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

试　　题

题号	一	二	总分
分数			

1. 考试形式：闭卷■　开卷□；2. 本试卷共两大题，满分 100 分；

3. 考试日期：　　年　　月　　日；（答题内容请写在装订线外）

一、简答题（共 56 分）

（1）简述典型数据类型与属性类型（三种，6 分）

（2）简述数据约简的主要策略，并进行解释（至少两种 6 分）

（3）对比朴素贝叶斯、KNN(K-近邻)、决策树算法的优缺点（6 分）

（4）描述 GINI 系数、信息熵、最大错误率，并说明其区别 (8 分)

(5) 描述 DBSCAN 算法过程，并证明其时间复杂性（10 分）

（6）Aprior 算法采用到剪枝策略，利用到频繁项集的反单调性，以{A,B,C,D}-频繁四项集为例，陈述**候选规则**反单调性并加以证明(10 分)

（7）分层聚类算法需要度量集合与集合之间的关系，给定两个集合 U，V 阐述单链、全链、组平均定义，并比较其优缺点（10 分）

二、计算题（要求写出必要的步骤，无过程不积分，共 44 分）

1. 给定如下数据，完成下述任务(12 分)：

（1）以 O1 与 O5 为初始质心，利用 K-均值算法进行聚类分析 （8 分）

（2）如何解决 K-均值聚类结果不稳定的问题 （4 分）

	特征 1	属性 2	属性 3
O1	-1	1	5
O2	-1	-1	5
O3	1	-1	5
O4	4	-1	5
O5	4	1	5

2. 给定事务数据库 X，支持度阈值为 50%，（16 分）

（1）构建频繁模式树（FP-Tree）（6 分）

（2）利用 FP-tree 挖掘频繁项集（按照 Aprior 计算不得分） （10 分）

	购物清单
1	A,B
2	A,C,D,E
3	B,C,D
4	A,B,C
5	A,D,E
6	A
7	A,B,C,D
8	B,C,E
9	A,B,C
10	A,B,D

装

订

3 给定如下数据，利用 ID3 算法(信息熵与信息收益)构建决策树 （16 分）

(特别说明：考虑到计算复杂性，只需要完成根节点选择，计算过程中列出表达式，无结果不扣分,需要详细计算过程)

时间	天气	气温	湿度	风度	运动
D1	晴	热	高	弱	N
D2	晴	热	高	强	N
D3	阴	热	高	弱	Y
D4	雨	温	高	弱	Y
D5	雨	冷	低	弱	Y
D6	雨	冷	低	强	N
D7	阴	冷	低	强	Y
D8	晴	温	高	弱	N
D9	晴	冷	低	弱	Y
D10	雨	温	低	弱	Y
D11	晴	温	低	强	Y
D12	阴	温	高	强	Y
D13	阴	热	低	弱	Y
D14	雨	温	高	强	N