《商务智能与数据挖掘》教学大纲

课程编号：070663B

课程类型：□通识教育必修课 □通识教育选修课

□学科基础课 □专业核心课

☑专业提升课 □专业拓展课

总学时：48 讲课学时：32 实验（上机）学时：16

学　　分：3

考试类型：□考试 ☑考查

适用对象：计算机科学与技术专业

□是 ☑否 适合作为其他专业学生的个性化选修课

先修课程：数据库原理与应用、概率与数理统计

一、教学目标

《商务智能与数据挖掘》是计算机技术、数据库技术与现代数理统计方法综合应用的课程，为决策支持系统的建设提供有效的信息资源，对研究现代信息管理技术与方法具有重要的作用，而且在各行各业中得到广泛的应用。

目标1：了解数据仓库系统整体结构、相关技术和产品等。

目标2：理解数据仓库和数据挖掘的基本概念和基本方法。

目标3：掌握数据仓库的实现技术与应用、OLAP以及数据挖掘方法与技术。

通过本课程的学习，学生应该能够根据企业的现状与需求，选择合适的技术、工具与实施方案，以适应计算机科学技术新的发展趋势，并为实际应用打下坚实的基础。

思政教学目标

本着“润物细无声”的理念，结合计算机科学与技术专业特点，在《商务智能与数据挖掘》课程教学中融入社会主义核心价值观，加强大学生理想信念、价值取向、政治信仰、社会责任的教育。在课堂教学中挖掘思政元素，引导学生端正学习态度，树立正确的人生观、价值观、世界观；在实验教学环节中引入思政元素，塑造学生的工匠精神、创新精神；通过综合的课程设计训练，培养学生的团队协作、纪律观念以及良好的职业道德修养。通过课程思政目标的实现，使得学生在掌握扎实的专业知识的过程中，能获得良好的科学素养和职业道德修养，完成思政教育、专业教育、能力培养的有机统一。

二、教学内容及其与毕业要求的对应关系

本课程的核心内容分为两大方面：数据仓库和数据挖掘。数据仓库方面包括数据仓库的基本概念，数据立方体，联机分析处理等内容；数据挖掘方面包括数据的预处理技术、关联分析、数据分类分析、数据聚类分析、文本挖掘以及Web挖掘等。为有效学习本课程，要求学生生已具备数据库系统的基本知识和数据库应用系统的开发经验。

根据教学目标，拟采用的教学方法：课堂讲解、案例教学与上机操作相结合。课堂讲解基本概念和核心知识，讲授和讨论相结合领会知识要点，课堂讲解以方法的应用为主线，系统叙述数据仓库和数据挖掘的有关概念和基础知识，使学生尽快掌握数据仓库和数据挖掘的基本概念，基本方法和应用背景；案例教学训练解决问题的能力；上机实践环节主要以Microsoft SQL Server 2012、SPSS Clementine和weka为上机环境进行实践操作，以加深学生对教学内容的理解和掌握。上机实践应掌握数据仓库中数据抽取、转换与装载过程、多维数据集的建立与使用、各种数据挖掘方法与技术。

学生学习时应根据大纲提纲挈领地掌握数据仓库和数据挖掘的内容，随后，还要参考大量的参考文献，通过课后思考和练习题进行多角度和层次的反复学习。

本课程涉及的内容广泛，很多内容需要学生课后学习，以提高学生的自学能力；同时，尽可能做到把理论知识与实际应用相结合，提高学生的动手能力与解决实际问题的能力。

三、各教学环节学时分配

以表格方式表现各章节的学时分配，表格如下：

**教学课时分配**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | 章节内容 | 讲课 | 实验 | 其他 | 合计 |
| 1 | 第一章 商务智能概述 | 2 | 2 |  | 4 |
| 2 | 第二章 数据仓库 | 2 |  |  | 2 |
| 3 | 第三章 联机分析处理 | 2 | 2 |  | 4 |
| 4 | 第四章 Power BI数据可视化 | 2 | 2 |  | 4 |
| 5 | 第五章 数据挖掘概述 | 2 |  |  | 2 |
| 6 | 第六章 数据预处理 | 2 |  |  | 2 |
| 7 | 第七章 关联分析 | 3 | 2 |  | 5 |
| 8 | 第八章 聚类分析 | 5 | 4 |  | 9 |
| 9 | 第九章 分类分析 | 8 | 4 |  | 12 |
| 10 | 第十章 文本与Web挖掘 | 4 |  |  | 4 |
| **合计** |  | **32** | **16** |  | **48** |

