**课程教学大纲参考模板-专业教育课程**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **课程代码** | **DATA0031132027.01** | **课程性质** | | **专业选修** |
| **课程名称：** | **数据挖掘** | | | |
| **英文名称** | **Data mining** | | | |
| **学时/学分** | **72学时/3学分** | **其中实验/实践学时** | **36学时** | |
| **开课单位** | **数据科学与工程学院** | **适用专业：** |  | |
| **先修课程** | **算法基础、编程、统计方法与机器学习** | | | |
| **大纲撰写人** | **周昉** | **大纲审核人** | **陈宇皓** | |
| **课程网址** | **(没有请填写“无”)** | **授课语言** | **中文** | |

注：课程性质选择下列类别之一：学科基础、大类平台、专业必修、专业选修、教师教育

一、课程说明

数据挖掘旨在从大量的、不完全的、有噪声的、模糊的、随机的数据中提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识。随着大数据时代的到来，人们积累的数据量急剧增长，动辄以TB计，如何从海量的数据中提取有用的知识成为当务之急。数据挖掘就是为顺应这种需要应运而生发展起来的数据处理技术，是知识发现的关键步骤。本课程主要介绍数据清洗与预处理、关联分析、异常检测等算法，并针对一个实际应用需求，了解数据的获取、清洗、特征提取和建模过程。通过本课程的学习，学生能够具备以下能力：1）掌握数据分析、设计和实现的基本方法；2）了解大数据应用中需求分析、数据和应用建模、模型设计、开发和实施的过程；3) 理解数据对于社会经济发展的赋能作用，了解相关应用场景的前沿技术。

二、课程目标

通过对本门课程的学习,使学生系统地获得数据挖掘算法的基本知识和必要的基础理论，认识数据、了解数据预处理、掌握关联规则分析、异常检测等智能挖掘算法,并注重培养学生数据分析能力和熟练的运用各类算法能力，从而使学生学会利用数据挖掘的知识去分析和解决应用中的实际问题。

目标1：掌握数据挖掘中关联规则和异常检测基本模型算法（支持毕业要求6）

目标2：了解异常检测在金融领域中的应用并掌握主要解决方法（支持毕业要求8）

目前3：深刻理解数据的获取、预处理、特征提取和建模过程（支持毕业要求7）

目标4：能够针对应用问题，应用开源软件进行数据处理和分析（支持毕业要求10）

三、课程目标与毕业要求的对应关系

|  |  |  |
| --- | --- | --- |
| **毕业要求** | **指标点** | **课程目标** |
| 专业技能扎实 | 掌握数据分析和机器学习的基本模型和算法 | 目标1 |
| 学科理念先进 | 深刻理解数据的获取、建模、管理、利用的全生命周期，深刻理解数据科学与工程相关技术发展与社会经济发展的关系 | 目标2、3 |
| 深刻理解数据对于社会经济发展的赋能作用，了解金融、物流、零售、制造等领域的典型应用的技术问题并掌握主要解决方法 |
| 工程能力全面 | 了解大数据应用中需求分析、数据和应用建模、系统选型、应用设计、开发和实施的过程，具备合作进行系统和应用研发能力 | 目标4 |

四、教学内容与学时安排

（重难点请标注🞛）

第一章 数据（支撑课程目标3）

学时：4（理论）+ 2（实验）

1.数据的类型

2.数据的质量

3.数据的预处理🞛

要求学生：了解数据的类型和预处理方法

第二章 关联规则（支撑课程目标1、3、4）

学时：6（理论）+ 4（实验）

1.频繁项集的产生

2.规则的产生🞛

3.频繁项集的紧凑表示🞛

4.FP增长算法🞛

要求学生：熟悉关联规则挖掘的频繁集、频繁项集、支持度和可信度的基本概念，理解关联规则的运用方法，掌握关联规则的Apriori算法

第三章：关联规则（支撑课程目标1、3、4）

学时：6（理论）+ 4（实验）

1.处理分类属性和连续属性

2.序列模式🞛

3.子图模式🞛

4.非频繁模式

要求学生：掌握关联规则在序列数据和图数据中的应用

第四章：无监督异常检测(（支撑课程目标1、2、3、4）

学时：6（理论）+ 5（实验）

1.基于距离或密度的无监督异常检测

2.基于隔离的异常检测

3.基于重构误差的异常检测🞛

要求学生：了解无监督异常检测算法以及在实际中的应用

第五章：半监督异常检测（支撑课程目标1、2、3、4）

学时：6（理论）+ 5（实验）

1.基于one-class的异常检测算法🞛

2.基于生成模型的异常检测算法🞛

3.基于距离的半监督异常检测算法🞛

要求学生：掌握半监督异常检测算法以及在实际中的应用

第六章：数据分析（支撑课程目标1、2、3、4）

学时：8（理论）+ 16（实验）

1.实际应用需求

2.数据的清洗

3.特征的提取🞛

4.模型的选择🞛

5.模型的评估

要求学生：了解大数据应用中需求分析、数据清洗、模型设计、开发和实施的过程

实验内容：

通过在公开数据集上完成特定目标任务，使学生逐步掌握使用所学习的数据挖掘方法，使用开源工具，通过深入的研究逐步完善目标任务。

实验1：数据预处理练习

实验2：对序列数据集进行频繁子序列挖掘

实验3：对图数据进行频繁子图挖掘

实验4：任选一种无监督异常检测方法，对数据集进行异常检测

实验5：任选一种半监督异常检测方法，对数据集进行异常检测

数据分析项目：

针对一个实际应用需求，对数据进行清洗、特征提取、特征分析、并根据数据特点调用2个以上的模型，进行模型间的比较与分析。

五、教学方法

本课程采用以讲授法为主的教学手段，在课程中采用提问的方式引导学生共同思考，让学生进行头脑风暴，锻炼学生批判性思考的能力；同时，采用案例教学，让学生理解数据挖掘的原理，并结合实际应用需求，进行实践练习。

