1. （必填）自己提出的问题的理解（罗列全部）：
2. 提出的问题1：如何检测outlier

讨论后的理解：

【1】基于统计模型的方法：首先建立一个数据模型，异常是不显著属于任何簇的对象；要求预先知道数据的分布和分布的参数（例如均值和方差）。

【2】基于邻近度的方法：通常可以在对象之间定义邻近性度量，异常对象是那些远离其他对象的对象。一般情况下,在低维空间用距离来度量能效果较好,但在高维空间中效果并不好。不适合大数据集，不能处理具有不同区域密度的数据集

【3】基于密度的方法：仅当一个点的局部密度显著低于它的大部分近邻时才将其分类为离群点。（1）给出了对象是离群点的定量度量，并且即使数据具有不同的区域（密度分布及不均匀）也能够很好的处理；2）参数选择是困难的。虽然LOF算法通过观察不同的k值，然后取得最大离群点得分来处理该问题，但是，仍然需要选择这些值的上下界。

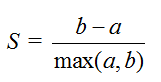
【4】基于聚类的方法：聚类分析用于发现局部强相关的对象组，而异常检测用来发现不与其他对象强相关的对象。因此，聚类分析非常自然的可以用于离群点检测。聚类算法产生的簇的质量对该算法产生的离群点的质量影响非常大。

1. 提出的问题2：如何确定最佳簇数

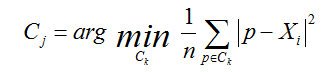
讨论后的理解：

【1】手肘法：根据合理推导，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。

【2】轮廓系数法：该方法的核心指标是轮廓系数（Silhouette Coefficient），某个样本点Xi的轮廓系数定义如下：



其中，a是Xi与同簇的其他样本的平均距离，称为凝聚度，b是Xi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p是某个簇Ck中的样本。事实上，简单点讲，就是用Xi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Xi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。

1. （必填）别人提出的问题的理解（选择几个问题罗列，并给出理解）：
2. 问题3：为什么层次聚类中为什么合并聚类比分裂聚类的使用要广泛？两种方法分别有什么优势和劣势，分别会在什么情况下出现？两种方法一般会分别用在什么情况下？

自己的理解：

矩阵更新过程中，总是将两个距离最近的聚类合并，那么我们只要加入一个阈值判断，当这个距离大于阈值时，就说明不需要再合并了，此时算法结束。这样的阈值引入可以很好的控制算法结束时间，将层次截断在某一层上。

分裂可能会分错的概率大？分几份也是个问题。

1. 问题4在混合属性的处理中，为什么将一个具有多余两个状态的名词性属性或者非对称布尔属性转换为区间度量函数是没有意义的？

自己的理解：多于两个状态的名词性属性很难用区间度量（难以判断离谁近），不平衡的布尔值权重不均衡

1. （必填）读书计划

1、本周完成的内容章节：4.1-4.7

2、下周计划：读完第四章，第三章能读多少读多少