装 订 线

任课教师：

装 订 线

学号：

姓名：

装 订 线

班级：

西 安 电 子 科 技 大 学

考试时间 120 分钟

《数据仓库与数据挖掘》试题 B 卷

题号	一	二	三	总分
分数				

1. 考试形式：闭卷√ 开卷□ ；2. 本试卷共三大题，满分 100 分；
3. 考试日期： 年 月 日；（答题内容请写在装订线外）

一、填空题（每小题 2 分，共 20 分）

1、两个文档向量 X=（3，2，0，5，0，0，0，2，0，0）Y=（1，0，0，0，0，0，0，1，0，2）的余弦相似性为：_____。

2、一组数据 9，1，2，2，3，4，12，0，33 的极差为：_____。

3、解决分类问题时，当决策树节点的规模过大后，模型的训练误差在降低，但检验误差却随之升高，这种现象称为：_____。

4、频繁项集发现中，FP 增长算法将每个事务数据映射到 FP 树的一条_____。

5、一个无向网络中包含 10 个节点和 30 条边，则该网络中边的密度为：_____。

6、假设属性 income 的最大最小值分别是 12000 元和 98000 元。利用最大最小规范化的方法将属性的值映射到 0 至 1 的范围内。对属性 income 的 73600 元将被转化为：_____。

7、设 X={1，2，3}是频繁项集，则可由 X 产生_____个关联规则。

8、在基本 K 均值算法里，当邻近度函数采用欧氏距离的时候，合适的中心点是簇中各点的 _____。

9、Ward 方法将两个簇的邻近度定义为两个簇合并时导致的 _____的增量，它是一种凝聚层次聚类技术。

10、数据挖掘的主要任务包括：_____、_____、_____。

二、问答题（共 30 分）

1、（10 分）考虑一个文档-词矩阵，其中 f_{ij} 是第 *i* 个词出现在第 *j* 个文档中的频率，*m* 是文档总数目。考虑如下逆文档频率变换：

$$f'_{ij} = f_{ij} \times \log \frac{m}{d_i}$$

其中，d_i 是出现第 *i* 个词的文档数目。

（a）如果一个词仅仅出现在一个文档中，该变化的结果是什么？如果出现在所有文档中结果又是什么？

（b）该变换的目的是什么？

2、（5 分）在采用抽样来减少需要可视化的数据对象时，简单随机抽样（无放回）是一种有效的方法吗？为什么是？为什么不是？

3、（15 分）领导者算法基本思想：用一个数据点（称作领导者）代表一个簇，并将每个点指派到最近的领导者对应的簇，除非距离大于用户指定的阈值。如果一个点到最近的领导者的距离大于阈值时，该点成为一个新簇的领导者。

(a) 与 K 均值算法比较，领导者算法的优点和缺点是什么？如何改进？

(b) 对 M 维空间中的 N 个数据点进行聚类，分析领导者算法的时间复杂度。

装

订

线

三、计算题（每题 10 分，共 50 分）

1、（10 分）考虑如下二元分类问题的数据集：

A	B	类
T	F	-
T	T	+
T	T	-
T	F	-
T	T	+
F	F	-
F	F	+
F	F	-
T	T	+
T	F	-

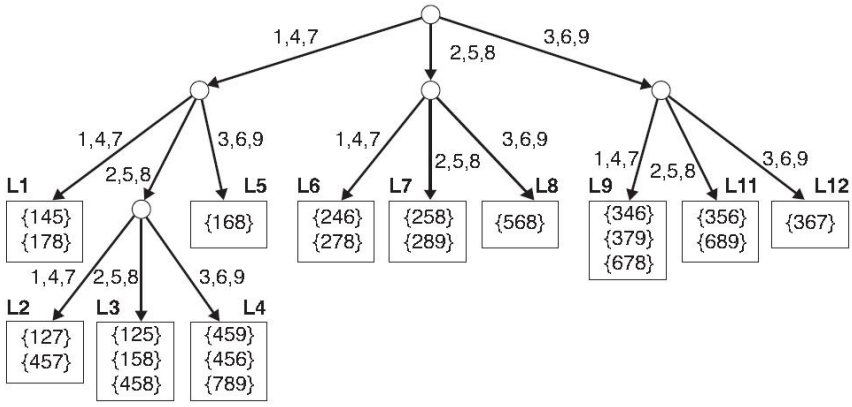
(a)计算按照属性 A 和 B 划分时的信息增益。决策树归纳算法会选择哪个属性？
(b)计算按照属性 A 和 B 划分时的 Gini 系数。决策树归纳算法将会选择哪个属性？

