**简答题：**

**1、简述商务智能的概念**

答：（答案来自课件）

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。

**2、举例说明支持度、置信度的概念及其计算方法**

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。

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、说说你对分类与预测概念的理解**

答：（答案来自百度）

分类和预测是两种数据分析的形式，可用于提取描述重要数据类的模型或预测未来的数据趋势。

分类：用于预测数据对象的分类标号（或离散值），如，通过构造分类模型对银行贷款进行风险评估（安全或危险）

预测：用于预测数据对象的连续取值，如，建立预测模型利用顾客收入与职业（参数）预测其可能用于购买计算机设备的支出大小。

**4、说明分类与聚类概念的不同之处**

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。

**5、说明数据仓库与数据库的不同之处**

答：(同2016上半年考题)

1) 面向主题

操作型数据库的数据组织面向事务处理任务，各个业务系统之间各自分离，而数据仓库中的数据是按照一定的主题域进行组织。主题是一个抽象的概念，是指用户使用数据仓库进行决策时所关心的重点方面，一个主题通常与多个操作型信息系统相关。

2) 集成的

面向事务处理的操作型数据库通常与某些特定的应用相关，数据库之间相互独立，并且往往是异构的。而数据仓库中的数据是在对原有分散的数据库数据抽取、清理的基础上经过系统加工、汇总和整理得到的，必须消除源数据中的不一致性，以保证数据仓库内的信息是关于整个企业的一致的全局信息。

3) 相对稳定的

操作型数据库中的数据通常实时更新，数据根据需要及时发生变化。数据仓库的数据主要供企业决策分析之用，所涉及的数据操作主要是数据查询，一旦某个数据进入数据仓库以后，一般情况下将被长期保留，也就是数据仓库中一般有大量的查询操作，但修改和删除操作很少，通常只需要定期的加载、刷新。

4) 反映历史变化

操作型数据库主要关心当前某一个时间段内的数据，而数据仓库中的数据通常包含历史信息，系统记录了企业从过去某一时点(如开始应用数据仓库的时点)到目前的各个阶段的信息，通过这些信息，可以对企业的发展历程和未来趋势做出定量分析和预测。

**6、举例说明多维数据分析的主要操作类型有哪些**

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。

**1. 简述商务智能系统的构成。**

商务智能系统六个主要组成部分:

1）数据源

企业内部的操作型系统，即支持各业务部分日常运营的信息系统，企业的外部，如人口统计信息、竞争对手信息等;

2)数据仓库

各种数据源的数据经过抽取、转换之后需要放到一个供分析使用的环境，以便对数据进行管理，这就是数据仓库。

数据集市：通常针对单个部门的数据仓库，区别于企业范围内的数据仓库。

3）在线分析处理

在线分析处理：业务性能度量可以通过多个维度、多个层次进行多种聚集汇总，通过交互方式发现业务运行的关键性能指标的异常之处。

4）数据探查

包括灵活的查询、即时报表以及统计方法等，属于被动分析方法。

5）数据挖掘

数据挖掘是从大量数据中自动发现隐含的信息和知识的过程，属于主动分析方法，不需要分析者的先验假设，可以发现未知的知识。常用的分析方法包括分类、聚类、关联分析、数值预测、序列分析、社会网络分析等。

6）业务性能管理

是对企业的关键性能指标，如销售、成本、利润以及可盈利性等，进行度量、监控和比较的方法和工具。这些信息通常通过可视化的工具如平衡积分卡和仪表盘等进行展示。

**2. 说明提升度的概念及其计算方法。**

提升度（Lift）：表示“包含A的事务中同时包含B事务的比例”与“包含B事务的比例”的比值。公式表达：Lift=(P(A&B)/P(A))/P(B)=P(A&B)/P(A)/P(B)。

