## 数据挖掘

填空题：

1. 分类器设计阶段包括三个过程：划分数据集、分类器构造、分类器测试
2. 数据仓库就是一个面向主题的、集成的、相对稳定的、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、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，回答查询的种类事实表
5. 关联规则的经典算的包括Apriori算 法和FP-growth算法，其中FP-growth算法的效率最高
6. 分类问题中常用的评价准则有精确率、查全率和查准率、F-measure、几何均值
7. 支持向量机常用的核函数有多项式核函数、径向基核函数、S型核函数
8. 层次聚类方法包括凝聚型层次聚类**、**分散型层次聚类两种层次聚类方法
9. 聚类分析包括连续型属性**、二值离散型属性、多值离散型、**混合类型属性四种类型描述属性的相似度近似方法划分聚类对数据集进行聚类包含三个要点：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

简述题：

1. 简述你对数据仓库未来发展趋势的看法：
2. 数据仓库技术的发展包括数据抽取、存储管理、数据表现和方法论等方面。
3. 在数据抽取方面，未来的技术发展将集中在系统集成化方面。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或数据源可能的变化，使系统更便于管理**和维护。
4. 在数据管理方面，带有决策支持扩展的并行关系数据库将最具发展潜力。
5. 在数据表现方面，数理统计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。
6. 数据仓库仍是一个方兴未艾的领域，有着良好的发展前景。
7. 请列出三种数据仓库产品，并说明其优缺点
8. IBM 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：（VW），Essbase/DB2 OLAP Server 5.0、IBM DB2 UDB。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。
9. Oracle 数据仓库解决方案主要包括Oracle Express 和Oracle Discoverer 两个部分。Oracle Express ，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 Discoverer即查询工具是专门为最终用户设计的，使用方便简洁分为最终用户版和管理员版。
10. 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。 BI 平台是BI解决方案的基础。
11. 什么是数据仓库的数据ETL过程

是负责将操作性数据转换成调和数据的过程。数据调和是构建一个数据仓库中最难的和最具技术挑战性的部分。数据调和可分为两个阶段：一是企业级数据仓库首次创建时的原始加载；二是接下来的定期修改，以保持数据仓库当前有效性和扩展性。整个过程由四个步骤组成：抽取、清晰、转换、加载和索引。

1. 什么是消息包图法？它为什么适用于数据仓库的概念模型设计？

信息包图法，也称用户信息需求表，就是在一张平面表格上描述元素的多维性，其中每一个维度用平面表格的一列表示。创建信息包图需要确定最高层和最低层的信息需求，以便最终设计出包含各个层次需要的数据仓库。它是一种自上而下的数据建模方法，几乎考虑了所有的信息源，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是什么？

分类：

1. 基于规则中涉及到的数据的维数，可以分为单维和多维的；
2. 基于规则中数据的抽象层次，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类型不同，分为布尔型和数值型。

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。
3. 请描述K-means聚类算法的操作步骤：

(1)从数据中选择k个对象作为初始聚类中心;

　　(2)计算每个聚类对象到聚类中心的距离并划分类别；

(3)计算划分的每个类别的均值作为新的聚类中心

(4)计算每个聚类对象到新的聚类中心的距离并划分到距离最近的类别中；

(5)重复34直到各个聚类不再发生变化

1. 贝叶斯网络的三个重要议题是什么？

预测、诊断和历史数据训练

1. 怎样从历史数据中训练出结点之间的条件概率或联合条件概率？

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A发生的数据中B发生的次数T(A,B)，条件概率P（B|A）= T(A,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)，然后统计A、B共同发生的数据中C发生的次数T(A,B,C)，条件概率P（C|A,B）= T(A,B,C)/ T(A,B)。

1. 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能作为网络的输入

1. 简单遗传算法包括哪些步骤？
2. 初始化，产生初始种群；
3. 个体评价，即计算种群中每个个体的适应度；
4. 按选择概率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）
7. 简述数据挖掘种关联规则Apriori算法的思想

Apriori算法的基本思想是通过对数据库的多次扫描来计算项集的支持度，发现所有的频繁项集从而生成关联规则。Apriori算法对数据集进行多次扫描。第一次扫描得到频繁1-项集的集合L1，第K（k>1）次扫描首先利用第(k-1)次扫描的结果Lk-1来产生候选集k-项集的集合Ck，然后在扫描的过程中确定Ck的支持度。最后，在每次扫描结束时计算频繁k-项集的集合Lk，算法在候选集k-项集的集合Ck为空时结束。

