《商务智能》2018年7月期末考真题及答案

**一、单选题（每小题2分，共计36分）**

**1. 某超市通过分析销售就数据后发现，买啤酒的人很大概率也会同时购买尿布，这种属于数据挖掘的那类问题？（）**

A. 关联规则发现 B. 聚类 C. 分类  D. 预测

**答：A**

**2. 以下两种描述分别对应哪两种对分类算法的评价标准？（）**

(a)警察抓小偷，描述警察抓的人中有多少个是小偷的标准。

(b)描述偶多少比例的小偷给警察抓了的标准。

A. Precision,Recall  B. Recall, Precision  C. Precision, ROC  D. Recall, ROC

**答：A**

**3. 将原始数据进行集成、变换、维度规约、数值规约是一下哪个任务的主要工作？（）**

A.  频繁模式挖掘  B. 分类和预测  C. 数据预处理  D. 数据流挖掘

**答：C**

**4. 当不知道数据所带标签时，可以使用哪种技术促使带同类标签的数据与带其他标签的数据相分离？（）**

A.分类  B.聚类  C.关联分析  D.隐马尔可夫链

**答：B**

**5. 建立一个模型，通过这个模型根据一致的变量值来预测其他某个变量值属于数据挖掘的哪一类任务？（）**

A.根据内容检索  B.建模描述  C.预测建模  D.寻找模式和规则

**答：C**

**6. 假设12个销售价格记录组已经排序如下：5，10，11，13，15，35，50，55，72，92，204，215，事用如下每种方法将它们划分成四个箱。等频（等深）划分时，15在第几个箱子内？（）**

A.第一个  B. 第二个  C. 第三个  D. 第四个

**答：B**

**7. 上题中，等宽划分时（宽度为50），15又在哪个箱子里？（）**

A.第一个  B. 第二个  C. 第三个  D. 第四个

**答：A**

**8. 下面那个不属于数据的属性类型（）**

A.标称  B.序数  C.区间  D.相异

**答：D**

**9. 假定属性income的最大最小值分别是12000元和98000元。利用最大最小规范化的方法将属性的值映射到0至1的范围内。对属性income的73600元将被转化为（）**

A.0.821  B.1.224 C.1.458  D.0.716

**答：D**

**10. 假定用于分析的数据包含属性age。数据元组中age的值如下（按递增序）：13，15，16，16，19，20，20，21，22，22，25，25，25，30，33，33，35，35，36，40，45，46，52，70，问题：使用按箱平均值平滑方法对上述数据进行平滑，箱的深度为3.第二个箱子值为（）**

A.18.3  B.22.6 C.26.8  D.27.9

**答：A**

**11. 一岁大学内的各年级人数分别为：一年级200人，二年级160人，三年级130人，四年级110人。则年纪属性的众数是（）**

A.一年级  B.二年级  C.三年级  D.四年级

**答：A**

**12. 数据仓库是随时间变化的，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. 数据仓库随时间不断删去旧的数据内容；

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变化不断地进行重新综合。

**答：C**

**13. OLAP技术的核心是（）**

A.在线性  B.对用户的快速响应  C.互操作性  D.多维分析

**答：D**

**14. 设X={1，2，3}是频繁项集，则可由X产生的关联规则的个数是（）**

A.4  B.5  C.6  D.7

**答：C**

**15. 频繁项集、频繁闭项集、最大频繁项集之间的关系是（）**

A.最大频繁项集⊆频繁项集⊆频繁闭项集

B.频繁项集⊆频繁闭项集⊆最大频繁项集

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

**答：B**

**16. 考虑下面的频繁3-项集的集合：{1，2，3}{1，2，4}{1，2，5}{1，3，4}{1，3，5}{2，3，4}{2，3，5}{3，4，5}假定数据集中只有5个项，采用合并策略，由候选产生过程得到4-项集不包含（）**

A.1，2，3，4，  B.1，2，3，5 C.1，2，4，5  D.1，3，4，5

**答：C**

**17. 以下算法属于分类算法的是（）**

A.DBSCAN  B.C4.5 C.K-means  D.EM

**答：B**

**18. 简单的将数据对象集划分为不重叠的子集，使得每个数据对象恰好在一个子集中，这种聚类类型称作（）**

A.层次聚类  B.划分聚类  C.非互斥聚类  D.模糊聚类

**答：B**

**二、简答题（每小题10分，共计40分）**

**1. 简述商务智能的概念。**

**答：(同2017.7考题)**

**（课件）**

商务智能（Business intelligence）

工业界

- 商务智能可以被看作是一类技术或工具，利用它们可以对大量的数据进行收集、管理、分析和挖掘，以改善业务决策水平，增强企业的竞争力

学术界

- 商务智能是一套理论、方法和应用，通过它们可以快速地发现海量数据中隐含的各种知识，有效地解决企业面临的管理和决策问题，支持企业的战略实施。

**（其他参考）**

数据仓库（或数据集市）、查询报表、数据分析、数据挖掘、数据备份和恢复等部分组成的、以辅助企业决策为目的一类技术及其应用。

商务智能可以被看作是一类技术或工具，利用它们可以对大量的数据进行收集、管理、分析和挖掘，以改善业务决策水平，增强企业的竞争力。

商务智能是一套理论、方法和应用，通过它们可以快速地发现海量数据中隐含的各种知识，有效地解决企业面临的管理和决策问题，支持企业的战略实施。

商务智能是将数据转换成信息的过程，  
然后通过发现将信息转化为知识。

**2. 举例说明支持度、置信度的概念机器计算方法。**

**答：(同2016.7、2017.7考题)**

支持度：X和Y同时出现在数据集D中的概率。

置信度：X在数据集D中出现时，Y同时出现的条件概率。

举例：

总共有10000个消费者购买了商品，其中购买尿布的有1000人，购买啤酒的有2000人，购买面包的有500人，同时购买尿布和啤酒的有800人，同时购买尿布的面包的有100人。

