《数据仓库与数据挖掘》教学大纲

课程编号：070223A

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：数据科学与大数据技术专业

□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

《数据仓库与数据挖掘》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。

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Python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 数据仓库概述 | 2 | 2 |  | 4 |
| 2 | 第二章 数据仓库开发模型 | 2 |  |  | 2 |
| 3 | 第三章 联机分析处理 | 2 | 2 |  | 4 |
| 4 | 第四章 数据仓库系统设计 | 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

本课程包含数据仓库和数据挖掘两大部分内容，在数据仓库部分不仅全面介绍数据仓库的基本概念和体系结构，而且详细阐述了数据仓库的实现技术；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  数据仓库概述

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. 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

**课程思政切入点**：

从数据仓库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。播放《厉害了我的国》中相关视频片段、图片和故事，并和学生一起讨论对新技术发展的认识。



用电影中、现实中的事例告诉学生，中国在信息技术行业取得了很大的进步，从追赶者变为领先者。

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、数据仓库的系统结构、数据仓库的数据组织，其中对数据仓库特点的理解是本章的难点。

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。

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。

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。

应用：数据抽取转换装载的过程。

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. 为什么不能依靠传统的业务处理系统进行决策分析？
3. 什么是数据仓库？简述数据仓库的基本特征。
4. 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。
5. 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？
6. 数据仓库的信息使用者与知识挖掘者对数据仓库的应用有何不同？

第二章 数据仓库的开发模型

第一节 数据仓库的各种数据模型

第二节 数据仓库的概念模型

1. 概念模型

2. 规范的数据模型

3. 星型模型

4. 雪花模型

第三节 中间层逻辑模型

1. 事实表模型设计

2. 维模型设计

第四节 物理数据模型

1. 数据仓库物理模型的存储结构

2. 数据仓库物理模型的索引构建

3. 数据仓库物理模型的优化

第五节 元数据模型

1. 元数据的类型与组成

2. 元数据在数据仓库中的作用

3. 元数据的收集、存储、管理与维护

第六节 数据仓库的粒度模型和聚集模型

1. 数据粒度的划分

2. 确定粒度的级别

3. 聚集模型的确定

4. 聚集模型的处理

**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，是一个由数据驱动、以技术支撑并满足应用需求的不断增长和完善的开发过程。数据仓库包含了数据、技术、应用三方面的要求，只有把良好的数据模型、合理的技术和准确的应用设计结合起来才能建立起一个成功的数据仓库系统。

(1) 数据仓库系统的建立需要数据、技术、应用等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。

(2) 数据仓库系统的建立过程中会不断遇到困难和挫折，从而引出及时有效沟通的重要性以及化压力为动力的重要性（当前很多大学生缺乏沟通和抗压能力）。

(3) 从数据仓库系统的建立过程让学生深刻理解解决复杂问题做常用的方法之一是分而治之，从而引导学生对自己的学习生活和将来的职业生涯进行合理规划，稳步前行。

教学重点、难点：数据仓库的概念模型、逻辑模型、元数据模型和粒度模型。

课程的考核要求：通过本章的学习，使学生

了解：数据仓库的各种数据模型以及和数据库中各种数据模型的差异。

理解：数据仓库的概念模型、逻辑模型、物理模型、元数据模型和粒度模型。

掌握：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：概念模型、逻辑模型和物理模型的设计。

复习思考题：

* 1. 在一般的信息管理中采用那些概念模型来描述信息处理的对象？这些概念数据模型是否适合数据仓库的开发环境？
  2. 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。
  3. 要建立第2题中的数据仓库，需要那些元数据？这些元数据在不同的阶段应发挥什么作用？
  4. 第2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

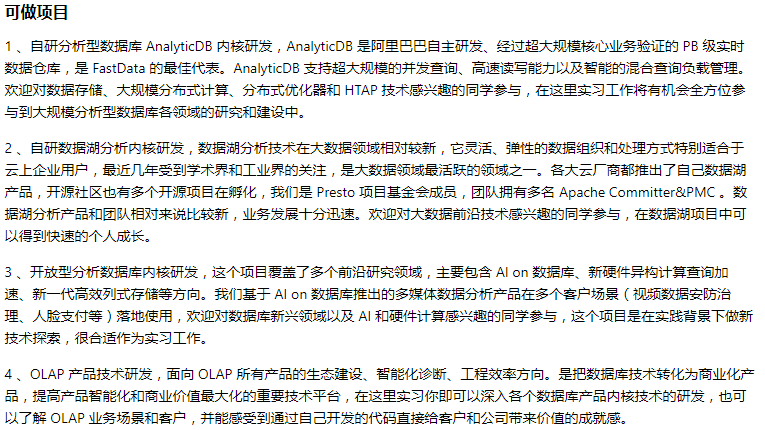
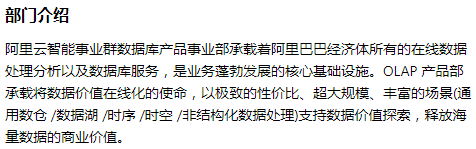
5. ROLAP的多维表示方法

