## 《数据挖掘方法与应用》教学大纲

**一、基本信息**

|  |  |  |  |
| --- | --- | --- | --- |
| **课程代码**： |  | **课程性质**： | 专业选修课、专业核心课 |
| **课程名称**： | 数据挖掘方法与应用 | **英文名称**： | Data Mining Methods and Applications |
| **学时/学分**： | 48/3 | **开课时间**： | 大三（上、下） |
| **适用对象**： | 经济统计学、应用统计学、大数据、数据科学等专业 | | |
| **先修课程**： | 概率论、数理统计学、统计学 | | |
| **大纲执笔人**： |  | **大纲审核人**： |  |
| **修订时间**： |  | **当前版本**： |  |

**二、课程描述**

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。数据挖掘出现于20世纪80年代后期，随着信息化的发展，不断汲取统计学、机器学习、数据库技术、人工智能、模式识别和数据可视化等多学科领域的知识，无可争议地成为当今数据分析获取知识的核心利器。数据挖掘方法与应用课程以学生深入理解并掌握数据挖掘的基本方法、了解相关的应用环境、熟练运用相关软件进行海量数据挖掘为培养目标。

本课程坚持以应用为导向、以方法为基础，要求学生掌握关联挖掘、决策树、贝叶斯分析和神经网络等数据挖掘常用方法，并能至少熟练使用一种数据挖掘分析工具分析解决实际问题。通过本课程的学习，同时希望能培养学生良好的分析问题、解决问题的能力。

**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和基本流程。

2、熟悉数据挖掘常用工具和常用数据源平台。

3、掌握数据预处理的常用方法，能够运用IBM SPSS Modeler或R对实际数据进行预处理。

4、掌握关联挖掘的经典算法，能够运用IBM SPSS Modeler或R对实际数据进行分析。

5、掌握决策树分析的经典算法，能够运用IBM SPSS Modeler或R对实际数据进行分析。

6、掌握贝叶斯分析的经典算法，能够运用IBM SPSS Modeler或R对实际数据进行分析。

7、掌握神经网络的经典算法，能够运用IBM SPSS Modeler或R对实际数据进行分析。

8、养成良好的数据分析思维能力，能够对实际问题运用多种方法进行综合分析。

**四、课程目标对毕业要求的支撑（结合统计学专业培养方案）**

|  |  |  |
| --- | --- | --- |
| **毕业要求** | **指标点** | **课程目标** |
| （二）具备良好的统计思维和应用分析能力 | 3.通过统计工具软件的学习，掌握SAS、SPSS等常用统计软件以及R、Python等计算机语言，提高统计编程计算能力。 | 教学目标1-8 |
| 4.通过大数据方向的专题课程，掌握一定的大数据采集技术、大数据建模、数据挖掘和可视化分析能力。 | 教学目标1-8 |
| （三）具有一定的经济学、管理学的商业管理能力，成为复合型“统计+”人才 | 2.综合利用统计知识和统计工具，面向经济管理问题，具备统计设计、数据采集与处理、统计建模与分析等的基本能力，解决经济管理领域的实际问题。 | 教学目标1-8 |

