南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

课程详述
COURSE SPECIFICATION

以下课程信息可能根据实际授课需要或在课程检讨之后产生变动。如对课程有任何疑问，请联系授课教师。

The course information as follows may be subject to change, either during the session because of unforeseen circumstances, or following review of the course at the end of the session. Queries about the course should be directed to the course instructor.

1.	**课程名称 Course Title**	大数据分析与实战 Big Data Analysis and Application
2.	**授课院系 Originating Department**	信息系统与管理工程系 Department of Information Systems & Management Engineering
3.	**课程编号 Course Code**	MIS 301
4.	**课程学分 Credit Value**	3
5.	**课程类别 Course Type**	专业核心课 Major Core Course
6.	**授课学期 Semester**	春季 Spring
7.	**授课语言 Teaching Language**	English
8.	**授课教师、所属学系、联系方式（如属团队授课，请列明其他授课教师） Instructor(s), Affiliation& Contact （For team teaching, please list all instructors）**	何翘楚，信息系统与管理工程系, heqc@sustech.edu.cn Qiaochu He, Department of Information Systems & Management Engineering Email: heqc@sustech.edu.cn
9.	**实验员/助教、所属学系、联系方式 Tutor/TA(s), Contact**	待公布 To be announced
10.	**选课人数限额(可不填) Maximum Enrolment （Optional）**	

11. 授课方式 Delivery Method	讲授 Lectures	习题/辅导/讨论 Tutorials	实验/实习 Lab/Practical	其它(请具体注明) Other（Please specify）	总学时 Total
学时数 Credit Hours	32	0	32	0	64

12. 先修课程、其它学习要求 Pre-requisites or Other Academic Requirements	MA 212 概率论与数理统计
13. 后续课程、其它学习规划 Courses for which this course is a pre-requisite	MIS 310 社交网络模型及应用
14. 其它要求修读本课程的学系 Cross-listing Dept.	无 None

教学大纲及教学日历 SYLLABUS

15. 教学目标 Course Objectives

大数据分析是数据科学与人工智能具体运营的关键，其涵盖的领域广泛，包括概率统计、最优化与运筹学、机器学习、深度学习等，是未来学术研究与产学合作的必备技术之一。数据分析技术是分析和处理大数据的手段和方法，当今大数据作为信息的重要载体在信息化社会扮演着重要的角色。通过研究有关信息获取、信息传输、信息处理与信息控制等核心基础，掌握运用现代计算机工具高效求解科学与工程问题的数学理论与方法。本课程将从实际案例切入，逐步引入大数据分析关键概念。

Big data analysis is the key to the specific operations of data science and artificial intelligence. It covers a wide range of fields, including probability statistics, optimization and operations research, machine learning, deep learning, etc. It is one of the necessary technologies for future academic research and industry-university cooperation One. Data analysis technology is a means and method for analyzing and processing big data. Today, big data, as an important carrier of information, plays an important role in the information society. By studying the core foundations of information acquisition, information transmission, information processing and information control, master the mathematical theories and methods for efficiently solving scientific and engineering problems with modern computer tools.

16. 预达学习成果 Learning Outcomes

基础理论方面，本课程介绍数据驱动的随机模型，并强调其在商业中的应用。这门课程的目标是让你对随机模型的理论和最重要的类别有一个熟悉和直观的理解：对应用随机过程所需的数学基础有很好的理解；熟悉并了解最重要的一类随机过程的性质，特别是泊松过程、离散和连续时间马尔可夫链、排队理论和服务操作应用；知道如何在商业中应用随机过程建模和解决问题。您将进一步发展将所学理论应用于各种现实世界问题的能力，并欣赏随机模型作为一个强大的工具，适用于多个业务领域。

在实践部分通过介绍不同的编程、统计及计量经济学分析模型，并结合 R 和 Python 的实际操作，深入讲解如何通过数据库、Stata、python 进行大数据分析，以此帮助学生更加深入的了解大数据分析，并提升其实际操作能力。

Methodological-wise, this course introduces stochastic models and puts an emphasis on their applications to business. The objective of this course is that you obtain familiarity and an intuitive understanding for the theory and most important classes of stochastic processes. In particular: (1) have a good understanding of the mathematical foundations needed to apply stochastic processes;(2) are familiar with and know the properties of the most important classes of stochastic processes, in particular Poisson processes, discrete- and continuous-time Markov chains, queueing theory and service operations applications; (3) know how to apply stochastic processes to model and solve problems in Business. You will further develop the ability to apply the learned theory to various real world problems and appreciate

stochastic processes as a powerful tool applicable to several business domains.

This course will start from actual cases and gradually introduce the key concepts of big data analysis. In the practice part, by introducing different programming, statistics and econometric analysis models, combined with the actual operation of Stata and Python, in-depth explanation of how to conduct big data analysis through databases, Stata, and python, so as to help students have a deeper understanding of big data Analyze and improve its practical operation capabilities.

