读书报告

71117203

袁佳怡

1. 自己提出的问题的理解：
2. 提出的问题1：4.3.1中提到的one can use the set of rules to evaluate the cluster to see whether they conform to some existing domain knowledge or initution可以利用规则集合来评估聚类是否符合某些已经存在的领域知识或常识？

讨论后的理解：利用规则集合来评估聚类是否符合某些已经存在的领域知识或常识，举例说明，某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的男生需要传一个尺寸的衣服，但根据领域知识来讲，这显然是错误的，所以这就是利用已经存在的领域知识或常识进行判断。

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：在硬盘实现的版本中，是类似于累加的手法，值扫描全部数据一次就能够进行分类和计算，在普通的k-均值实现中，对于每一次迭代至少扫描数据两遍，一遍是分配聚类，另一遍是算均值。

1. 别人提出的问题的理解：
2. 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？

自己的理解：在进行聚类的时候，选取了一个阈值，与其他距离都很远的异常点设置一个阈值提前删去这些，但是呢因为第一次的质心是随机选择的，所以monitor these possible outliers over a few iterations and then decide whether to remove them，remove一个点不是一次判断就得出来的结果。，在多次循环中进行监视，通过多次迭代将可能的outlier去除。

1. 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：因为每次选取的初始质心是不一样的，所以最后聚类的结果也可能是不同的，局部最小值指的在选取这种初始点的情况下能达到的最佳分类，全局最优指的是按照要求能得到的最小值，由于初始点是随机的，所以只能达到一个局部最小值，很难达到一个全局最小，多次取值去逼近全局最小。

1. 问题5：SSE的计算为什么要有平方？是为了保证SSE可以增量来计算吗？

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，第五章读过之后没有很理解

2、下周计划：第五章继续加深理解，《统计学习》第一二章