提升度反映了关联规则中的A与B的相关性，提升度>1且越高表明正相关性越高，提升度<1且越低表明负相关性越高，提升度=1表明没有相关性。

例子，已知有1000名顾客买年货，分为甲乙两组，每组各500人，其中甲组有500人买了茶叶，同时又有450人买了咖啡；乙组有450人买了咖啡，如表所示：

买茶叶的人数 买咖啡的人数

甲组（共500人） 500 450

乙组（共500人） 0 450

求解"茶叶→咖啡"的提升度。

分析：

设X= {买茶叶}，Y={买咖啡}，则规则"茶叶→咖啡"表示"即买了茶叶，又买了咖啡"，于是，"茶叶→咖啡"的置信度为

Confidence(X→Y)= 450 / 500 = 90%

"茶叶→咖啡"的提升度为

Lift(X→Y)= Confidence(X→Y) / P(Y) = 90% / ((450+450) / 1000) = 90% / 90% = 1

由于提升度Lift(X→Y)=1，表示X与Y相互独立，即是否有X，对于Y的出现无影响。也就是说，是否购买咖啡，与有没有购买茶叶无关联。即规则"茶叶→咖啡"不成立，或者说关联性很小，几乎没有，虽然它的支持度和置信度都高达90%，但它不是一条有效的关联规则。

满足最小支持度和最小置信度的规则，叫做“强关联规则”。然而，强关联规则里，也分有效的强关联规则和无效的强关联规则。

如果Lift(X→Y)>1，则规则“X→Y”是有效的强关联规则。

如果Lift(X→Y)<=1，则规则“X→Y”是无效的强关联规则。

特别地，如果Lift(X→Y)=1，则表示X与Y相互独立。

**3. 说明数据预处理都包含那些主要任务**。

1）数据离散化（discretization）

数据规范化又称标准化（standardization），通过将属性的取值范围进行统一，避免不同的属性在数据分析的过程中具有不平等的地位

常用方法

最小-最大法（min-maxnormalization）

z-score

2）数据规范化（normalization）

分箱离散化

基于熵的离散化

离散化方法ChiMerge

3） 数据清洗（datacleaning）

处理数据的缺失、噪音数据的处理以及数据不一致的识别和处理

处理数据的缺失：

如果数据集含有分类属性，一种简单的填补缺失值的方法为，将属于同一类的对象的该属性值的均值赋予此缺失值；对于离散属性或定性属性，用众数代替均值。更复杂的方法，可以将其转换为分类问题或数值预测问题。

4）特征提取与特征选择

介绍面向分类的特征选择方法。有效地特征选择不仅降低数据量，提高分类模型的构建效率，有时还可以提高分类准确率。

特征选择方法有很多，总结它们的共同特点，其过程可以分为以下几步：

根据一定的方法选择一个属性子集；

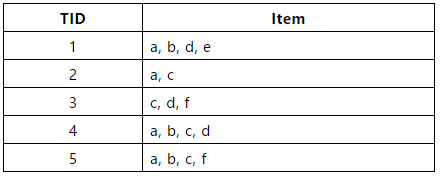
衡量子集的相关性；

判断是否需要更新属性子集，若是，转第1步继续，若否，进入下一步；

输出最终选取的属性子集。

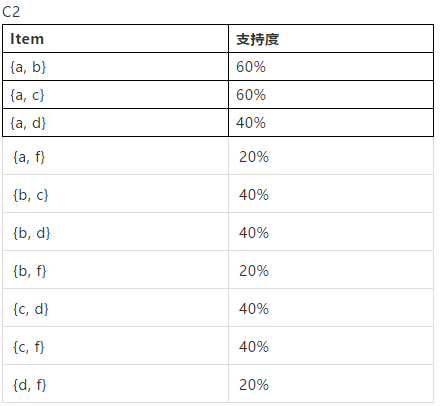
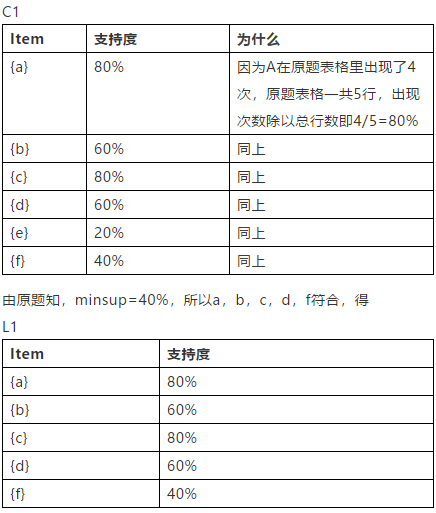
计算题：