四、教学内容

本课程包含商务智能和数据挖掘两大部分内容，在商务智能部分主要介绍商务智能的核心技术，包括数据仓库、联机分析处理以及数据可视化；在数据挖掘部分介绍各种数据分析方法，并基于案例介绍数据挖掘技术的应用。

第一章  商务智能概述概述

第一节 商务智能的产生与发展

1. 商务智能产生的原因

2. 商业决策需要商务智能

3. 企业智能化管理需要商务智能

4. 商务智能的发展

第二节 商务智能的概念

1. 数据、信息与知识

2. 商务智能的定义

3. 商务智能的特点

4. 商务智能的过程

第三节 商务智能基础

1. 商务智能的基本架构

2. 商务智能的功能

第四节 商务智能的关键技术

1. 数据仓库

2. 联机分析处理

3. 数据可视化

4. 数据挖掘

第五节 商务智能的相关应用

1. 商务智能在金融业的应用

2. 商务智能在保险业的应用

3. 商务智能在零售业的应用

4. 商务智能在电子商务领域的应用

5. 商务智能在制造业的应用

教学重点、难点：商务智能的基本架构和商务智能的关键技术是本章的重点。

课程的考核要求：通过本章的学习，使学生

了解：商务智能的产生与发展、商务智能的相关应用。

理解：商务智能的概念和特点。

掌握：商务智能的基本架构、商务智能的关键技术。

课程思政切入点：

从商务智能技术的产生引入当前IT领域新技术的产生与发展，讨论新技术产生的原因，激发学生的创新精神；同时引出我国在IT领域的智慧创造，引导学生认识我们国家取得的成就，培养学生的爱国情怀，激发学生的民族自豪感以及对社会主义核心价值观的认同感

复习思考题：

1. 什么是商务智能？商务智能有哪些特点？
2. 如何构建商务智能？它具备怎样的功能？
3. 商务智能系统架构是什么？它由哪些要素构成？
4. 简述商务智能的关键技术。

第二章  数据仓库

第一节 数据仓库的产生

1. 决策支持技术的发展

2. 数据库技术的发展

3．数据仓库与数据库的比较

4. 分析型数据处理与操作型处理相分离

第二节 数据仓库的定义与特点

1. 数据仓库的定义

2. 数据仓库的基本特性

第三节 与数据仓库相关的概念

1. 外部数据源

2. 抽取转换装载ETL

3. 维度

4. 多维数据集

第四节 数据仓库的数据组织

1. 数据仓库的数据组织结构

2. 数据仓库的数据组织形式

3. 数据仓库的数据追加

第五节 数据仓库的体系结构

1. 数据仓库系统的层次结构

2. 数据仓库系统的构造模式

第六节 数据仓库的各种数据模型

1. 数据仓库的概念模型

2. 中间层逻辑模型

3. 数据仓库的物理模型

4. 数据仓库的粒度模型

教学重点、难点：数据仓库的定义和特点、数据仓库与传统数据库的主要区别、数据的ETL操作、数据仓库的数据组织、数据仓库的概念模型、逻辑模型和粒度模型，其中对数据的ETL操作以及数据仓库的模型设计是本章的难点。

课程的考核要求：通过本章的学习，使学生

了解：数据仓库的发展过程、数据仓库的外部数据源、数据仓库的数据组织形式及数据追加方法、数据仓库的各种数据模型和数据库中各种数据模型的差异。

理解：数据仓库与传统数据库的主要区别、多维数据集、元数据等与数据仓库相关的概念、数据仓库的体系结构、数据仓库的概念模型、逻辑模型、物理模型和粒度模型。

掌握：数据仓库的定义与特性、数据抽取转换装载的概念、作用及工具、数据仓库的数据组织结构、概念模型中的星型结构和雪花结构事实表模型设计、维模型设计。

应用：数据抽取转换装载的过程。

课程思政切入点：

构建数据仓库系统是一项非常复杂的任务，需要各方面团队成员等经过长时间不断沟通完善，不是一个人单枪匹马在短期内就可以完成的，从而引入团结协作精神，引导学生在平时的技术学习和实践中组成研究小组，协同攻关、合作共赢；数据仓库系统的建立过程中会不断遇到困难和挫折，鼓励同学们在学习过程中不畏艰难、协同攻关、及时沟通、合理规划、稳步前行。

