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（每小题 6 分，共 42 分）

（1）简述数据对象和数据特征的概念以及两者之间的关系。

（2）简述 k 折交叉验证的步骤及其应用。

（3）简述分位数图（盒图）和散点图（散布图）的构造及用途。

（4）简要概述如何计算被如下属性描述的对象的相异性：

（a）标称属性。

（b）非对称的二元属性。

（c）数值属性。

（d）词频向量。

（5）阐述频繁项集、极大频繁项集、闭频繁项集的定义，它们的关系是什么？

（6）对于下面的向量 x 和 y，计算指定的相似度或距离度量。

(a)x=(2，-1，3，0)，y=(-1，4，2，1)计算欧几里得距离

(b)x=(0，1，1，0，1，1)，y=(1，0，0，0，1，0) 计算 Jaccard 系数

(c)x=(0，2，1，-1)，y=(3，-1，1，-2) 计算余弦相似性

（7）简述 DBSCAN 聚类算法的主要思想和优缺点。

二、计算题（要求写出必要的步骤，共 58 分）

1. 聚类（12 分）

假设数据集 D 含有 9 个数据对象（用 2 维空间的点表示）：A1(3,2), A2(3,9),A3(8,6), B1(9,5), B2(2,4), B3(3,10), C1(2,6), C2(9,6), C3(2,2)．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及其质心。

（2）请简述 K-均值聚类的优缺点。

装

订

线

2. 频繁模式挖掘（14 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	方便面，尿布，驱蚊水，面包，雨伞
2	驱蚊水，果汁，洗衣液
3	方便面，尿布，果汁
4	方便面，尿布，面包
5	方便面，果汁，洗衣液

支持度阈值为 60%，置信度阈值为 80%

（1）使用 Apriori 算法找出 X 中的所有频繁项集。

装 **（2）找出 X 中的强关联规则。**

（3）构建频繁模式树（FP-Tree）。

（4）说明支持度-置信度关联模式评估的局限性，并阐述一种其他的评估关联模式的客观度量

订

线

3. 决策树（20 分）

考虑下表中的二元分类问题的训练样本集($\log_2 3 = 1.58\ \log_2 5 = 2.32$)

（a） 计算整个训练样本集关于类属性（最后一列）的熵 Entropy；

（b） 分别计算属性 a1、a2 的信息增益 Gain；

（c） 根据信息增益进行二分，如何确定属性 a3 的最佳划分点？(给出求解思路，并列出所有可能的划分点)

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

（e） 根据 Gini 指标，按哪个属性（a1、a2 中）划分更佳？

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

4. 以下两个题目选做一个即可，两道都做以得分高者记录成绩（12 分）

A. 对于下表中给出的数据集：

(1)分别使用其 3-近邻、5-近邻和 7-近邻计算点 A(2，3)的分类结果。

(2)KNN 对样例进行分类时一般使用多数表决方法来确定，请简述一种其他的表述方法。

点	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
x	(1, 1)	(8, 6)	(3, 5)	(2, 2)	(9, 1)	(4, 5)	(6, 7)	(2, 6)	(7, 3)	(6, 6)
y	-	-	+	+	+	-	+	-	+	-

B. 考虑下表中的二元分类训练集：

（1）估计条件概率 P(A=1|+)、P(A=1|-)、P(B=1|+)和 P(B=1|-)。

（2）使用朴素贝叶斯方法预测样本（A=0,B=0,C=0）的类别。

Id	A	B	C	Class
1	1	0	1	-
2	0	0	1	-
3	1	1	1	-
4	1	0	1	-
5	0	0	1	-
6	1	1	0	+
7	0	0	1	+
8	1	0	0	+
9	0	1	1	+
10	1	0	1	+