1. 比较PCA和LDA的区别：

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。

1. 请分析特征选择和特征提取有何区别：

特征选择定义为congN个特征的集合种选出具有M个特征的子集，并满足条件M<=N。特征选择能够为特定的应用在不是去数据原有价值的基础上选择最小的属性子集，去除不相关的和冗余的属性。

特征提取广义上指的是一种变换，将处于高维空间的样本通过映射或变换的方式转换到低维空间，达到降维的目的。它可以从一组特征中去除冗余和不相关的特征来降维。

1. 聚类和分类有什么区别和联系：

区别：分类是事先定义好类别 ，类别数不变。聚类则没有事先预定的类别，类别数不确定。分类适合类别或分类体系已经确定的场合，聚类则适合不存在分类体系、类别数不确定的场合。

1. 什么是关联规则？应用有哪些？

用规则的形式表示出物品之间隐含的关联关系。

应用包括文本挖掘、商品广告邮寄分析、网络故障分析等。

1. TF-IDF算法是什么？有什么实际意义？

表示一个词在文档中的重要程度的算法。

用于信息检索和文本挖掘的统计度量，用于评估术语对文档集合中特定文档的重要程度。可以用于衡量整个网络上特定属于的平均使用频率。

1. 数据挖掘与统计的区别与联系？

统计学主要利用概率论建立数学模型，是研究随机现象的常用数学工具之一。

数据挖掘分析大量数据，发现其中的内在联系和知识，并以模型或规则表达这些知识。

两者采用的某些分析方法（如回归分析）是相同的

1. 什么是分类？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？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。

聚类分析在数据挖掘、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

计算题：

1. P107
2. P115
3. P151

论述题：

列举几项你知道的数据挖掘应用，并讨论数据挖掘在其中的作用。

**一：风险控制（贷款偿还预测和客户信用评价）**

有很多因素会对货款偿还效能和客户信用等级计算产生不同程度的影响。数据挖掘的方法，如特征选择和属性相关性计算，有助于识别重要的因素和非相关因素。例如，与货款偿还风险相关的因素，包括货款率、资款期限、负债率、偿还与收入(payment—to—income)比率、客户收入水平、受教育程度、居住地区、信用历史等等。而其中偿还与收入比率是主导因素，受教育水平和负债率则不是。银行可以据此调整货款发放政策，以便将货款发放给那些曾被拒绝、但根据关键因素分析、其基本信息显示是相对低风险的申请者。

**二：交叉销售（业务关联分析）**

通过关联分析可找出数据库中隐藏的关联网，银行存储了大量的客户交易信息，可对客户的收入水平、消费习惯、购买物品等指标进行挖掘分析，找出客户的潜在需求；并对各个理财产品进行交叉分析，找出关联性较强的产品，从而对客户进行有针对性的关联营销，提高银行业绩

**三：客户市场细分**

根据银行大量的客户资料以及客户存储款情况，利用有效的聚类或者协同过滤，将客户有效地划分为不同的组，使得具有相同存储和贷款行为的客户分为一组，从而可以对每一组总结各自每个组的特点，对每个组开展有针对性活动。

此外，针对不同的客户类型(例如大客户类型，潜在价值高，但是忠诚度很难保持)设计出量体裁衣的产品组合、沟通方式，以及客户服务，从而达到提高客户忠诚度、实现关联销售、最优化定价、产品直销、产品再设计，以及渠道管理的目的。而这些目标的实现，致使客户管理总体成本降低，客户关系得以改善，最终成功实现零售业务块利润率的提高。

**四：客户价值分析**

根据“二八原则”，找出重点客户，即对银行创造80%价值的20%客户实施最优质的服务，通过对客户金融产品的使用频率以及持续性等指标判断客户的忠诚度。

**五：客户流失预警**

根据客户属性特征、存储款、贷款、金融产品使用等数据，运用数据挖掘技术，找到流失客户的共同特征，从而针对具有相似特征的客户还未流失前，进行有针对性的弥补或者营销活动，从而起到避免客户流失到其他公司的作用，起到稳定本企业客户的作用。