**读书报告**

姓名：赖泽升

1. 自己提出的问题的理解（罗列全部）：

**提出的问题1：**使用k-均值算法的时候，为什么全局最小值对于大规模数据集来说在计算上是不可行的？

**讨论后的理解：** 局部最小值指的是在选取某种初始点的情况下能达到的最佳分类，而全局指的是按照要求能得到的最佳分类。因为并不能遍历所有的初始点组合，所以对于大规模数据集来说取不到全局最小值。

1. 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2：** 在4.3.1节的最后，One can use the set of rules to evaluate the clusters to see whether they conform to some existing domain knowledge or intuition.怎么理解，怎么操作呢？

**自己的理解：** 用相应规则完成聚类后看看是否存在已有的知识模型可以应用到上面，去检验这个聚类是否合适。

**问题3：** 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：** 因为它最有可能是潜在的聚类，这样可以更快迭代出结果。

**问题4：** 为何距离函数需要有平方？是为了保证SSE可以用增量来计算吗？

**自己的理解：** 在数据量大的情况下，开根号会造成误差较大，而且平方也便于运算。

**问题5：**在k-均值算法的硬盘版本中，是如何实现了每次for循环中，算法只是简单地扫描全部数据一次，普通版本的k-均值算法呢，扫描了多少次，是如何扫描的？

**自己的理解：**用了类似于累加的手法，让内存每次都读了一个数据。而在普通版本中，至少要两次。一次是分配聚类，一次是计算均值。

**问题6：**To be safe, we may want to monitor these possible outliers over a few iterations and then decide whether to remove them. It is possible that a very small cluster of data points may be outliers. Usually, a threshold value is used to make the decision.其中，threshold value指的是什么？

**自己的理解：**指的是一个阈值，这个值数量要很少而且与其他聚类距离要远。

读书计划

1、本周完成的内容章节：4.1-4.3

2、下周计划：5.1和回顾前面所读的部分