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	数据科学中的统计学习 Statistical Learning for Data Science
2.	授课院系 **Originating Department**	电子与电气工程系 Department of Electrical and Electronic Engineering
3.	课程编号 **Course Code**	EE340
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**	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**）	唐晓颖，助理教授，电子与电气工程系，Email: tangxy@sustc.edu.cn Xiaoying Tang, Assistant Professor, Department of Electrical & Electronic Engineering, Email: tangxy@sustc.edu.cn
9.	实验员/助教、所属学系、联系方式 **Tutor/TA(s), Contact**	已确定的实验员/助教联系方式 Please list all Tutor/TA(s) 杨慧琳，助教，电子与电气工程系，Email: yanghuilin102@hotmail.com 陈敏华，助教，电子与电气工程系，Email:11849296@mail.sustc.edu.cn Huilin Yang, TA, Department of Electrical & Electronic Engineering, Email: yanghuilin102@hotmail.com Minhua Chen, TA, Department of Electrical & Electronic Engineering, Email: 11849296@mail.sustc.edu.cn
10.	选课人数限额(可不填) **Maximum Enrolment** （**Optional**）	40

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				48

12.	先修课程、其它学习要求 Pre-requisites or Other Academic Requirements	MA103A 线性代数 I-A MA103A Linear Algebra I-A
13.	后续课程、其它学习规划 Courses for which this course is a pre-requisite	
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

统计学习旨在利用不同的统计工具来对复杂的数据集进行建模和分析。这是一门随着计算机科学，尤其是机器学习，的发展而并行兴起的统计领域里的新学科。随着"大数据"问题的与日俱增，急需具有统计学习技能的专业人员。考虑到这门学科的复杂度和深度，我们期望进行两个学科的授课。这门课主要是统计学习的入门课，学生不需要特别的数学背景。

Statistical learning refers to modeling and analyzing complex datasets using a variety of statistical tools. It is a recently developed area in statistics and blends with parallel developments in computer science and especially in machine learning. With the advent of ever-growing "Big Data" problems, professional people with statistical learning skills are in high demand. Due to the complexity and depth of this discipline, we expect to offer two consecutive courses (I and II) spanning two semesters. This course is designed as an introductory class to statistical learning, with no background in mathematical science required.

16. 预达学习成果 Learning Outcomes

通过这门课的学习，学生能够：

1)理解统计学习的基础知识，线性回归（简单线性回归，多变量线性回归），分类（逻辑回归，线性判别分析，二次判别分析，k-近邻算法），重采用（交叉验证，置换检验，自助抽样），以及线性模型的选择和正则化（子集选择，特征缩减技术，维度降低，高维度统计）。

2)将统计方法运用于工程实践中。

3)打好在统计学习领域进行深入研究的基础。

After completing this course, the students will be able to

1) understand the basic principles of statistical learning, linear regression (simple linear regression, multiple linear regression), classification (logistic regression, linear discriminant analysis, quadratic discriminant analysis, k-nearest neighbor), resampling (cross-validation, permutation test, bootstrapping), and linear model selection and regularization (subset selection, shrinkage method, dimension reduction, high dimensional statistics).

2) apply statistical methods to engineering practice.

3) lay a solid foundation on conducting further research in statistical learning.

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.)

1. 统计学习的基础 : (3学时)
   - 什么是统计学习？
   - 评估模型的准确度
2. 线性回归: (10 学时)
   - 简单线性回归
   - 多变量线性回归
   - 回归模型中的其他考虑
   - 市场计划
   - 比较线性回归和K-近邻算法
3. 分类 : (10 学时)
   - 分类方法总览
   - 为什么不用线性回归？
   - 逻辑回归
   - 线性判别分析
   - 比较不同的分类方法
4. 重采样: (10 学时)
   - 交叉验证
   - 自助抽样
5. 模型选择和正则化: (10 学时)
   - 子集选择
   - 特征缩减技术
   - 降低维度方法
   - 高维度中的考虑
6. 不仅仅是线性: (5 学时)
   - 多项式回归
   - 阶梯函数
   - 基函数
   - 回归样条函数
   - 平滑样条函数
   - 局部回归
   - 广义加性模型

1. Basics of statistical learning : (3 credit hours)
   - What Is Statistical Learning?
   - Assessing Model Accuracy
2. Linear regression : (10 credit hours)
   - Simple Linear Regression
   - Multiple Linear Regression
   - Other Considerations in the Regression Model
   - The Marketing Plan
   - Comparison of Linear Regression with K-Nearest Neighbors
3. Classification : (10 credit hours)
   - An Overview of Classification
   - Why Not Linear Regression?
   - Logistic Regression
   - Linear Discriminant Analysis
   - A Comparison of Classification Methods
4. Resampling : (10 credit hours)
   - Cross-Validation
   - The Bootstrap
5. Model selection and regularization : (10 credit hours)
   - Subset Selection
   - Shrinkage Methods
   - Dimension Reduction Methods
   - Considerations in High Dimensions
6. Moving Beyond Linearity: ~ 2 weeks (5 credit hours)
   - Polynomial Regression

- Step Functions
- Basis Functions
- Regression Splines
- Smoothing Splines
- Local Regression
- Generalized Additive Models

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

Required Textbook: *"An Introduction to Statistical Learning: with Applications in R" by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani*

Further Reading: *"The Elements of Statistical Learning: Data Mining, Inference, and Prediction" by Trevor Hastie, Robert Tibshirani, Jerome Friedman*

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 Class Performance		10%		
小测验 Quiz		10%		
课程项目 Projects		10%		
平时作业 Assignments		20%		
期中考试 Mid-Term Test		20%		
期末考试 Final Exam		20%		
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)**			
		5	

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