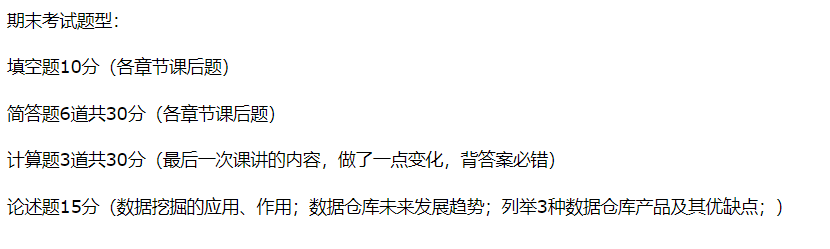
**数据挖掘复习**



1. **填空题（5小题，没空1分，15分）**
2. **数据仓库是一个面向对象、集成的、相对稳定、反映历史变化的数据集合。**
3. **数据处理通常分为两大类：联机事务处理和联机分析处理。**
4. **使用星型模式可以从 一定程度上提高查询效率。因为星型模式中数据的组织已经经过预处理，主要数据都在庞大的事实表中。**
5. **分类的过程包括获取数据、预处理、分类器设计、分类决策。**
6. **分类器设计阶段包含三个过程：划分数据集、分类器构造、分类器测试。**
7. **分类问题中常见的评价准则有精确度、查全率和查准率、几何均值。**
8. **支持向量机中常用的核函数有多项式核函数、径向基核函数和S型核函数。**
9. **聚类分析包括划分连续性属性、二值离散型属性、多值离散型属性、混合类型属性四种类型描述属性的相似度计算方法。**
10. **连续型属性的数据样本之间的距离有欧氏距离、曼哈顿距离、明考斯基距离。**
11. **划分聚类方法对数据集进行聚类时包含三个要点：选定某种距离作为数据样本间的相似性度量、选择评价聚类性能的准则函数、选择某个初始分类，之后用迭代的方法得到聚类结果，使得评价聚类的准则函数取得最优值。**
12. **层次聚类方法包括凝聚型、分解型两种层次聚类方法。**
13. **关联规则经典算法包括Apriori和FP-growth，其中FP-growth的效率更高。**
14. **简答题（5小题，每题5分，25分）**
15. **简述你对数据仓库未来发展趋势的看法。P23 （论述题）**

**数据仓库的发展将从数据抽取、存储管理、数据表现和方法论四个方面的发展。**

**数据抽取方面，数据库技术集中体现在系统集成化的方面，它将互连、转换、复制、调度、监控纳入标准化的统一管理，以适应数据仓库本身和数据源的变动，更有利于系统的管理与维护。**

**存储管理方面，未来的发展将会使数据库厂商生产数据仓库引擎，作为数据仓库服务器产品和数据仓库服务器并驾齐驱。**

**数据表现方面，统计分析的算法和功能将普遍集成到联机分析产品中，并与Internet/Web技术紧密结合。**

**方法论方面，数据仓库实现过程中的方法论将更加普及，将成为数据库设计的一个明确的分支，成为管理信息系统设计的必备。**

**在未来大规模的定制经济环境下，数据仓库将成为企业获得竞争优势的关键武器，也是机器学习、人工智能应用的重要数据来源及数据存储方式。总之，数据仓库是基于数据管理和数据应用的综合性技术和解决方案，他是数据库市场上新一轮的增长点，同时也是企业未来应用系统的重要组成部分。**