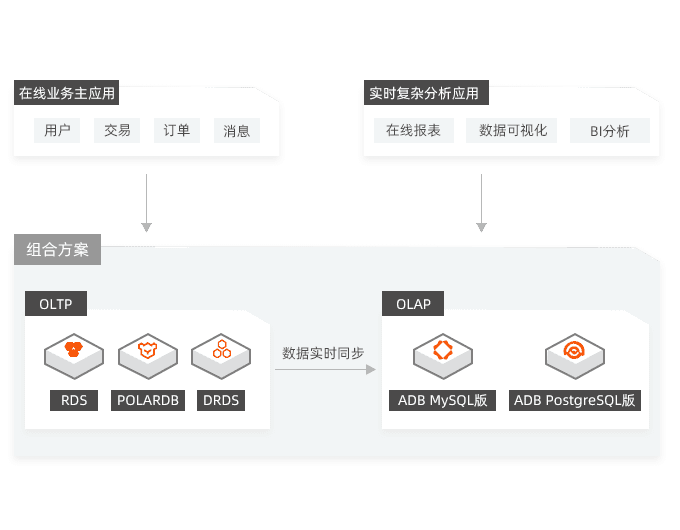
6. ROLAP的创建与功能

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。

**课程思政切入点**：

使用多媒体课件展示阿里云OLAP产品的相关技术与解决方案。







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。

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  数据仓库系统的设计与开发

第一节 SQL Server 2012数据仓库开发工具及应用

第二节 SQL Server 2012的数据仓库创建

1. 创建数据库

2. 创建表

第三节 SQL Server 2012中的数据提取与加载

1. 用SSIS进行数据导入导出

2. 数据转换

第四节 SQL Server 2012中的数据仓库访问与操纵

1. 用SSAS创建维和多维数据集

2. 用Microsoft Excel远程连接并访问多维数据库，进行多维分析。

第五节 数据仓库的开发实例

1. 超市销售数据仓库的规划与分析

2. 用SQL Server 2012创建数据仓库

3. 用SQL Server 2012建立数据仓库的多维数据集

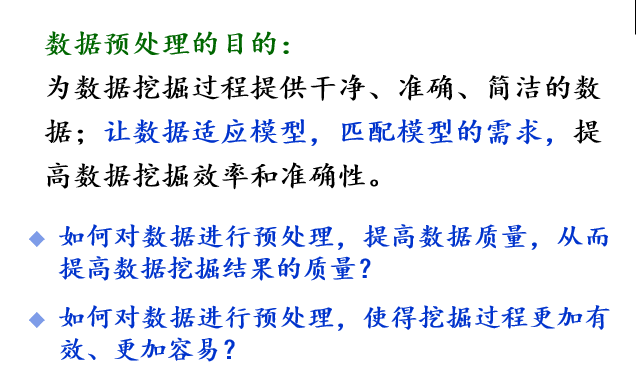
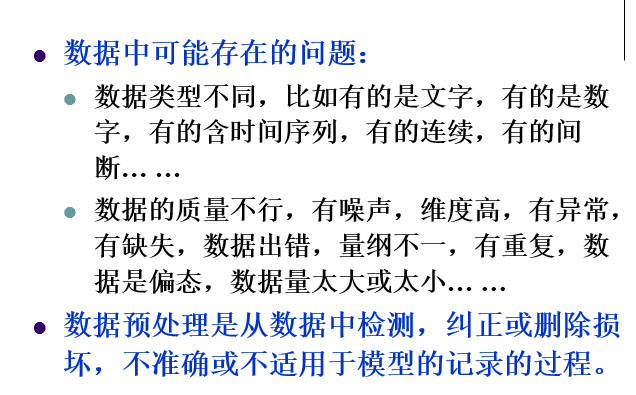
**课程思政切入点**：

