广东工业大学考试试卷（A）

2018 -- 2019 学年度第 二 学期

课程名称：数据挖掘 学分 2 试卷满分 100 分

考试形式：闭卷

题 号	一	二	三	四	五	六	七	八	九	十	总分
评卷得分											
评卷签名											
复核得分											
复核签名											

说明：因为格式要求，第一页不会自动换页，但第二页起可以自动换页，页码也会自动生成。（请教师出题时把此说明删除！）

1. 简答题（共计 20 分，每小题 5 分）

(1) 请列举 5 种数据预处理方法，并简要说明。

(2) 举例说明某企业如何应用数据挖掘，给出其挖掘过程。

(3) 请回答关联分析中为何要首先寻找频繁项集，在非频繁项集中可以发现规则吗？

(4) 给出三种离群点分析方法，并加以简要叙述。

2. 对下列成绩表进行最小-最大规范化，将数据变换到[0, 1]区间。找出 001, 002, 003 三位同学中成绩最接近的两个同学，用曼哈顿距离计算。（10 分）

学号	C 语言	C 语言课程设计	JAVA 语言
001	75	4	80
002	76	2	82
003	75	3	70
004	56	1	78
005	86	5	65

30 4 17

3. 客户收入属性income的值（人民币元）如下：（共计5分）

4500, 800, 1000, 1500, 1500, 1800, 2300, 4800, 2000, 2500, 2800, 3000, 1200, 4000, 5000, 3500

(1) 请按等深度为4进行分箱（3分）；

(2) 按均值对分箱后的数据平滑，写出分箱后的结果和平滑后的结果（2分）。

4. 某商场为做定点营销，想要根据已有客户信息（具体如下表所示）建立一个分类树，从而有效选出目标客户，假如该商场使用ID3算法建立决策树，完成下列各题（10分）。

(1) 首次选取的分类属性为哪个？为什么？

(2) 并画出首次建立的树结构。

（参考：$\log_2 3=1.585$, $\log_2 5=2.322$）

年龄	收入	性别	是否有车	目标客户
<=30	High	Female	No	否
<=30	High	Female	Yes	否
31~50	High	Female	Yes	是
>50	Medium	Female	No	是
>50	Low	Male	Yes	否
31~50	Low	Male	Yes	是
<=30	Medium	Female	No	否
31~50	Medium	Male	No	否
<=30	Medium	Male	Yes	是
>50	Medium	Male	Yes	是
31~50	Medium	Female	Yes	是
>50	Low	Male	Yes	否

5. 已知各种类型的水的矿物质含量情况，如下表所示（6分）：

Mg+浓度	Na+浓度	Ca+浓度	类型
高	中	低	矿物质水
中	高	高	矿物质水
低	中	低	矿泉水
高	中	中	纯净水
中	低	高	矿泉水
高	高	中	矿物质水
低	低	中	纯净水
中	高	中	纯净水
低	高	低	矿泉水
低	中	高	矿泉水
低	低	高	纯净水
中	中	高	矿物质水

请使用Bayes分类算法确定某含量为（Mg+：中，Na+：中，Ca+高）的水的类型？

6. 已知某学校某学院的教师的等级，工作年限等信息，如下表（9 分）：

Name	RANK	YEARS	Tenured
Milk	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Charly	Professor	2	yes
Jim	Associate Prof	8	yes
Jack Lee	Associate Prof	3	no
Frank	Associate Prof	5	no
Bill	Assistant Prof	10	yes
Mary2	Professor	4	yes
Yi Liu	Professor	1	no
Anna	Associate Prof	7	no

对于该学院的教师信息，请使用 k-summary 算法将其划分为 2 个簇。并选择 Jack Lee 和 Mary 这两位教师为初始的簇中心，请问：

（1）第一次划分后的每个簇的 CSI 信息；

（2）计算 Milk，Mary 和 Charly 三位老师在第二次划分后分别分到哪个簇？(给出计算过程)

（距离计算使用 Mahattan 距离，参考：$dif(p_l,C_l)=1-\frac{Freq_{C|D_l}(p_l)}{|C|}$ 和

$d(p,C)=\left(\sum_{l=1}^{m} dif(p_l,C_l)^x\right)^{1/x}$）。

7. 一事务数据库如下表，假设 min-support = 2/9，min-confidence = 70%（共计 12 分，每小题 6 分）：

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

(1) 用 apriori 算法挖掘所有频繁项集，并给出挖掘过程。

(2) 给出形如下列形式的强关联规则，其中 p，q，和 r 为购买项：

{p, q}→{r}；　　{p}→{q, r}

8. 在 10000 个人中，有 6000 个人买产品 A；7500 个人买产品 B；4000 个人既买产品 A 又买产品 B，如下表所示（共计 8 分，每小题 4 分）：

	购买产品 A	不买产品 A	合计
购买产品 B	4000	3500	7500
不买产品 B	2000	500	2500
合计	6000	4000	10000

考察下列关联规则：

购买产品 A → 购买产品 B

购买产品 A → 不买产品 B

(1) 分别计算上述关联规则的支持度，置信度，提升度。

(2) 在上述关联规则中，那条关联规则更有意义。

9. 假设当地银行有一个数据挖掘系统。该银行正在研究你的信用卡使用模式。注意到你在家庭装修店有多笔交易，银行决定与你联系，提供有关家居改善方面的特别贷款信息。回答下列问题（共计 20 分，每小题 5 分）。

(1) 这是否与你的隐私权相冲突。

(2) 给出另外一个使你感到数据挖掘侵犯你的隐私权的情况。

(3) 描述一种保护隐私的数据挖掘方法，它可以允许银行进行客户模式分析，而不侵犯顾客的隐私权。

(4) 举出 2 个数据挖掘对社会有帮助的例子，同时分别给出它们可用来危害社会的方法。