**1、简述商务智能的概念**

**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、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、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

**4、说说你对分类与预测概念的理解**

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。

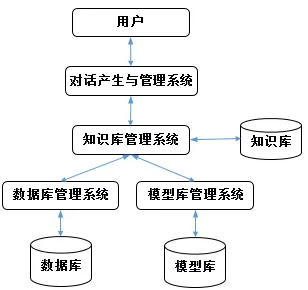
**5、说明分类与聚类概念的不同之处**

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。

**6、说明数据仓库与数据库的不同之处**

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。

**7、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。

**8、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。

**9、商务智能系统的开发有哪些主要阶段？各阶段主要任务是什么？**

**10、商务智能系统成功的关键因素要哪些？**

业务驱动、高层支持、业务人员和IT人员的合作、循序渐进、培训

3.商务智能系统与决策支持系统的关系是什么？

决策支持系统是一种交互式的基于计算机的系统，用于协助决策者使用数据和模型解决非结构化的问题。其用户主要是管理人员和业务分析人员，主要目的是辅助决策者进行科学决策。

决策支持系统由数据库管理系统、模型库管理系统和对话产生与管理系统三部分组成。后来又引入了知识库管理系统。

· DSS中数据库的数据集成功能较弱，而数据仓库技术，具有良好的数据集成、转换等功能

· 决策支持系统的知识库通常是设置好的，知识很少发生变化，知识的类型和范围很窄。商务智能系统则能从大量的数据中发现新颖有用的知识，提供更加灵活的查询和报表功能以及多维分析功能，可以对决策支持系统的知识库进行动态更新

· 商务智能系统与决策支持系统相比，在数据分析和知识发现方面具有更强的功能，但是它只提供对决策有帮助的信息，并不提供可能的决策方案

1.分类的概念？常用的分类方法有哪些？

分类（classification）是通过对具有类别的对象的数据集进行学习，概括其主要特征，构建分类模型，根据该模型预测对象的类别的一种数据挖掘和机器学习技术。

例如，电信公司的客户可以分为两类，一类是忠诚的，一类是流失的。根据这两类客户的个人特征方面的数据以及在公司的消费方面的数据，利用分类技术可以构建分类模型

分类概念：分类任务就是通过学习得到一个目标函数（分类模型）f，把每个属性集x映射到一个预先定义的类标号y。总结已有类别的对象的特点并进而进行未知类别对象的类别预测的过程

分类方法：

决策树 （decision tree）

朴素贝叶斯（Naïve Bayes）

K近邻（K nearest Neighbors）

基于关联的分类

支持向量机（Support Vector Machines ）

人工神经网络

Logistic Regression

数据预处理的主要任务有哪些？

（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-max normalization）

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（data cleaning）

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。