(1) 数据挖掘的数据源可以是关系数据库、数据仓库、事务数据库、文本数据库、Web网站等，引导学生在使用数据的时候一定要遵守数据的安全保密性要求，同时警示与启示学生自主知识产权的重要性。

(2) 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%，引导学生一定要一定要有耐心和毅力。

先介绍流行的数据库管理系统（如Oracle、MySQL、SQL Server等）的优点，然后介绍近几年受美国制造的中兴和华为事件。为了不让霸权国家卡脖子，从目前的国际大环境和国家发展的角度考虑，上机环境采用国内与国外相结合的方式，一方面可以让学生了解和掌握当前国际上优秀的数据库产品，增强自己的业务能力；另一方面，必须让学生了解并熟悉我国自主知识产权的数据库产品，如由蚂蚁金服和阿里巴巴完全自主研发的金融级分布式关系数据库OceanBase、华为公司研发的企业级AI-Native分布式数据库GaussDB、由中国人民大学研发的Kingbase ES金仓数据库管理系统等等，不仅让学生们看到国家的进步，也有利于进一步推动基础平台国产化。

通过对一个高维数据集进行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。



教学重点、难点：数据仓库创建、访问与操纵，ETL的设计与实施，本章的难点：如何根据实际情况创建数据仓库、建立数据仓库的多维数据集。

课程的考核要求： 通过本章的学习，使学生

了解：数据仓库开发应用的特点、数据仓库系统建立的步骤及过程。

掌握：利用SQL Server 2012进行数据仓库设计的步骤、进行数据的ETL过程、创建维度和多维数据库以及使用Excel访问多维数据库的过程。

应用：能根据给定的源数据库实现ETL过程、并创建数据仓库和多维数据库。

复习思考题：

1. 如何利用SQL Server的数据仓库创建工具完成数据仓库的创建。
2. 如何利用SQL Server工具创建维和多维数据库？
3. 如何对业务系统中的数据进行清理，筛选数据的条件是销售地区、销售产品的限制。数据仓库在从业务系统中析取数据时，如何利用SQL Server进行数据转换？
4. 使用SQL Server 2012对给定的源数据Foodmart2000.mdb进行数据SSIS和SSAS操作。

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

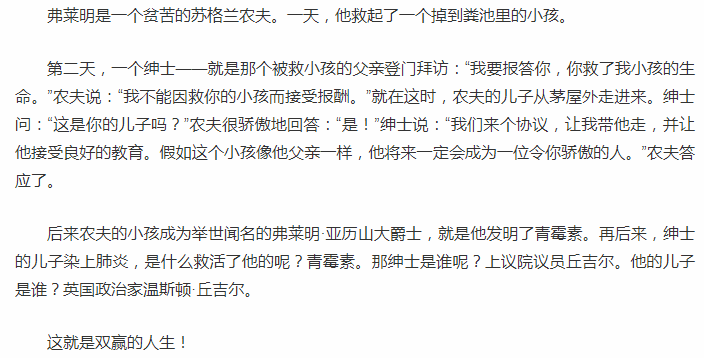
4．市场营销中的应用

5．数据挖掘的其他应用

**课程思政切入点**：

关联分析是一种简单、实用的分析技术，就是发现存在于大量数据集中的关联性或相关性。尿布和啤酒的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。

由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。



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。

复习思考题：

1. 简述KDD与数据挖掘。
2. 简述数据挖掘的分类方法。
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

**课程思政切入点**：

(1) 聚类的概念可以用中国古语“物以类聚、人以群分”（出自《战国策.齐策三》）来形象地解释，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。

(2) 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。

（四）教学手段与方法

(1) 用物以类聚、人以群分与聚类分析进行类比，引出聚类的原理，并使用往届学生的课程成绩进行聚类分析探索；

(2) 讲述闵科夫斯基距离大师的励志小故事。



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。

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

**课程思政切入点**：

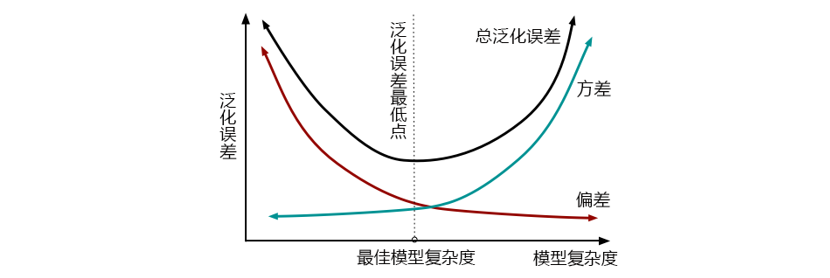
(1) 奥卡姆剃刀理论即简单有效原理。对这个理论的解释，还有一位更著名的科学家和一段更著名的话——“Everything should be made as simple as possible but no simpler”，这句话出自爱因斯坦，意思是做理论的时候一定要尽可能的简单，但又不能过于简单。当模型本身过于复杂时，特征和类别之间的关系中所有的细枝末节都被捕捉，主要的趋势反而在乱花渐欲迷人眼中没有得到应有的重视，这就会导致过拟合（overfitting）的发生。反过来，如果模型过于简单，它不仅没有能力捕捉细微的相关性，甚至连主要趋势本身都没办法抓住，这样的现象就是欠拟合（underfitting）。

