窦一蒲-2020年2月23日-读书报告

读书报告内容：

1. （必填）自己提出的问题的理解（罗列全部）：
2. 提出的问题1：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？

讨论后的理解：用那些rule给出的的一些判别标准是否适用于已有的经验规律。比如服装厂对人身高进行聚类 某一聚类结果把男生身高160和190聚类在了一起，因此得出结论160和190的男生需要穿同一个尺寸的衣服 但是根据领域知识来讲 这显然是错误的。

1. 提出的问题2：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？

讨论后的理解：用了类似于累加的手法，让内存每次都读了一个数据，但每个数据都是要读的。至少两次，一次是分配聚类，一次是算均值。

1. 提出的问题3：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？能具体讲一下怎么监视与这个方法的整体吗？

讨论后的理解：这个阈值应该就是一个评判标准吧更倾向于一点到所有质心的和的阈值，在多次循环中监视，大于阈值的被认为是outliers，将被删掉。

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4： 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？

自己的理解：局部最小值指的是在选取这种初始点的情况下能达到的最佳分类，因为每次选取的初始点不一样，所以最后聚类的结果不同。全局指的是按照要求能得到的最小值，由于初始点是随机的，很难保证能达到全局最小。

5、问题5：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：选取这个的原因是它最有可能是潜在的聚类，再靠后面多次迭代来判断。

6、问题6：SSE的公式里面的平方是怎么回事？

自己的理解：开根号运算本身误差会很大的，在数据量大的情况下，平方便于运算。

1. （必填）读书计划

1、本周完成的内容章节：第四章4.1-4.3

2、下周计划：第五章5.1节