1. 自己提出的问题的理解：

提出的问题1：如何判断离聚点中心最远的点是否为异常值？

讨论后的理解：常用的异常值检测方法有四种，基于统计模型的方法、基于邻近度的方法、基于密度的方法、基于密度的方法以及基于聚类的方法。基于统计模型的方法：首先建立一个数据模型，异常是那些同模型不能完美拟合的对象；如果模型是簇的集合，则异常是不显著属于任何簇的对象；在使用回归模型时，异常是相对远离预测值的对象。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。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。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。

讨论后我们还认为，异常值检测出来后，**不一定要直接丢弃，要分析具体的应用**。大部分的时候，异常值这种差异信息被视为噪声而丢弃，然而在**一些应用中，罕见的数据可能蕴含着更大的研究价值**。

提出的问题2：如果数据点很适合被分成2个聚类，但用户给的K值为3怎么办？或者适合分成3类，k值为2，没有很理想的分法怎么办？可以改变K值吗？

讨论后的理解：我认为K值的选择是用户在我们算法开始执行前需要去做的事情。一般来说，K值会取一个比较小的数值，然后采用交叉验证法取选取最优的K值。而当用户把K值交付给我们后，我们要做的就是根据某个距离函数，反复地把数据分入K个聚类中，而不能去改变K值了。

k值的选择有一些经典的方法，我们查阅资料了解手肘法。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. （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）：

社交网站的手机截图

描述已自动生成电脑屏幕的截图

描述已自动生成问题3：书中提到k-均值算法存在一个空聚类的问题（即在聚类过程中可能有的聚类中心没有被分配任何数据点而使得某些聚类变空，这些聚类被称为空聚类），但是聚类中心本身就是某一个聚类的中心点位置，为什么还会存在空聚类问题呢

自己的理解：空聚类的出现就像图中一样，是在算法执行的过程中，一个聚类的数据点被归到了其他的聚类，导致该聚类成为空聚类，而且如果我们不加处理的话，空聚类就会一直存在。所以我们可以选择一个数据点代替原有的聚类中心，例如在一个含有大数据的聚类中离中心最远的点。

问题4：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？

自己的理解：合并聚类是自下而上的方法，分裂聚类是自上而下。从容错率的角度来看，分裂聚类一旦一步分错，可能会导致下面分出的聚类都产生问题，合并聚类可能出现错误以后产生的影响更小。还有我觉得合并聚类更加灵活，如果用户更改了需求，合并聚类要更改的要少，分裂聚类就要全改。

问题5：K-meas算法由于其本身的特性，只能在数据为球形簇的情况下具有良好的效果，那么有什么好的聚类思路可以解决非球形簇的聚类问题呢？

自己的理解：我们组查阅了资料，有一种DBSCAN算法可以解决。它是一种基于密度的聚类算法，所以它无需知道聚类簇的数量，也可以发现任意形状的聚类簇

问题6：在混合属性的处理中，为什么将一个具有多于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：1.多余两个状态的名词属性，比如书中示例的苹果、橘子、梨，苹果的值为1，橘子和梨设定为0，最后结果是0.3，只能知道30%是苹果，不能知道不是苹果的到底是里还是橘子2.非对称布尔属性是不能转化成区间度量属性的，因为区间度量属性要求在整个线性度量空间上相同长度的区间具有相同的重要性。

1. （必填）读书计划

1、本周完成的内容章节：第四章

2、下周计划：第三章