1. 自己提出的问题的理解：
2. 提出的问题1：为什么层次聚类中合并聚类比分裂聚类使用得要广泛？

讨论后的理解：

1. 分裂聚类的特性是越在分裂的开始所分得的聚类的范围要广，因为分裂之后的聚类比较广，所以一个数据点被分到错误聚类中的可能性更大一些，而合并聚类则是从少到多，一点一点地将一个节点合并到所属的聚类中，所以犯错的可能性相对较小。
2. 同时，合并聚类的原理和主要操作方法和传统的聚类方法如K-means等相同，所以操作更为方便。
3. 别人提出的问题的理解：

1、问题2：空聚类是如何产生的？为什么解决空聚类的时候，选择离一个含有大量数据的聚类的聚类中心最远的数据点？

自己的理解：在聚类过程中，因为聚类中心不断变化，所以有可能会出现某个聚类中的点离另一个聚类的聚类中心更加近的情况，此时将此节点分到别的聚类中，如果一个聚类中所有的节点都被重新分配到别的聚类中，那么此时就会产生空聚类。

将一个聚类中心选为某个含有大量数据的聚类的聚类中心的最远的数据点，下一次聚类的时候可以将此大数据聚类分成两个聚类，新产生的聚类是一个自然聚类的可能性是把聚类中心放到别的地方的所有方法中最大的，此时原来的空聚类消失。

2、问题3：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？

自己的理解：k值本来就是由用户选定，我们开始不会知道有多少个自然聚类，那么出现k值不符合实际情况的情况就肯定会出现。

3、问题4：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：同问题2相同，因为点会被“抢走”，所以会出现空聚类。

1. 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

四、读书摘要及理解或伪代码的具体实现

1、读书摘要及理解

4.2 k-均值聚类

（1）之所以在聚类中会出现空聚类，是因为聚类中心不一定是一个存在的点，在每次重新选取聚类中心后，有可能有新的聚类中心把当前的聚类中的点都给抢走了。

（2）k-模算法其实就是用属性出现频率来代替值，使用出现频率最大的属性来代替均值。

（3）处理异常点时的随机采样是先从总体数据中随机采样一部分数据，对这部分数据进行聚类，然后再采用那三种方法对剩下的数据进行聚类。

（4）K-均值的缺点：

数据的属性值必须是能够取得均值的，而不能是类别（当然k均值存在k-模算法的变体改进）

必须人工指定k值。

对异常点十分敏感，可能会产生不理想的结果

对初始种子十分敏感，可能会直接影响最后结果的好坏。

对于不是超维椭圆体的形状的聚类不适用。

4.3 聚类的表示方法

4.3.1 聚类的一般表示方法

聚类中心表示。

使用类别表示，比如决策树产生的规则集合。

最常出现的值（关键字就是这样的）

4.5 距离函数

（1）平方欧几里得距离相比于传统的欧几里得距离加大了距离较远的数据点的权重。（在损失函数中可能会用到这个思想）