数据科学与大数据分析

课题大纲

01
课程背景及简介

适合人群：本课程属于通识课程，适合对数据科学、统计学、计算机科学和机器学习感兴趣的同学。

联合国胜任力：不断学习，创新

数据是对事物运行规律的刻画和衡量。对数据的深入理解和运用有助于我们认识世界和改造世界。本课程的开设目的是使学生掌握数据科学所涵盖的各领域的专业知识、技能和相关算法。

本课程将会系统性地讲解数据科学的来源及其研究的重要性，将结合各个领域丰富案例，讲解数据科学的应用技巧。通过本课程，学生将了解数据科学、其应用领域和发展趋势，掌握大数据时代背景下数据密集型科学研究方法，熟悉完整的数据科学工程流程和未来数据科学所面临的发展和挑战。在面对未来数据爆炸的背景下，着重帮助学生运用崭新的思维模式，应用数据科学体系性知识，结合数据科学各种相关技巧和工具，提升解决实际工程问题的能力。

02
学习目标

1. 知悉和理解数据科学的基本概念和范畴
2. 理解数据科学研究的重要性
3. 知悉和理解数据预处理、分析和可视化的过程和原理
4. 利用数据预处理理论及方法，解决数据清洗、去冗余、一致性等问题

03
导师信息

杨教授

北京理工大学计算机学院副教授，硕士生导师。科技部脑科学重大项目论证专家组成员，中国疫苗行业协会疫苗与预防接种大数据应用专委会常委，中国社区卫生协会信息化与网络安全专委会常委，中华预防医学会健康大数据与人工智能应用专委会委员。研究方向包含人工智能、数据科学、并行计算。

主持或参与包括国家“十四五”首批科技创新 2030——“脑科学与类脑研究”重点项目、国家自然科学基金重大研究计划集成项目、国家自然科学基金青年项目、北京市自然科学基金面上项目、国家发改委重点项目、国家核高基项目在内的多项国家及省部级纵向项目。

发表论文 80 余篇，以第一/通讯作者被 SCI 检索近 40 余篇，申请或授权国家发明专利近 20 余项，参与撰写的两项团体标准《大型人群队列研究数据处理规范 T/CPMA 001-2018》与《大型人群队列研究数据安全规范 T/CPMA 002-2018》获得中国标准创新贡献奖标准项目奖提名。

04
课程设置

模块 1：数据科学概述

学习目标：

介绍数据的概念，大数据的概念和定义，大数据的典型特征，数据的发展历史，数据科学和大数据分析的重要性。

模块 2：大数据预处理的方法

学习目标：

数据预处理主要包括数据清洗（Data Cleaning）、数据集成（Data Integration）、数据转换（Data Transformation）和数据消减（Data Reduction）。

模块 3：大数据的并行处理和可视化技术

学习目标：

在大数据分析的应用过程中，可视化通过交互式视觉表现的方式来帮助人们探索和理解复杂的数据。大规模数据的可视化应用主要是基于并行算法设计的技术，合理利用有限的计算资源，高效地处理和分析特定数据集的特性。

模块 4：大数据分析的常用算法

学习目标：

越来越多的应用涉及到大数据，而这些大数据的属性，包括数量、速度和多样性等都是呈现了大数据不断增长的复杂性，此时大数据的分析算法能力就尤为重要，包括：回归、聚类等。

模块 5：数据科学和大数据分析的经典案例

学习目标：

通过案例分析讲解如何运用数据科学相关知识和技术来开展相关分析、趋势预测、决策支持和模式创新。

05
延伸阅读

1. 《数据科学导论》3.0 版本，杨旭、丁刚毅主编，北京理工大学出版社
2. Doing Data Science, written by Cathy O’Neill, published by O’Reilly Media, Inc.
3. Data Science, written by Sandya Mannarswamy, published by Apress.