**往年考题**

**填空**

1. 数据仓库的四个基本特征是指数据仓库中的数据是 面向主题的、集成的、不可更新的和随时间不断变化的 。
2. 数据仓库的设计中存在三级数据模型，分别是： 概念模型设计、逻辑模型设计、物理模型设计 。
3. 从操作型数据库到数据仓库的过程中，要进行的步骤有：数据抽取、清洗、转换、加载。
4. 粒度是数据仓库的重要概念，粒度越小，数据的细节程度越 高 ，可以回答查询的种类就越 多 ，但是查询效率将会很低；提高粒度将会提高查询效率，在数据仓库中通常采用多重粒度。
5. OLAP服务器的实现方法有： ROLAP、MOLAP、HOLAP 。
6. OLAP的实现方式有以下两种： 基于关系数据库系统的实现和基于多维数据库系统的实现 。
7. 数据分割是指： 将数据分布到各自的物理单元中，以便能分别独立处理，提高数据分析效率 ，数据分割后的数据单元称为分片，数据分片的类型有 水平分片、垂直分片 、混合分片和导出分片 等。
8. 数据从操作型环境到数据仓库过程中，通常需要进行的处理操作有 抽取 、 转化 、 分析 和 模式化处理
9. 数据仓库系统是多种技术的综合体，它是由 数据仓库,管理部分和分析工具 三部分组成。
10. 聚集函数分为三种类型，分别是 分布的 、 代数的 和 整体的 。
11. 从操作型数据库到数据仓库的过程中，要进行的步骤有 数据抽取， 数据存储 ， 管理 ， 数据表现 。

**名字解释**

1. **数据仓库体系化环境**。在一个企业或组织内, 由各面向应用的OLTP数据库、以及各级面向主题的数据仓库所组成的完整的数据环境；并在这个数据环境上建立和进行一个企业或部门的从联机事务处理到企业管理决策的所有应用。
2. **粒度。**对数据仓库中的数据的综合程度高低的一个度量。
3. **维层次**。维层是一种分类方法。把每种分类方法叫做一个层次。
4. **广义索引**。广义索引是为了提高数据仓库中的数据访问速度而采用的一种索引技术。主要是处理一些最值问题。
5. **星型模型**。是最常用的数据仓库设计结构的实现模式。使数据仓库形成了一个集成系统，为用户提供分析服务对象。核心是事实表，围绕事实表的是维度表。通过事实表将各种不同的维度表连接起来，各个维度表都连接到中央事实表
6. **雪花模型**。雪花模型是对星型模型的扩展，每一个维度都可以向外连接到多个详细类别表。雪花模型是对维度表的进一步标准化，对星型模型的维度表进行的规范化处理。对维度表进行规范后形成，用多张维度表描述一个复杂维，支持对不同层上的相同属性查询，易于维护而且节省存储空间。执行查询时需要进行较多的链接操作，可能影响系统的性能。
7. **分类**。分类是指把数据样本映射到一个事先定义的类中的学习过程，即给定一组输入的属性向量及其对应的类，用基于归纳的学习算法得出分类。通过学习得到一个目标函数（分类模型）f，把每个属性集x映射到一个预定义的类标号y。总结已有类别的对象的特点从而对未知类别对象的类别预测的过程。常用的分类方法有：决策树；朴素贝叶斯；K-近邻；基于关联的分类；支持向量机；人工神经网络；逻辑回归。
8. **聚类分析**。聚类分析是将物理的或者抽象的数据集合划分为多个类别的过程，聚类之后的每个类别中任意两个数据样本之间具有较高的相似度，而不同类别的数据样本之间具有较低的相似度。
9. **数据集市**。是一种更小、更集中的数据仓库，为公司提供分析商业数据的一条廉价途径。
10. **数据仓库的元数据**。是指在数据仓库建设过程中产生的有关数据源定义、目标定义、转换规则等关键数据，是定义数据仓库对象的数据。如传统数据库中的数据字典就是一种元数据。
11. **数据仓库的主题**。是在较高层次上将企业信息系统中的数据进行综合、归类和分析利用的一个抽象概念,每一个主题基本对应一个宏观的分析领域。
12. **序列模式分析**。发现在时间序列上，一个项目集之后的项目集是什么，即找到时间上连续的事件。
13. **关联分析**。适合于从关系中挖掘知识。包含关联发现、序列模式发现和类似的时序发现等。