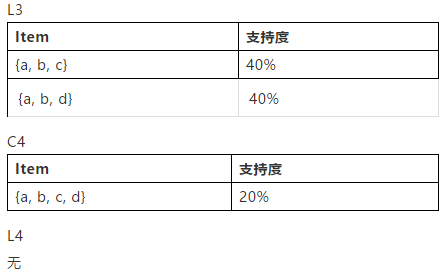
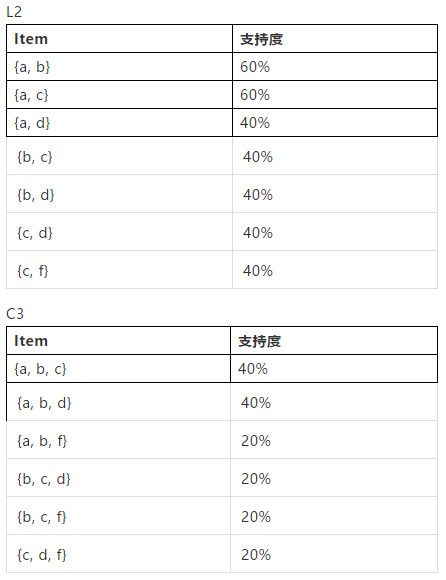
**1、假设如下表中的数据为购物交易数据，设minsup=40% 和minconf=60%，请找出所有的频繁项集以及关联规则。**



答：（同2016上半年考题，只是数据换了）

频繁项集：





关联规则：

{a, b, c}所有非空子集有

{a}， {b}， {c}， {a, b}， {a, c}， {b, c}

{a}→{b, c} 2/4=50%

{b}→{a, c} 2/3=66.67%

{c}→{a, b} 2/4=50%

{a, b}→{c} 3/3=100%

{a, c}→{b} 3/3=100%

{b, c}→{a} 2/3=66.67%

由原题知，minconf=60%，所以{b}→{a, c}，{a, b}→{c}，{a, c}→{b},，{b, c}→{a}符合

对于{a, b, d}同理

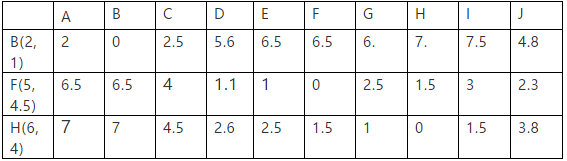
**2、假设10个对象的两个属性取值分别为A(1,2)、B(2,1)、C(3,2.5)、D(4,4.6)、E(4.5,5)、F(5,4.5)、G(6,3)、H(6,4)、I(7,3.5)、J(3,4.8)。用K均值法将其聚为3个类。**

