**第十四章读书报告**

09118223 吴亦珂

读书进度：十四章读完

1. **问题列表**

（我提出）（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。

讨论结果：14.21其实表达的是一个组合问题，其公式是第二类斯特林数，将其根据容斥原理展开可以得到式（14.21）。

（我提出）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？

讨论结果：这个应该是没有考虑迭代的次数的，因为一次迭代的时间复杂度就为o(mnk)。我觉得迭代次数不好定量地用数字进行度量，因而值写出来异步迭代的时间复杂度。

（别人提出）为什么p261页聚合聚类的时间复杂度是O(n³m)？我分析出来的结果只有n²m(第(1)步计算距离的时间约为O(m)，两两距离共有O(n²)组，这一步总共的时间复杂度是O(n²m).第(1)步的时间约为O(n).第(3)步涉及一次扫描，时间约为O(n).第(4)步重算的距离共有O(n)组.(3)(4)步共要经历(n/2+n/4+...+1)=O(n),总共的时间复杂度是O(n²))

讨论结果：（4）中说只有类的个数为1是才会中止计算，所以如果有n个样本应该需要（n-1）次迭代，所以时间复杂度为o()。

（别人提出）样本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？

讨论结果：因为单单选择欧式距离还需要开根号，而k均值聚类时比较的是距离大小，同时平方并不会影响结果。

（别人提出）k均值聚类算法与EM算法是否有相同之处？

讨论结果：我觉得最大的差别就是层次聚类会将类别进行一个分层，也就是可以得到一个阶梯式的分类结果。而k均值聚类是直接将样本分为k个类。得到结果是平坦的，非层次化的。而且，k均值聚类k必须实现确定。同时k均值聚类必须用欧式距离来度量，而层次聚类可以选择多种度量方式。

1. **下周读书计划**

下周读书计划是读完第十五章前半部分内容，SVD在上学期期末考试中也有所涉及。所以对于这个概念也不是特别陌生。但是，这也是我第一次正式学习这一概念。所以，我准备结合自己所学知识，进一步巩固所学到的知识。

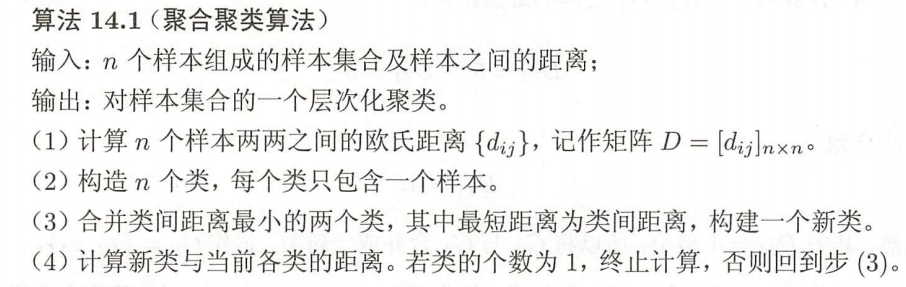
1. **读书收获**

第十四章后半部分主要介绍了层次聚类以及k均值聚类。

层次聚类有两种，一种是自下而上的聚合聚类，另一种是自上而下的分类聚类。本书中主要介绍了聚合聚类。

1. 聚合聚类：聚合聚类要预先确定下面三个要素：（1）距离或相似度（2）合并规则（3）停止条件。

聚合聚类算法：

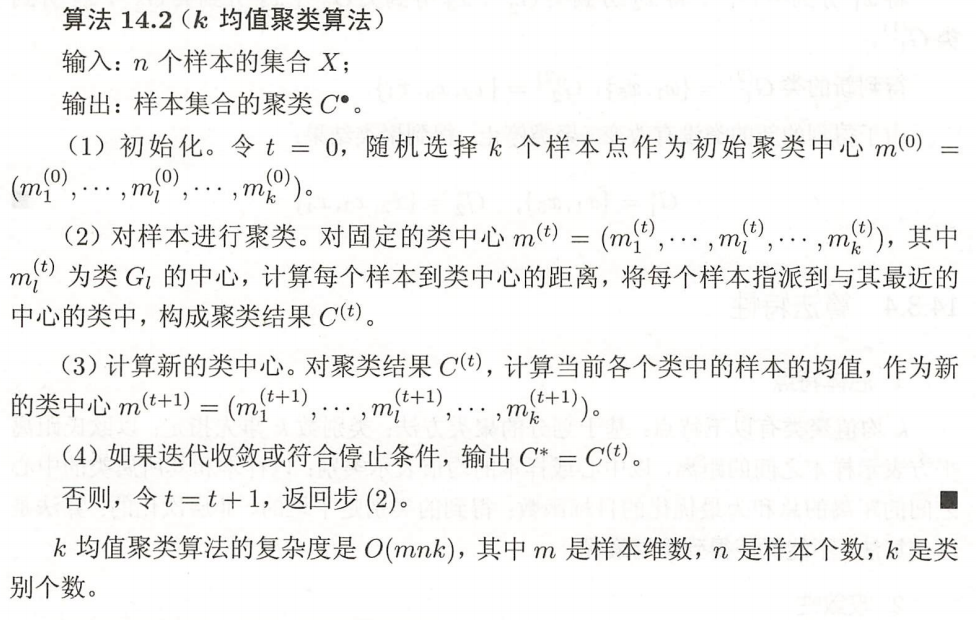


1. k均值聚类

欧式距离的平方表示样本之间的距离，中心或样本的均值表示类别。

简单来说就是将数据点分到离其最近的样本中心，之后不断更新样本中心，直至收敛。

算法内容：



k值选择：可以采用类的平均直径衡量。