1. （必填）自己提出的问题的理解（罗列全部）：

1、提出的问题：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？

讨论后的理解：

大概有三种解决的思路或办法。

1. Kernel-Kmeans

可以使用核函数，将低维空间的不规则图形转换到高维空间的规则图形。但核函数难以确定，实际使用效果可能不理想。

1. dbscan

一种基于密度，对噪声鲁棒的空间聚类算法。直观效果上看，DBSCAN算法可以找到样本点的全部密集区域，并把这些密集区域当做一个一个的聚类簇。

dbscan有很多优点：基于密度，对异常点有较高鲁棒性；无需确定参数K；可以进行任意形状的聚类。

dbscan的主要核心思想和概念如下：1个核心思想：基于密度；2个算法参数：领域半径R和最少点数目minpoints；3种点的类别：核心点，边界点和噪声点；四种点的关系：密度直达，密度可达，密度相连和非密度相连。

dbscan大概分为两步：

1. 寻找核心点形成临时聚类簇

扫描全部样本点，如果某个样本点R半径范围内点数目>=MinPoints，则将其纳入核心点列表，并将其密度直达的点形成对应的临时聚类簇。

1. 合并临时聚类簇得到聚类簇

对于每一个临时聚类簇，检查其中的点是否为核心点，如果是，将该点对应的临时聚类簇和当前临时聚类簇合并，得到新的临时聚类簇。重复此操作，直到当前临时聚类簇中的每一个点要么不在核心点列表，要么其密度直达的点都已经在该临时聚类簇，该临时聚类簇升级成为聚类簇。继续对剩余的临时聚类簇进行相同的合并操作，直到全部临时聚类簇被处理。

1. 谱聚类

谱聚类是从图论中演化出来的算法，后来在聚类中得到了广泛的应用。它的主要思想是把所有的数据看做空间中的点，这些点之间可以用边连接起来。距离较远的两个点之间的边权重值较低，而距离较近的两个点之间的边权重值较高，通过对所有数据点组成的图进行切图，让切图后不同的子图间边权重和尽可能的低，而子图内的边权重和尽可能的高，从而达到聚类的目的。

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. 对空聚类的处理

如果所有的点在指派步骤都未分配到某个簇，就会得到空簇。如果这种情况发生，则需要某种策略来选择一个替补质心，否则的话，平方误差将会偏大。一种方法是选择一个距离当前任何质心最远的点。这将消除当前对总平方误差影响最大的点。另一种方法是从具有最大SSE的簇中选择一个替补的质心。这将分裂簇并降低聚类的总SSE。如果有多个空簇，则该过程重复多次。

1. 层次聚类中的合并聚类与分裂聚类

分裂聚类有着很明显的缺点：每次分裂的k难以确定；每次分裂如果出现错误，将累积传递到下一次聚类；繁琐，不易实现。而合并聚类始终基于“最近”这一准则，不需要人为规定谁与谁合并，故而犯错的概率较小，同时合并聚类易于实现。

1. k值的确定
2. 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。
3. 轮廓系数法

该方法的核心指标是轮廓系数，某个样本点Xi的轮廓系数定义：S=(b-a)/max(a,b)。其中，a是Xi与同簇的其他样本的平均距离，称为凝聚度，b是Xi与最近簇中所有样本的平均距离，称为分离度。求出所有样本的轮廓系数后再求平均值就得到了**平均轮廓系数**。平均轮廓系数的取值范围为[-1,1]，且簇内样本的距离越近，簇间样本距离越远，平均轮廓系数越大，聚类效果越好。那么，很自然地，平均轮廓系数最大的k便是最佳聚类数。

1. （必填）读书计划

1、本周完成的内容章节：web data mining第四章。

2、下周计划：web data mining第三章