1）支持度（support）

支持度：{X,Y}同时出现的概率，例如：{尿布，啤酒}同时出现的概率

support=同时购买{X,Y}的人数/总人数

{尿布，啤酒}的支持度=800 / 10000 = 0.08

{尿布，面包}的支持度=100/ 10000 = 0.01

2）置信度（confidence）

置信度：购买X的人，同时购买X,Y的概率，例如：购买尿布的人，同时购买啤酒的概率，而这个概率就是购买尿布时购买啤酒的置信度

confidence（X−>Y）=同时购买{X,Y}的人数/购买X的人数

confidence（Y−>X）=同时购买{X,Y}的人数/购买Y的人数

( 尿布 -> 啤酒)的置信度 = 800 / 1000 = 0.8

( 啤酒 -> 尿布)的置信度 = 800 / 2000 = 0.4

**3. 说明分类与聚类概念的不同之处。**

**答：(同2016.7、2017.7考题)**

分类：通过学习得到一个目标函数f，把每个属性集x映射到一个预先定义的类标号y。分类是有监督学习。

分类模型用于预测未知记录的类标签。

聚类：将数据对象分到各个簇中，聚类是无监督学, 作为一种深入了解数据分布的独立工具。作为其他算法的预处理步骤

**4. 举例说明多维数据分析的主要操作类型有哪些？**

**答**：**(同2016.7、2017.7考题)**

有**切片，切块，旋转，下钻，上卷**。

**切片**：在数据方体的某一维上选定一个维成员的动作。

**切块**：在数据方体的某一维上选定某一区间的维成员的动作。

**旋转**：改变数据方体维的次序的动作。

**下钻**：在某个分析的过程中，用户需要从更多的维或者某个维的更细层次上观察数据。操作类型有两种，第一种为在现有的维上钻取到更细一层的数据；另一种是增加更多的维。

**上卷**：在某个分析的过程中，用户需要从更少的维或者某个维的更粗层次上观察数据。操作类型有两种，第一种为上卷到现有的某个维的更高层次去进行分析；另一种是减少一个维来进行分析。

**三、计算题（每小题12分，共计24分）**

假设10个对象的两个属性取值分别为A(1,2)、B(2, 1)、C(3, 2.5)、D(4, 4.6)、E(4.5, 5)、F(5,4.5)、G(6, 3)、H(6, 4)、I(7, 3.5)、J(3, 4.8)。

1. 用K均值法将其聚为3个簇。

2. 用凝聚层次聚类算法将其聚类，输出3个簇。

答：（**第一问同2017.7考题**）

1. 由题意知K=3，所以选取三个初始中心B，F，H

曼哈顿距离为|x1-x2|+|y1-y2|

第一次迭代（横屏看）

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | A | B | C | D | E | F | G | H | I | J |
| B(2,1) | 2 | 0 | 2.5 | 5.6 | 6.5 | 6.5 | 6. | 7. | 7.5 | 4.8 |
| F(5,4.5) | 6.5 | 6.5 | 4 | 1.1 | 1 | 0 | 2.5 | 1.5 | 3 | 2.3 |
| H(6,4) | 7 | 7 | 4.5 | 2.6 | 2.5 | 1.5 | 1 | 0 | 1.5 | 3.8 |

得出三个分组为

C1:   A B C

C2:   D E F J

C3:   G H I

算出新的中心为

C1的中心：((1+2+3)/3,(2+1+2.3)/3)=(2,1.8)

C2的中心：((4+4.5+5+3)/4,(4.6+5+4.5+4.8)/4)=(4.1,4.7)

C3的中心：((6+6+7)/3,(3+4+3.5)/3)=(6.3,3.5)

第二次迭代（横屏看）

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | A | B | C | D | E | F | G | H | I | J |
| (2,1.8) | 1.2 | 0.8 | 1.7 | 4.8 | 5.7 | 5.7 | 5.2 | 5.2 | 6.7 | 4 |
| (4.1,4.7) | 5.8 | 5.8 | 3.3 | 0.2 | 0.7 | 1.1 | 3.6 | 2.6 | 5.1 | 1.2 |
| (6.3,3.5) | 6.8 | 6.8 | 4.3 | 3.4 | 3.3 | 2.3 | 0.8 | 0.8 | 0.7 | 4 |

得出三个分组为

C1:   A B C

C2:   D E F J

C3:   G H I

表明已收敛，所以3个簇为

ABC，DEFJ，GHI

2. 不会