0252 应用统计硕士专业学位研究生核心课程指南

01 统计学基础

一、课程概述

本课程是统计类课程体系中最重要的专业基础课之一。课程内容主要包括三大部分——概率论基础、探索性数据分析、数理统计基础。概率论基础为数理统计以及后续统计课程学习提供必要的理论基础，内容涉及事件与概率运算、Bayes 公式、随机变量及其概率分布、数字特征、随机向量及其联合分布与边缘分布、独立性、条件分布与条件期望、大数定律、中心极限定理、随机过程初步等。探索性数据分析技术已被用于数据挖掘，也用于大型数据分析，是统计思维的启蒙课程和数据处理的基础课程，内容包括数据预处理、描述统计量、数据可视化等。数理统计部分为统计方法应用与后续统计课程学习提供基础，内容涉及总体与样本、参数估计、假设检验、方差分析与回归分析等。

二、先修课程

数学分析（或高等数学）、线性代数。

三、课程目标

通过本课程的学习，学生不仅能理解和掌握概率论与统计学的基本概念、理论与方法，为后续各门课程的学习打下必要的理论基础，而且能至少使用一种软件（R、Python 等），准确地进行数据预处理、描述分布特征和绘制可视化图形，为后续数据分析工作提供“干净”、简洁和符合模型要求的数据，并为数据分析结果的展示提供优质的可视化素材。

四、适用对象

本课程适用于应用统计硕士专业学位研究生。

五、授课方式

1. 课堂讲授。以课堂讲授为主，并结合计算机和多媒体等教学手段。课堂讲授注重概念、方法、理论和实际应用相结合。

2. 计算机实验。根据教学内容，进行计算机模拟实验和实际数据分析。

3. 案例讨论及学生汇报展示。运用实际数据进行完整的探索性数据分析等。

六、课程内容

统计学基础的内容主要包括事件与概率、随机变量及其概率分布、随机变量数字特征、随机向量、大数定律和中心极限定理、随机过程初步、探索性数据分析、总体与样本、参数估计、假设检验、方差分析与回归分析等。

内容模块	知识点	重点与难点
1. 概率论基础	1.1 事件与概率:样本空间、事件域、概率、加法公式、乘法公式、全概率公式、Bayes 公式	概率空间的基本思想、事件与概率的运算公式、全概率与 Bayes 公式
	1.2 随机变量:离散型随机变量(两点分布、二项分布、泊松分布、几何分布),连续型随机变量(均匀分布、指数分布、正态分布),指数分布族,随机变量函数的分布函数,随机变量的期望和方差	随机变量的概念与定义、随机变量的类型、指数分布族的概念、随机变量函数分布的计算
	1.3 随机向量:离散型随机向量,连续型随机变量(多元正态分布),条件分布、条件密度与条件期望,次序统计量,协方差和相关系数	多元正态分布及其相关性质、条件分布、条件密度与条件期望的概念与计算
	1.4 极限定理:大数定律,中心极限定理	极限定理的理解与证明
	1.5 随机过程初步:泊松过程,马尔科夫链和平稳序列,Monte Carlo 模拟方法简介	随机过程的概念与定义、Monte Carlo 模拟方法的实践
2. 探索性数据分析	2.1 数据预处理:数据清洗、数据集成、数据变换、数据规约	特征工程(变量构建、变量筛选等)、缺失数据处理、数据变换、离群值识别与处理
	2.2 描述统计:单变量特征描述、变量相关性描述	描述性和相关性统计量的计算、适用范围及其结果解读
	2.3 数据可视化:数据图表显示、可视化组件与技巧等	各种图形的制作及适用范围,提升图表可读性和美观性的技巧
3. 数理统计基础	3.1 统计学基本概念:总体与样本,统计量及抽样分布,三大抽样分布(卡方分布、t 分布、F 分布),统计量的充分性和完备性(选讲)	总体与样本的概念、抽样、三大抽样分布的定义与相关计算

续表

内容模块	知识点	重点与难点
3. 数理统计基础	3.2 参数估计：点估计与区间估计的概念，矩估计与极大似然估计，点估计的评价准则，各类置信区间的构造	点估计方法、置信区间的构造方法
	3.3 假设检验：假设检验的基本思想与概念，正态总体参数的假设检验，非正态总体参数的假设检验，卡方拟合优度检验，正态性检验	假设检验的基本思想与基本步骤、各类检验的构造方法

七、考核要求

本课程进行平时考核和期末考试。考试方法包括笔试和计算机实验。

八、编写成员名单

郑明（复旦大学）、林路（山东大学）、杨瑛（清华大学）、汪荣明（上海对外经济贸易大学）、王立洪（南京大学）

02 统计调查与数据采集

一、课程概述

本课程主要培养学生针对特定问题制定统计调查、数据采集方案和科学收集数据的能力，包括统计调查方法、抽样技术、试验设计方法以及网络数据与大数据采集技术。本课程属于应用统计硕士专业学位研究生核心课程，为培养学生的专业素养奠定基础。