**简答**

1. 简要说明数据仓库中元数据的内容

元数据是描述数据的数据。元数据描述数据的结构、来源、抽取和转换规则、存储，描述操纵数据的进程和应用程序的结构、功能，等等。其主要目的是提供数据资源的全面指南，使得数据仓库管理员和开发人员可以方便的了解数据仓库中有什么数据，数据在什么地方，它们来源于哪里，以及数据仓库系统中是如何利用这些数据、如何管理这些数据的。

元数据 ：“关于数据的数据”，是指在数据仓库建设过程中产生的有关数据源定义、目标定义、转换规则等关键数据，是定义数据仓库对象的数据。如传统数据库中的数据字典就是一种元数据。

第一种：从操作型环境向数据仓库环境转换而建立的元数据。包含：所有源数据项名、属性及其在数据仓库中的转换。

第二种：与终端用户的多维商业模型／前端工具之间建立映射的DSS元数据。

1. 数据挖掘的步骤是什么

确定挖掘对象，准备数据，建立模型，数据挖掘，结果分析，知识应用阶段

1. 星星模式

星型模型：是最常用的数据仓库设计结构的实现模式。使数据仓库形成了一个集成系统，为用户提供分析服务对象。核心是事实表，围绕事实表的是维度表。通过事实表将各种不同的维度表连接起来，各个维度表都连接到中央事实表

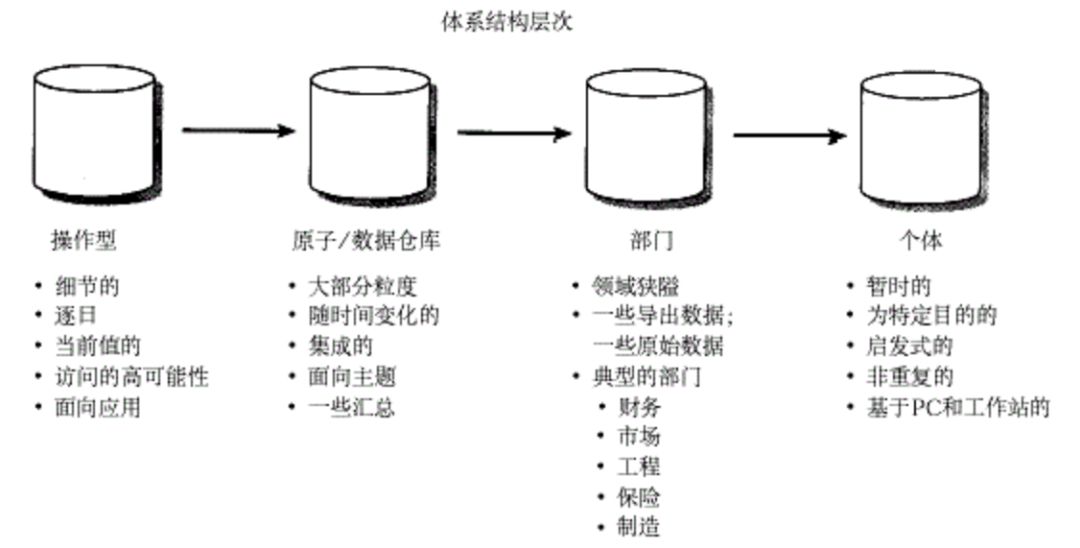
1. 你是如何理解数据仓库的数据是不可更新的，又随时间不断变化的

从数据的使用方式上来看，数据仓库的数据是不可更新的，这是指当数据被存放到数据仓库中之后，最终用户只能通过分析工具进行查询、分析，而不能修改数据仓库中存储的数据。数据仓库的不可更新并不是说数据从进入到数据仓库中就永远不变，是随时间变化而定期的被更新，从而保证前端分析结论的时间有效性。