六、考核方式

本课程的考核采用考察方式，成绩采用五级评分制，考核方式以及各考核方式所占比例如下：

考勤:10%

关联规则相关作业：20%

异常检测相关作业：20%

数据分析项目和汇报：50%

七、推荐教材和参考资料

教材：

数据挖掘导论，范明等译. 人民邮电出版社，2011年 .

参考书目：

1. 数据仓库与数据挖掘教程, 陈文伟. 清华大学出版社，2006年.

2. 数据仓库与数据挖掘，陈志泊. 清华大学出版社，2009年.

3. 数据挖掘概念与技术，Jiawei Han等. 机械工业出版社, 2012年第3版.

4. Ester M, Kriegel H P, Sander J, et al. “A density-based algorithm for discovering clusters in large spatial databases with noise”. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1996. 96(34): 226-231.

5. Breunig M M, Kriegel H P, Ng R T, et al. “LOF: identifying density-based local outliers” In *Proceedings of the ACM SIGMOD international conference on Management of data*. 2000. pp. 93-104.

6. Liu F T, Ting K M, Zhou Z H. “Isolation forest”. In *Proceddings of IEEE International Conference on Data Mining.* 2008. pp. 413-422.

7. Liu F T, Ting K M, Zhou Z H. “Isolation-based anomaly detection”. In *Proceedings of ACM Transactions on Knowledge Discovery from Data (TKDD)*. 2012. 6(1): 1-39.

8. Audibert J, Michiardi P, Guyard F, et al. “Usad: Unsupervised anomaly detection on multivariate time series”. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020 pp. 3395-3404.

9. Zheng P, Yuan S, Wu X, et al. “One-class adversarial nets for fraud detection”. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 2019. 33(01): 1286-1293.

10. Samet Akcay, Amir Atapour-Abarghouei, and Toby P. Breckon. “GANomaly: Semi-supervised anomaly detection via adversarial training”. In*Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision.* 2018. pp. 622-637.

11. Tian B, Su Q, Yin J. “Anomaly detection by leveraging incomplete anomalous knowledge with anomaly-aware bidirectional gans”. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*. 2022.

12. Guansong Pang, Chunhua Shen, and Anton van den Hengel. “Deep anomaly detection with deviation networks”. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019. pp. 353-362.

13. Ruff L, Vandermeulen R A, Görnitz N, et al. “Deep semi-supervised anomaly detection”.In *Proceedings of The International Conference on Learning Representations.* 2019. pp. 353-362.

## 九、评分标准【请按照本门课程采用的课程考核方式选择下表之一填写】（具体分段可以根据实际情况调整）

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **课程目标** | **评分标准** | | | | |
| **优（A）** | **良（B+）** | **中（B-）** | **及格（C）** | **不合格（F）** |
| 学生能够掌握数据挖掘中关联规则和异常检测基本模型算法，了解异常检测在金融领域中的应用并掌握主要解决方法；深刻理解数据的获取、预处理、特征提取和建模过程；能够针对应用问题，应用开源软件进行数据处理和分析。 | 学生能够全面掌握数据挖掘中关联规则和异常检测基本模型算法，熟悉异常检测在金融领域中的应用并掌握主要解决方法；深刻理解数据的获取、预处理、特征提取和建模过程；针对应用问题，能够灵活应用开源软件进行数据处理和分析。 | 学生能够较全面掌握数据挖掘中关联规则和异常检测基本模型算法，较为熟悉异常检测在金融领域中的应用并较为掌握主要解决方法；理解数据的获取、预处理、特征提取和建模过程；针对应用问题，能够应用开源软件进行数据处理和分析。 | 学生能够了解数据挖掘中关联规则和异常检测基本模型算法，了解异常检测在金融领域中的应用并较为掌握主要解决方法；了解数据的获取、预处理、特征提取和建模过程；针对应用问题，能够选择合适的开源软件进行数据处理和分析。 | 学生能够基本了解数据挖掘中关联规则和异常检测基本模型算法，较为了解异常检测在金融领域中的应用并较为掌握主要解决方法；较为了解数据的获取、预处理、特征提取和建模过程；针对应用问题，可以选择相关的开源软件进行数据处理和分析。 | 学生无法了解数据挖掘中关联规则和异常检测基本模型算法，不了解异常检测在金融领域中的应用并较为掌握主要解决方法；不太了解数据的获取、预处理、特征提取和建模过程；针对应用问题，无法选择相关的开源软件进行数据处理和分析。 |