二、先修课程

统计学基础（概率论基础、探索性数据分析、数理统计基础等）。

三、课程目标

通过本课程的学习，使学生们掌握统计调查的方法和技术，理解抽样理论并掌握常用抽样技术，掌握试验设计中常用的设计方法与建模技术，了解网络数据和大数据的获取方式。使学生能够针对实际问题，设计抽样或试验方案，能够根据具体需求和成本，选择合适的数据采集方法，培养学生灵活运用所学统计知识的能力和和应变能力。

四、适用对象

本课程适用于应用统计硕士专业学位研究生。

五、授课方式

课堂讲授与案例教学相结合,配合上机实习,使用计算机完成抽样、试验方案构造、数据收集、数据分析等。

六、课程内容

本课程每一部分的重点在于让学生掌握在何种情况下使用以及如何使用每一种方法,难点在于提供机会让学生去实践使用每一种方法。

(标“*”为根据学生情况选讲部分)

内容模块	知识点	重点与难点
1. 统计调查	1.1.1 数据类型	统计调查、数据采集与分析中的数据类型,问卷调查的一般程序、各调查法的特点和适用场合
	1.1.2 问卷调查设计:问卷调查法的类型、适用范围;问卷的设计与实施	
	1.1.3 几种常用方法:访谈法、小组座谈会、深访法;各自的特点、适用范围以及每一种方法的实施过程与技巧;访谈纲要	
	*1.1.4 其他方法:观察法和实验法;适用范围,实施过程与技巧	
	1.2.1 抽样调查基本概念、一般程序包括总体和总体单位,样本、抽样单位与抽样框,抽样调查的工作程序,概率抽样和非概率抽样的概念	有限总体概率抽样的概念和实施,估计量及其评价
	1.2.2 基于调查的估计:简单估计、部分估计、比率估计、回归估计,了解估计方法的特点和适用范围,估计量的分布和特征数,估计量的偏差和精度、准确度以及区间估计的构造	
	1.2.3 样本量的确定:估计总体均值时样本量的确定、估计总体比例时样本量的确定	

续表

内容模块	知识点	重点与难点
1. 统计调查	1.3.1 简单随机抽样：定义和实施步骤；总体均值、总体比率估计量及性质、子总体的估计量	各种抽样技术的应用场合与实施过程，估计量的计算及其性质，子总体的估计，样本量的确定
	1.3.2 分层抽样：定义和分层原则；总体均值、总体比率估计及性质、子总体的估计；各层样本量的分配、总样本量的确定	
	1.3.3 等距抽样：定义；总体均值的估计量及性质；与简单随机样本的关系	
	1.3.4 单级整群抽样：定义及优点；群大小相等的单级整群抽样的总体估计及设计效果、群大小不相等的单级整群抽样总体估计及设计效果	
	1.3.5 两级抽样：定义及优点；群大小相等的两级抽样的总体估计、群大小不相等的两级抽样的总体估计及设计效果	
	*1.3.6 非概率抽样常用方法：系统抽样的定义和实施方法、其他抽样方法的定义和实施方法	
2. 实验设计	2.1.1 试验设计的概念与意义：试验设计方法的目的、内容、发展概况及应用；试验设计的基本原则与拟定	试验设计的研究内容、试验设计的基本原则
	2.1.2 试验设计的常用术语及统计模型：试验考察指标、试验因素、试验水平、交互作用、准确性和精确性；抽样分布、假设检验与参数估计、简单回归分析	准确性和精确性、假设检验与参数估计
	2.1.3 试验设计的一些典型应用：食品安全、生物、医学、质量等领域的应用	试验设计的应用
	2.1.4 试验数据的误差分析：误差的基本概念、来源及分类、误差的估计与检验	误差的基本概念、随机误差的估计、系统误差和过失误差的检验
	2.1.5 试验数据的方差分析：单因素方差分多因素试验的方差分析的基本步骤和计算	单因素、双因素试验的方差分析的基本步骤和计算
	2.2.1 正交试验设计：基本原理及构造；多指标、混合水平、有交互作用的正交试验设计；正交试验设计的方差分析	正交设计的原理、正交表的构造、正交试验设计的方差分析

