## 《数据挖掘实务》教学大纲

**一、基本信息**

|  |  |  |  |
| --- | --- | --- | --- |
| **课程编号**： | 3042003 | **课程性质**： | 专业选修课 |
| **课程名称**： | 数据挖掘实务 | **英文名称**： | Data mining practice |
| **学时/学分**： | 32/2 | **开课时间**： | 一（下） |
| **适用对象**： | 应用统计学专业硕士研究生 | | |
| **大纲执笔人**： | 徐雪琪 | **当前版本**： | 2023-08 |

**二、课程概述**

大数据时代不仅意味着大数据的存储和管理，更是意味着可从大数据中获取丰富的知识。**数据挖掘无可争议地成为当今大数据分析获取知识的核心利器**。《数据挖掘实务》课程是应用统计学专业硕士研究生的专业选修课程，以深入理解并掌握数据挖掘的经典方法、了解相关的应用环境、熟练运用相关软件分析实际问题为培养目标。

本课程**坚持以应用为导向**、以方法为基础，通过深入浅出的讲解，使学生了解数据存储环境、分析平台，掌握常用的数据预处理及特征工程方法，深入理解关联挖掘、决策树、集成学习、贝叶斯分析、神经网络、深度学习等常用方法的原理及相应的经典算法，**并能至少熟练使用一种分析工具分析解决实际问题**。通过本课程的学习，同时希望能提升学生数据化思维的能力，使学生具有较强的分析问题和解决问题的能力，**为将来从事大数据分析相关领域的工作打下坚实的基础**。

本课程教学过程中，**力求将思政教育内容以润物细无声的方式渗透到专业授课过程中，以期实现知识传授与价值引领的有机统一**。具体目标如下：一、培养学生的爱党、爱国情怀和家国情怀；二、培养学生的宪法法治意识；三、培养学生奋勇攀登的学习精神和为人民服务意识；四、培养学生的责任意识和职业素养；五、培养学生的创新理念和创新精神。

**三、教学目标**

通过本课程的理论教学和相关实验训练，使学生具备如下能力：

1、掌握数据挖掘的基本思想，熟悉数据挖掘的常用工具。

2、了解数据存储环境、大数据平台。

3、掌握数据预处理及特征工程的常用方法，能够运用R或Python对原始数据实现预处理及特征工程。

4、掌握关联挖掘的基本原理及Apriori、FP-Tree算法，能够运用R或Python对相关数据进行分析。

5、掌握决策树分析的基本原理及ID3、C5.0、CART算法，能够运用R或Python对相关数据进行分析。

6、掌握集成学习的基本原理及随机森林、AdaBoost、Gradient Boosting等相关算法，能够运用R或Python对相关数据进行分析。

7、掌握贝叶斯分析的基本原理及朴素贝叶斯、TAN贝叶斯分类算法，能够运用R或Python对相关数据进行分析。

8、掌握神经网络的基本原理及BP算法，能够运用R或Python对相关数据进行分析。

9、掌握深度学习的基本原理及CNN等算法，能够运用Python对相关数据进行分析。

10、养成良好的数据化思维能力，能够对实际问题运用多种方法进行综合分析。

**四、课程内容**

**第1讲 数据挖掘概述**

**重点内容：**数据挖掘定义，数据挖掘过程、数据挖掘过程模型和数据挖掘工具。

**难点内容：**数据挖掘概念、过程模型、R与数据挖掘、Python与数据分析。

**教学内容：**理解数据挖掘的概念，了解数据挖掘系统的发展历程、数据挖掘当前热点和未来发展趋势、数据挖掘常见功能与使用技术、数据挖掘应用领域等。

**前言：大数据概念、特征、应用现状及发展前景**

1.1数据挖掘的产生与发展

1.2数据挖掘实务过程

1.3数据挖掘功能与使用技术

1.4数据挖掘应用现状与工具

1.5我们可以做什么？我们需要准备什么？

**课程思政教学设计：**结合我国各部门使用大数据进行高效管理的相关案例，培养学生的爱党、爱国情怀和家国情怀；结合大数据时代爬虫技术、隐私保护等问题培养学生的宪法法治意识；结合全国优秀教师代表“教育家精神”2024巡回宣讲事例，引导学生寻找“心中有爱、眼里有光“的职业目标；分析我们可以做什么，我们需要准备什么，以已经毕业的学长为例，培养学生奋勇攀登的学习精神和为人民服务意识。

