中 山 大 学

二00七 年港澳台人士攻读博士学位研究生入学考试试题

科目代码：522

科目名称：数据挖掘

考试时间： 4 月 22 日 上午

考 生 须 知

全部答案一律写在答题纸上，答在试题纸上的不得分！答题要写清题号，不必抄题。

一、 名词解释（20 分，每个 5 分）

数据仓库

关联分析

概念分层

支持度(support)、置信度(confidence)

二、看以下表格：

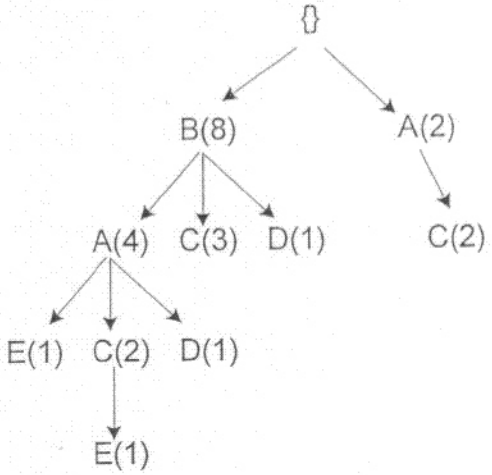
Age	Income	Class
< 35	high	no
[35,50]	high	yes
[35,50]	medium	yes
[35,50]	low	yes
[35,50]	low	no
[35,50]	low	yes
< 35	medium	no
< 35	low	yes
< 35	medium	yes
[35,50]	medium	yes
[35,50]	high	yes
[35,50]	medium	no

用信息增益(Gain)作为属性选择的度量，为以上训练数据集建立一棵完整的判断树（decision tree），写出过程。（20 分）

（$\log_2\frac{1}{12}$=-3.585； $\log_2\frac{2}{12}$=-2.585； $\log_2\frac{3}{12}$=-2； $\log_2\frac{4}{12}$=-1.585； $\log_2\frac{5}{12}$=-1.263； $\log_2\frac{6}{12}$=-1；

$\log_2\frac{7}{12}$=-0.7776； $\log_2\frac{8}{12}$=-0.585； $\log_2\frac{9}{12}$=-0.415； $\log_2\frac{10}{12}$=-0.263； $\log_2\frac{11}{12}$=-0.12553）

三、给定一棵 FP-Tree 如下所示：



假设绝对最小支持度为 2，给出所有的：
频繁模式（frequent patterns），并用图示说明频繁集生成的过程（15 分）

四、给定两个对象，分别用元组（22, 1, 42, 10）和（20, 0, 36, 8）表示。（6 分）
a) 计算两个对象之间的欧几里德距离（Euclidean Distance）；
b) 计算两个对象之间的曼哈坦距离（Manhattan Distance）；
c) 计算两个对象之间的明考斯基距离（Minkowski Distance），$q=3$。

五、假设数据挖掘的任务是将如下的 8 个点（用（x,y）代表位置）聚类为 3 个簇：
$A_1(2,10), A_2(2,5), A_3(8,4), B_1(5,8), B_2(7,5), B_3(6,4), C_1(1,2), C_2(4,9)$
距离函数是欧几里德距离。假设初始选择 A_1, B_1, C_1 分别为每个聚类的中心，用 k-平均算法来给出（15 分）
（a） 在第一次循环执行后三个聚类中心
（b） 最后的三个簇

六、*partition* 是关联规则挖掘中一种求解频繁模式的算法思想，它采取分而治之的策略，将大小为 N 的事务数据 D 库分为 p+1 个部分，满足 $D=\bigcup_{i=0}^{p} D_i, \forall i \neq j, D_i \cap D_j = \phi$，其中 D_i 是第 i 个部分，$N_i = |D_i|$ 是分区 D_i 中的事务数。*partition* 算法首先计算局部频繁模式，也就是在每个分区中频繁的模式，第二步将所有的局部频繁模式进行整合，计算整个数据库中实际的频繁模式（全局频繁模式）。请证明如果一个模式是全局频繁模式，那么它一定至少在一个分区中是频繁的。（9 分）

七、 描述朴素贝叶斯分类的工作过程。（15 分）