装 订 线 ____ 任课教师：____ 装 订 线 ____ 学号：____ 装 订 线 ____ 姓名：____ 装 订 线 ____ 班级：____

西 安 电 子 科 技 大 学

考试时间 120 分钟

2017 年春

《数据仓库与数据挖掘》试题

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、判断题，正确的在题后的线上写 T，错的写 F。（共 10 分，每题 1 分）

1、数据中的噪声点或离群点总是被丢弃的 _____

2、发现复杂网络中的模块或者社团是数据挖掘任务 _____

3、对公司的销售额进行汇总统计是数据挖掘_____

4、按姓名给来宾排座位是数据挖掘_____

5、两个对象余弦度量的取值可以大于 1 _____

6、使用历史记录数据预测某地区未来的降雨量是数据挖掘 _____

7、K-means 算法擅长处理不规则分布的数据聚类分析 _____

8、DBSCAN 算法可以同时发现高密度和低密度的簇 _____

9、根据客户交易活动对客户进行划分不是数据挖掘 _____

10、发现 DNA 序列中频繁出现的子序列不是数据挖掘_____

二、简答与计算（共 40 分）

1.（5 分）描述数据集、数据对象、属性三者的含义与区别。说明数据的类型和属性的类型，并举例。

2.（5 分）对比率属性 x 使用取倒数变换，得到一个新属性 x*，在 x*的取值区间（a,b)内，x*与另一个属性 y 具有线性负相关关系，回答下面的问题，

（1）换算成 x，（a,b)的对应区间是什么？

（2）给出 y 关联 x 的方程。

装

订

3.（5 分）简述分类模型的两种误差，并简述什么是模型的过分拟合。

线

4.（5 分）画出下面给定数据集 D，计算其 25%分位数、中位数和 75%分位数，并说明其特点及应用。

数据集 D：{10, 20, 30, 32, 40, 44, 49, 50, 60, 65, 71, 88, 90, 11, 92, 95, 100, 8, 17,65, 0.001,10000}

装

订

5. (5 分) 简单描述 2 种自己熟悉的数据可视化技术。

线

6.（10 分）对于下面的向量 x 和 y，计算指定的相似性或距离度量

(a) x=(2,-1,0,2,0,-3)，y=(-1,1,-1,0,0,-1) 欧几里得距离

(b) x=(1,1,0,1,0,1)，y=(1,1,1,0,0,1) 余弦

(c) x=(1,1,1,1),y=(2,2,2,2) 相关系数

(d) x=(0,1,1,1,0,1,0,0,0,1)，y=(0,1,0,1,1,1,0,0,0,1) Jaccard

(e) x=(1,1,0,1,0,1)，y=(1,1,1,0,0,1) 简单匹配系数

装

订

线

三、综合题（共 50 分）

1. （20 分）考虑下表中的二元分类问题的训练样本集：

实例	a1	a2	a3	目标类
1	F	F	1.0	+
2	F	T	4.0	+
3	T	T	8.0	-
4	F	F	5.0	+
5	F	T	7.0	-
6	T	T	3.0	-
7	T	F	6.0	-
8	T	F	2.0	+
9	F	T	9.0	-

(1) 计算整个训练样本集关于类属性的熵 Entropy （描述不清）；

(2) 分别计算属性 a1、a2 的信息增益 Gain；

(3) 确定二分的条件下，根据信息增益，属性 a3 的最佳划分点在哪里？给出计算过程。

(4) 根据分类错误率（Classification error），按哪个属性（a1、a2 中）更佳？

(5) 根据 Gini 指标，按哪个属性（a1、a2 中）更佳？

装

订

线

2.（10 分）已知一个简单的事务数据库 X，如下表所示：

记录号	购物清单
1	巧克力，尿布，驱蚊水，面包，雨伞
2	尿布，驱蚊水
3	巧克力，尿布，果汁
4	尿布，巧克力，洗衣粉
5	巧克力，果汁，啤酒

使用支持度为 20%，置信度为 60%的阈值找出：

（1）数据库 X 中所有的最大频繁项集

（2）X 中强关联规则

（3）画出该事务的 FP 树

装

订

线

3.（10 分）給出一组样本 A，B，C，D，E 之间的距离矩阵

$$\begin{pmatrix} & A & B & C & D & E \\ A & 0 & 3 & 2 & 3 & 1 \\ B & 3 & 0 & 4 & 2 & 2 \\ C & 2 & 4 & 0 & 1 & 5 \\ D & 3 & 2 & 1 & 0 & 3 \\ E & 1 & 2 & 5 & 3 & 0 \end{pmatrix}$$

分别采用基于单链接（最大 or 最小 用哪个？）相似度量标准、基于全链接 (平均值) 相似度量标准进行凝聚的层次聚类。

（1）说明两种方案的基本策略；

（2）绘制两种解决方案的树状图，展示聚类结果。

装

订

线

4.（10 分）描述 K-均值聚类算法的思想、指出算法的不足之处，并简述二分 K 均值的思想及其与基本 K-均值聚类算法的区别。

假设数据集 S 含有 12 个数据对象（用 2 维空间的点表示）：A1(1,1), A2(2,2),A3(1,2),A4(2,1), B1(2,4), B2(2,6), B3(1.5,5), B4(2.5,5), C1(5,4.5), C2(5,5.5), C3(4,5), C4(6,5). 采用 K-均值方法进行聚类，距离函数采用欧几里德距离，取 k=3，假设初始的三个簇质心为 A1,B1,和 C1，求：

（1）第一次循环结束时的三个簇的质心。

（2）最后求得的三个簇。

装

订

线