(2) 人工神经网络是上世纪人类在科学技术领域所取得的重大科技成果之一，是人类认识自然和利用科技发展道路上的一座重要里程碑。小故事“神经网络的几起几落”告诉人们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。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。

(3) 二十世纪伟大的统计学家李奥.布瑞曼老当益壮的故事。

（四）教学手段与方法

(1) 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。

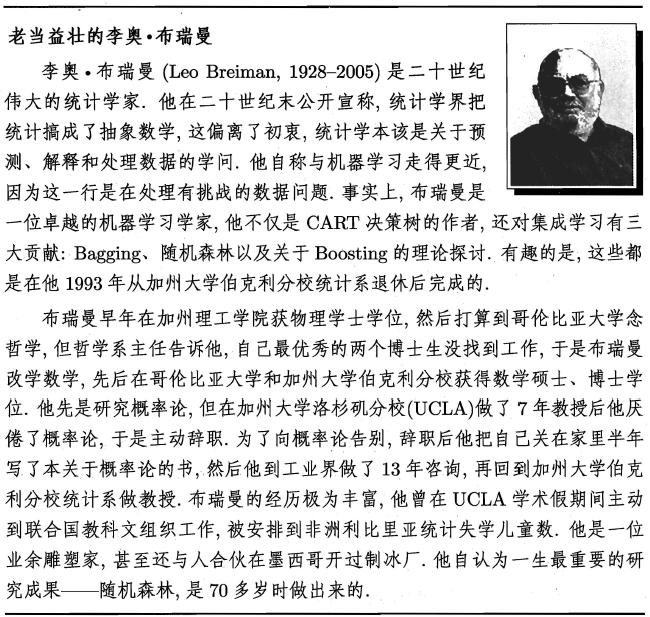


(2) 通过神经网络发展的小故事引导学生树立不畏艰难，积极探索科学技术的勇气和决心。



注：闵斯基于1969年获图灵奖

(3) 通过李奥.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。



[奥卡姆剃刀](https://www.seotest.cn/yunying/23792.html)原理让学生明白在科学研究任务中应该优先使用较为简单的公式或者原理，而不是复杂的，应用到数据挖掘任务中，可以通过减小[模型](https://www.seotest.cn/wenzhang/moxing/)的复杂度来降低过拟合的风险，即模型在能够较好拟合训练集（经验风险）的前提下，尽量减小模型的复杂度（结构风险）。

两个小故事激励学生在以后的学习的工作中要不畏艰难勇攀高峰。

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。

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 基于模型的聚类

1．模型聚类

2．EM算法

**课程思政切入点**：

(1) 中文分词可以选择一些当前比较有影响力、正能量的事件作为教学案例与教学素材，从而激发学生的正能量。

(2) 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。

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。

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。

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。

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，平时成绩占40%，期末考试成绩占60%。平时考核采用课下撰写分析报告、课堂案例讨论和上机操作实践等方式，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

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

1. 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
2. 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
3. [美][Pang-Ning Tan](https://book.jd.com/writer/Pang-Ning%20Tan_1.html)著，[段磊，张天庆等](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（第二版）．北京：机械工业出版社．2019年8月
4. 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
5. 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
6. 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
7. [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
8. 谢邦昌．SQL Server数据挖掘与商业智能基础及案例实战．北京：中国水利水电出版社．2015年8月
9. 张文彤，钟云飞．IBM　SPSS　数据分析与挖掘实战案例精粹．北京：清华大学出版社．2013年2月
10. 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
11. [[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)，[Jian Pei](https://book.jd.com/writer/Jian%20Pei_1.html) 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
12. [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)），[迈克尔·斯坦巴赫（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)，[Stein](https://book.jd.com/writer/Stein_1.html)） 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： 教研室主任：　　　　　系教学主任审核签名：