**《数据仓库与数据挖掘》作业**

2、数据集A=｛1，1，3，3，4，5，5，5，6，7，7，7，9，10，12，15｝

（1）请用等深分箱法将其分成k=4个箱子

（2）请用等宽分箱法将其分成k=4个箱子

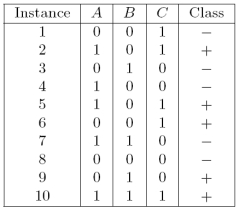
3、设某公司现有15名员工的基本信息，包括其个子为高个、中等、矮个的分类标识。



公司现刚招进一位名叫刘萍的新员工Z1，令k=5，试采用k-最近邻分类算法判断员工刘萍的个子属于哪一类？



4、依据表中数据集，使用朴素贝叶斯方法预测样本(A=0,B=0,C=1)的类标号。



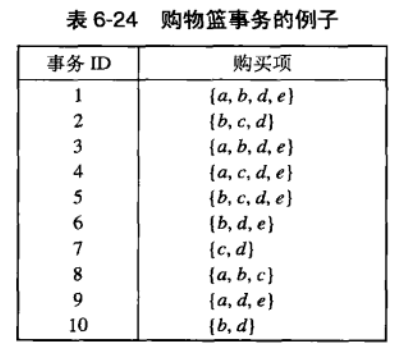
5、给定如图所示的事务表和格结构，用如下标记直接在图上标识出每一个结点，假定支持度阈值为30%。（请先统计每个结点的支持度，支持度写在左上角，类别字母写在右下角）

M：如果结点是极大频繁项集

C：如果结点是闭频繁项集

Ｎ：如果结点是频繁的，但既不是极大的也不是闭的

Ｉ：如果结点是非频繁的



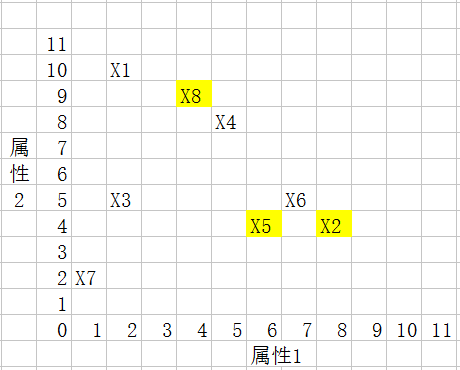


1. 支持度-置信度框架存在什么样的局限性？

7、请用k-均值算法把表中S的8个样本数据聚为3个簇，并给出每个簇的平均值点，假设初始迭代时选择X2、X5、X8作为初始簇中心点。

表：样本数据集S

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 数据点 | 属性1 | 属性2 | 数据点 | 属性1 | 属性2 |
| X1 | 2 | 10 | X5 | 6 | 4 |
| X2 | 8 | 4 | X6 | 7 | 5 |
| X3 | 2 | 5 | X7 | 1 | 2 |
| X4 | 5 | 8 | X8 | 4 | 9 |



8、请对表中6个点的坐标求欧几里得距离，分别用单链（min）和全链（max）算法进行层次聚类，并绘制树状图。

表：样本数据集S

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 数据点 | x坐标 | y坐标 | 数据点 | x坐标 | y坐标 |
| X1 | 2 | 10 | X4 | 6 | 4 |
| X2 | 8 | 4 | X5 | 1 | 2 |
| X3 | 2 | 5 | X6 | 4 | 9 |

