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**	大数据导论与实践　　Introduction to Theoretical and Practical Data Science
2.	**授课院系 Originating Department**	数学系 Mathematics
3.	**课程编号 Course Code**	MA234
4.	**课程学分 Credit Value**	4
5.	**课程类别 Course Type**	专业选修课 Major Elective Courses
6.	**授课学期 Semester**	春季 Spring
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、zhangz@sustech.edu.cn Zhang Zhen, Department of Mathematics, Associate Professor
9.	**实验员/助教、所属学系、联系方式 Tutor/TA(s), Contact**	实验员：龙欢、数学系、longh@sustech.edu.cn Huan Long (longh@sustech.edu.cn), Department of Mathematics
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	48		32		80

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 MA212 或数理统计 MA204 Probability and Statistics MA212 or Mathematical Statistics MA204
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，并完成数据分析报告 Learn programming and processing real data with python, including data collection,

data extraction, data integration , data cleansing, and data mining, and write reports for data analysis.

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.)

每个 Lecture 按 2 学分计算

Lecture 1 大数据简介：数据挖掘和机器学习；python 语言程序简介 Introduction to big data science: Data mining and machine learning; python programming

Lecture 2 数据预处理：数据收集、提取和清理 Data preprocessing: Data collection, data extraction, and data cleansing

Lecture 3 分类模型：支持向量机 Classification: Support vector machine

Lecture 4 分类模型：k-近邻 Classification: k-nearest neighbours

Lecture 5 决策树 Decision trees

Lecture 5 朴素贝叶斯法则 Naïve Bayes rules

Lecture 6 逻辑回归和线性判别分析 Logistic regression and Linear discriminant analysis

Lecture 7 线性回归模型 Linear regression

Lecture 8 正则化方法 Regularization methods: Ridge and Lasso

Lecture 9 集成算法：袋装，随机森林 Ensemble learning: Bagging, Random forests

Lecture 10 提升，AdaBoost 算法 Boosting, AdaBoost

Lecture 11 梯度提升决策树 Gradient boosting decision tree (GBDT), XGBoost

Lecture 12 聚类模型：k 均值方法， Clustering models: k-means

Lecture 13 层次聚类 hierarchical clustering

Lecture 14 图算法和谱聚类 Graphical algorithms and spectral clustering

Lecture 15 特征与模型选择：偏差-方差分解 Feature and model selection: Bias-variance decomposition

Lecture 16 评价指标，交叉验证 Evaluation indices, cross-validation

Lecture 17 降维：线性判别分析，主成分分析 Dimension reduction: Linear discriminant analysis (LDA), principle component analysis (PCA)

Lecture 18 核主成分分析，流形学习 Kernel PCA, manifold learning

Lecture 19 EM 算法和高斯混合模型 Expectation-Maximization (EM) methods and Gaussian mixed models

Lecture 20 社交网络分析：谷歌 PageRank 算法 Social network analysis: Google PageRank

Lecture 21 神经网络 Neural networks

Lecture 22 深度学习：卷积神经网络和循环神经网络 Deep learning: CNN and RNN

Lecture 23 受限玻尔兹曼机和生成模型 Restrictive Boltzmann machine and generative model

Lecture 24 推荐系统 Recommender systems

实验课（32 学时）部分

章节 1 软件安装与数据预处理：展示 jupyter notebook 安装过程。分析案例 1--青少年市场细分数据集预处理，分析案例 2--高血压数据分析，分析如何进行变量标准化、离散化、缺失值处理、异常值检测。（2 学时）。

Chapter 1: Software installation and data preprocessing: show the installation process of jupyter notebook. Analysis case 1 - pretreatment of youth market segmentation data set, analysis case 2 - hypertension data analysis, and analysis of how to standardize and discretize variables, deal with missing values and detect abnormal values. (2 hours)

章节 2 分类模型：分析案例--使用 SVM 进行光学字符识别。分析不同的核函数得到的不同的效果，不同的核函数处理不同的数据，优化调参。（2 学时）。

Chapter 2: Classification model: analysis case -- optical character recognition using SVM. Different effects are obtained by analyzing different kernel functions. Different kernel functions process different data and optimize parameters. (2 hours)

章节 3 分类模型：分析案例--*K*近邻算法构建乳腺癌自动诊断模型，分析案例--使用决策树建立个人信用风险评估模型。调参分析，不同的 K 值与分割点等对应不用效果。（2 学时）。

Chapter 3: Classification model: analysis case - neighbor neighbor algorithm to build an automatic breast cancer diagnosis model, analyze the case - use the decision tree to establish a personal credit risk assessment model. Through parameter adjustment analysis, different K values correspond to segmentation points. (2 hours)

章节 4 回归模型：分析案例--预测医疗费用的模型，分析一元回归、多元回归，分析参数估计（最小二乘估计、极大似然估计），调参分析正则化解决过拟合和多重共线性等问题。（2 学时）。

Chapter 4: Regression model: analyze the case - the model for predicting medical expenses, analyze univariate regression and multiple regression, analyze parameter estimation (least squares estimation and maximum likelihood estimation), adjust parameter analysis regularization, and solve the problems of over fitting and multicollinearity. (2 hours).

