中国海洋大学本科生课程大纲

课程名称	数据挖掘 Data Mining	课程代码	075103301339
课程属性	专业知识	课时/学分	32/3
课程性质	选修	实践学时	32
责任教师	宋博文	课外学时	96

课程属性：公共基础/通识教育/学科基础/专业知识/工作技能，课程性质：必修、选修

一、 课程介绍

1.课程描述：

在互联网时代，人类的各种行为所产生的信息呈指数级爆炸式增长，如何从海量多样的数据中提取出可以加以利用的有效信息，在过去的20年中越来越受到人们的重视。数据挖掘的产生和发展一直是分析和理解数据的实际需求推动的。本课程针对数学类高年级学生开设，课程包括数据挖掘的若干基本内容：数据的本身的基本分析方法以及统计方法，监督学习以及无监督学习算法的介绍，关联分析，异常检测等。通过课程学习，要求学生掌握数据挖掘的基本思路框架，基本理论和方法，同时能够利用这些理论方法并借助计算机软件对实际问题进行分析建模以及提出解决方案，进而提升对应用数学的理解。

2.设计思路：

本课程引导数学类专业高年级学生通过数据挖据以及机器学习来探讨、理解由实际问题所驱动的数学在理论和应用两方面的发展途径。课程内容的选取基于学生“掌握了线性代数，概率统计，回归分析以及优化的基本内容”。课程内容包括五个模块：数据理解，有监督学习，无监督学习，关联算法以及异常检测。这几个模块基本涵盖了数据处理过程中的基本步骤以及经常用到的算法，能够体现数据挖据的基本特征。

数据理解是数据挖掘实践的第一步也是最重要的部分，对数据正确有效的理解可

以指导后续的建模分析。该部分包括理解数据类型、数据质量、数据预处理、相似性和相异性度量、汇总统计、可视化、多维数据分析等，每个部分都配有具体的案例分析。

机器学习是数据挖据内容里面重要的组成部分，根据所处理问题因变量的性质，机器学习算法可分为有监督学习和无监督学习，其中包括回归模型，决策树，贝叶斯模型，k 临近算法，支持向量机，随机森林算法等，每个部分都会配有具体的案例分析。

关联算法以及异常检测算法被广泛应用在商业模型中，主要内容包括：Apriori 算法，FP 增长算法，高级关联算法概念，正态分布中的离散点检测，以及基于邻近度、密度、聚类的离散点检测等。我们在课程中通过对实际案例的讲解来加深对这些方法的理解。

3. 课程与其他课程的关系：

先修课程：高等代数 I&II、概率论、数理统计、回归分析；后置课程：机器学习。

二、课程目标

本课程目标是为数学类专业高年级学生提供一个数学应用的窗口，引导并培养学生用数学语言和数学思维来描述和解决实际问题的能力，增强沟通能力和团队合作意识。

到课程结束时，学生应能：

（1）对实际问题中的数据的性质进行合理的分析,针对特定问题建立合理模型(有监督模型，无监督模型，关联模型，异常分析模型等)，并理解这些模型的求解算法，对小规模的问题给出书面的计算过程;

（2）提高数学理论分析能力，理解有监督、无监督算法，关联算法以及异常检测算法在对应问题中的应用，同时利用这些理论进行实践;

（3）利用计算机软件（Python、R、MATLAB 等）对所建立的算法模型进行求解、并对结果进行合理分析、提供合理的决策依据；

（4）针对实际问题开展小组研究（包括数据的合理处理、建模、选择合适模型、结果分析等），并通过口头报告或书面研究报告形式提供研究结果；激发同学深入理解数据挖掘所表达的人们处理实际问题时所遵循的理念，提升提出问题并解决问题的能

力。

三、学习要求

要完成所有的课程任务，学生必须：

（1）按时上课,上课认真听讲，积极参与课堂讨论、随堂练习和测试。本课程将包含较多的随堂练习、讨论、小组作业展示等课堂活动，课堂表现和出勤率是成绩考核的组成部分。

（2）按时完成常规练习作业。这些作业要求学生按书面形式提交，只有按时提交作业，才能掌握课程所要求的内容。延期提交作业需要提前得到任课教师的许可。

（3）完成教师布置的一定量的阅读文献和背景资料、案例分析、理论探讨和算法软件应用等作业，其中大部分内容要求以小组合作形式完成。这些作业能加深对课程内容的理解、促进同学间的相互学习、并能引导对某些问题和理论的更深入探讨。