1. 企业的数据仓库体系化结构的4个层次是什么，它们之间的关系是什么。

数据库的体系化环境，是一个企业或者组织内部，由各面向应用的OLAP数据库及各级面向主题的数据仓库所组成的完整的数据环境。

四层体系化环境：操作型环境-OLAP，全局级-数据仓库，部门级-局部仓库，个人级-个人仓库，用于启发式的分析。



1. 简要说明数据仓库的设计步骤

数据仓库的设计可以分为以下几个步骤：

1. 明确主题
2. 概念模型设计。所要完成的工作：确定系统边界，确定主要的主题域及其内容
3. 技术准备工作。包括：技术评估，技术环节准备。形成技术评估报告，软硬件配置方案，系统（软、硬件）总体设计方案。
4. 逻辑模型设计。所要完成的工作：分析主题域，确定当前要装载的主题，确定粒度层次划分，确定数据分割策略，关系模式定义。
5. 物理模型设计。所要完成的工作：确定数据的存储结构-RAID技术，确定索引策略-B树索引、位图索引，确定数据存放位置-磁带，磁盘等，确定存储分配优化。
6. 数据仓库生产。所要完成的工作：通过专用的数据抽取工具或者自行编程实现数据抽取、转换和装载。
7. 数据仓库运行与维护。所要完成的工作：建立DSS应用，使用数据仓库理解需求，调整和完善系统，维护数据仓库。
8. 简要介绍决策树分类

决策树是以实例为基础的归纳学习算法。它从一组无次序、无规律的元组中推理出决策树表示形式的分类规则。它采用自顶向下的递归方式，在决策树内部节点进行属性值比较，并根据不同的属性值从该节点向下分支，叶节点是要学习规划的类。从根到叶节点的一条路径就对应着一条合取规则，整个决策树对应着一组析取表达式规则。

1. 异常点挖掘常用方法
2. 基于统计模型。确定替代分布，固有替代分布，混合替代分布，滑动替代分布。
3. 基于距离模型。基于索引的算法、嵌套循环算法、基于单元的算法。
4. 基于密度模型。
5. 基于偏离模型。序列异常技术、OLAP技术。
6. 简要回答数据挖掘的步骤

确定数据挖掘的对象，准备数据，建立模型，数据挖掘，结果分析，知识应用阶段。

1. 在数据仓库追加数据是，捕捉数据变化的常用途径有哪些

数据仓库的数据初装完成以后，在向数据仓库输入数据的过程称为数据追加。捕捉数据变化的常用途径有：

1. 时标法：加标识
2. DELTA法：对更新做记录
3. 前后映像法：两次快照的对比
4. 日志法：利用DBMS的日志，需改进
5. 举例说明多维数据操作：上卷，下钻，切片，切块，旋转

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。

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。

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。

1. 试述常用的数值属性离散化方法

离散化方法的研究，已经提出了等频划分、等宽划分和适应离散法等。

（1）等宽划分：在最小值和最大值之间平均划分成N个区间（N用户给定）,假定A和B分别是最大值和最小值，则每个区间的宽度为W=（B-A）/N，区间的边界线分别为A，A+W，A+2W，......，A+（N-1）W，A+NW=B。

（2）等频划分：把整个区域分为N个区间，每个区间有大约相同数目的例子。如：N=10，则每个区间中有大约10%的例子。（3）这两种离散化方法都与分类信息无关，当属性值分布不均匀时，不能很好表示数据分布。

1. 简要说明梳理处理环境不适宜DSS应用的原因

事务处理和分析处理的性能特性不同；数据集成问题；数据动态集成问题；历史数据问题；数据的综合问题；

1. 常用的聚类方法

常见的k-means、层次聚类、SOM、FCM等四种聚类算法

k-means是划分方法中较经典的聚类算法之一。由于该算法的效率高，所以在对大规模数据进行聚类时被广泛应用。目前，许多算法均围绕着该算法进行扩展和改进。

层次聚类算法：根据层次分解的顺序是自底向上的还是自上向下的，层次聚类算法分为凝聚的层次聚类算法和分裂的层次聚类算法。

SOM神经网络算法假设在输入对象中存在一些拓扑结构或顺序，可以实现从输入空间(n维)到输出平面(2维)的降维映射，其映射具有拓扑特征保持性质,与实际的大脑处理有很强的理论联系。

