1. （必填）自己提出的问题的理解（罗列全部）：
2. 提出的问题1：如何保证聚类算法一定收敛呢？换言之，在算法实现时，如何判断迭代过程收敛呢？因为如果按照P100(中文版第二版)提到的三个收敛条件终止迭代，但是一个条件都无法满足，那么这时候要怎么去判断终止（收敛）条件呢？如果通过设置迭代次数终止，阈值该如何考虑？如果通过判断两次迭代结果的差别，又该从哪些角度考虑呢？

讨论后的理解：可以通过多次反复调整阈值（调参）的方式，通过对算法结果和运行时间等多方面进行评估，选择合适的阈值。

1. 提出的问题2：在算法迭代计算过程中，要怎么处理空cluster的问题呢？如果再随机给定一个新质心，对最后的聚类结果肯定会有影响；如果在计算过程中改变k值，是否可行？；给定一个无穷远的点作为新的质心，貌似可行？

讨论后的理解：尽管选择一个无穷远的虚拟点作为质心在理论上是可行的，但是这样就相当于肯定了前面的聚类结果是正确的，而实际上由于前面的聚类操作并非是启发式的，所以不能因为局部就肯定整体，即不能因为空聚类就做k-1的操作，同时也不能选择一个无穷远的虚拟结点。正如书上所说，可以选择一个最大聚类中距离质心最远的点作为新的聚类中心。

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：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：空聚类可能是由于在迭代过程中K值设置不合理或存在异常值并且在初始化时将其选为了质心。选择最远的数据点，是为了将数据量最多的聚类，将其一分为二的操作，选择距离数据中心最远的数据点作为新的聚类中心，这样能使得SSE下降最快。

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：对于只有两个状态的属性，可以将其转化为布尔类型，最后结果类似0.7代表一种可能性，即该数据的该属性有70%的可能性是属于布尔类型定义为1的状态，但是如果有大于等于3个状态的属性，很明显这样是行不通的。而之所以非对称属性不能转换，是因为区间度量函数要求属性值在区间的权重是相同的，但是非对称布尔属性本质上就是一种权重不同的属性，所以这样的转化也是没有意义的。

1. （必填）读书计划

1、本周完成的内容章节：第四章全部

2、下周计划：第三章（3.1-3.7）

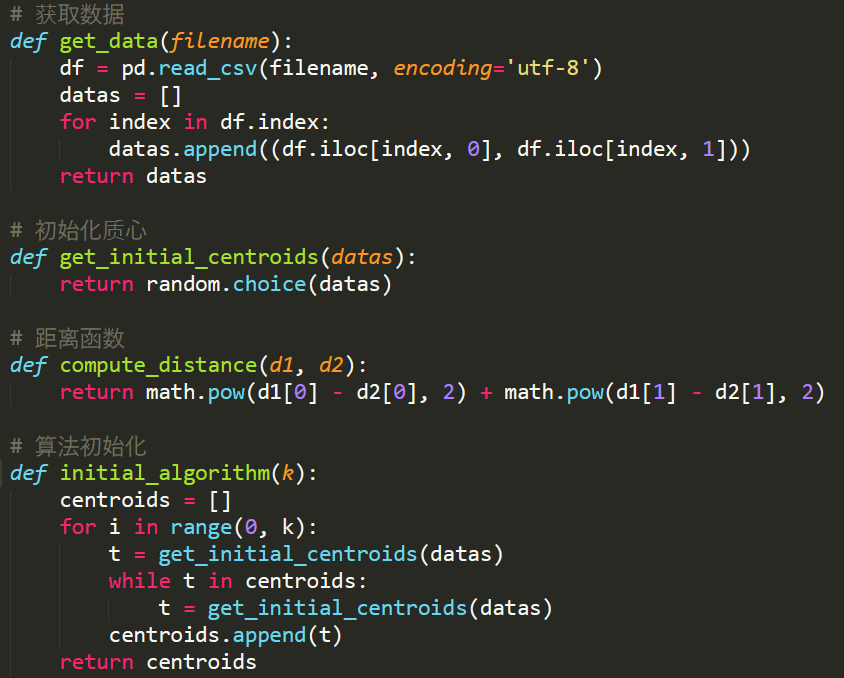
四、（选做）读书摘要及理解或伪代码的具体实现（读书摘要、伪代码的具体实现代码等可以写到这个部分）

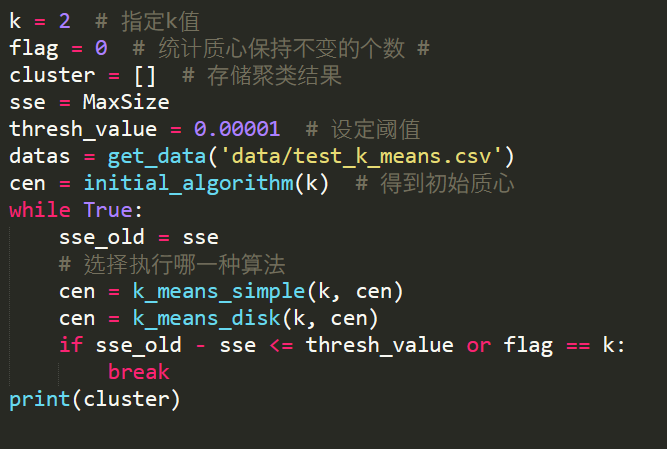
1、读书摘要及理解（选做）

2、伪代码的具体实现(选做)

实现了k-means的一般版本与disk版本：两种算法大部分的处理都是一致的，只有在重新计算质心的时候，disk版本不需要再重新计算一次每一个聚类中数据点的和，因为在划分数据到最近聚类的时候就已经计算过一次了

算法的初始化操作：





一般版本：



disk版本：