**第2讲 数据与数据平台**

**重点内容：**数据类型，多维数据模型，大数据平台。

**难点内容：**大数据环境下的数据仓库架构、大数据平台，NoSQL数据库。

**教学内容：**理解数据形态与数据类型、数据环境与数据类型，了解关系型数据库的优缺点、掌握传统数据仓库与多维数据模型，了解NoSQL数据库常见类型及典型代表，了解大数据环境下的数据仓库架构、大数据环境下的数据仓库、大数据平台及其核心组件。

2.1数据类型

2.2关系型数据库

2.3传统数据仓库

2.4 NoSQL数据库

2.5大数据平台

**课程思政教学设计：**结合我国企业数据仓库、数据分析等相关岗位需求分析，帮助学生了解本专业的就业岗位及企业用人标准，规划职业发展。

**第3讲 数据预处理与特征工程实务**

**重点内容：**原始数据中存在的问题，各种数据预处理与特征工程的方法。

**难点内容：**特征工程。

**教学内容：**了解原始数据中存在的问题，掌握空缺值的处理方法、噪声数据的处理方法，明确数据集成时可能存在的问题及解决方法，掌握数据变换的常用方法和数据归约的常用方法，能熟练运用R或Python语言实现数据预处理与特征工程。

3.1数据预处理与特征工程概述

3.2数据清洗理论与实现

3.3数据集成理论与实现

3.4数据变换理论与实现

3.5数据归约理论与实现

**课程思政教学设计：**唯有高质量的数据，才有可能挖掘到有用的知识，结合数据预处理等数据挖掘过程各个阶段的特点、注意事项，培养学生的责任意识和职业素养。

**第4讲 关联挖掘及基于协同过滤的推荐理论与实务**

**重点内容：** Apriori算法，FP-tree算法的建立过程，基于协同过滤的推荐。

**难点内容：**基于协同过滤的推荐。

**教学内容：**掌握关联规则挖掘的基本过程，理解Apriori算法、FP-tree算法、基于协同过滤的推荐算法基本原理，能熟练运用R或Python语言进行相关实务分析。

4.1关联分析概述

4.2 Apriori算法

4.3 FP-tree算法

4.4强关联规则的悖论

4.5基于R语言的应用（Groceries数据集）

4.6思考与拓展：观影数据关联规则挖掘与推荐

4.7基于协同过滤的推荐理论与实务

**课程思政教学设计：**结合大数据时代特征，关联挖掘及协同过滤推荐方法与算法的发展过程，启发与培养学生的创新理念和创新精神。

**第5讲 决策树及集成学习理论与实务**

**重点内容：**决策树，信息增益，信息增益率； ID3算法，C5.0算法，CART算法，集成学习。

**难点内容：**ID3算法，C5.0算法，CART算法，集成学习之基于Bagging和Boosting的相关算法（随机森林、Adaboost、GBDT、Xgboost）。

**教学内容：**理解决策树分析的核心问题，掌握ID3算法、C5.0算法、CART算法及集成学习相关算法基本原理，能熟练运用R或Python语言进行相关实务分析。

5.1决策树概述

5.2 ID3算法

5.3 C5.0算法

5.4 CART算法

5.5基于R语言的应用（credit数据集）

5.6集成学习理论与实务（Python 实现）

**课程思政教学设计：**结合大数据时代特征，决策树及集成学习方法与算法的发展过程，启发与培养学生的创新理念和创新精神。

**第6讲 贝叶斯分类理论与实务**

**重点内容：**贝叶斯定理，贝叶斯信念网络，朴素贝叶斯算法，拉普拉斯平滑，TAN贝叶斯算法。

**难点内容：**TAN贝叶斯算法、文本分类预测实务。

**教学内容：**理解贝叶斯定理与贝叶斯信念网络，掌握朴素贝叶斯分类、TAN贝叶斯分类相关原理与算法实现，能熟练运用R或Python语言进行相关实务分析。

6.1贝叶斯分类概述

6.2朴素贝叶斯分类

6.3 TAN贝叶斯分类

6.4基于R或Python语言实现文本分类（短信数据集用于识别垃圾短信）

**课程思政教学设计：**结合大数据时代特征，贝叶斯分类方法与算法的发展过程，启发与培养学生的创新理念和创新精神。

**第7讲 神经网络及深度学习理论与实务**

**重点内容：**神经元，神经网络，激活函数，梯度下降法，BP算法，深度学习之卷积神经网络等算法。

**难点内容：**梯度下降法，BP算法，卷积神经网络等深度学习算法。

**教学内容：**理解神经元、神经网络及激活函数的作用，掌握BP算法、卷积神经网络基本原理，能熟练运行R或Python语言进行相关实务分析。

7.1神经网络概述