复习思考题：

1. 比较数据仓库与数据库的相同点和不同点。
2. 什么是数据仓库？简述数据仓库的基本特征。
3. 数据仓库中数据是如何组织的？简述数据仓库的数据组织结构和组织形式。
4. 在将源数据加载到数据仓库之前需要完成哪些工作？为什么要进行这些工作？
5. 航空公司希望分析在其服务旅客中的常客旅行趋势，可为公司正确定位航空市场中的常客市场，并且希望跟踪不同航线上旅客随季节的变化情况和增长；跟踪在不同航班上所消费的食品和饮料情况，帮助航空公司安排不同航线上的航班和食品供应。请为其设计一个数据仓库的概念模型、逻辑模型和物理数据模型。
6. 第5题中航空公司希望将旅客数据至少保持三年，公司每天有100条航线，共300架次飞行，每架次的旅客平均为100人。每架次的食品种类50种，前后共采购过1000种。食品受季节影响较大，每年的食品价格呈现一种周期性变化。食品的详细数据只需保持一年。请为航空数据仓库设计一个合适的数据粒度模型。

第三章  联机分析处理

第一节 OLAP概述

1. OLAP的发展

2. OLAP的特性

第二节 OLAP与多维分析

1. 多维基本概念

2. 多维分析

3. 维的层次关系与维的类关系

4. OLAP与数据仓库的关系

第三节 OLAP的实施

第四节 多维OLAP与关系OLAP

1. 多维数据库

2. 多维数据库的数据存储

3. 多维数据库与数据仓库

4. MOLAP的创建与功能

5. ROLAP的多维表示方法

6. ROLAP的创建与功能

教学重点、难点： OLAP与多维分析、多维OLAP与关系OLAP。

课程的考核要求：通过本章的学习，使学生

了解：OLAP的发展和特点、对OLAP技术的评价及OLAP实施方法。

理解：OLAP的定义和特性。

掌握：OLAP和OLTP的区别、OLAP的多维分析方法、MOLAP与ROLAP的概念及两者的差异。

应用：OLAP的多维分析方法。

课程思政切入点：

阿里云是学生心目中心仪的互联网公司，由阿里云 OLAP 团队的招聘广告给学生树立一个明确的学习目标，引导学生认真学习专业知识，努力提升自己的专业技能，争取去互联网大厂实现自己的人生价值。

复习思考题：

* 1. 什么是OLAP？OLAP的系统结构有哪几种？不同的数据在进行在线分析时各有什么特点？
  2. OLAP的数据切片是如何实现的？
  3. MOLAP和ROLAP在OLAP的数据存储中各有什么特点？在什么情况下，选择MOLAP？在什么情况下，选择ROLAP？
  4. OLAP中的钻取操作可以用来为哪些决策提供帮助？
  5. 请用一种OLAP工具完成对航空公司数据仓库的数据进行多维分析：总航班数与利润数，飞行时间、旅客数量与消耗食品，旅客服务成本与旅客的机票收入。

第四章  Power BI数据可视化

第一节 Power BI组件与报表设计流程

1. 产品组成

2. 产品应用流程

3. 配置Power BI应用环境

4. 安装Power BI Desktop

5. Excel PowerPivot组件加载方法

第二节 PowerQuery数据清洗

1. PowerQuery数据导入

2. 数据清洗

3. 数据结构调整

4. 多表格数据合并

5. Excel.PowerQuery数据查询应用

第三节 数据分析模型

1. DAX表达式

2. 事实表和维度表

3. 建立表关系方法

4. 日期表作用与编辑

5. 度量值计算

6. 新建计算表

第四节 可视化报表设计

1. 常规可视化报表设计

2. 探索报表交互分析与动态呈现功能

3. 其他可视化图表设计

第五节 Power BI在线报表服务

1. Power BI在线服务

2. 应用Power BI Pro工作区

3. 数据集维护

教学重点、难点：使用Power BI对数据进行可视化。

课程的考核要求： 通过本章的学习，使学生

了解：使用Power BI进行数据分析的流程。

掌握：使用Power BI完成数据清洗、建立数据分析模型、可视化报表设计、在线报表发布等工作。

应用：将Power BI与实际应用数据相结合，按照数据分析流程完成数据导入、清洗、建立数据分析模型、可视化报表设计、在线报表发布。

课程思政切入点：

课程案例和实验素材选择当前比较有影响力、充满正能量的事件数据，完成数据清洗、建立数据分析模型、可视化报表设计、在线报表发布等工作，希望培养青年学生积极向上的学习和工作态度，不负青春韶华。