1. **请列车3种数据仓库产品，并说明其优缺点。（论述题）**

**（1）IBM 公司提供了一套基于可视化数据仓库的商业智能（BI）解决方案，包 括：Visual Warehouse（VW）、 Essbase/DB2 OLAP Server 5.0、IBM DB2 UDB，以及来自第 三方的前端数据展现工具（如 BO）和数据挖掘工具（如 SAS）。其中，VW 是一个功能很 强的集成环境，既可用于数据仓库建模和元数据管理，又可用于数据抽取、转换、装载和调 度。Essbase/DB2 OLAP Server 支持“维”的定义和数据装载。Essbase/DB2 OLAP Server 不 是 ROLAP（Relational OLAP）服务器，而是一个（ROLAP 和 MOLAP）混合的 HOLAP 服 务器，在 Essbase 完成数据装载后，数据存放在系统指定的 DB2 UDB 数据库中。它的前端 数据展现工具可以选择 Business Objects 的 BO、Lotus 的 Approach、Cognos 的 Impromptu 或 IBM 的 Query Management Facility；多维分析工具支持 Arbor Software 的 Essbase 和 IBM （与 Arbor 联合开发）的 DB2 OLAP 服务器；统计分析工具采用 SAS 系统。 （2）Oracle 数据仓库解决方案主要包括 Oracle Express 和 Oracle Discoverer 两个部分。Oracle Express 由 四个工具组成：Oracle Express Server 是一个 MOLAP(多维 OLAP)服务器，它利用多维模型， 存储和管理多维数据库或多维高速缓存，同时也能够访问多种关系数据库；Oracle Express Web Agent 通过 CGI 或 Web 插件支持基于 Web 的动态多维数据展现；Oracle Express Objects 前端数据分析工具（目前仅支持 Windows 平台）提供了图形化建模和假设分析功能，支持可视化开发和事件驱动编程技术，提供了兼容 Visual Basic语法的语言，支持 OCX 和 OLE； Oracle Express Analyzer 是通用的、面向最终用户的报告和分析工具（目前仅支持 Windows 平台）。Oracle Discoverer 即席查询工具是专门为最终用户设计的，分为最终用户版和管理员版。在 Oracle 数据仓库解决方案的实施过程中，通常把汇总数据存储在 Express多维数据库中，而将详细数据存储在 Oracle 关系数据库中，当需要详细数据时，Express Server 通过 构造 SQL 语句访问关系数据库。**

**（3）Microsoft 将 OLAP 功能集成到 SQL Server 数据库中，其解决方案包括 BI 平台、BI 终端工具、BI 门户和 BI 应用四个部分。① BI 平台是 BI 解决方案的基础，包括 ETL 平台 SQL Server 2005 Integration Service(SSIS)、数据仓库引擎 SQL Server 2005 RDBMS 以及多维分析和数据挖掘引擎 SQL Server 2005 Analysis Service、报表管理引擎 SQL Server 2005 Reporting Service。 ② BI 终端用户工具，用户通过终端用户工具和 Analysis Service 中的 OLAP 服务和数据挖掘服务进行交互来使用多维数据集和数据挖掘模型，终端用户通 常可使用预定义报表、交互式多维分析、即席查询、数据可视化、数据挖掘等多种方法。 ③ BI 门户提供了各种不同用户访问 BI 信息的统一入口。BI 门户是一个数据的汇集地，集成了 来自不同系统的相关信息。用户可以制定个性化的个人门户，选择和自己相关性最强的数据， 提高信息访问和使用的效率。 ④ BI 应用是建立在 BI 平台、BI 终端用户工具和 BI 统一门 户这些公共技术手段之上的满足某个特定业务需求的应用，例如零售业务分析、企业项目管 理组合分析等。**

**3.什么是数据仓库的数据ETI过程？**

**答：ETI过程就是负责将操作型数据转换成调和数据的过程。**

**ETI过程有四个步骤组成，即抽取、清洗、转换、加载和索引，这些步骤也可进行不同的组合。**

**数据抽取：从源文件和源数据库中获取相关数据用于填充数据仓库，称为数据抽取。数据抽取的两个常见类型是静态抽取和增量抽取。静态抽取是用于最初用于填充数据仓库，增量抽取是用于进行数据仓库的维护，仅仅获取那些从上一次获取之后源数据所发生的变化。**

