武汉大学计算机学院

2014 级研究生“数据仓库和数据挖掘”课程期末考试试题

要求：所有的题目的解答均写在答题纸上，需写清楚题目的序号。每张答题纸都要写上姓名和学号。

一、**单项选择题（每小题 2 分，共 20 分）**

1. 下面列出的条目中，（ ）不是数据仓库的基本特征。B

A.数据仓库是面向主题的　　B.数据仓库是面向事务的

C.数据仓库的数据是相对稳定的　　D.数据仓库的数据是反映历史变化的

2. 数据仓库是随着时间变化的，下面的描述不正确的是（ ）。

A.数据仓库随时间的变化不断增加新的数据内容

B.捕捉到的新数据会覆盖原来的快照

C.数据仓库随事件变化不断删去旧的数据内容 C

D.数据仓库中包含大量的综合数据，这些综合数据会随着时间的变化不断地进行重新综合

3. 以下关于数据仓库设计的说法中（ ）是错误的。A

A.数据仓库项目的需求很难把握，所以不可能从用户的需求出发来进行数据仓库的设计，只能从数据出发进行设计

B.在进行数据仓库主题数据模型设计时，应该按面向部门业务应用的方式来设计数据模型

C.在进行数据仓库主题数据模型设计时要强调数据的集成性

D.在进行数据仓库概念模型设计时，需要设计实体关系图，给出数据表的划分，并给出每个属性的定义域

4. 以下关于 OLAP 的描述中（ ）是错误的。A

A.一个多维数组可以表示为（维 1，维 2，…，维 *n*）

B.维的一个取值称为该维的一个维成员

C.OLAP 是联机分析处理

D.OLAP 是数据仓库进行分析决策的基础

5. 多维数据模型中，下列（ ）模式不属于多维模式。D

A.星型模式　　B.雪花模式　　C.星座模式　　D.网型模式

6. 通常频繁项集、频繁闭项集和最大频繁项集之间的关系是（ ）。C

A.频繁项集⊂频繁闭项集⊂最大频繁项集

B.频繁项集⊂最大频繁项集⊂频繁闭项集

C.最大频繁项集⊂频繁闭项集⊂频繁项集

D.频繁闭项集⊂频繁项集⊂最大频繁项集

7. 决策树中不包含（ ）结点。C

A.根结点　　B.内部结点　　C.外部结点　　D.叶结点

8. 下面选项中 t 不是 s 的子序列的是（ ）。C

A.s=<{2,4},{3,5,6},{8}> t=<{2},{3,6},{8}>

B.s=<{2,4},{3,5,6},{8}> t=<{2},{8}>

C.s=<{1,2},{3,4}> t=<{1},{2}>

D.s=<{2,4},{2,4}> t=<{2},{4}>

9. 前馈神经网络用于分类时，以下（ ）是不合理的迭代结束条件。D

A.前一周期所有的 Δw_{ij} 都很小，小于某个指定的阈值

B.前一周期未正确分类的样本百分比小于某个阈值

C.超过预先指定的周期数

D.学习率小于某个阈值

10. 以下叙述中，（ ）是错误的。D

A.逻辑回归用于分析二分类或有次序的依变量和自变量之间的关系

B.SVM 是一种基于分类边界的方法

C.朴素贝叶斯算法和树增强朴素贝叶斯算法是按照描述属性是否独立来划分的

D.以上都不对

二、（20 分）假设某大型人事部门已有一个人事管理系统，包含如下数据表：

职工（编号，姓名，出生日期，工作地点，月工资，备注）

现要设计一个人事数据仓库，用于分析各地区（华北、华中、华东、…）、各年龄层次（老、中、青）的工资水平（高、中、低）等。

回答以下问题：

（1）根据你的思考设计该数据仓库的模式图，包含每个维表和事实表的结构。（10 分）

（2）指出你设计的数据仓库属于哪种模式。（5 分）

（3）由[出生日期，工作地点，月工资]的基本方体开始，求华东地区的青年职工中高收入的人数，应当执行哪些 OLAP 操作？（5 分）

三、（20 分）有一个如表 1 所示的事务数据库，设最小支持度为 40%，最小置信度为 80%。

表 1　一个事务数据库

TID（编号）	Itemset（项集）
1	1，3，4
2	2，3，4，5
3	1，3，5，7
4	2，5
5	1，2，4，6，7
6	2，4，6

回答以下问题：

（1）采用 Apriori 算法求出所有的频繁集。要求给出求解过程。（15 分）

（2）求出所有与元规则“$item_1 \wedge item_2 \rightarrow item_3$”相匹配的强关联规则。（5 分）

四、（15 分）对于如表 2 所示的决策表$(U, C \cup D)$，$C=\{a, b, c, d\}$，$D=\{e\}$，回答以下问题：

（1）求 U/C 和 U/D。（5 分）

（2）求 $POS_C(D)$，该决策表是否为一致（或协调）决策表？ （5 分）

（3）采用分辨矩阵求其所有条件属性约简和核。（5 分）

表 2 一个决策表

U	a	b	c	d	e
1	1	0	2	1	1
2	1	0	2	0	1
3	1	2	0	0	2
4	1	2	2	1	0
5	2	1	0	0	2
6	2	1	1	0	2
7	2	1	2	1	1

五、（25 分）回答以下关于聚类的问题：

（1）k–中心点算法和 k–均值算法相比有什么优点？（5 分）

（2）BIRCH 算法是什么类型的聚类算法？通常采用簇的聚类特征为 $CF=(N, LS, SS)$，设置这样的聚类特征有什么好处？（10 分）

（3）什么是离群点？简述将 DBSCAN 算法用于离群点检测的基本过程。（10 分）