装

订

线

2、（10 分）Apriori 算法使用 Hash 树数据结构有效地计算候选项集的支持度，下图为一个候选 3-项集的 Hash 树。

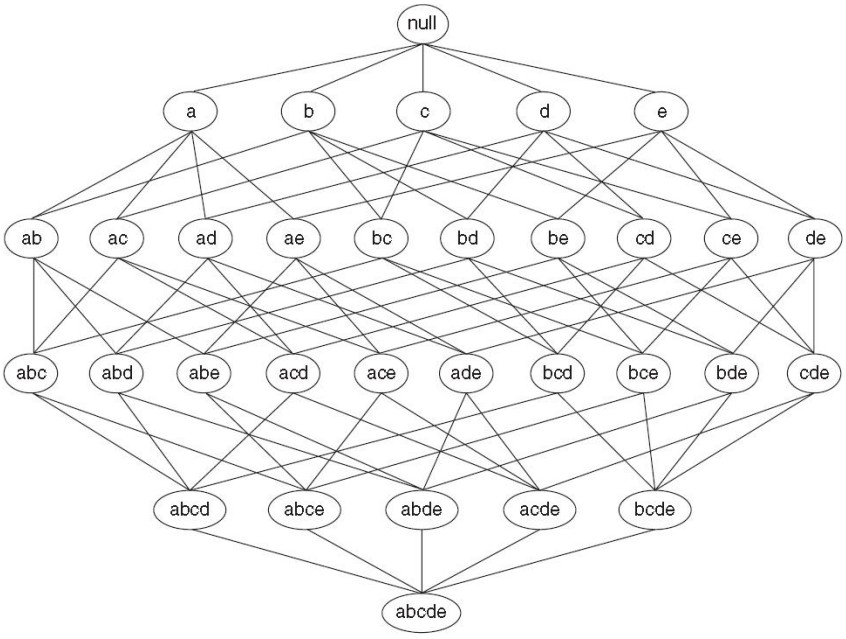


（a） 给定一个包含项{1，4，5，8，9}的事务，在寻找该事物支持的候选项集时，访问了该 Hash 树的哪些叶子节点？

（b） 使用（a）中访问过的叶子节点确定事务{1，4，5，8，9}包含的候选项集。

3、（10 分）给定如下事务数据，最小支持度 30%，在格结构图中用字母 M、C、I 分别标记：极大频繁项集、闭频繁项集、非频繁项集。

事务 ID	购买项
1	***{a, b, d, e}***
2	***{a, b, c, d}***
3	***{ b, d, e}***
4	***{a, c, d, e}***
5	***{b, c, d}***
6	***{a, d, e}***
7	***{c, d}***
8	***{a, b, c}***
9	***{a, d, e}***
10	***{b, d, e}***



4、（10 分）计算如下混淆矩阵的熵和纯度。

簇	娱乐	财经	国外	都市	国内	体育	合计	熵	纯度
#1	1	1	0	22	4	3	31		
#2	3	5	4	5	4	5	26		
#3	0	10	1	1	12	2	26		
合计	4	16	5	28	20	10	83		

5、（10 分）给定如下簇标号集和相似度矩阵：

簇标号集

数据点	簇标号
P1	1
P2	2
P3	2
P4	2

相似度矩阵

数据点	P1	P2	P3	P4
P1	1	0.1	0.2	0.1
P2	0.1	1	0.9	0.8
P3	0.2	0.9	1	0.7
P4	0.1	0.8	0.7	1

（a）计算该相似度矩阵与理想的相似度矩阵之间的相关度。如果两个对象 *i*，*j* 属于同一个簇，则理想的相似度矩阵的第 *ij* 项为 1，否则为 0。

（b）计算每个点、每个簇的轮廓系数。

装

订

线