装 订 线 装 订 线 装 订 线

任课教师：_____ 学号：_____ 姓名：_____ 班级：_____

西 安 电 子 科 技 大 学

考试时间 120 分钟

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、判断题（每小题 1 分，共 10 分）

1、DBSCAN 算法处理高维数据有优势 ( )

2、决策树通常用于分类和预测 ( )

3、将一组数据点根据其特征分成几个明显区分的类是一项数据挖掘任务 ( )

4、使用历史记录数据预测某公司未来的股票价格是一项数据挖掘任务 ( )

5、数据仓库的数据量越大，其应用价值也越大 ( )

6、KNN（K Nearest Neighbors）是一种常用的基于距离的分类方法，KNN 技术假设整个训练集不仅包含数据集，而且包含每个样本期望的类别标签，实际上训练数据就成为模型 ( )

7、两个对象的余弦度量的取值可以大于 1 ( )

8、所谓过拟合是指对训练数据拟合度过高，导致训练误差较大 ( )

9、决策树分类算法的对大规模数据集分类的运行效率很高 ( )

10、分类是有指导的学习，聚类是无指导的学习 ( )

二、简答题（每小题 6 分，共 30 分）

（1）描述数据集、数据对象、属性三者的含义与区别。

（2）简述 K-means 算法的输入、输出及聚类过程(流程)。

（3）按要求计算数据点之间相似性

（a）给定向量 a=(1, 1, 0, 1, 0, 0), b=(1, 0, 0, 1, 1, 1)，计算 Jaccard 系数

（b）给定向量 a=(2, 1, 0, 3), b=(0, 3, 1, -1)，计算余弦相似度

（c）给定向量 a=(0, 2, 1, 3), b=(1, 0, 1, 0)，计算欧几里得距离

（4）什么是决策树？如何用决策树进行分类?

（5）根据下表中的一维数据集，依照其 1-近邻、3-近邻、5-近邻分别对数据点 x=6.8 进行分类。分类结果体现了 KNN 的什么缺点？可以如何改进？

x	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.9	7.0	7.2
y	-	-	+	+	+	+	+	+	-	-

三、综合计算题（要求写出必要的步骤，共 60 分）

1. 聚类（12 分）

给出一组样本 A，B，C，D，E 之间的相似度矩阵：

	A	B	C	D	E	F
A	**1**	**0.13**	**0.25**	**0.55**	**0.14**	**0.67**
B	**0.13**	**1**	**0.22**	**0.58**	**0.38**	**0.42**
C	**0.25**	**0.22**	**1**	**0.16**	**0.85**	**0.33**
D	**0.55**	**0.58**	**0.16**	**1**	**0.26**	**0.47**
E	**0.14**	**0.38**	**0.85**	**0.26**	**1**	**0.28**
F	**0.67**	**0.42**	**0.33**	**0.47**	**0.28**	**1**

装

分别采用基于单链接相似度量标准、基于全链接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，展示聚类结果并给出计算过程。

订

线

2. 朴素贝叶斯（12 分）

给定数据如下表所示，10 个样本，两个类别，三个属性，利用朴素贝叶斯方法计算:

（1）估计条件概率 P(A1|C=0)和 P(A3|C=1)。

（2）用朴素贝叶斯分类法预测下面样本{A1=2,A2=1, A3=1}的类别。

	属性 A1	属性 A2	属性 A3	类 C
1	1	2	3	0
2	0	0	1	0
3	2	3	2	0
4	1	2	1	0
5	0	1	3	0
6	2	2	2	1
7	1	0	3	1
8	2	1	3	1
9	1	2	1	1
10	1	1	0	1

装

订

线

3. 决策树（18 分）

考虑下表中的二元分类问题的训练样本集

（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？(说明求解思路即可，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	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	-

4. 频繁模式挖掘（18 分）

利用频繁模式树算法(FP-Tree)挖掘交易事务数据，根据下列事务表，使用支持度为 40%，置信度为 70%的阈值：

(1) 构建频繁模式树（FP-Tree）。
(2) 找出所有的频繁项集以及强关联规则。
(3) 若要发现所有以 D 结尾的频繁项集，请画出该过程中 D 的条件 FP 树。
(4) 说明支持度-置信度关联模式评估的局限性。

TID	Items
T01	ABCD
T02	ACDE
T03	AB
T04	ABD
T05	A
T06	BC
T07	ABDE
T08	BCDE
T09	BD
T10	ABC