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**	大数据导论 Introduction to Big Data Science
2.	**授课院系 Originating Department**	数学 Mathematics
3.	**课程编号 Course Code**	MA333
4.	**课程学分 Credit Value**	3
5.	**课程类别 Course Type**	专业选修课 Major Elective Courses
6.	**授课学期 Semester**	秋季 Fall
7.	**授课语言 Teaching Language**	中英双语 English & Chinese
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）**	张振，数学系，副教授 慧园 3 栋 417 zhangz@sustc.edu.cn Zhang Zhen, Mathematics, Associate Professor Room 417, Block 3, Wisdom Valley zhangz@sustc.edu.cn
9.	**实验员/助教、所属学系、联系方式 Tutor/TA(s), Contact**	无 NA
10.	**选课人数限额(可不填) Maximum Enrolment （Optional）**	50

11. **授课方式**
Delivery Method

学时数
Credit Hours

	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	48				48

12.	先修课程、其它学习要求 Pre-requisites or Other Academic Requirements	概率论与数理统计(或概率论) Probability and Statistics (or Probability Theory)
13.	后续课程、其它学习规划 Courses for which this course is a pre-requisite	数据挖掘 Data Mining，统计机器学习 Statistical Machine Learning，大数据计算 Big Data Computing
14.	其它要求修读本课程的学系 Cross-listing Dept.	

教学大纲及教学日历 SYLLABUS

15. 教学目标 Course Objectives

1. 介绍大数据科学的基本概念和研究对象 Show the basic concepts and objectives of big data research

2. 传授大数据科学的基本方法论以及数学模型 Teach basic methodology of big data science, including mathematical modeling

3. 引导学生用 Python 语言编程处理数据，解决实际问题 Guide students to programming and data processing with R and solving real problems

16. 预达学习成果 Learning Outcomes

通过本课程学习，学生将能够：

By the end of the semester, the students will be able to:

1. 描述现实生活中的大数据问题 Describe the big data problems in real life

2. 将大数据问题转化为数学和可计算模型 Turn the big data problems into mathematical and computational models

3. 掌握大数据科学的基本方法论,如分类模型、回归模型、聚类模型、模型选择和降维等方法 Master the basic methodology of big data science, e.g., Classification, regression, clustering, model selection, dimension reduction, etc.

4. 了解热门应用问题的算法机理，如自然语言处理、文本分析、社交网络分析、神经网络和深度学习、分布式计算，推荐系统和在线学习等 Get to know hot topics in applications, e.g., Natural language processing (NLP), text analysis, social network analysis, neural network and deep learning, distributed computing, recommender systems, online learning, etc.

5. 学会用 Python 语言编程以及对实际数据的处理，包括数据收集、提取、集成和清洁，以及数据挖掘 Learn programming and processing real data with python, including data collection, data extraction, data integration , data cleansing, and data mining

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.)

第一章：大数据简介：数据挖掘和机器学习（2 学时）

Chapter 1: Introduction to big data science: Data mining and machine learning (2 hours);

第二章：数据预处理：数据收集、提取和清理，python 语言程序简介（4 学时）

Chapter 2: Data preprocessing: Data collection, data extraction, and data cleansing, python programming (4 hours)

第三章：回归模型：线性回归和正则化方法（4 学时）

Chapter 3: Supervised learning: Linear regression, regularization methods: Ridge and Lasso (4 hours)

第四章：分类模型：k-近邻，决策树，支持向量机，逻辑回归和朴素贝耶斯法则（8 学时）

Chapter 4: Classification: k-nearest neighbours, decision trees, support vector machine, logistic regression and Naïve Bayes rules (8 hours)

第五章：集成算法：袋装，随机森林，提升，AdaBoost 算法，梯度提升决策树（4 学时）

Chapter 5: Ensemble learning: Bagging, Stochastic forests, boosting, AdaBoost, gradient boosting decision tree (GBDT) (4 hours)

第六章：聚类模型：k 平均方法，层次聚类（4 学时）

Chapter 6: Clustering models: k-means, hierarchical clustering, association rules (4 hours)

第七章：特征与模型选择：偏差-方差分解，评价指标，交叉核实（2 学时）

Chapter 7: Feature and model selection: Bias-variance decomposition, evaluation indices, cross-validation (2 hours)

第八章：降维：线性判别分析，主成分分析（4 学时）

Chapter 8: Dimension reduction: Linear discriminant analysis (LDA), principle component analysis (PCA) (4 hours)

第九章：EM 算法和高斯混合模型（2 学时）

Chapter 9: Expectation-Maximization (EM) methods and Gaussian mixed models (2 hours)

第十章：概率图模型，图算法与社交网络分析（4 学时）

Chapter 10: Probabilistic graphical models, graphical algorithms and social network analysis (4 hours)

第十一章：神经网络与深度学习（4 学时）

Chapter 11: Neural networks and deep learning (4 hours)

第十二章：自然语言处理和文本分析（4 学时）

Chapter 12: Natural language processing (NLP) and text analysis (4 hours)

第十三章：推荐系统（2 学时）

Chapter 13: Recommender systems (2 hours)

第十四章：其他专题：在线学习，大规模数据与分布式计算（如有时间）

Chapter 14: Online learning, large scale data and distributed computing (if time permit)

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

参考教材 Textbook：

1. 数据科学导引，欧高炎等著，高等教育出版社，2017.
2. 机器学习，周志华 著，清华大学出版社，2016.
3. An Introduction to Statistical Learning with Applications in R, by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, Springer, 2013.
4. Machine Learning, by Tom Mitchell, McGraw Hill, 1997.
5. Pattern Recognition and Machine Learning, by Christopher M. Bishop, Springer, 2006.

其他参考资料 Supplementary Readings：

The Elements of Statistical Machine Learning: Data mining, Inference and Prediction, 2nd Edition, by Trevor Hastie, Robert Tibshirani, and Jerome Friedman, Springer, 2009.

Pattern Classification, 2nd Edition, by Richard O. Duda, Peter E. Hart, and David G. Stork, John Wiley & Sons, Inc., 2000.

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				
课堂表现 Class Performance				
小测验 Quiz		10%		
课程项目 Projects				
平时作业 Assignments		30%		
期中考试 Mid-Term Test		30%		
期末考试 Final Exam		30%		
期末报告 Final Presentation				
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