**数据清洗：是一种使用模式识别和其他技术，在将原始数据转换和转入数据仓库之前来升级原始数据质量的技术。**

**数据转换：数据转换在整个过程中处于中心位置，它把数据从源操作型业务系统转换到企业数据仓库的数据格式。可能是简单的数据格式的转换，也可能是高强度的数据组合的变化。**

**数据加载和索引：即是把数据加载到数据仓库或数据集市的过程。**

1. **什么是信息包图法？它为什么适用于数据仓库的概念模型的设计？P57**

**答：信息包图定义一个主题内容和主要性能指标之间的关系，其目标就是在概念层满足用户需求。信息包图拥有三个重要对象，（度量）指标、维度和类别。**

**信息包图法也叫用户信息需求表，就是在一张平面表格上描述元素的多维性，其中的每一个维度用平面表格的一列表示。**

**总之信息包图法是一种自上而下的数据建模方法，即从用户的观点开始设计（用户的观点是通过与用户交流得到的），站在管理者的角度把焦点集中在企业的一个或几个主题上，着重分析主题所涉及数据的多维性，这种自上而下的方法几乎考虑了所有的信息以及这些信息源影响业务活动的方式。**

1. **关联规则的分类有哪些？关联规则挖掘的步骤包括什么？P43.**

**关联规则的分类：**

**（1）基于规则中涉及到的数据的维数，关联规则可以分为单维的和多维的。**

**（2）基于规则中数据的抽象层次，可以分为单层关联规则和多层关联规则。**

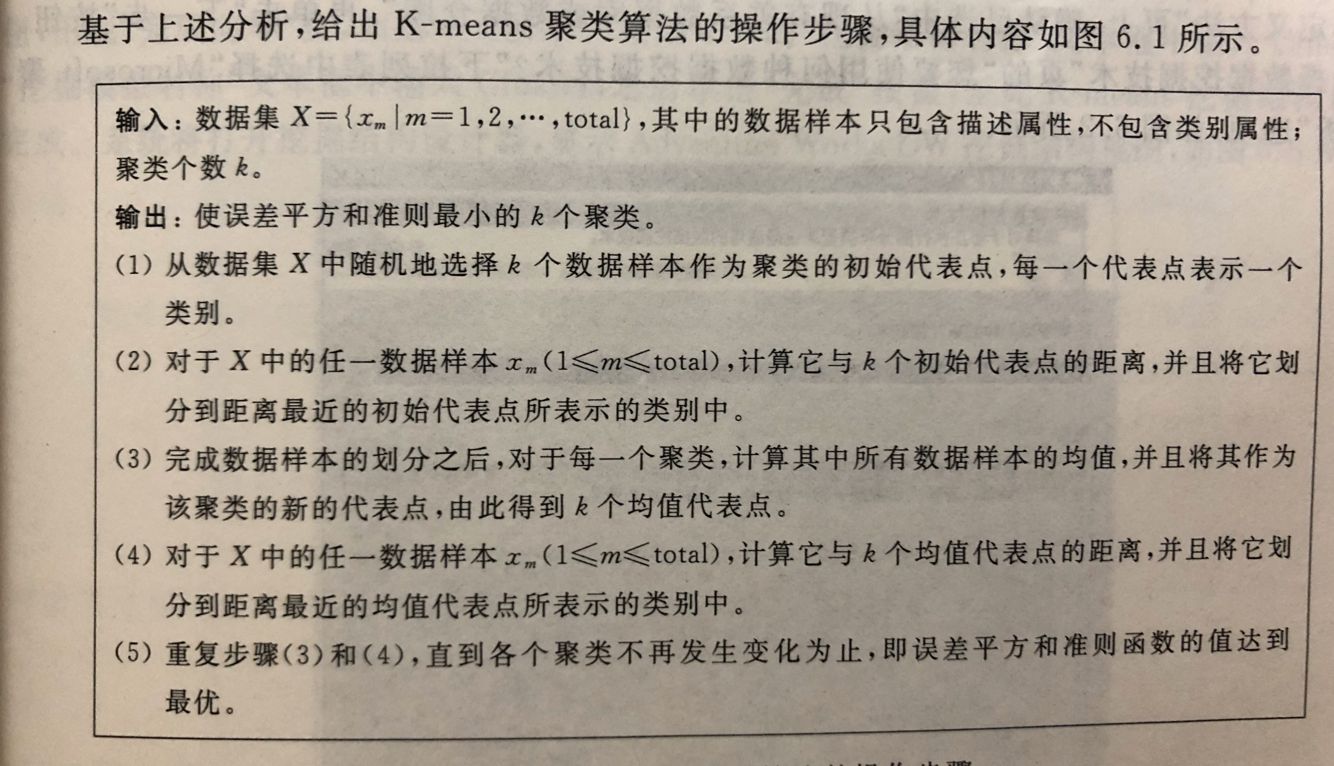
**（3）基于规则中处理的变量的类型不同，关联规则可以分为布尔型和数值型。**

**关联规则挖掘的步骤：**

1. **找出交易数据库中所有大于或等于用户指定的最小支持度的频繁项集；**

**（2）利用频繁项集生成所需要的关联规则，根据用户设定的最小可信度进行取舍，产生强关联规则。**

**5.请描述K-means聚类算法的操作步骤。P139**



**6.贝叶斯网络的三个主要议题是什么？P153**

**答：①贝叶斯网络预测：是指从起因推测一个结果的推理，也称为由顶向下的推理。