FCM算法是一种以隶属度来确定每个数据点属于某个聚类程度的算法。该聚类算法是传统硬聚类算法的一种改进。

1. 举例说明如何确定数据仓库的主题，如何确定主题所包含的数据内容

数据仓库的主题(Subject)是在较高层次上将企业信息系统中的数据进行综合、归类和分析利用的一个抽象概念,每一个主题基本对应一个宏观的分析领域。

1. 神经网络分类过程

人类大脑的思维分为抽象（逻辑）思维、形象（直观）思维和灵感（顿悟）思维三种基本方式。

逻辑性的思维是指根据逻辑规则进行推理的过程；它先将信息化成概念，并用符号表示，然后，根据符号运算按串行模式进行逻辑推理；这一过程可以写成串行的指令，让计算机执行。然而，直观性的思维是将分布式存储的信息综合起来，结果是忽然间产生想法或解决问题的办法。这种思维方式的根本之点在于以下两点:1.信息是通过神经元上的兴奋模式分布储在网络上;2.信息处理是通过神经元之间同时相互作用的动态过程来完成的。

人工神经网络就是模拟人思维的第二种方式。这是一个非线性动力学系统，其特色在于信息的分布式存储和并行协同处理。虽然单个神经元的结构极其简单，功能有限，但大量神经元构成的网络系统所能实现的行为却是极其丰富多彩的。

1. 朴素贝叶斯为什么朴素

**CSDN 上的帖子**

朴素贝叶斯（naive Bayes classifiers）是一种分类器，在机器学习中有着广泛的应用。相信很多人知道贝叶斯定理,即

所以，当我们有一组事件，我们想通过这组事件去估计一个事件C发生的概率，比如我们想估计水果的种类，如果有一组事件分别 黄，长，弯...那我们就可以判断这是一个香蕉。 尽管黄，长，弯几个事件之间可能会相互依赖，但是在朴素贝叶斯模型中，我们假设它们相互独立，这就是他的朴素之处。值得注意的是，这里的朴素对应英语单词naive，单纯的意思，所以也可以理解为很天真单纯的估计（把数据中的每个特征看作独立分布）。

OK，现在我们具体看一下朴素贝叶斯的概率模型，其实就是一个条件概率的模型，在发生的条件下去估计事件C的概率，即计算 

根据贝叶斯定理，我们有

我们可以看到, 如果每个特征F的取值至少有两个，然后有100个特征，那给定一组数据后，想要计算是几乎不可能的，因为你的联合分布中有种可能，这样计算机是无法扫描完所有的概率空间的，即使可以，在会出现大量的0（即不存在的某种组合）。为了解决这个问题，我们假设所有特征F之间相互独立，这样一来，我们的等式，即可写成



这样即可算出我们想要的结果。

**另一种答案：**

朴素贝叶斯分类器是一种应用基于独立假设的贝叶斯定理的简单概率分类器，更精确的描述这种潜在的概率模型为独立特征模型。这种方法采用忽略了因数之间的关系，简化处理,把各属性彼此独立。

三、计算题

**1.请根据能找出的cluster的形状、预先指定的参数、所存在的缺陷这三个方面对如下聚类方法进行评价：K-Means、BIRCH、DBSCAN、OPTICS**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | K-Means | DBSCAN | OPTICS | BIRCH |
| 参数 | K |  |  |  |
| 形状 | 凸状的簇 | 任意形状 | 任意形状 |  |
| 密度变化 | 无法应对 | 无法应对 | 可应对 |  |
| 鲁棒 | 对异常值敏感 | 对异常值不敏感 | 对异常值不敏感 |  |

