华南理工大学计算机科学与工程学院
2012—2013 学年度第二学期期末考试
《数据仓库与数据挖掘技术》试 卷(假的)

专业：计算机科学与技术　年级：2010　姓名：　　学号：

注意事项：1. 本试卷共四大题，满分 100 分，考试时间 120 分钟；
2. 所有答案请直接答在试卷上；

题号	一	二	三	四	总分
得分					

一. 填空题（每空 1 分，共 20 分）

1. 数据仓库的特征包括_面向主题_____、___集成_____、__时变_____和非易失性。
2. 数据仓库的三种数据模式包括_星形模式_、__雪花形模式_____、___事实星座形模式_____。
3. 仓库数据库服务器、_OLAP 服务器_____、__前端客户_____为数据仓库的多层结构。
4. OLAP 技术多维分析过程中，多维分析操作包括 __上卷___、__下钻____、___切片____、__切块_____、__转轴_____等。
5. 知识发现过程的主要步骤有：数据清理、__数据集成_____、__数据选择___、数据交换、_数据挖掘_____、___模式评估_____、__知识表示_____。
6. 数据仓库的视图的分类有:自顶向下视图、_数据源视图_____、数据仓库视图、_商务视图_____。

二. 简答题(每题 6 分，共 42 分)

1. 简述处理空缺值的方法。
   1、忽略该记录
   2、手工填写空缺值
   3、使用默认值
   4、使用属性平均值
   5、使用同类样本平均值
   6、使用最可能的值
2. 挖掘的知识类型。
   1、概念/类描述：特征化和区分
   2、挖掘频繁模式、关联和相关
   3、分类和预测
   4、聚类分析
   5、离群点分析
   6、演变分析

3. 何为 OLTP 与 OLAP 及他们的主要区别。

联机事务处理 OLTP (on-line transaction processing)；联机分析处理 OLAP (on-line analytical processing)：
OLTP 和 OLAP 的区别：
用户和系统的面向性:OLTP 面向顾客，而 OLAP 面向市场；
数据内容：OLTP 系统管理当前数据，而 OLAP 管理历史的数据；
数据库设计：OLTP 系统采用实体-联系（ER)模型和面向应用的数据库设计，而 OLAP 系统通常采用星形和雪花模型；
视图：OLTP 系统主要关注一个企业或部门内部的当前数据，而 OLAP 系统主要关注汇总的统一的数据；
访问模式：OLTP 访问主要有短的原子事务组成，而 OLAP 系统的访问大部分是只读操作，尽管许多可能是复杂的查询。

4. 在数据挖掘之前为什么要对原始数据进行预处理？
数据预处理对于数据仓库和数据挖掘都是一个重要的问题，因为现实中的数据多半是不完整的、有噪声的和不一致的。数据预处理包括数据清理、数据集成、数据交换和数据规约。

5. 为什么需要构建单独隔离的数据仓库？
使得操作数据库与数据仓库都获得高性能
DBMS—OLTP：访问方法，索引，并发控制，数据恢复。
Warehouse—OLAP：复杂 OLAP 查询，多维视图，整理。
对数据与功能的要求不同：
丢失的数据：决策支持需要历史数据，而传统数据库并不一定维护历史数据。
数据整理： 决策支持需要对异构数据源进行数据整理 。
数据质量：不同的数据源常常具有不一致的数据表示，编码结构与格式。

6. 关联规则的确定性度量与实用性度量的分类及定义。
支持度和置信度是关联规则的确定性度量与实用性度量。
（1）支持度：事务包含 XUY 的概率，即 support=P(XUY)
支持度计算：
Support(X→Y) = P(X U Y)
={XUY}的支持度计数（模式或项集在 DB 中出现的频率）/事务表中总的事务数
（2）置信度：事务同时包含 X 与 Y 的条件概率：confidence=P(Y|X)
置信度计算：
Confidence(X→Y) = P(Y|X)
=P(XUY)/P(X) = {XUY} 支持度计数/X 支持度计数

7. 简述分箱平滑的方法。
对数据进行排序，然后把它们划分到箱，然后通过箱平均值，箱中值或者箱边界值进行平滑。
分箱的方法主要有：
① 等深分箱法 ② 等宽分箱法

数据平滑的方法主要有：平均值法、边界值法和中值法

三. 计算题（共 38 分）

1.一个食品连锁店每周的事务记录如下表所示，其中每一条事务表示在一项收款机业务中卖出的项目，假定 supmin=40%，confmin=40%，使用 Apriori 算法计算生成的关联规则，标明每趟数据库扫描时的候选集和大项目集。（10 分）

事务	项目
T1	面包、果冻、花生酱
T2	面包、花生酱
T3	面包、牛奶、花生酱
T4	啤酒、面包
T5	啤酒、牛奶

解：（1）由 I={面包、果冻、花生酱、牛奶、啤酒}的所有项目直接产生 1-候选 C_1，计算其支持度，取出支持度小于 sup_{min} 的项集，形成 1-频繁集 L_1，如下表所示：

项集 C_1	**支持度**	**项集 L_1**	**支持度**
{面包}	4/5	{面包}	4/5
{花生酱}	3/5	{花生酱}	3/5
{牛奶}	2/5	{牛奶}	2/5
{啤酒}	2/5	{啤酒}	2/5