**

**②贝叶斯网络诊断：是指从结果推测一个起因的推理，也称为由底至上的推理。**

**③贝叶斯网络学习：是指由先验的贝叶斯网络得到后验的贝叶斯网络的过程。先验的贝叶斯网络是根据用户的先验知识构造的贝叶斯网络，后验贝叶斯网络是把先验贝叶斯网络和数据相结合而得到的贝叶斯网络。**

**7.怎样从历史数据中训练出节点之间的条件概率或联合条件概率？**

**答：要训练条件概率 P(B|A)，可以在历史数据中统计 A 发生的次数 T(A)，然后统计在 A 发生的数据中 B 发生的次数 T(A,B)，条件概率 P(B|A) = T(B)/T(A)。要训练联合条件概率 P(C|A,B)，可以在历史数据中统计 A、B 共同发生的次数 T(A,B)，然后在 A、B 共同发生的数据中统计 C 发生的次数T(A,B,C)，联合条件概率 P(C|A,B)=T(A,B,C)/T(A,B)。以上的符号 A、B、C 可以表示某个事件，也可以表示该事件的相反事件。**

**8.前馈网络和递归网络有什么本质区别？**

**前馈网络和递归网络的本质区别是网络的某些输出是不是循环作为网络的输入。 前馈网络的所有输出都不能作为输入，而递归网络的某些输出可以循环作为网络的输入。**

**9.简单遗传算法包括哪些步骤？P200**

**（1）初始化，产生初始种群。**

**（2）个体评价，即计算种群中每个个体的适应度。**

**（3）按选择概率 Ps，执行选择算子，从当前种群中选择部分个体进入下一代种群。**

**（4）按交叉概率 Pc，执行交叉算子。**

**（5）按变异概率 Pm， 执行变异算子。**

**（6）若满足设定的终止条件，则输出种群中适应度最优的个体作为问题的最优解或满意解，否则执行（2）**

**10.请简述数据挖掘中关联规则Apriori算法的思想。**

**基本思路是层次搜索的迭代方法。通过多次扫描交易记录集，产生不同长度的频繁集，首先产生1频繁集，在此基础上经过连接修剪产生2频繁集，直到无法产生频繁集。在第n次循环后产生n频繁集中，首次产生n-候选频繁集的集合C，简称候选集。c中的每一个项集是对两个只有一个项不同的属于（n-1）频繁项集连接产生。**