7.2 BP神经网络算法

7.3卷积神经网络等深度学习理论

7.4基于R语言的应用（BP：葡萄酒数据集）

7.5 基于Python的应用（CNN：手写数字或图片识别等）

**课程思政教学设计：**结合大数据时代特征，神经网络及深度学习方法与算法的发展过程，启发与培养学生的创新理念和创新精神。

**五、教学安排**

该课程每周2学时，16周，22学时为课堂授课教学时间，10学时为课内实验教学时间。

建议教学进度如下：

|  |  |
| --- | --- |
| **章节** | **学时数** |
| **第1讲 数据挖掘概述** | **4** |
| **第2讲 数据与数据平台** | **3** |
| **第3讲 数据预处理与特征工程实务** | **3** |
| **第4讲 关联挖掘及基于协同过滤的推荐理论与实务** | **3** |
| **第5讲 决策树及集成学习理论与实务** | **4** |
| **第6讲 贝叶斯分类理论与实务** | **2** |
| **第7讲 神经网络及深度学习理论与实务** | **3** |
| **实验1：关联挖掘相关实验** | **2** |
| **实验2：基于协同过滤的推荐相关实验** | **2** |
| **实验3：决策树与集成学习相关实验** | **2** |
| **实验4：贝叶斯分类相关实验** | **2** |
| **实验5：神经网络与深度学习相关实验** | **2** |

**六、授课方式与手段**

以课堂理论教学为主，实验实践为辅。课堂理论教学以多媒体课件为主。在教学过程中注重数据分析思维能力的培养，以解决实务问题为例，提高理论教学实用性，激发学生学习兴趣，提高学生分析和解决实务问题的能力。同时鼓励学生课后多练习相关工具，提高工具使用能力，多参加相关的竞赛，提升实务处理能力。

**七、考核方式及成绩评定**

**考核方式**：课程论文。

**成绩评定标准：**平时课堂表现(30%) + 学期论文(70%)；其中，平时课堂表现包括：出勤率、课堂讨论、课内回答问题、课内实验练习等。

**八、教材及主要参考书**

|  |
| --- |
| **指定教材：无** |
| **参考书目：** |
| [1] 徐雪琪编著，数据挖掘方法与应用，清华大学出版社，2020。 |
| [2] JiaweiHan, MichelineKamber著，数据挖掘概念与技术（第3版），机械工业出版社，2012。 |
| [3] 周志华著，机器学习，清华大学出版社，2016。 |
| [4] Jeffrey David Ulman, Anand Rajaraman, Jure Leskovec著，斯坦幅数据挖掘教程（第3版），人民邮电出版社，2021。 |
| [5] Mehmed Kantardzic著 ，数据挖掘概念、模型、方法和算法（第3版），清华大学出版社，2021。 |

**教学大纲编写说明**

1. 课程性质：指学位课、必修课和选修课。
2. 适用对象：指适用年级、学科类别、具体专业。

4、修订时间：大纲修订的年月，格式为年-月，如2015-12。

5、修订版本：与培养方案的版本相一致。

6、课程概述：对课程的性质、面向对象、教学目标、教学要求等做概要性描述。

7、教学目标：描述通过本课程的学习，使学生掌握的知识或达到的能力要求。

8、课程内容：明确教学内容（含“课程思政”教学设计）、重点内容和难点内容。

9、教学安排：说明课程总课时、教学周、课堂授课教学时间、课内实验教学时间、实践教学时间及对应的学时数等内容。

10、授课方式与手段：说明本课程采用哪些教学方法与手段。若采用线上教学或线上线下混合式教学，请明确标注。

11、考核方式及成绩评定：说明课程的考核方式、成绩比例及平时的纪律要求等。

14、教材及主要参考书：列出该课程相关的推荐教材和参考文献（包括书目和网络资源），格式为：作者，《书名》（版别），出版社，出版时间。