章节 5 逻辑回归和朴素贝叶斯法则：分析案例--基于朴素贝叶斯算法的手机垃圾短信过滤及中文人名性别预测，分析连续型变量处理方法（离散化、概率分布函数）。分析案例--使用逻辑回归进行鸢尾花品种分类，分析极大似然估计。（2 学时）。

Chapter 5: Logistic regression and naive Bayes rule: analysis of cases -- mobile phone spam message filtering and gender prediction of Chinese names based on Naive Bayes algorithm, and analysis of continuous variable processing methods (discretization and probability distribution function). Case analysis - use logistic regression to classify iris varieties and analyze maximum likelihood

estimation.. (2 hours).

章节 6 集成算法：分析案例--使用随机森林进行红酒品质分类。调参分析不同的抽样方法，分析不同的分类器。（4 学时）。

Chapter 6: Integrated algorithm: case analysis - red wine quality classification using random forest. Different sampling methods and different classifiers are analyzed by parameter adjustment analysis. (4 hours).

章节 7 提升，AdaBoost 算法，梯度提升决策树 ，分析案例--红酒品质分类，与随机森林进行对比分析。（2 学时）。

Chapter 7: Lifting, AdaBoost algorithm, gradient lifting decision tree, analysis case - red wine quality classification, comparative analysis with random forest. (2 hours).

章节 8 聚类模型：k 均值方法，层次聚类 ，分析案例-- k 均值方法对青少年信息和兴趣爱好分类，分析案例--层次聚类对汽车型号聚类，调参分析不同簇的数目、簇间距离度量、不同的 K 值、不同的质心，改进 K-Means，（2 学时）。

Chapter 8: K-means method, hierarchical clustering, case analysis - k-means method classifies teenagers' information and interests, case analysis - hierarchical clustering clusters automobile models, adjusts parameters to analyze the number of different clusters, distance measurement between clusters, different K values and different centroids, and improves k-means. (2 hours).

章节 9 图算法和谱聚类 ，分析案例--青少年信息和兴趣爱好分类，与 k 均值方法比较分析。（2 学时）。

Chapter 9: Graph algorithm and spectral clustering, case analysis -- Classification of teenagers' information and interests, and comparative analysis with k-means method. (2 hours).

章节 10 特征与模型选择：偏差-方差分解，评价指标，交叉验证 ，分析案例-对汽车在 11 个指标上特征分析。（2 学时）。

Chapter 10: Feature and model selection: deviation variance decomposition, evaluation index, cross validation, case analysis - feature analysis of automobile on 11 indexes. (2 hours).

章节 11 降维：线性判别分析，主成分分析 ， 分析案例-- PCA 在人脸识别任务中的应用，分析案例--降维方法在光学字符数据集中的应用。（2 学时）。

Chapter 11: Dimensionality reduction: linear discriminant analysis, principal component analysis, analysis case -- Application of PCA in face recognition task, analysis case -- Application of dimensionality reduction method in optical character dataset. (2 class hours).

章节 12 用 EM 算法处理缺失值。（2 学时）。

Chapter 12: Processing missing values with EM algorithm. (2 hours).

章节 13 社交网络分析，案例分析--使用 Gephi 发现社区。（2 学时）。

Chapter 13: Social network analysis, case study -- using gephi to discover communities. (2 hours).

章节 14 神经网络与深度学习 ，分析案例，课程项目分析讨论（2 学时）。

Chapter 14: Neural network and deep learning, case analysis, course project analysis and discussion. (2 hours).

章节 15 课程项目分析讨论（2 学时）。

Chapter 4: Course project analysis and discussion. (2 hours).

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

参考教材 Textbook：

数据科学导引，欧高炎等著，高等教育出版社，2017.

其他参考资料 Supplementary Readings：

机器学习，周志华 著，清华大学出版社，2016.

An Introduction to Statistical Learning with Applications in R, by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, Springer, 2013.

Pattern Recognition and Machine Learning, by Christopher M. Bishop, Springer, 2006.

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

Understanding Machine Learning, by Shai Shalev-Shwartz and Shai Ben-David, Cambridge University Press, 2018.

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		0%		

课堂表现 **Class Performance**		0%		
小测验 **Quiz**		15%		
课程项目 **Projects**		20%		
平时作业 **Assignments**		30%		
期中考试 **Mid-Term Test**		0%		
期末考试 **Final Exam**		35%		
期末报告 **Final Presentation**		0%		
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