第一章导言

1.什么是商务智能？

数据仓库（或数据集市）、查询报表、数据分析、数据挖掘、数据备份和恢复等部分组成的、以辅助企业决策为目的一类技术及其应用。

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。

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。

商务智能是将数据转换成信息的过程，

然后通过发现将信息转化为知识。

2.商务智能系统的主要组成要素是什么？

六个主要组成部分

**- 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

**- 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。

**- 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。

**- 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

**- 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

**- 业务性能管理**

业务绩效管理（business performance management），简称BPM，又称为企业绩效管理（corporate performance management），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。

第二章商务智能的过程

1.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？

2.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

第三章关联分析

1.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。

2.关联规则的概念及其生成方法。

3.   P30练习题中的第   2，3题。

要求：根据给定 minsup找出所有频繁项集，根据  minconf找出所有关联规则

题型与作业相同，变换数据。

第四章分类

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

2.信息熵的概念及其计算方法。

所谓信息熵，是一个数学上颇为抽象的概念，在这里不妨把信息熵理解成某种特定信息的出现概率。

**计算公式**

数据集D及类别集合C={c1, c2, …, ck}

count(ci)：类别ci在D中出现的次数，

p(ci)：ci在D中出现的相对频率

p(ci)=count(ci)/|D|

|D|代表D中的数据行数

3.   p48页练习题第2题。

第五章数值预测

1.数值预测的概念及其主要预测方法。

**方法：**

**回归分析**

一元线性回归

多元线性回归

非线性回归

**回归树，模型树**

模型树的构建

模型树的剪枝

**K近邻**

2.计算题略。

第六章聚类分析

1.聚类的概念及其与分类的不同？

**聚类分析**是根据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（cluster）。

简单地说，**分类**(Categorization or Classification)就是按照某种标准给对象贴标签(label)，再根据标签来区分归类。

简单地说，聚类是指事先没有“标签”而通过某种成团分析找出事物之间存在聚集性原因的过程。

**区别是，**分类是事先定义好类别，类别数不变 。分类器需要由人工标注的分类训练语料训练得到，属于有指导学习范畴。聚类则没有事先预定的类别，类别数不确定。 聚类不需要人工标注和预先训练分类器，类别在聚类过程中自动生成。分类适合类别或分类体系已经确定的场合，比如按照国图分类法分类图书；聚类则适合不存在分类体系、类别数不确定的场合，一般作为某些应用的前端，比如多文档文摘、搜索引擎结果后聚类(元搜索)等。

分类的目的是学会一个分类函数或分类模型(也常常称作分类器 ),该模型能把数据库中的数据项映射到给定类别中的某一个类中。 要构造分类器，需要有一个训练样本数据集作为输入。训练集由一组数据库记录或元组构成，每个元组是一个由有关字段(又称属性或特征)值组成的特征向量，此外，训练样本还有一个类别标记。一个具体样本的形式可表示为：(v1,v2,...,vn;c)；其中vi表示字段值，c表示类别。分类器的构造方法有统计方法、机器学习方法、神经网络方法等等。

聚类(clustering)是指根据“物以类聚”原理，将本身没有类别的样本聚集成不同的组，这样的一组数据对象的集合叫做簇，并且对每一个这样的簇进行描述的过程。它的目的是使得属于同一个簇的样本之间应该彼此相似，而不同簇的样本应该足够不相似。与分类规则不同，进行聚类前并不知道将要划分成几个组和什么样的组，也不知道根据哪些空间区分规则来定义组。其目的旨在发现空间实体的属性间的函数关系，挖掘的知识用以属性名为变量的数学方程来表示。聚类技术正在蓬勃发展，涉及范围包括数据挖掘、统计学、机器学习、空间数据库技术、生物学以及市场营销等领域，聚类分析已经成为数据挖掘研究领域中一个非常活跃的研究课题。常见的聚类算法包括：K-均值聚类算法、K-中心点聚类算法、CLARANS、 BIRCH、CLIQUE、DBSCAN等。

2.聚类的的主要方法有哪些？

**划分法（Partitioning approach）：**

k均值（k-means）、k中心点（k-medoids）等方法。

**层次法（Hierarchical approach）：**

凝聚层次聚类（agglomerative hierarchical clustering）和分裂层次聚类（divisive  hierarchical clustering）。