**2.在下面的事务数据库中，关联规则A→C的置信度和支持度是什么？关联规则C→A的置信度和支持度又是什么？如果min sup = 60%，min conf = 80%，则频繁项集有哪些？**

**答：支持度：**指项集X和Y在数据库D中同时出现的概率。(符合集数量／总样本数量)

**置信度：** 指在项集X出现的情况下，项集Y在数据库D中同时出现的条件概率，即Pr(Y/X)=Pr(XUY)/Pr(X)。(符合集数量／X样本数量)

（1）假设顾客购物的交易数据库为例，总的交易数为5，给定关联规则**A→C，置信度2/4；支持度2/5。**

（2）假设顾客购物的交易数据库为例，总的交易数为5，给定关联规则**C→A，置信度2/3；支持度2/5。**

**题库**

1. 数据处理的两种基本类型
2. 操作型（与业务活动关联，事务处理）

操作型处理主要完成数据的收集、整理、存储，查询和增删改等操作，由一般工作人员和几层管理人员完成。

操作型数据处理主要用于企业的日常事务处理工作，存放的是细节的数据。数据的组织方式以方便事务处理，提高事务处理的性能为主要指标。

数据存储操作频率高而每次操作处理时间短，主要针对本部门业务相关数据，对企业范围内集成应用考虑较少。

1. 分析型（与决策活动关联）

分析型处理是对数据的再加工，往往要访问大量的历史数据，进行复杂的统计分析，从中获取信息，因此也称为信息型处理，主要有高级管理人员完成。

分析型处理主要用于企业的管理工作，存放的是历史和综合数据。

运行时间长，消耗系统资源多，数据全面集成，相关数据收集越完整越可靠。

1. 操作型数据和分析型数据的区别

|  |  |
| --- | --- |
| 操作型数据 | 分析型数据 |
| 细节的 | 综合的，或提炼的 |
| 在存取瞬间是准确的（当前数据） | 代表过去的数据（历史数据） |
| 可更新 | 不可更新 |
| 操作需求事先可知道 | 操作需求事先不知道 |
| 生命周期符合SDLC（系统生命周期） | 完全不同的生命周期 |
| 对性能要求高 | 对性能要求宽松 |
| 一个时刻操作一单元 | 一个时刻操作一集合 |
| 事务驱动 | 分析驱动 |
| 面向应用（业务处理） | 面向分析 |
| 一次操作数据量小，计算简单 | 一次操作数据量大，计算复杂 |
| 支持日常操作 | 支持管理需求 |

**小**：一次操作数据量小

**事**：事务驱动

**常**：支持日常操作

**细**：细节的

**心**：可更新

**先**：操作事先知道

**生**：生命周期符合SDLC

**一**：一时刻操作一个单元

**准**：存取瞬间是准确的

**姓**：性能要求高

**阴**：面向应用

1. 数据库系统的局限性

数据的分散，“蜘蛛网”问题，数据不一致问题，数据动态集成问题，历史数据问题，数据的综合问题。

1. 数据的分散，即事务处理应用分散。由于多种原因，如经费有限，公司地点分散或企业兼并等，当前企业内部各事务处理的应用实际上几乎是独立的，设计方案仅对于当前问题有效，没有良好的可扩展性。
2. “蜘蛛网”问题。不加限制的连续抽取导致企业的数据间形成了错综复杂的网状结构。时间基准不统一，抽取算法不相同，抽取级别不相同，参考不同的外部数据。
3. 数据不一致。同一字段不同的数据类型，同一字段不同名字，同名字段不同含义。
4. 数据的动态集成问题。
5. 历史数据问题。联机事务处理系统中的数据往往是当前的，或近几年的。但对于决策分析，历史数据相当重要，许多分析以历史数据为依托。
6. 数据的综合问题。联机事务处理系统中的数据太细节，而决策系统中并不需要对细节数据进行分析。
7. 数据仓库的4个基本特征

面向主题，集成，不可更新，随时间不断变化。

已有数据记录不可更新，整体数据仓库数据跟随时间变化不断增加新的数据，并且去除某一时间点之前的数据，只有读取，删除，插入操作，没有修改操作。

1. 面向主题。根据要求将数据组织成一个完备的分析领域，即主题域。
2. 集成。数据是从许多其他数据源中抽取得到的，然后再进行清洗，转换，整合，统一，最后加载到数据仓库中。
3. 不可更新。数据的主要操作是查询，原始数据一旦进入数据仓库后，一般不允许再修改，且会被长期保留。只需要定期的加载和刷新。
4. 随时间不断变化。已有数据记录不可更新，随时间不断增加新的数据，将不需要的数据从数据仓库中卸出，转储到其他设备，随着时间的变化不断的重新综合数据。
5. 数据仓库体系结构

