1. （必填）自己提出的问题的理解（罗列全部）：

提出的问题1：书中提到k-均值算法存在一个空聚类的问题，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

讨论后的理解：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。

提出的问题2：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：有一个基于模型的方法：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。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。

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）：

问题3：分裂聚类较于合并聚类的局限性和缺点有哪些？

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4：k均值算法中的K怎么取？

自己的理解：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。

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：3.1-3.5

四、（选做）读书摘要及理解或伪代码的具体实现（读书摘要、伪代码的具体实现代码等可以写到这个部分）

1、读书摘要及理解（选做）

本章主要介绍了两个聚类方法：K均值算法以及层次聚类

k均值聚类算法（k-means clustering algorithm）是一种迭代求解的[聚类分析](https://baike.baidu.com/item/%E8%81%9A%E7%B1%BB%E5%88%86%E6%9E%90/3450227)算法，其步骤是，预将数据分为K组，则随机选取K个对象作为初始的聚类中心，然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个聚类。每分配一个样本，聚类的聚类中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。终止条件可以是没有（或最小数目）对象被重新分配给不同的聚类，没有（或最小数目）聚类中心再发生变化，误差平方和局部最小。

层次聚类，是一种很直观的算法。顾名思义就是要一层一层地进行聚类，可以从下而上地把小的cluster合并聚集，也可以从上而下地将大的cluster进行分割。

层 次聚类较大的优点，就是它一次性地得到了整个聚类的过程，只要得到了上面那样的聚类树，想要分多少个cluster都可以直接根据树结构来得到结果，改变 cluster数目不需要再次计算数据点的归属。层次聚类的缺点是计算量比较大，因为要每次都要计算多个cluster内所有数据点的两两距离。另外，由 于层次聚类使用的是贪心算法，得到的显然只是局域最优，不一定就是全局最优，这可以通过加入随机效应解决，这就是另外的问题了。