Diana、 Agnes、BIRCH、ROCK、CAMELEON等。

**基于密度的方法（Density-based approach）**

DBSCAN、OPTICS和 DenClue等。

**基于模型的方法（Model-based）**

EM、SOM和COBWEB等。

3.   P76页练习题第 8题。

要求：掌握基本概念及其功能

题型：简答题

第七章数据预处理

1.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。

2.掌握最小-最大规范化和z-score规范化方法。

**最小-最大规范化**

假设需要映射到目标区间为[L,R]。原来的取值范围为[l,r]，则根据等比例映射的原理，一个值x映射到新区间后的值v的计算方法如下：

例如，对于描述客户的属性“年收入（万元）”，如果原来的取值范围为[3，200]，新的取值范围为[0，1]，则若某客户的年收入为60万元，规范化后为(60-3)/(200-3)=0.29

**z-score规范化**

z-score，又称零均值规范化（zero-mean     normalization）。给定一个属性A，设其取值的均值为*mA*，标准差为*sA*，A的某个取值*x*规范化后的值*v*计算如下：

- 均值为*mA*和标准差为*sA*通过已有样本的属性值进行计算。规范化后的属性A取值的均值为零

- 例如，年收入属性的均值为82，标准差为39，则年收入60万规范化后为-0.31

3.掌握分箱离散化及基于熵的离散化方法。

**分箱离散化**

等距离分箱、等频率分箱

**等距离（equal- distance）分箱**

- 又称为等宽度分箱（equal-width binning），是将每个取值映射到等大小的区间的方法

- 给定属性A的最小和最大取值分别为min和max，若区间个数为k，则每个区间的间距为I=(max-min)/k，区间分别为[min，min+I)、[min+I,min+2I)、…、[min+(k-1)I, min+kI]

- 等距离分箱可能导致属于某些的取值非常多，而某些又非常少

**等频率（equal-frequency）分箱**

- 称等深度分箱（equal-depth binning）。它将每个取值映射到一个区间，每个区间内包含的取值个数大致相同

**例如：**

- 假设14个客户的属性“年收入”的取值按顺序为：

20，40，50，58，65，80，80，82，86，90，96，105，120，200

- 利用等距离分箱，区间的个数为4，则区间间距为(200-20)/4=45，则4个箱的区间分别为[20，65)，[65，110)，[110，155)，[155，200]

- 利用等频率分箱，每箱3个值，则4个箱分别为[20，40，50]，[58，65，80，80]，[82，86，90]，[96，105，120，200]

**基于熵的离散化**

分箱离散化由于是一种无监督离散化方法

基于熵的离散化方法是常用的有监督的离散化方法

给定一个数据集D及分类属性的取值，即类别集合C={c1,     c2, …, ck}，数据集D的信息熵entropy(D)的计算公式

其中p(ci)=count(ci)/|D|，count(ci)表示类别ci在D中出现的次数，|D|代表D中的数据行数，即对象个数。信息熵的取值越小，类别分布越纯，反之越不纯

首先将D中的行按照属性A的取值进行排序。

分割的方法是利用条件A£v，*v*是A的一个取值。相应地，数据集D按照此条件分裂为两个子数据集：*D*1, *D*2，综合这2个子数据集的信息熵作为衡量这种分割优劣的度量，entropy(D,     v)，

一个数据集D按A£v分裂前后信息熵的差值称为信息增益，记为gain(D,v)

gain(D,v)=entropy(D)-entropy(D,v)

4.   p89页练习题的第 2、3（仅做分箱及基于熵的离散化）题。

要求：数据规范化方法，等深度法和等宽度法数据离散化方法，欧氏距离计算方法

考试内容以作业题型为主，变换数据。

第八章数据仓库

1.数据仓库的基本概念及与业务数据库系统的不同。

数据仓库是一个面向主题的（subject-oriented）、集成的（integrated）、随时间变化的（time-var-ing）、稳定的（non-volatile）用于支持组织决策的数据集合。

2.数据仓库系统的体系结构。

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。

3.多维数据模型的概念及主要的多维数据模型有哪些？

**多维数据模型**是一种从业务分析的角度来对数据进行逻辑建模的方法，具有简单、易于理解、方便查询等优点，因而是一种常用的数据仓库建模方法。又称维度数据模型（dimensional data model），由维度表（dimension table）和事实表（fact     table）两种类型的表构成。

**模型**

**星型模式（star schema）**

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放在这一个表中，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。

**雪片模式（snowflake schema）**

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。

**事实星座（fact constellation）**

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。

第九章在线分析处理

4.在线分析处理的基本概念。

在线分析处理是一类软件技术，利用它可以使分析人员、管理人员以及主管从多种信息视角通过快速、一致和交互的访问数据，达到对数据的洞察。这些视角是从原始数据转换过来的，反映了企业的真实维度，易于被用户理解。

5.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