答：

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）



得出三个分组为

C1: A B C

C2: D E F J

C3: G H I

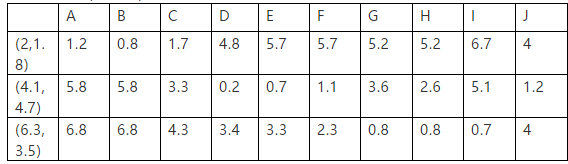
算出新的中心为

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）



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

**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。

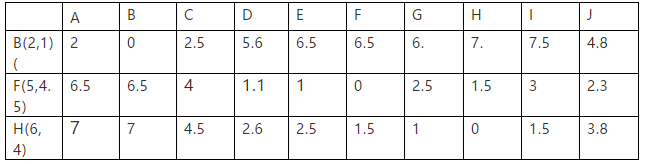
2. 用凝聚层次聚类算法将其聚类，输出3个簇。

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）



得出三个分组为

C1: A B C

C2: D E F J

C3: G H I

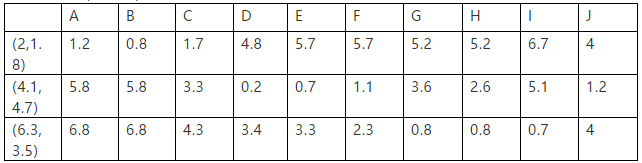
算出新的中心为

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）



得出三个分组为

C1: A B C

C2: D E F J

C3: G H I

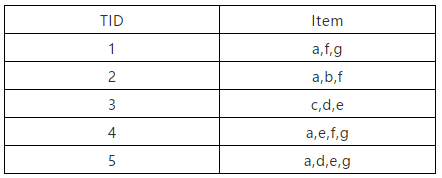
表明已收敛，所以3个簇为

ABC，DEFJ，GHI

2. 不会

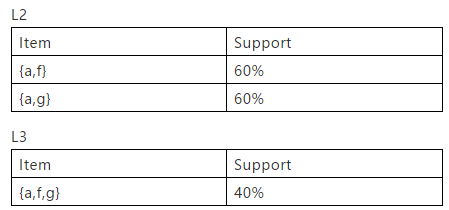
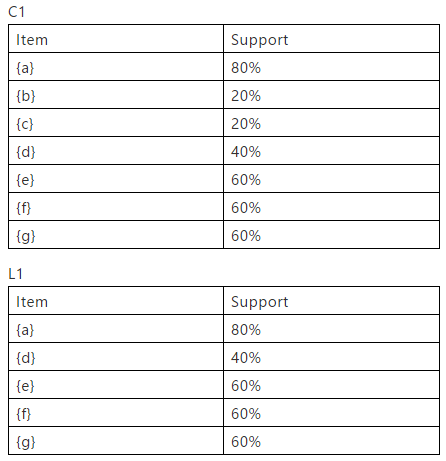
二、计算题（每小题20分，共计40分）

1. 假设如下表中的数据为购物交易数据，设minsup=40％，和minconf=60%，请找出所有的频繁项集以及关联规则



答：

频繁项集：



关联规则：

{a, f,g}所有非空子集有

{a}，{f}， {g}， {a, f}， {a, g}， {f, g}

{a}→{f,g} 2/4=50%

{f}→{a,g} 2/3=66.67%

{g}→{a,f} 2/3=66.67%

{a, f}→{g}2/3=66.67%

{a, g}→{f}2/3=66.67%

{f, g}→{a}2/4=50%

由原题知，minconf=60%，所以{f}→{a,g}，{g}→{a, f}，{a, f}→{g}，{a, g}→{f}符合。

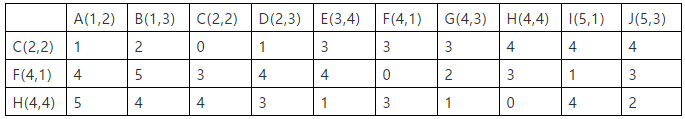
2. 假设10个对象的两个属性取值分别为A(1，2)、B(1，3)、C(2,2)、D(2, 3)、 E(3, 4)、F(4，1)、G(4, 3)、H(4, 4)、I(5，1)、J(5, 3)。用K均值法将其聚为3个类。

答：

由于K=3，所以初始聚类中心选择C(2,2)，F(4,1)与H(4,4)

曼哈顿距离为|x1-x2|+|y1-y2|

第一次迭代（横屏看）



新的组为

C1：A，B，C，D

C2：F，I

C3：E，G，H，J

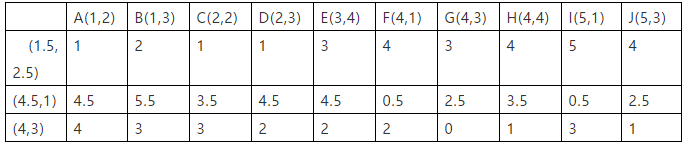
计算出新的聚类中心为

C1的中心：((1+1+2+2)/4,(2+3+2+3)/4)=(1.5,2.5)

C2的中心：((4+5)/2,(1+1)/2)=(4.5,1)

C3的中心：((3+4+4+5)/4,(4+3+4+1)/4)=(4,3)