四、参考教材与主要参考书

1、选用教材：

《数据挖掘导论》（完整版），英文书名：Introduction to Data Mining，陈封能（Pang-Ning Tan）、斯坦巴赫（Michael Steinbach）、库玛尔（Vipin Kumar）著，范明、范宏建等译，人民邮电出版社，2011 年 1 月出版。

2、主要参考书：

[1]《统计学习方法》李航著，清华大学出版社，2010 年 3 月出版。

[2]《统计学习基础：数据挖掘、推理和预测》，黑斯蒂（Travor Hastie）等著，范明等译，电子工业出版社，2004 年 01 月出版。

[3] 《模式识别和机器学习》，英文书名：Pattern Recognition and Machine Learning, 毕肖普（Christopher Bishop）著，出版社：Springer，2006 年 08 月出版。

[4]《利用 Python 进行数据分析》,英文书名：Python for Data Analysis，麦肯尼（Wes

Mackinney）著，机械工业出版社，2014 年 01 月出版。

[5] 《互联网大规模数据挖掘与分布式处理》，英文书名：Mining of Massive Datasets，罗家罗曼（Anand Rajaraman）等著，王斌译，人民邮电出版社，2012 年 09 月出版。

五、进度安排

序号	专题	主题	计划课时	主要内容概述	实验实践内容
1	绪论	什么是数据挖掘（Data Mining）	1	数据挖掘要解决的问题，起源和任务	
2	数据	数据基本特征分析	3	数据类型；数据质量；数据预处理；数据相似度	
3	探索数据	数据汇总统计以及可视化	2	数据的汇总统计度量；可视化以及多维数据分析;鸢尾花数据案例分析	运用 Python 进行可视化实践
4	有监督学习	有监督学习以及回归分析	2	机器学习分类；基本线性回归分析	
5	分类	基本概念、决策树与模型分析	4	分类问题的一般方法；决策树模型；模型过度拟合；分类器性能评估；分类器比较方法	
		其他技术	10	基于规则分类器；最近邻分类器；贝叶斯分类器；人工神经网络算法；支持向量机算法；随机森林算法；不平衡问题以及多类问题分类	
6	务监督学习	主成分分析介绍	2	主成分分析算法及其应用	
7	聚类分析	聚类分析算法及其评估	2	K 均值算法；凝聚层次聚类；DBSCAN；簇评估	
		聚类分析方法介绍及其选取	2	基于原型、密度、图的聚类；可伸缩的聚类算法；如何选取聚类方法	
8	关联分析	关联分析：基本概念和算法	2	什么是关联分析；频繁项集相关介绍；如何评估关联模式	

9	异常检测	什么是异常检测	2	异常检测成因；统计检测方法；离群点检测	

六、成绩评定

（一）考核方式 A ：A.闭卷考试 B.开卷考试 C.论文 D.考查 E.其他

（二）成绩综合评分体系：

成绩综合评分体系	比例%
1.课下作业、课堂讨论及平时表现	30
2.平时测验成绩	20
3.期末考试成绩	50
总计	100

附：作业和平时表现评分标准

1）作业的评分标准

作业的评分标准	得分
1.严格按照作业要求并及时完成，基本概念清晰，解决问题的方案正确、合理，能提出不同的解决问题方案。	90-100 分
2.基本按照作业要求并及时完成，基本概念基本清晰，解决问题的方案基本正确、基本合理。	70-80 分
3.不能按照作业要求，未及时完成，基本概念不清晰，解决问题的方案基本不正确、基本不合理。	40-60 分
4.不能按照作业要求，未及时完成，基本概念不清晰，不能制定正确和合理解决问题的方案。	0-30 分

2）课堂讨论及平时表现评分标准

课堂讨论、平常表现评分标准	得分
1.资料的查阅、知识熟练运用，积极参与讨论、能阐明自己的观点和想法，能与其他同学合作、交流，共同解决问题。	90-100 分
2.基本做到资料的查阅、知识的运用，能参与讨论、能阐明自己的观点和想法，能与其他其他同学合作、交流，共同解决问题。	70-80 分
3.做到一些资料的查阅和知识的运用，参与讨论一般、不能阐明自己的观点和想法，与其他同学合作、交流，共同解决问题的能力态度一般。	40-60 分

4.不能做到资料的查阅和知识的运用，不积极参与讨论，不能与其他同学合作、交流，共同解决问题。	0-30 分

七、学术诚信

学习成果不能造假，如考试作弊、盗取他人学习成果、一份报告用于不同的课程等，均属造假行为。他人的想法、说法和意见如不注明出处按盗用论处。本课程如有发现上述不良行为，将按学校有关规定取消本课程的学习成绩。

八、大纲审核

教学院长：　　院学术委员会签章：