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**	机器学习和大数据分析导论 Introduction to Machine Learning and Big Data Analytics
2.	授课院系 **Originating Department**	信息系统与管理工程系 Department of Information Systems and Management Engineering
3.	课程编号 **Course Code**	MIS110
4.	课程学分 **Credit Value**	3
5.	课程类别 **Course Type**	通识选修课-专业导论类 General Education (GE) Elective Courses – Introduction to Majors
6.	授课学期 **Semester**	秋季 Fall
7.	授课语言 **Teaching Language**	英文 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）**	Dr. Sandro Lera Department of Information Systems and Management Engineering Institute of Risk Analysis, Prediction & Management leras@sustech.edu.cn
9.	实验员/助教、所属学系、联系方式 **Tutor/TA(s), Contact**	TA: Yishan Luo Department of Information Systems and Management Engineering 12031121@mail.sustech.edu.cn
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				48

12.	先修课程、其它学习要求 Pre-requisites or Other Academic Requirements	无 None
13.	后续课程、其它学习规划 Courses for which this course is a pre-requisite	无 None
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

本课程概述了数据科学相关的最常用的方法和概念，以及从生物到金融的各种例子和应用。课程材料包括 Python 的实践编程练习，教学生如何使用数据科学来解决各学科的问题。

This course provides an overview of the most commonly used methods and concepts related to data science, along with a wide variety of examples and applications ranging from biology to finance. The course material includes hands-on coding exercises in Python and teaches students how to use data science to solve problems across disciplines.

16. 预达学习成果 Learning Outcomes

- 对不同的数据类型进行分类，知道如何对数据进行预处理和探索
- 建立概率分布模型，重点关注重尾分布
- 知道如何制定和评估一个统计假设
- 如何对一个目标进行数值优化
- 实现和评估多变量回归
- 了解统计学和机器学习之间的区别
- 了解机器学习的核心概念（交叉验证、偏差-方差权衡等）。
- 为机器学习任务训练一个随机森林
- 解释结构化数据和非结构化数据之间的区别，以及为什么神经网络在非结构化数据中表现出色
- 应用无监督的机器学习方法
- 通过基本的自然语言处理，对文本数据进行数据分析
- 上述所有的任务都用 python 实现

- classification of different data types and know how to preprocess and explore them
- model probability distributions, with emphasize on heavy-tailed ones
- know how to formulate and evaluate a statistical hypothesis
- how to numerically optimize an objective

- implement and evaluate a multivariate regression
- understand the difference between statistics and machine learning
- understand the core concepts of machine learning (cross-validation, bias-variance trade-off etc.)
- train a random forest for machine learning tasks
- explain the difference between structured and unstructured data, and why neural nets excel in the later
- apply unsupervised machine learning methods
- perform data analysis on textual data by means of basic natural language processing
- all of the above tasks are to be implemented in python

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.)

理论（48 学时）

第一章（5 学时）：数据科学概述

- 数据科学、大数据和统计的定义和特点
- 统计学和机器学习的区别
- 人工智能和机器学习的区别
- 数据的不同类别
- 大数据分析的机会和陷阱

第二章（4 学时）：数据探索

- 数据可视化
- 降维

第三章（5 学时）：概率分布

- 概率分布的基础知识
- 重尾分布的重要性

第四章（5 学时）：概率分布

- 概率分布的基础知识

- 重尾分布的重要性

第五章（5 学时）：数值优化

- 分析优化与数值优化

- 梯度下降法

- 超越梯度下降法

第六章（5 学时）：多变量线性回归

- 如何估计回归系数

- 如何评价统计学和机器学习中的回归

- 线性回归的范围和限制

第七章（5 学时）：机器学习的基础知识

- 分类与回归

- 逻辑回归

- 性能评估

- 过度拟合的危险

第八章（5 学时）：合集方法

- 决策树

- 决策森林

- 提升树

第九章（5 学时）：神经网络

- 神经网络的历史

- 结构化与非结构化数据

- 深度学习

第十章（4 学时）：文本挖掘

- 文本的预处理

- 词嵌入

- 变换器

Lecture (48 hours)

Section 1 (5 credit hours): Overview of Data Science

- definition and characteristics of data science, big data and statistics
- difference between statistics and machine learning
- difference between artificial intelligence and machine learning
- different categories of data
- opportunities and pitfalls of big data analysis

Section 2 (4 credit hours): Data Exploration

- data visualization
- dimensionality reduction

Section 3 (5 credit hours): Probability Distributions

- basics of probability distributions
- the importance of heavy-tail distributions

Section 4 (5 credit hours): Statistical Testing

- how to form a statistical hypothesis
- how to evaluate a statistical hypothesis
- beware of p-value hacking

Section 5 (5 credit hours): Numerical Optimization

- analytical optimization vs. numerical optimization
- gradient descent methods
- beyond gradient descent

Section 6 (5 credit hours): Multivariate Linear Regression

- how to estimate the regression coefficients
- how to evaluate a regression in statistics and machine learning
- scope and limitation of linear regression

Section 7 (5 credit hours): The basics of Machine Learning

- classification vs. regression
- logistic regression
- performance evaluation
- dangers of overfitting

Section 8 (5 credit hours): Ensemble Methods

- decision trees
- decision forests
- boosted trees

Section 9 (5 credit hours): Neural Networks

- history of neural networks
- structured vs unstructured data
- deep learning

Section 10 (4 credit hours): Text Mining:

- pre-processing of text
- word embeddings

- transformers

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

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. New York: springer.

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**		20%		
课堂表现 **Class Performance**		20%		
小测验 **Quiz**		20%		
课程项目 **Projects**		40%		
平时作业 **Assignments**				
期中考试 **Mid-Term Test**				
期末考试 **Final Exam**				
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