第二次迭代（横屏看）



新的组为

C1：A，B，C，D

C2：F，I

C3：E，G，H，J

与第一次迭代比没有变化，表明已收敛。所以三个类即为

C1：A，B，C，D

C2：F，I

C3：E，G，H，J

**二、计算题（每小题20分，共计40分）**

**1. 假设如下表中的数据为购物交易数据，设minsup=40％，和minconf=60%，请找出所有的频繁项集以及关联规则**

|  |  |
| --- | --- |
| TID | Item |
| 1 | a,f,g |
| 2 | a,b,f |
| 3 | c,d,e |
| 4 | a,e,f,g |
| 5 | a,d,e,g |

**答：**

频繁项集：

C1

|  |  |
| --- | --- |
| Item | Support |
| {a} | 80% |
| {b} | 20% |
| {c} | 20% |
| {d} | 40% |
| {e} | 60% |
| {f} | 60% |
| {g} | 60% |

L1

|  |  |
| --- | --- |
| Item | Support |
| {a} | 80% |
| {d} | 40% |
| {e} | 60% |
| {f} | 60% |
| {g} | 60% |

L2

|  |  |
| --- | --- |
| Item | Support |
| {a,f} | 60% |
| {a,g} | 60% |

L3

|  |  |
| --- | --- |
| Item | Support |
| {a,f,g} | 40% |

关联规则：

{a, f,g}所有非空子集有

{a}，{f}， {g}， {a, f}， {a, g}， {f, g}

{a}→{f,g} 2/4=50%

{f}→{a,g} 2/3=66.67%

{g}→{a,f} 2/3=66.67%

{a, f}→{g}2/3=66.67%

{a, g}→{f}2/3=66.67%

{f, g}→{a}2/4=50%

由原题知，minconf=60%，所以{f}→{a,g}，{g}→{a, f}，{a, f}→{g}，{a, g}→{f}符合。

**2. 假设10个对象的两个属性取值分别为A(1，2)、B(1，3)、C(2,2)、D(2, 3)、 E(3, 4)、F(4，1)、G(4, 3)、H(4, 4)、I(5，1)、J(5, 3)。用K均值法将其聚为3个类。**

**答：**

由于K=3，所以初始聚类中心选择C(2,2)，F(4,1)与H(4,4)

曼哈顿距离为|x1-x2|+|y1-y2|

第一次迭代（横屏看）

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | A(1,2) | B(1,3) | C(2,2) | D(2,3) | E(3,4) | F(4,1) | G(4,3) | H(4,4) | I(5,1) | J(5,3) |
| C(2,2) | 1 | 2 | 0 | 1 | 3 | 3 | 3 | 4 | 4 | 4 |
| F(4,1) | 4 | 5 | 3 | 4 | 4 | 0 | 2 | 3 | 1 | 3 |
| H(4,4) | 5 | 4 | 4 | 3 | 1 | 3 | 1 | 0 | 4 | 2 |

新的组为

C1：A，B，C，D

C2：F，I

C3：E，G，H，J

计算出新的聚类中心为

C1的中心：((1+1+2+2)/4,(2+3+2+3)/4)=(1.5,2.5)

C2的中心：((4+5)/2,(1+1)/2)=(4.5,1)

C3的中心：((3+4+4+5)/4,(4+3+4+1)/4)=(4,3)

第二次迭代（横屏看）

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | A(1,2) | B(1,3) | C(2,2) | D(2,3) | E(3,4) | F(4,1) | G(4,3) | H(4,4) | I(5,1) | J(5,3) |
| (1.5,2.5) | 1 | 2 | 1 | 1 | 3 | 4 | 3 | 4 | 5 | 4 |
| (4.5,1) | 4.5 | 5.5 | 3.5 | 4.5 | 4.5 | 0.5 | 2.5 | 3.5 | 0.5 | 2.5 |
| (4,3) | 4 | 3 | 3 | 2 | 2 | 2 | 0 | 1 | 3 | 1 |

新的组为

C1：A，B，C，D

C2：F，I

C3：E，G，H，J

与第一次迭代比没有变化，表明已收敛。所以三个类即为

C1：A，B，C，D

C2：F，I

C3：E，G，H，J