**五、教学内容**

**第1章 数据挖掘概述 （支撑课程目标1、8）**

**重点内容：**大数据与云计算，数据挖掘定义，数据挖掘过程和数据挖掘过程模型。

**难点内容：**数据挖掘概念和过程模型。

**教学要求：**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、数据挖掘过程模型每一环节的具体要求、数据挖掘常见功能与使用技术、数据挖掘应用领域。

1.1数据挖掘的产生与发展

1.2数据挖掘过程

1.3数据挖掘功能与使用技术

1.4数据挖掘应用

**第2章 数据挖掘工具 （支撑课程目标2）**

**重点内容：**IBM SPSS Modeler，R。

**难点内容：**R与数据挖掘，Python与数据分析。

**教学要求：**了解Weka各运行界面、IBM SPSS Modeler主界面及功能，熟悉R语言可用于数据挖掘的各种软件包，熟悉数据分析中常用的Python库。

2.1 Weka

2.2 IBM SPSS Modeler

2.3 R语言

2.4 Python语言

**第3章 数据与数据平台 （支撑课程目标2）**

**重点内容：**数据类型，多维数据模型。

**难点内容：**大数据平台，NoSQL数据库。

**教学要求：**理解数据环境与数据类型、数据形式与数据类型，掌握数据仓库与多维数据模型，了解大数据平台及其核心组件、NoSQL数据库常见类型及典型代表。

3.1数据类型

3.2关系型数据库

3.3 NoSQL数据库

3.4数据仓库与大数据平台（大数据平台可视教学时间与学生基础，选讲）

**第4章 数据预处理 （支撑课程目标3、8）**

**重点内容：**原始数据中存在的问题，各种数据预处理的方法。

**难点内容：**数据归约的主要内容和常用方法。

**教学要求：**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IBM SPSS Modeler或R语言进行数据预处理。

4.1数据预处理概述

4.2数据清洗

4.3数据集成

4.4数据变换

4.5数据归约

**第5章 关联分析 （支撑课程目标4、8）**

**重点内容：**支持度，置信度，提升度，Apriori算法的建立过程。

**难点内容：**Apriori算法实现。

**教学要求：**掌握关联规则挖掘的基本过程，理解Apriori算法基本原理，能熟练运用IBM SPSS Modeler分别对事实表数据集和事务表数据集进行相关案例分析，能熟练运用R语言进行相关案例分析。

5.1关联分析概述

5.2 Apriori算法

5.3强关联规则的悖论

5.4基于IBM SPSS Modeler的应用

5.5基于R语言的应用

**第6章 决策树 （支撑课程目标5、8）**

**重点内容：**决策树，信息增益，信息增益率，ID3算法，C5.0算法。

**难点内容：**ID3算法，C5.0算法，集成学习之Bagging和Boosting。

**教学要求：**理解决策树分析的核心问题，掌握ID3算法基本原理及C5.0算法的基本原理，能熟练运用IBM SPSS Modeler与R语言分析相关案例数据集。

6.1决策树概述

6.2 ID3算法

6.3 C5.0算法

6.4基于IBM SPSS Modeler的应用

6.5基于R语言的应用

**第7章 贝叶斯分类 （支撑课程目标6、8）**

**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贝叶斯分类过程。

**难点内容：**TAN贝叶斯分类过程。

**教学要求：**理解贝叶斯定理与贝叶斯信念网络，掌握朴素贝叶斯分类原理与算法实现，TAN贝叶斯分类原理与过程，能熟练运行IBM SPSS Modeler或R语言进行贝叶斯分类相关案例分析。

7.1贝叶斯分类概述

7.2朴素贝叶斯分类

7.3 TAN贝叶斯分类

7.4基于IBM SPSS Modeler的应用

7.5基于R语言的应用

**第8章 神经网络 （支撑课程目标7、8）**

**重点内容：**神经元，神经网络，激活函数，梯度下降法，BP算法。

**难点内容：**梯度下降法，BP算法。

**教学要求：**理解神经元、神经网络及激活函数的作用，掌握BP算法基本原理，能熟练运行IBM SPSS Modeler或R语言进行相关案例分析。

8.1神经网络概述

8.2 BP神经网络

8.3卷积神经网络

8.4基于IBM SPSS Modeler的应用

8.5基于R语言的应用

**六、教学安排**

该课程每周3学时，16周，33学时为课堂授课教学时间，15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章 数据挖掘概述** | 5 |
| **第2章 数据挖掘工具** | 3 |
| **第3章** 数据与数据平台 | 3 |
| **第4章 数据预处理** | 4 |
| **第5章 关联分析** | 4 |
| **第6章 决策树** | 5 |
| **第7章 贝叶斯分类** | 4 |
| 第8章 神经网络 | 5 |
| 课内实验 | 15 |

**七、课内实验内容、要求及学时**

**（一）实验内容**

**实验一：数据预处理及关联挖掘（3课时）**

使用IBM SPSS Modeler工具实现各类型数据读入、合并、抽样、填补等预处理；使用IBM SPSS Modeler进行简单事实表与事务表购物数据集关联规则挖掘练习。

**实验二：关联挖掘（3课时）**

使用IBM SPSS Modeler进行事务表观影数据集关联规则挖掘练习，使用R语言进行Groceries数据集关联规则挖掘及可视化练习。

**实验三：决策树（3课时）**

使用IBM SPSS Modeler进行药物数据集分类练习，使用R语言进行credit数据集分类练习。

**实验四：贝叶斯分类（3课时）**

使用IBM SPSS Modeler进行蘑菇数据集蘑菇分类练习，使用R语言进行短信数据集短信分类练习（文本分类）。

**实验五：神经网络（3课时）**

使用IBM SPSS Modeler进行电信客户数据集客户流失、套餐类型预测练习，使用R语言进行白葡萄酒数据集白葡萄酒质量预测练习。

**（二）要求及学时**

本实验需要的设备是台式电脑及IBM SPSS Modeler软件、R软件。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **序号** | **实验名称** | **实验类型** | **实验要求** | **实验时数** |
| 1 | 数据预处理及关联挖掘 | 综合性 | 必做 | 3课时 |
| 2 | 关联挖掘 | 综合性 | 必做 | 3课时 |
| 3 | 决策树 | 综合性 | 必做 | 3课时 |
| 4 | 贝叶斯分类 | 综合性 | 必做 | 3课时 |
| 5 | 神经网络 | 综合性 | 必做 | 3课时 |
| 合 计 |  |  |  | 15课时 |

**八、教学方法与手段**

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，以解决实际问题为例，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际问题的能力。另外，本课程在线视频建议即将完成，学生可在2021-2022学年第2学期开始进行线上自主学习。同时鼓励学生课后多练习相关工具，提高工具使用能力。

**九、考核方式及成绩评定**

**考核方式**：期末考试闭卷，平时：课堂互动，出勤，课堂纪律，课堂上机实验，课后练习与复习，大作业。

**成绩评定标准**：总成绩（百分制）＝平时成绩×30％＋期末考试成绩×70％。

**十、教材及主要参考书**

|  |
| --- |
| **指定教材：** |
| [1]徐雪琪编著，《数据挖掘方法与应用》，清华大学出版社，2020。 |
| **参考书目：** |
| [1]Jiawei Han; Micheline Kamber，《数据挖掘概念与技术》(第3版)，机械工业出版社，2012。  [2] 薛薇编著，《SPSS Modeler数据挖掘方法与应用》(第3版)，电子工业出版社，2020。 |