复习思考题：

1. 简述使用Power BI进行数据分析的流程。
2. 根据给定的数据文件使用Power BI完成数据导入、清洗、建立数据分析模型、可视化报表设计、在线报表发布。

第五章 数据挖掘概述

第一节 数据挖掘技术概述

1．数据挖掘技术的发展

2．数据挖掘的定义

3．数据挖掘与数据仓库的关系

第二节 数据挖掘技术与数据挖掘工具

1．数据挖掘的过程

2．常用的数据挖掘技术

3．常用的数据挖掘工具

第三节 数据挖掘的功能

1．概念描述：定性与对比

2．关联分析

3．分类与预测

4．聚类分析

5．异类分析

6．演化分析

第四节 数据挖掘技术的应用

1．客户的细分应用

2．客户盈利能力分析

3．客户的获取与保持分析

4．市场营销中的应用

5．数据挖掘的其他应用

教学重点、难点：数据挖掘的定义与步骤、数据挖掘的常用方法。

课程的考核要求：通过本章的学习，使学生

了解：数据挖掘技术的产生与发展、数据挖掘的常用工具与技术、数据挖掘的应用领域和发展趋势。

理解：数据挖掘的定义及步骤、数据挖掘的基本任务。

课程思政切入点：

从数据挖掘的产生和发展入手，引出任何一门影响巨大、意义深远的科学技术，其发展过程必然揭示了科学技术发展的基本规律以及影响其发展的主要因素，引导学生努力学生，掌握先进的科学技术，为祖国的发展贡献力量。

复习思考题：

1. 简述KDD与数据挖掘。
2. 简述数据挖据的分类方法。
3. 从数据挖掘与数据库、统计学、机器学习的关系讨论什么是数据挖掘。
4. 给出一个数据挖掘在商务处理中的应用，这种数据挖掘是什么类型的？具有什么功能？能否利用通常的数据查询技术或统计分析来替代？
5. 举例说明在现实中哪些人需要使用数据挖掘技术帮助他的工作？

第六章 数据预处理

第一节 数据清理

1．消除不完整数据

2．消除噪声

3．消除不一致数据

第二节 数据集成

第三节 数据变换

1．最小-最大规格化

2．零-均值规格化

3．小数定标规格化

第四节 数据规约

1．属性归纳

2．记录归纳

第五节 属性概念分层的自动生成

1．离散属性概念分层的自动生成

2．连续属性概念分层的自动生成

教学重点、难点：各种数据预处理的方法，难点为：概念分层的生成。

课程的考核要求：通过本章的学习，使学生

了解：数据预处理的重要性。

理解：数据离散化和概念分层的自动生成方法。

掌握：数据清洗、集成、转换以及数据消减的方法。

课程思政切入点：

数据挖掘项目中最费力的事就是数据获取和预处理，占用项目的时间一般能达到80%，告诫学生做任何事情都需要有足够的耐心和毅力；通过预处理操作的实例让学生理解数据挖掘是一个循序渐进的过程，引导学生欲速则不达，见小利则大事不成。

复习思考题：

1. 数据预处理主要有哪些方法和技术？
2. 假设某属性的值为28、29、16、32、36、31、18、24、26、21、22、34，请采用分箱技术平滑数据消除噪声。分布规则为等深，深度为4，平滑规则为边界平滑。
3. 假设某属性的取值区域为[-400，600]，分别采用最小-最大规格化和小数定标规格化将200规格化至区间[-1，1]。

第七章  关联分析

第一节 关联规则概述

第二节 基本概念

1．项、项集、事务

2．关联规则

3．支持度、可信度

4．频繁项集

第三节 关联规则的分类

第四节 关联规则算法

1．APRIORI算法挖掘频繁项集

2．FP-Growth算法挖掘频繁项集

3. 利用频繁项集生成关联规则

第五节 序列关联规则

1．序列数据集

2．序列合并规则

第六节 关联规则的评价

1．负关联

2．兴趣度

教学重点、难点：Apriori算法及生成关联规则的方法。

课程的考核要求：通过本章的学习，使学生

了解：关联分析的目的、关联规则的评估以及如何判断有用的关联规则。

理解：与关联规则相关的概念、Apriori算法的两个性能瓶颈。

掌握：规则的支持度和可信度计算方法、频繁项集、关联规则挖掘算法的两个步骤； Apriori性质、Apriori算法产生频繁项集的过程、FP-Growth算法产生频繁项集的过程；生成关联规则的过程和方法。