**Apriori算法的基本原理是用支持度表示关联规则的强度，把具有关联规则的商品看做一个集合。从最小的集合开始，筛选出支持度大于某个值的集合，然后合并集合，再循环，直到找不集合为止。**

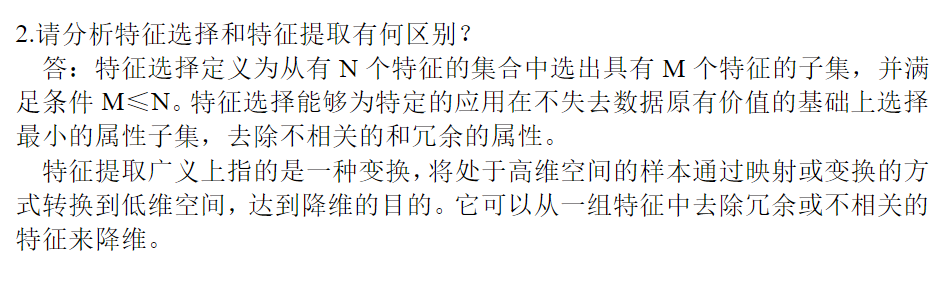
**priori算法的优点：结构简单、易于理解。**

**Apriori算法的缺点：产生大量的候选项集，I/O开销较大。**

**11.请比较PCA和LDA的区别。**

**PCA（主成成分分析）和LDA（线性判别分析）都是经典的降维算法。PCA是无监督的，也就是训练样本不需要标签；LDA是有监督的，也就是训练样本需要标签。PCA是去除掉原始数据中冗余的维度，而LDA是寻找一个维度，使得原始数据在该维度上投影后不同类别的数据尽可能分离开来。**

**12.请分析特征选择和特征提取有何区别。**



**13.聚类和分类有什么区别和联系？**

**聚类与分类异同：**

**聚类和分类都是为了对数据进行划分，将相似度高的数据样本划分到一个集合中。**

**聚类要划分的类是未知的，分类则可按已知规则进行；**

**聚类是一种无指导学习，它不依赖预先定义的类和带类标号的训练实例，属于观察式学习，分类则属于有指导的学习，是示例式学习。**

1. **TF-IDF算法是什么？有什么实际意义？**

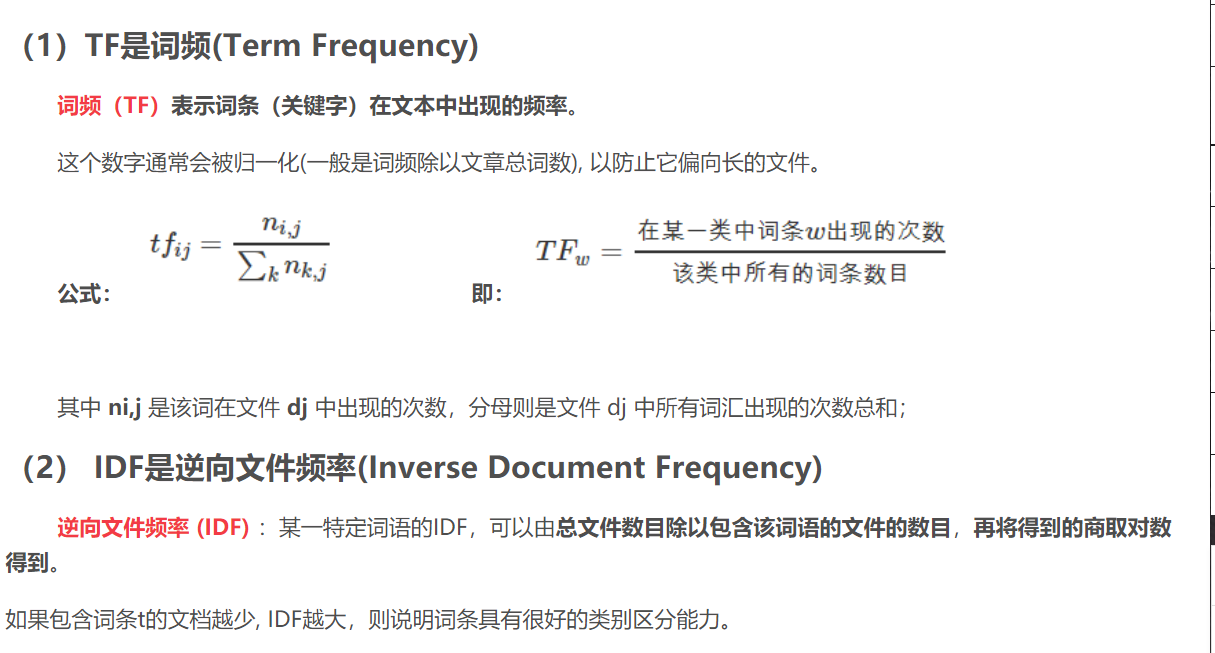
TF-IDF是一种用于信息检索与数据挖掘的常用加权技术。TFIDF的主要思想是：如果某个词或短语在一篇文章中出现的频率TF高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。