17. 课程内容及教学日历 （如授课语言以英文为主，则课程内容介绍可以用英文；如团队教学或模块教学，教学日历须注明主讲人）
Course Contents (in Parts/Chapters/Sections/Weeks. Please notify name of instructor for course section(s), if this is a team teaching or module course.)

下面的大纲是本课程的初步计划，可根据实际情况修改。

讲座及实验(64 小时)

讲座(32 小时)

模块 1 随机模型的基本理论

授课:概率论 I 课程描述与复习(第一周（2022/2/15）2 小时)

课程描述/概率空间/随机变量/期望值/矩生成函数/特征函数/拉普拉斯变换

讲座:概率论复习(第二周（2022/2/22） 2 小时)

条件期望/指数分布和缺乏记忆性质/危险率函数/概率不等式/极限定理

讲座:随机过程概论(第三周（2022/3/1） 2 小时)

定义随机过程/平稳性/遍历性/例子和应用

讲座:泊松过程(第四周（2022/3/8） 2 小时)

泊松过程的定义/到达间隔和等待时间分布/到达时间的条件分布/ M/G/1 队列和业务应用

讲座:更新理论(第五周（2022/3/15）2 小时)

介绍和初步/沃尔德方程和极限定理/更新定理和应用/交替更新过程

讲座:离散时间马尔可夫链(第六周（2022/3/22）2 小时)

介绍和例子/ Chapman-Kolmogorov 方程和状态分类/极限定理

讲座:连续时间马尔可夫链(第七周（2022/3/29）2 小时)

介绍连续时间马尔可夫链/生灭过程/柯尔莫哥洛夫微分方程/极限概率

单元 2:数据驱动的随机模型、大数据应用与实践

讲座:数据驱动的随机模型(第八周（2022/4/5）2 小时)

介绍数据驱动的随机模型/报童模型/布朗运动

讲座:统计学学习概论(第九周（2022/4/12）2 小时)

介绍统计学概论/常见的推断方法/p-value

讲座:大数据驱动的统计与仿真训练营(第十周（2022/4/19）2 小时)

介绍大数据驱动的统计与仿真/ANOVA/联合分析

讲座:大数据领域的机器学习(第十一周（2022/4/26）2 小时)

介绍大数据领域的机器学习/LDA/SVM/深度学习

讲课内容:规定分析中的随机方法(第十二周（2022/5/3）2 小时)

介绍规定分析/预测性分析/仿真/优化

讲座:数据驱动强化学习(第十三周（2022/5/10）2 小时)

介绍数据驱动强化学习/贝尔曼方程/DQN

讲座:端到端决策分析框架(第十四周（2022/5/17）2 小时)

介绍端到端分析框架/深度学习/CNN

讲座:大数据前瞻 1(第十五周（2022/5/24）2 小时)

介绍大数据前沿/行为分析

讲座:大数据前瞻 2(第十六周（2022/5/31）2 小时)

介绍大数据前沿/商业大数据

实验室(32 小时)

单元 3:实验室和项目

基本教程，R(第一周（2022/2/17）2 小时)，

基本教程，Python(第二周（2022/2/24）2 小时)

实现技术:实验室:Hadoop 简介，Spark(第三周（2022/3/3）2 小时)

实验:统计/模拟/机器学习(第四周（2022/3/10）2 小时)

实验:机器学习在 R(第五周（2022/3/17）2 小时)

实验:实际操作 Python+Gurobi(第六周（2022/3/24）2 小时)

实验:Python 示例:Dijkstra(第七周（2022/3/31）2 小时)

实验:Python 示例:GA for TSP(第八周（2022/4/7）2 小时)

实验:Python 示例:Q-learning for TSP(第九周（2022/4/14）2 小时)

实验室:Deep Q-network(第十周（2022/4/21）2 小时)

案例分析:Uber 数据分析(第十一周（2022/4/28）2 小时)

案例分析:摩拜单车数据分析(第十二周（2022/5/5）2 小时)

案例分析:Amazon 数据分析(第十三周（2022/5/12）2 小时)

项目 1:阿里巴巴数据挑战赛(第十四周（2022/5/19）2 小时)

项目二:网易音乐数据挑战赛(第十五周（2022/5/26）2 小时)

项目三:超市展示数据挑战赛(第十六周（2022/6/2）2 小时)

The outline below represents a tentative roadmap for the course. We may deviate from it depending on interest and time.

Lecture and Lab (64 hours)

Lecture (32 hours)

Module 1 Fundamental Theory in Stochastic Models

Lecture: Course description and review of probability theory I (Week 1(2022/2/15)2 hours)

Course description / probability space / random variables / expected value / moment generating function / characteristic function / Laplace transform

Lecture: Review of probability II (Week 2(2022/2/22)2 hours)

Conditional expectation / exponential distribution and lack of memory property / hazard rate functions / probabilistic inequalities / limit theorems

Lecture: Introduction to stochastic processes (Week 3(2022/3/1)2 hours)

Definition of stochastic processes / stationarity / ergodicity / examples and applications

Lecture: Poisson process (Week 4(2022/3/8)2 hours)