应用：使用Apriori算法和FP-Growth算法对实际数据进行关联分析。

课程思政切入点：

关联分析是一种简单、实用的数据挖掘技术，其中典型的尿布和啤酒故事告诉人们有时候风马牛不相及的事务之间也可能存在关联性，合理地加以利用就有可能产生双赢的结果，并引用英国首相丘吉尔与青霉素的发明者弗莱明的传奇故事让学生体会什么是双赢的人生，如何实现双赢的人生，希望通过正能量的名人效应对青年学生的学习和生活等方方面面都产生深远的影响。

复习思考题：

1. 简述支持度与可信度。
2. 简述APRIORI算法的缺陷。
3. 给定事务数据集如下表所示，假设最小支持度MinSupport=50%，最小可信度MinConf=85%，给出挖掘出的所有强关联规则。

|  |  |
| --- | --- |
| TID | Item set |
| 1 | *b, c, d, e* |
| 2 | *a, c, e* |
| 3 | *a, b, c, e* |
| 4 | *b, c, e, f* |
| 5 | *a, b, c, d, e, f* |

1. 数据库中有如下5个事务，设最小支持度计数为2，计算项集*abc*和*bce*的支持度。如果（*bcd*）是频繁三项集，计算由此获得的全部可能关联规则及其置信度。

|  |  |
| --- | --- |
| TID | 购买项 |
| T1 | {*a*, *d*, *e*} |
| T2 | {*a*, *b*, *c*, *d*} |
| T3 | {*a*, *b*, *d*, *e*} |
| T4 | {*a*, *b*, *c*, *d*, *e*} |
| T5 | {*b*, *c*, *e*} |

第八章  聚类分析

第一节 聚类概述

1．聚类分析的原理

2．聚类的应用

3．聚类算法的评价

第二节 数据的相似性度量

1．类间距

2．点间距

第三节 基于划分的聚类

1．K均值

2．K中心点

第四节 基于层次的聚类

1．凝聚的层次聚类

2．分裂的层次聚类

第五节 基于密度的聚类

1．DBSCAN的相关概念

2．DBSCAN算法

第六节 基于网格的聚类

1．网格聚类

2．STING算法

第七节 基于模型的聚类

1．模型聚类

2．EM算法

教学重点、难点：基于划分的聚类方法、基于层次的聚类方法、基于密度的聚类方法和基于模型聚类的方法，其中密度聚类和模型聚类是本章的难点。

课程的考核要求：通过本章的学习，使学生

了解：聚类分析常用算法分类，了解网格聚类和模型聚类。

理解：聚类分析原理、聚类分析中的相似性度量方法。

掌握：基于划分聚类中k-平均算法和k-中心点算法的原理和过程、基于层次聚类中AGNES算法和DIANA算法的原理和过程、与密度聚类相关的基本概念及DBSCAN算法的原理和过程。

应用：熟练使用几种主要的聚类方法对实际数据进行聚类分析。

课程思政切入点：

用物以类聚、人以群分与聚类分析进行类比，教导学生探究事物的本质，注重事物的主要矛盾，同时也引导学生多跟学习上努力勤奋、思想上积极向上的同学交朋友；用闵科夫斯基距离大师的励志小故事引导学生树立勇于探索的科学精神。

复习思考题：

1. 给定簇C1={*o*1，*o*2，*o*3}，C2={*o*4，*o*5，*o*6}，其中*o*1=（1，1），*o*2=（2，2），*o*3=（3，3），*o*4=（7，7），*o*5=（8，8），*o*6=（9，9）。计算簇C1和C2之间的类平均距离。
2. 用K-均值算法将表中的8个点聚为3个簇，假设第一次跌代选择序号2、5和8当作初始点，请给出聚类后最后的3个簇及簇中心（距离函数为欧基里得距离）。

|  |  |  |
| --- | --- | --- |
| 序号 | 属性1 | 属性2 |
| 1 | 2 | 3 |
| 2 | 2 | 5 |
| 3 | 6 | 4 |
| 4 | 5 | 2 |
| 5 | 1 | 5 |
| 6 | 3 | 4 |
| 7 | 1 | 2 |
| 8 | 4 | 1 |

1. 设*k*=2，初始簇中心点为B和E，根据下表所给的数据使用PAM算法计算用点A替换点B总代价TCBA。

|  |  |  |
| --- | --- | --- |
| 序号 | 属性1 | 属性2 |
| A | 3 | 1 |
| B | 1 | 2 |
| C | 4 | 2 |
| D | 2 | 3 |
| E | 1 | 3 |
| F | 3 | 2 |

