2020.02.23问题列表

侯宇捷：

问题1：如果一个聚类任务的数据中，其属性既有离散值又有连续值怎么办？是否可以将连续属性按照数据集中的数据离散化分为几个区间，将区间视为离散的继续做？

问题2:

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？

赖泽升：

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？

孙毅远：

1.除了人工选取外，有无合理的方法去选择恰当的初始点？

窦一蒲

1. 在4.3.1节，the centroid representation alone works well if the clusters are of the hyper-spherical shape. If clusters are elongated or are of other shapes, centroids may not be suitable.是因为可以计算出the radius，standard deviation of the cluster吗？我不太清楚。
2. 在4.3.1节的最后，One can use the set of rules to evalute the clusters to see whether they conform to some existing domain knowledge or intuition.怎么理解，怎么操作呢？

袁佳怡：

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 intution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
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？

黄一凡：

1. 是否可以通过多次重复k-means算法，每次选取不同的数据点作为初始centroids，最后取最小SSE，来对算法进行优化？
2. 对于出现outlier的情况，书上介绍的两种方法感觉是直接忽略它们，这样会影响最终cluster的形成吗，是否可以增加k的数量，也就是把outlier也作为一个聚类？

李鸿烨：

1. 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？
2. 在4.2.2节中提到，公式(2)的求和运算是分开计算的，那么这个分开计算是如何实现的呢？

方骏：

1. 文中所说的k-mode究竟是属于哪一类的模型，属性有什么特征？

2. 书上第129页：

这里究竟是什么特殊情况。

喻泽弘：

1. 对于4.2.3节中提出的计算good initial seed的方法，是否存在除了outliers以外的其它弊端，比如之后选取的seed和之前选取的seed重合或者距离很近？

闻浩：

1. 当使用SSE时，局部最小如何理解？为什么对大量数据集来说局部最小是不可行的？

2. 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？