Definition of Poisson processes / interarrival and waiting time distributions / conditional distribution of arrival times / the M/G/1 queue and business applications

Lecture: Renewal theory (Week 5(2022/3/15)2 hours)

Introduction and preliminaries / Wald's equation and limit theorems / the renewal theorem and applications / alternating renewal processes

Lecture: Discrete-time Markov chains (Week 6(2022/3/22)2 hours)

Introduction and examples / Chapman-Kolmogorov equations and classification of states / limit theorems

Lecture: Continuous-time Markov chains (Week 7(2022/3/29)2 hours)

Introduction to continuous-time Markov chains / birth and death processes / Kolmogorov differential equations / limiting probabilities

Module 2: Data-driven stochastic models, big data applications and practices

Lecture: Data-Driven Stochastic Models (Week 8 (2022/4/5) 2 hours)

Introduction to Data-Driven Stochastic Models / Newsboy Models / Brownian Motion

Lecture: Introduction to Statistical Learning (Week 9 (2022/4/12) 2 hours)

Introduction to Statistics / Common Inference Methods / p-value

Lecture: Big Data-Driven Statistics and Simulation Training Camp (Week 10 (2022/4/19) 2 hours)

Introduction to Big Data-Driven Statistics and Simulation/ANOVA/Joint Analysis

Lecture: Machine Learning in Big Data Field (Week 11 (2022/4/26) 2 hours)

Introduce machine learning/LDA/SVM/deep learning in the field of big data

Lecture Content: Stochastic Methods in Prescriptive Analysis (Week 12 (2022/5/3) 2 hours)

Introduction to Prescriptive Analysis / Predictive Analysis / Simulation / Optimization

Lecture: Data-Driven Reinforcement Learning (Week 13 (2022/5/10) 2 hours)

Introduction to Data-Driven Reinforcement Learning/Bellman Equation/DQN

Lecture: An End-to-End Decision Analysis Framework (Week 14 (2022/5/17) 2 hours)

Introduction to end-to-end analysis frameworks/deep learning/CNN

Lecture: Big Data Foresight 1 (Week 15 (2022/5/24) 2 hours)

Introduction to Big Data Frontiers/Behavioral Analytics

Lecture: Big Data Foresight 2 (Week 16 (2022/5/31) 2 hours)

Introduction to Big Data Frontiers/Business Big Data

Lab (32 hours)

Module 2: Data-Driven Stochastic Models, Big Data Applications and Practice

Lecture: Data-driven Stochastic Models (2 hours)

Lecture: Intro to statistical learning (2 hours)

Lecture: Big-Data Driven Statistics and Simulation Bootcamp (2 hours)

Lecture: Machine Learning in Big-Data Area (2 hours)

Lecture: Stochastic Method in Prescriptive Analytics (2 hours)

Lecture: Data-driven Reinforcement Learning (2 hours)

Lecture: the Framework of End-to-End Decision Analytics (2 hours)

Module 3: Lab and Projects

Basic tutorials, R (2 hours),

Basic tutorials, Python (2 hours)

Enabling Technology: Lab: Introduction to Hadoop、Spark (2 hours)

Lab: Statistics/Simulation/Machine learning EX (2 hours)

Lab: Maching learning in R (2 hours)

Lab: Hands-on with Python+Gurobi (2 hours)

Lab: Python example: Dijkstra (2 hours)

Lab: Python example: GA for TSP (2 hours)

Lab: Python example: Q-learning for TSP (2 hours)

Lab: Deep Q-network (2 hours)

Case Studies: Uber data analysis (2 hours)

Case Studies: Mobike data analysis (2 hours)

Case Studies: Amazon data analysis (2 hours)

Project 1: Alibaba data challenge (2 hours)

Project 2: NetEase music data challenge (2 hours)

Project 3: Supermarket display data challenge (2 hours)

18. **教材及其它参考资料 Textbook and Supplementary Readings**

Ross, Sheldon M. "Introduction to Probability Models." (2010).

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. "An Introduction to Statistical Learning with Applications in R." (2017).

课程评估 ASSESSMENT

19.

评估形式 Type of Assessment	评估时间 Time	占考试总成绩百分比 % of final score	违纪处罚 Penalty	备注 Notes
出勤 Attendance		10		
课堂表现 Participation		10		根据学生的课堂表现以及小组活动表现进行评比。 There will be opportunities to earn extra credit during the course through in- class assignments and group activities. These opportunities will be unannounced.
小测验 Quiz				
课程项目		10		
平时作业 Graded Team Assignments		10		
期中考试		30		

Mid-Term Quiz				
期末考试 Final Exam				
期末报告 Final Presentation		30		
其它（可根据需要改写以上评估方式） Others (The above may be modified as necessary)				

20. **记分方式 GRADING SYSTEM**

☑ A. 十三级等级制 **Letter Grading**
☐ B. 二级记分制（通过/不通过） **Pass/Fail Grading**

课程审批 REVIEW AND APPROVAL

21. **本课程设置已经过以下责任人/委员会审议通过**
This Course has been approved by the following person or committee of authority