1. 自己提出的问题的理解

提出的问题1：分裂聚类是如何进行分裂？与合并聚类有什么差异？

讨论后的理解：分裂所有类别中到该类中心距离最大的样本，将其单独作为一类，按照最近邻分类，直到满足某个终止条件。

与合并聚类算法的差异：合并聚类对于类别较多的训练集分类较快，而分裂算法对于类别较少的训练集分类较快。

提出的问题2：如何对不同的比例度量属性进行标准化？比如，当某一属性的变化不遵从任何函数？

讨论后的理解：可以使用多项式进行拟合，选取最大次数，对属性进行开方，再作为区间度量属性标准化。

1. 别人提出的问题的理解

问题3：为什么会形成空聚类？

自己的理解：第一种情况，初始点选取不恰当，选择了异常值作为初始点，异常值一直是空聚类。或者k值选取不恰当，k值大于适合的聚类数目，导致空聚类。

问题4：k均值算法中k值如何选取？

自己的理解：通过算法预先求出k值，例如，设置最大迭代次数，再设置不同k值，列举出同一迭代次数之后不同k值的误差平方和，选取误差平方和最小的k值作为聚类的分类数。

资料查阅：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。

可以选择SSE开始趋于平缓的k值。

1. 读书计划

本周完成内容：第四章

下周计划：3.1-3.5

1. 读书摘要

**K均值算法**

1. 随机选取k个数据点作为初始的聚类中心

2. 计算每个数据点与各数据中心的距离，用距离最近划分聚类

3. 划分完成后重新计算聚类中心

4. 重复1、2、3直至没有数据点杯重新分配给不同聚类/没有聚类中心发生变化/误差平方和局部最小

劣势

只能适用于均值能被定义的数据集上。

解决方法：变体k-模算法。k-模算法用模来代替均值作为聚类中心。假设数据实例是用r个类别属性来描述，聚类C\_j 的模是一个元组m\_j=(m\_j1,m\_j2,m\_j3,〖…m〗\_jr) ,其中m\_ji 是聚类C\_j 中第i个属性出现次数最多的值。数据实例与模之间的相似程度是指它们之间相匹配的属性的数目

用户需要事项指定k的值。

算法对于异常值十分敏感。异常值是指数据中那些与其他数据点相隔很远的数据点。异常值可能是数据采集时产生的错误或者是一些具有不同值的特殊点。

**层级聚类**

分为合并与分裂。

采用树状图结构。

**距离函数**

把各种属性化为数值计算。

布尔属性→距离矩阵

符号属性→属性值匹配程度

数据标准化：区间度量属性、比例度量属性