1. 假如空间中的10个点，各点之间的距离关系如下图所示，根据给的数据使用DBSCAN算法实现密度聚类（设*n*=10，用户输入*ε*=2，MinPts=4）。

0

1

2

3

4

5

6

7

8

9

10

0

1

2

3

4

5

6

7

8

9

10

A

D

E

F

G

H

I

J

C

B

（1）给出所有的核心对象。

（2）对每个核心对象，写出从该点出发直接密度可达的点。

（3）对每个核心对象，写出从该点出发间接密度可达的点。

（4）写出最后的聚类结果。

第九章  分类分析

第一节 分类的基本概念

1．分类描述

2．分类的步骤

3．分类的常用方法

第二节 基于距离的分类算法

1．基于距离的分类

2．K最近邻居算法

第三节 基于决策树的分类算法

1．决策树概述

2．ID3算法

3．C4.5算法

4．CART算法

5．决策树剪枝

第四节 贝叶斯分类

1．贝叶斯定理

2．朴素贝叶斯分类算法

3．贝叶斯网络

第五节 神经网络

1．神经网络概述

2．神经网络模型

3．后向传播算法

第六节 遗传算法

1．遗传算法的基本原理

2．遗传算法的关键技术

3．遗传算法实例

教学重点、难点：决策树算法原理、常用决策树算法、神经网络BP算法，其中决策树算法和BP算法是本章的难点。

课程的考核要求：通过本章的学习，使学生

了解：决策树的应用、发展趋势以及决策树剪枝的方法；贝叶斯网络的组成与算法、BP神经网络在模式识别中的应用。

理解：分类的基本概念和步骤、决策树的基本知识、贝叶斯定理、神经元模型和神经网络的网络拓扑结构。

掌握：k-近邻分类方法、决策树算法的原理和常用的决策树算法的分类过程、朴素贝叶斯分类过程、神经网络BP算法的原理和过程

应用：熟练使用几种主要的分类方法对实际数据进行分类分析。

课程思政切入点：

由对分类模型的复杂度评估引导学生了解奥卡姆剃刀理论，这也是人们在获取科学知识的历程中最核心、最本质的思想；通过神经网络发展的小故事告诉学生科学技术的发展并不是一帆风顺的，体现了问题、理论和实践的辩证统一关系，引导学生树立不畏艰难，积极探索科学技术的勇气和决心；通过二十世纪伟大的统计学家李奥.布瑞曼的故事引导学生树立不畏艰难、努力奋斗的精神，坚持活到老学到老。

复习思考题：

1． 给定样本数据如下表所示，计算描述属性的信息增益、信息增益率及Gini值，并分别使用ID3、C4.5和CART算法构建决策树。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| U | A | B | C | D | E |
| 1 | Sunny | Hot | High | Weak | No |
| 2 | Sunny | Hot | High | Strong | No |
| 3 | Overcast | Hot | High | Weak | Yes |
| 4 | Rainy | Mild | High | Weak | Yes |
| 5 | Rainy | Cool | Normal | Weak | Yes |
| 6 | Rainy | Cool | Normal | Strong | No |
| 7 | Overcast | Cool | Normal | Strong | Yes |
| 8 | Sunny | Mild | High | Weak | No |

2．根据上表中给定的样本数据，使用朴素贝叶斯算法对样本X（Outlook=Sunny，Temperature=Mild，Humidity=Normal，Wind= strong）分类，判断X是否能打网球。

3．给定训练集为X={(*xi*，*yi*)|*i=*1,2,…,8}，其中每个训练样本*xi*是一个二维特征向量；*yi*∈{+1，-1}为*xi*的类标号，即训练集中的数据样本包含两个类别。现有*x*1=(1,0)T，*x*2=(0,1)T，*x*3=(0,-0.6)T，*x*4=(0,0.4)T，*x*5=(0,2.4)T，*x*6=(0,-1.6)T，*x*7=(-2,0.4)T，*x*8=(1,0.6)T，其中*y*1*=y*2*=y*3*=*+1，*y*4*=y*5*=y*6*= y*7*= y*8*=*-1。对于未知类标号的数据样本*x*=(0.4,0.4)T，利用3-近邻分类方法对*x*进行分类并给出分类结果。

