统计学习读书报告

09017244 郑健雄

1. 自己提出的问题的理解：
2. 层次聚类与k均值聚类的差别是什么？有什么情况是只有层次聚类能进行处理的？

讨论后的理解：

k均值聚类更适合数据具有扁平化特征的数据，其聚类的数量也是实现给定的。而层次聚类适合处理层次化的数据，比如说新闻的分类问题，大话题小话题的差异不是扁平化结构可以体现的，此外，层次聚类也可以帮助寻找合适的聚类中心作为初始点，让k均值聚类效果更好。

1. k均值聚类算法与EM算法是否有相同之处？

讨论后的理解：

从处理的步骤上来看，它们有着类似的结构。EM算法分别进行E和M两种步骤，有着交替的特征，让期望达到最优再调整模型参数。而聚类也有着这种特征，让数据点进行聚类之后，再更新它们的中心参数。它们有着很类似的思想，也容易出现局部最优的问题。

1. 能量函数的直观解释是什么？为什么最优化能量函数就可以保证得到最优的划分？

讨论后的理解：

能量函数实际上衡量了两个指标，距离的因素以及划分的因素，单纯追求让距离变小而不考虑划分的情况的话，能量函数本身也会很大。其最优化的结果就是给出一个最好的划分，同时让聚类内部的点之间距离的平方也达到最小。可以选中k个中心以及它们周边的聚类点，以保证找到最好的聚类。但实际上这是一个NP问题。

1. 别人提出的问题的理解
2. 为什么k均值聚类算法的复杂度是o(mnk)？这样是否没有考虑迭代的次数？

自己的理解:

实际上这个复杂度应该没有考虑迭代的次数，因为这个参数不是很确定。而mnk代表n个样本点计算k次距离，而每一次距离运算需要m次运算，因为样本维度是m。

1. 如何用二分查找快速找到最优的k？

自己的理解：这个可以直接设定一个比较大的k值，开始计算，然后进行二分操作，如果半径变化不明显，就向下二分，如果上升很明显就向上二分，这样可以很快确定k值。

1. 间距为什么选取欧式距离平方？如果选用欧式距离会有什么变化？

自己的理解：

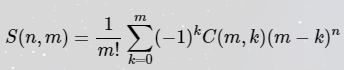
实际上应该是为了计算的方便，所以采用了平方，因为平方不用进行开放运算而且等价地表示了距离。

1. （14.21）式子表示的可能分法的数目是什么含义？不是很明白这个式子是如何表示出所有可能分发的数目。

自己的理解:



根据组合的性质，该等式是显然成立的。那么根据容斥定理，可以得出:

。

1. 读书计划

1、本周完成的内容章节：20章，21章第一节

2、下周计划：21章结束，22章

四、读书摘要及理解

1.层次聚类:

层次聚类分为聚合聚类和分裂聚类两种聚类，一种是从分散到集中，找最近的两个点合并新类，最后只有一个大类，另一种是从一个大类开始，分离最远的点，产生新类，直到结束。

2. 聚类的是三个要素:

距离和相似度:对于聚类，相似的指标是根本要素。

合并规则:管理如何合并，决定哪些类合并，比如层次聚类中就是距离最短的两个类合并。

停止条件：聚类需要一个停止的条件，比如聚类的数量或者最小的直径，满足要求就可以停止了。

3. k均值聚类：

k均值聚类的k是一个给定的值，其目标是给出一个样本点的最优划分，使得能量函数最小，也就是尽可能让样本的分布像一个类别分布而不是随机的。

直接求解最优分布是np问题，所以这里用近似的迭代算法来求解。该算法有着类似em算法的特征，也是两种优化策略交替使用，聚类然后计算中心，再次聚类。其有着比较好的效率，但是容易陷入局部最优，并且开始的中心点对于结果的影响很大。

为了处理k值问题，可以使用二分算法来估计一个合适的k。而为了更好地处理中心点的问题，可以采用层次聚类的方法给出初始的中心点。

思考:聚类作为无监督学习中的一种方法在生活中有着重要的作用，比如生产问题，模式问题等等，聚类方法可以挖掘出数据潜在的关系，给出它们的模式，对于人们进行分析与思考并且给出新的假设是很有意义的，比如说可以通过聚类的结果来产生有监督学习的标注以及挖掘数据之间的潜在联系。