机器学习第九章习题

罗帆靖 20214865

（一）试析k均值算法能否找到最小化(9.24)的最优解。

不能。最小化（9.24）问题本身是NP问题，且其为非凸的，容易陷入局部最优价是其固有的缺点，所以在使用K均值时找到的是其局部最优解，因此需要随机多次初始化中心店，然后挑选结果最好的一个。

（二）试析 AGNES 算法使用最小距离和最大距离的区别。

其中最小距离和最大距离是AGNES算法常用的两种策略。

- 最小距离：最小距离合并策略计算两个簇之间最相似的数据点之间的距离作为簇之间的相似度。具体来说，它计算两个簇中距离最近的两个数据点之间的距离，并将该距离作为簇之间的相似度。这种策略强调簇内数据点的紧密性，即两个簇合并后的新簇将包含最相似的数据点。
- 最大距离：最大距离合并策略计算两个簇中最不相似的数据点之间的距离作为簇之间的相似度。具体来说，它计算两个簇中距离最远的两个数据点之间的距离，并将该距离作为簇之间的相似度。这种策略强调簇之间的差异性，即两个簇合并后的新簇将具有最大的距离。

总的来说，最小距离合并策略强调簇内的紧密性，适用于处理非凸形状和噪声较多的情况。最大距离合并策略强调簇间的差异性，适用于处理分散较大和离散的数据。

（三）聚类结果中若每个簇都有一个凸包，且凸包不相交，则称为凸聚类。试析本章介绍的哪些聚类方法只能产生凸聚类，哪些能产生非凸聚类。

若在一个簇的凸包之内，有其他簇的样本，就说明凸包相交。

- 原型聚类：输出线性分类边界的聚类算法显然都是凸聚类，这样的算法有：K均值，LVQ；而曲线分类边界的也显然是非凸聚类，高斯混合聚类，在簇间方差不同时，其决策边界为弧线，所以高混合聚类为非凸聚类；

- 密度聚类：DBSCAN，如下图情况，显然当领域参数符合一定条件时，会生成两个簇，其中外簇会包括内簇，所以DBSCAN显然

也是非凸聚类；

- **层次聚类**：AGENS是凸聚类。

（四）试设计一个能用于混合属性的非度量距离

本人设计了一种“罗氏距离”

- 对于二元属性，相等则距离0，不等则距离1
- 对于名义属性，相等则距离0，不等则距离1
- 对于顺序属性：
  - 将属性值转化为数值，根据顺序关系进行映射
  - 计算属性值之间的差异，可以使用差值的绝对值
- 对于数值属性，计算属性值之间的差异，使用差值的绝对值

对于每个属性，根据属性的类型计算距离，并将不同属性类型的距离进行加权平均，得到最终的“罗氏距离”。