数据源，集成工具，数据仓库与数据仓库服务器，OLAP服务器，元数据与元数据管理工具，数据集市和前台分析工具等组成。

1. ODS

ODS主要是适应进行企业级全局应用的需求而产生的。这种全部应用还可以大致地划分为两类：一类是进行企业级的联机事务处理，另一类可以称为“即时OLAP”数据处理。

1. 数据粒度

两种形式：

1. 第一种粒度是对数据仓库中的数据的综合程度高低的一个度量；粒度大小影响数据仓库效率，能回答询问的种类。粒度越小，细节程度越高，综合程度越低，回答查询的种类就越多。数据仓库是多粒度的，不同粒度回答不同的查询。
2. 另一种特殊形式的粒度是样本数据库。根据采样率的高低来划分，以一定的采样率从细节数据或轻度综合数据中随机抽取一个子集来代替数据源进行模拟分析。
3. 数据分割（分片/数据分片）

将数据分布到各自的物理单元中，以便能够分别独立处理，提高数据分析效率。

数据分割后的数据的单元成为分片。

数据分割一般而言总应包括日期。

数据分割后要达到的目的是，使数据更易于重构、索引、重组、恢复、监控和顺序扫描。

1. 数据仓库中数据的追加

时标方法；DELTA文件；前后快照文件的方法；日志文件（推荐的方法）

1. 维的构成：维成员，维层，维层次，维属性

维是人们观察数据的特定角度，是某个事物的属性。维是商业活动中的一个基本要素。每个维都有一个唯一的名字。

维成员：维由一些维成员构成。维的一个取值称为维的一个成员，每个成员有一个名字。

维层：不同的细节程度称为不同的维层。

维层次：维层是一种分类方法，把每种分类方法叫做一个层次。

维属性：说明为成员所具有的特征。

1. 度量

要分析的目标或对象，输入度量好导出度量，有可累计型和不可累计型。

1. 常见的多维数据模型
2. 星形模型。由一个很大的中心表和一组较小的表组成，不支持维的层结构，实现时将所有的维层属性存在在这一个表中，没有进行规范化。每个层有自己的属性，有很多冗余。当不同的维层有相同属性时只能使用换名方法，影响查询。
3. 雪片模型。对维表进行规范化后形成，用多张维表描述一个复杂维，支持不同层上的相同属性查询，易于维护而且节省存储空间。执行查询时需要进行较多的链接操作，可能影响系统的性能。
4. 事实群模型（星系模型）。在复杂的应用中需要多个事实表共享维表，类似于星形模型集合。
5. 聚集函数的分类
6. 分布型聚集函数。将数据分成n份，对其中每一份应用该函数，可以得到n个聚集值，对这n个聚集值进行计算得到的结果和整个数据（不划分）应用该函数得出结果一致。具有可累计的特性。
7. 代数型具体函数。一个函数可以由若干个分布型函数进行代数运算得出。
8. 整体型聚集函数。一个函数不能由其他函数进行代数运算得出。
9. 常用多维分析操作
10. 切片。在数据方体的某一维上选定一个维成员的动作。
11. 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。
12. 旋转。改变数据方体维的次序的动作。
13. 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。
14. 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。
15. 数据方体的存储

基于多维数组存储的OLAP实现方式维称为MOLAP；

基于关系表存储的OLAP实现方式称为ROLAP，星型模型就是基于关系表存储的ROLAP；

基于混合数据组织的OLAP实现方式称为HOLAP。

1. 提高数据仓库效率的方法

合并表；建立数据序列；引入冗余；进一步细分数据；生成导出数据；建立广义索引；粒度划分；分割。

1. 支持度
2. 支持度。指项集X和Y在数据库D中同时出现的概率。（符合集数量/总样本数量）。
3. 置信度。指在项集X出现的情况下，项集Y在数据库D中同时出现的条件概率，即 Pr（Y/X）=Pr（YUY）/Pr（X）。（符合集数量/X样本数量）。
4. 预测模型：回归与分析

用于回归的预测模型：线性回归模型、非线性回归模型、分段线性模型；

用于分类的预测模型：判断模型、概率模型。

1. 聚类分析

聚类分析是给定一组对象的描述信息，发现具有共同特性的对象构成簇。

分类和聚类的不同之处：分类是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；聚类是将数据对象分到各个簇中，聚类是无监督学习，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。

1. 计算关联规则及频繁项集。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, b, d, e |
| 2 | a, c |
| 3 | c, d, f |
| 4 | a, b, c, d |
| 5 | a, b, c, f |