(2)组合连接 L_1 中的各项目，产生 2-候选集 C_2，计算其支持度，取出支持度小于 sup_{min} 的项集，形成 2-频繁集 L_2，如下表所示：

项集 C_2	**支持度**	**项集 L_2**	**支持度**
{面包、花生酱}	3/5	{面包、花生酱}	3/5

至此，所有频繁集都被找到，算法结束，

所以，confidence（{面包}→{花生酱}）=（4/5）/（3/5）=4/3>

$conf_{min}$

confidence（{花生酱}→{面包}）=（3/5）/（4/5）=3/4> $conf_{min}$

所以，关联规则{面包}→{花生酱}、{ 花生酱}→{面包}均是强关联规则。

2.给定以下数据集（2，4，10，12，15，3，21），进行 K-Means 聚类，设定聚类数为 2 个，相似度按照欧式距离计算。（10 分）

解：（1）从数据集 X 中随机地选择 k 个数据样本作为聚类的出示代表点，每一个代表点表示一个类别，由题可知 k=2，则可设 m1=2，m2=4：

（2）对于 X 中的任意数据样本 xm（1<xm<total），计算它与 k 个初始代表点的距离，并且将它划分到距离最近的初始代表点所表示的类别中：当 m1=2 时，样本（2 ，4，10，12，15，3，21）距离该代表点的距离分别为 2，8，10，13，1，19。

当 m2=4 时，样本（2 ，4，10，12，15，3，21）距离该代表点的距离分别为-2，6，8，11，-1，17。

最小距离是 1 或者-1 将该元素放入 m1=2 的聚类中，则该聚类为（2，3），另一个聚类 m2=4 为（4，10，12，15，21）。

（3）完成数据样本的划分之后，对于每一个聚类，计算其中所有数据样本的均值，并且将其作为该聚类的新的代表点，由此得到 k 个均值代表点：m1=2.5，m2=12：

（4）对于 X 中的任意数据样本 xm（1<xm<total），计算它与 k 个初始代表点的距离，并且将它划分到距离最近的初始代表点所表示的类别中：当 m1=2.5 时，样本（2 ，4，10，12，15，3，21）距离该代表点的距离分别为-0.5，0.5，1.5，7.5，9.5，12.5，18.5。

当 m2=12 时，样本（2 ，4，10，12，15，3，21）距离该代表点的距离分别为-10，-9，-8，2，3，9。

最小距离是 1.5 将该元素放入 m1=2.5 的聚类中，则该聚类为（2，3，4），另一个聚类 m2=12 为（10，12，15，21）。

（5）完成数据样本的划分之后，对于每一个聚类，计算其中所有数据样本的均值，并且将其作为该聚类的新的代表点，由此得到 k 个均值代表点：m1=3，m2=14.5：

（6）对于 X 中的任意数据样本 xm（1<xm<total），计算它与 k 个初始代表点的距离，并且将它划分到距离最近的初始代表点所表示的类别中：当 m1=3 时，样本（2 ，4，10，12，15，3，21）距离该代表点的距离分别为-1，1，7，9，12，18，。

当 m2=14.5 时，样本（2 ，4，10，12，15，3，21）距离该代表点的距离分别为-12.58，-11.5，-10.5，-4.5，-2.5，0.5，6.5。

最小距离是 0.5 将该元素放入 m1=3 的聚类中，则该聚类为（2，3，4），另一个

聚类 m2=14.5 为（10，12，15，21）。
至此，各个聚类不再发生变化为止，即误差平方和准则函数的值达到最优。

3. 表 3 提供了一个训练集的数据元组关于是否要打篮球。给定一个元组(天气=阳光明媚,温度= 凉快,湿度=高,风力=强),决定目标类 Playbasketball 是 YES 或 NO 使用贝叶斯朴素算法的分类器进行计算。 (18 分)

No.	Outlook	Temperature	Humidity	Wind	Playbasketball
1	Overcast	Hot	High	Weak	Yes
2	Sunny	Hot	High	Weak	No
3	Sunny	Hot	High	Strong	No
4	Overcast	Hot	Normal	Weak	Yes
5	Rain	Mild	High	Weak	Yes
6	Sunny	Cool	Normal	Weak	Yes
7	Rain	Cool	Normal	Weak	Yes
8	Rain	Mild	Normal	Weak	Yes
9	Rain	Cool	Normal	Strong	No
10	Overcast	Cool	Normal	Strong	Yes
11	Sunny	Mild	High	Weak	No
12	Overcast	Mild	High	Strong	Yes

表 3.

P(Outlook=sunny|yes)= 1/7
P(Outlook=sunny|no)= 3/5
P(temperature=cool|yes)=3/7
P(temperature=cool|no)=1/5
P(Humidity=high|yes)=2/7
P(Humidity=high|No)=4/5
P(wind=strong|yes)=2/7
P(Humidity=strong|No)=3/5
P(yes)=7/12
P(no) = 5/12
P(X|YES)=1/7 × 3/7 × 2/7 × 2/7 × 7/12 = 0.00292
P(X|NO)=3/5 × 1/5 × 4/5 × 3/5 × 5/12 = 0.024