4．假设某汽车销售公司希望对购买不同档次汽车的客户进行分类，分类的依据是客户的年龄、收入和学历。请根据这三个分类标准，用神经网络推断出客户会购买哪种档次的汽车。

5．现有某企业的员工数据库，数据已经概括处理，其中的合计数为对应所给定的部门、职务、年龄和工资数的人数。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 部 门 | 职 务 | 年龄(岁) | 工资(元) | 合计(人) |
| 销售 | 高级管理 | 31~35 | 4600~5000 | 30 |
| 销售 | 低级管理 | 26~30 | 2600~3000 | 40 |
| 销售 | 低级管理 | 31~35 | 3100~3500 | 40 |
| 生产 | 低级管理 | 21~25 | 4600~5000 | 20 |
| 生产 | 高级管理 | 31~35 | 6600~7000 | 5 |
| 生产 | 低级管理 | 26~30 | 4600~5000 | 3 |
| 生产 | 高级管理 | 41~45 | 6600~7000 | 3 |
| 财务 | 高级管理 | 36~40 | 4600~5000 | 10 |
| 财务 | 低级管理 | 31~35 | 4100~4500 | 4 |
| 行政 | 高级管理 | 46~50 | 3600~4000 | 4 |
| 行政 | 低级管理 | 26~30 | 2600~3000 | 6 |

设计一个遗传算法，分析员工的年龄、部门与工资的关系。

6．在超市中的商品价格都是大于等于零的，超市的总经理只关心如何利用送一件免费商品而带来1000元以上的总销售量。讨论如何挖掘这种商业模式。

第十章 文本挖掘与Web挖掘

第一节 文本挖掘技术

1．信息检索系统

2．文本分析和语义网络

3．文本挖掘

第二节 Web挖掘技术

1．Web的特点

2．Web内容挖掘

3．Web结构挖掘

4．Web使用记录的挖掘

5．Web数据挖掘的应用

教学重点、难点：文本挖掘方法、Web内容挖掘、Web使用挖掘、Web结构挖掘，其中Web使用挖掘、Web结构挖掘是本章的难点。

课程的考核要求：通过本章的学习，使学生

了解：文本挖掘的主要技术及一般过程、Web挖掘的意义、Web使用挖掘的常用技术。

理解：文本的预处理方法、文本聚类、文本分类及文本关联分析；Web结构挖掘、内容挖掘和使用挖掘。

掌握：Web页面重要性的评价方法，权威页面和中心页面的判断方法。

课程思政切入点：

讲解中文分词时可以选择一些当前比较有影响力、正能量的事件（如新冠肺炎疫情关于最美逆行者的报道）作为教学案例与教学素材，从而激发学生的正能量。使用爬虫抓取网页数据必须遵守互联网协议，并通过反面典型案例介绍互联网中不遵守道德规范付出的惨痛代价，以此引导学生理解信息世界中遵循道德规范的重要性，没有规矩不成方圆，从而培养学生责任意识、职业道德，增强学生的安全意识和法律意识。

复习思考题：

1. 电子邮件数据库中包含了大量的电子邮件信息，可以将其看成主要包含文本数据的半结构化数据库。如何将其转换成结构化的数据库，以便支持多维检索？例如按照发送者、接收者、主题和时间等检索。从这个数据库中可以挖掘什么信息？

2. 每个科学的学科都有其自身的主题索引分类标准，用于对学科的文档进行分类。试设计一个Web文档分类方法，可用学科的主题索引标准对Web文档进行自动分类。如何利用Web链接信息改进分类的质量？如何利用Web使用信息来改进分类的质量？

3. 由于因特网的动态性和海量存储数据，开发一个基于因特网的数据仓库，以帮助旅游者选择当地旅馆和餐厅。请设计一个能够提供汇总的、局部的和多维信息的数据仓库，来帮助旅游者选择旅馆和餐厅。假设每个旅馆和餐厅都有自己的Web页面，讨论如何使基于Web的旅游数据仓库大众化？如何查询这些页面？用什么方法从这些页面中抽取信息？能否提供一种数据挖掘方法，提供关联信息，如：去市郊旅游景点的旅游者有80%会在中心餐厅就餐一次。

五、考核方式、成绩评定

本课程的考核分为平时考核及期末考核相结合的方式，平时考勤占10%，平时作业占30%，期末考试成绩占60%。平时作业包括课后练习、课堂案例讨论和上机操作实践等，分配比例根据实际任务进行调整。

期末考核建议采用开卷形式，试题内容以考核学生基本理论、基本技能为主，要注重知识应用能力和解决问题能力的考核上，而知识点的记忆和理解要服务于学生能力的提高和巩固，考试题型包括简答题、计算题以及算法分析题等。