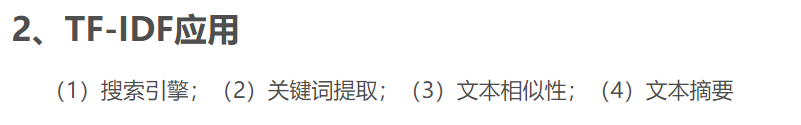
常用的领域：搜索引擎、文本相似性、文本摘要

**TF-IDF（term frequency–inverse document frequency，词频-逆向文件频率）是一种用于信息检索（information retrieval）与文本挖掘（text mining）的常用加权技术。**

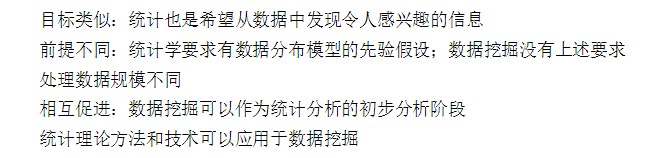
**TF-IDF是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。**

**TF-IDF的主要思想是：如果某个单词在一篇文章中出现的频率TF高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。**





1. **数据挖掘与统计的区别于联系？**



1. **什么是关联规则？关联规则的应用有哪些？**

**关联规则挖掘最初用来发现超级市场中用户购买的商品之间的隐含关联关系，并用规则的形式表示出来，称为关联规则(Association Rule)。关联规则的应用还包文本挖掘、商品广告邮寄分析、网络故障分析等。**

1. **什么是分类？分类的应用领域有哪些？**

**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。分类问题在商业、银行业、医疗诊断、生物学、文本挖掘、因特网筛选等领域都有广泛应用。**

1. **什么是聚类分析？聚类分析的应用领域有哪些？**

**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。**

**聚类分析是数据挖掘应用的主要技术之一，它可以作为一个独立的工具来使用， 将未知类标号的数据集划分为多个类别之后，观察每个类别中数据样本的特点，并且对某些特定的类别作进一步的分析。此外，聚类分析还可以作为其他数据挖掘技术（例如分类学习、 关联规则挖掘等）的预处理工作。聚类分析在科学数据分析、商业、生物学、医疗诊断、文本挖掘、Web 数据挖掘等领域都有广泛应用。**

1. **计算机（3小题，每题15分，45分）**

**1.Apriori算法。见P107页。**

**2.决策树算法。见P130页。**

**3.贝叶斯和全概率公式。P151.**

1. **论述题（1题，15分）**

**列举几项你知道的数据挖掘应用，并论述数据挖掘在其中的作用？**