**答：频繁项集**：

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **C1** | | | | | | minsup=40%，所以a，b，c，d，f符合，得L1 | | | |
| **Item** | **支持度** | | **为什么** | | | **Item** | **支持度** | | |
| {a} | 80% | | 因为A在原题表格里出现了4次，原题表格一共5行，出现次数除以总行数即4/5=80% | | | {a} | 80% | | |
| {b} | 60% | | {b} | 60% | | |
| {c} | 80% | | {c} | 80% | | |
| {d} | 60% | | {d} | 60% | | |
| {e} | 20% | | {f} | 40% | | |
| {f} | 40% | |  |  | | |
| C2 | | | | | L2 | | | |
| **Item** | | | **支持度** | | **Item** | | | **支持度** |
| {a, b} | | | 60% | | {a, b} | | | 60% |
| {a, c} | | | 60% | | {a, c} | | | 60% |
| {a, d} | | | 40% | | {a, d} | | | 40% |
| {a, f} | | | 20% | | {b, c} | | | 40% |
| {b, c} | | | 40% | | {b, d} | | | 40% |
| {b, d} | | | 40% | | {c, d} | | | 40% |
| {b, f} | | | 20% | | {c, f} | | | 40% |
| {c, d} | | | 40% | |  | | |  |
| {c, f} | | | 40% | |  | | |  |
| {d, f} | | | 20% | |  | | |  |
| **C3** | | | | **L3** | | | | |
| **Item** | | **支持度** | | **Item** | | | | **支持度** |
| {a, b, c} | | 40% | | {a, b, c} | | | | 40% |
| {a, b, d} | | 40% | | {a, b, d} | | | | 40% |
| {a, b, f} | | 20% | |  | | | |  |
| {b, c, d} | | 20% | |  | | | |  |
| {b, c, f} | | 20% | |  | | | |  |
| {c, d, f} | | 20% | |  | | | |  |

C4。 L4无

|  |  |
| --- | --- |
| **Item** | **支持度** |
| {a, b, c, d} | 20% |

**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

1. 计算聚类。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：** A1(1, 9) A2(3, 8)  A3(5, 5) A4(9, 2) B1(6, 8)  B2(10, 5)  B3(3, 8)C1(10, 2)  C2(4, 8)  C3(2, 6)  C4(7, 3)

由题意知 K=3。曼哈顿距离为|x1-x2|+|y1-y2|欧式=https://mmbiz.qpic.cn/mmbiz_png/4TQGCsqoianCuPNQhsuZXYErSSkibLBxQ505PKLwx1o7ulTZjbFicCBibeNXA7RxUawsiby7Tmk9hM1Yuj61ic7ZMjOA/640?wx_fmt=png&wxfrom=5&wx_lazy=1&wx_co=1

第一次迭代（横屏看）

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | A1 | A2 | A3 | A4 | B1 | B2 | B3 | C1 | C2 | C3 | C4 |
| A1 | 0 | 3 | 9 | 15 | 6 | 13 | 3 | 16 | 4 | 4 | 12 |
| B1 | 6 | 3 | 5 | 9 | 0 | 7 | 3 | 10 | 2 | 6 | 6 |
| C1 | 16 | 13 | 7 | 1 | 10 | 3 | 13 | 0 | 12 | 12 | 4 |

C1（第一个）:A1、A2、B3、C3；C2（第二个）: A3、B1、C2、C3（第三个）:A4、B2、C1、C4

第二次迭代，算出一个新中心

C1的中心：((1+3+3+2)/4, (9+8+8+6)/4)=(2,7)

C2的中心：((5+6+4)/3,(4+8+8)/3)=(5,6)

C3的中心：((9+10+10+7)/4,(2+5+2+3)/4)=(9,3)

（横屏看）

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | A1 | A2 | A3 | A4 | B1 | B2 | B3 | C1 | C2 | C3 | C4 |
| (2,7) | 3 | 2 | 6 | 12 | 5 | 10 | 2 | 13 | 2 | 1 | 9 |
| (5,6) | 7 | 4 | 2 | 8 | 3 | 6 | 4 | 9 | 3 | 3 | 5 |
| (9,3) | 14 | 11 | 5 | 1 | 8 | 3 | 11 | 2 | 10 | 10 | 2 |