六、主要参考书及其他内容

1. 陈晓红，寇纲，刘咏梅．商务智能与数据挖掘．北京：高等教育出版社．2018年9月
2. 张小梅，许桂秋．商务智能方法与应用．北京：人民邮电出版社．2019年5月
3. 蔡晓妍，杨黎斌，张晓婷，李梅．商务智能与数据挖掘(第2版) ．北京：清华大学出版社．2018年3月
4. [赵悦](https://book.jd.com/writer/%E8%B5%B5%E6%82%A6_1.html)，[王忠超](https://book.jd.com/writer/%E7%8E%8B%E5%BF%A0%E8%B6%85_1.html)．Power BI商务智能数据分析．北京：机械工业出版社．2020年11月
5. 陈文伟．数据仓库与数据挖掘教程(第3版）．北京：清华大学出版社．2021年1月
6. 韩家炜，[Micheling Kamber](https://book.jd.com/writer/Micheling%20Kamber_1.html)，裴健著，[范明，孟小峰等](https://book.jd.com/writer/%E6%AE%B5%E7%A3%8A%20%E5%BC%A0%E5%A4%A9%E5%BA%86%E7%AD%89_1.html)译．数据挖掘概念与技术（第三版）．北京：机械工业出版社．2018年12月重印
7. 刘三满，郭小磊著．数据仓库与数据挖掘技术及应用研究．北京：中国原子能出版社．2020年3月
8. [美] [陈封能](https://book.jd.com/writer/%E9%99%88%E5%B0%81%E8%83%BD_1.html)（[Pang-Ning Tan](https://book.jd.com/writer/Pang-Ning%20Tan_1.html)） 著，[范明，范宏建等](https://book.jd.com/writer/%E6%AE%B5%E7%A3%8A%20%E5%BC%A0%E5%A4%A9%E5%BA%86%E7%AD%89_1.html" \t "_blank)译．数据挖掘导论（完整版）．北京：人民邮电出版社．2020年9月
9. 李春葆，[蒋林](http://book.jd.com/writer/蒋林_1.html)，[陈良臣](http://book.jd.com/writer/陈良臣_1.html)，[喻丹丹](http://book.jd.com/writer/喻丹丹_1.html)，[曾平](http://book.jd.com/writer/曾平_1.html)．数据仓库与数据挖掘应用教程．北京：清华大学出版社．2016年9月
10. 王振武，徐慧编著．数据挖掘算法与实现．北京：清华大学出版社．2015年2月
11. 龙军，章成源著．数据仓库与数据挖掘．长沙：中南大学出版社．2018年3月
12. 陈志泊、韩慧、王建新、孙俏、聂耿青．数据仓库和数据挖掘（第二版）．北京：清华大学出版社．2017年11月
13. [张良均](https://book.jd.com/writer/%E5%BC%A0%E8%89%AF%E5%9D%87_1.html)，[王路](https://book.jd.com/writer/%E7%8E%8B%E8%B7%AF_1.html)，[谭立云](https://book.jd.com/writer/%E8%B0%AD%E7%AB%8B%E4%BA%91_1.html)，[苏剑林](https://book.jd.com/writer/%E8%8B%8F%E5%89%91%E6%9E%97_1.html). Python数据分析与挖掘实战. 北京：机械工业出版社，2016年1月
14. Xingdong Wu，Vipin Kumar编著，李文波，吴素研译．数据挖掘十大算法．北京：清华大学出版社．2013年5月
15. [[美] Jiawei Han](https://book.jd.com/writer/%E7%BE%8E%20Jiawei%20Han_1.html)，[Micheling Kamber](https://book.jd.com/writer/Micheling%20Kamber_1.html" \t "_blank)，[Jian Pei](https://book.jd.com/writer/Jian%20Pei_1.html" \t "_blank) 著.  DATA MINING CONCEPTS AND TECHNIQUES (英文版.第3版) ．北京：机械工业出版社．2012年3月
16. [美][Pang-Ning Tan](https://book.jd.com/writer/Pang-Ning%20Tan_1.html)，[Michael](https://book.jd.com/writer/%E8%BF%88%E5%85%8B%E5%B0%94%C2%B7%E6%96%AF%E5%9D%A6%E5%B7%B4%E8%B5%AB%EF%BC%88Michael_1.html" \t "_blank)，[Stein](https://book.jd.com/writer/Stein_1.html" \t "_blank)著. Introduction to Data Mining（英文版·原书第2版）. 北京：机械工业出版社．2019年11月

执笔人：覃爱明 教研室主任：　高静　　　　系教学主任审核签名：