以用于生成新的样本点。

3.自组织神经网络映射（SOM）

可以用于聚类、高维可视化、数据压缩、特征提取。
设计SOM及其需要的训练参数：
（1）输出层神经元的数量
（2）输出层节点的排列
（3）初始化权值
（4）设计拓扑邻域
（5）设计学习率

4.聚类算法的评估

常见数据簇的特点：
（1）以中心定义的数据簇，倾向于球形分布，通常中心被定义为质心，即此数据簇中所有点的平均值
（2）以密度定义的数据簇，数据集合呈现和周围数据簇明显不同的密度，或稀疏或稠密。当有噪声和离群点时常常使用基于密度的簇定义
（3）以连通度定义的数据簇，整个数据簇表现为图结构，对不规则形状或者缠绕的数据簇有效
（4）以概念定义的数据簇，这类数据集合中的所有数据点具有某种共同性质

聚类评估任务可以分为三个子任务：
（1）估计聚类趋势，即检测数据分布中是否存在非随机的簇结构，即聚类是否有意义
（2）判定数据簇数。判定的方法有很多，例如Elbow Method, Gap Statistic
（3）测定聚类质量，常用的度量指标有轮廓系数，均方根标准偏差（RMSSTD），R方