续表

内容模块	知识点	重点与难点
2. 实验设计	2.2.2 区组设计:基本原理、应用及优缺点;随机化完全区组设计、拉丁方设计、不完全区组设计	区组设计的原理及应用、随机化完全区组设计、拉丁方设计
	2.2.3 因子试验设计:基本原理及应用、2k 因子设计、3k 因子设计	因子试验设计的基本原理、2k 因子设计的构造及应用
	2.2.4 均匀设计:基本原理及应用场景、等水平均匀设计表和混合水平均匀设计表	均匀设计的基本原理及应用、等水平均匀设计表和混合水平设计表
	*2.2.5 响应曲面分析方法:基本原理、二阶响应曲面设计及分析、拟合响应曲面的设计及分析、应用场景	响应曲面分析方法的基本原理及应用
3. 网络数据收集	3.1 网络调查:网络调查的特点及适用范围;网络调查与传统调查的区别;Web 问卷的逻辑设置	掌握网络调查的特点
	3.2 文本和数据库搜索:了解各种开放数据库;使用 SQL 检索数据库;多种格式文本数据整合、转换	SQL 数据库检索、多种文本数据的整合转换
	3.3 文档和文件抽取文本和元数据:掌握从文档和文件中批量抽取文本和元数据;从 PDF 文件中抽取可编辑文本;编辑文件及删除元数据	各类文档和文件中批量抽取文本数据
	3.4 图片、声音中的文本识别:OCR 库概述;了解光学识别软件 Tesseract;从音频数据中识别、抽取文本	了解 OCR 库、如何从音频数据中抽取文本
	3.5 网络信息采集/网络爬虫:建立网络爬虫、利用 Python 或 R 实现网络 API 数据收集	网络爬虫建立、API 数据收集
	3.6 R 或 Python 语言实现:网络数据的抓取、数据清洗、数据整合、数据的可视化	数据的收集、预处理和简单的数据可视化
	*3.7 数据采集、存储平台概况:主要数据采集、存储平台的架构、特点及使用;调查平台和网络抓取平台的差别	熟悉采集、存储平台的使用

七、考核要求

采用笔试和实践作业相结合的考核方式。考核标准是既要考核学生是否从理论上掌握了统计调查和数据采集的方法,更要从实践层面上考核学生是否真正掌握了数据采集的能力与技术。实践作业要求学生独立完成一项统计调查或网络数据获取并撰写报告(包括题目确定、方

案设计、数据收集过程描述、数据整理、数据分析及报告撰写)。

八、编写成员名单

邹长亮(南开大学)、董麓(天津财经大学)、陈华峰(北京益派市场咨询有限公司)

03 统计计算

一、课程概述

本课程以概率论为基础,通过样本推断总体的统计特性,内容极其丰富,并且随着计算机的普及与发展,从事统计工作和实际工作的人都很关心如何利用计算机来更好地完成统计数据的分析工作,从而出现"统计计算"这个研究方向。统计计算是当今统计学的一个重要分支,是数据驱动与计算机相结合的产物,是数理统计、计算数学和计算机科学的交叉学科。本门课程主要由统计软件、统计模拟、云计算与并行计算三部分组成,既注重统计计算算法的讲解,又兼顾统计软件、并行计算等现代计算技术的介绍。本门课程是将统计理论方法与实际数据分析相结合的一门专业课程。

二、先修课程

数理统计、多元统计分析、贝叶斯统计等。

三、课程目标

通过本课程的学习使学生熟练掌握统计软件的使用,理解和掌握统计计算算法的基本原理与基本理论,掌握大数据背景下并行算法、云计算的计算技能,能用所学的基本理论进行实际数据分析,提高学生解决实际问题的能力。

四、适用对象

本课程适用于应用统计硕士专业学位研究生。

五、授课方式

课堂讲授与案例教学相结合,配合上机实习,使用计算机完成抽样、试验方案构造、数据收集、数据分析等。

六、课程内容

本课程主要讲授统计计算相关算法的基本原理和基本理论及算法实现、统计软件 R 的基本操作、最优化理论的基本思想与算法、大数据背景下的并行计算、云计算的计算技能等。

内容模块	知识点	重点与难点
1. 统计计算	1.1 统计计算简史、课程框架	
	1.2 统计计算算法简介	
	1.3 云计算、并行计算的优良性	
2. 统计软件 R 基本操作及生成随机数	2.1 R 基本操作 2.1.1 R 软件及相关软件包的安装 2.1.2 R 基本命令操作 2.1.3 R 变量类型的定义、操作 2.1.4 R 数据的导入和存储 2.1.5 R 软件的画图操作 2.1.6 R 子函数的编写	1. 熟练掌握统计软件 R 的基本操作 2. 掌握生成随机数的机理并能够进行算法实现
	2.2 随机数的生成 2.2.1 均匀分布随机数的生成 2.2.2 非均匀分布随机数的生成 2.2.3 生成随机数的 R 实现	
3. 蒙特卡洛方法	3.1 积分的模拟近似	1. 熟练掌握定积分的蒙特卡洛近似方法与 Bootstrap 方法,理解并运用重要性抽样原理 2. 利用 MCMC 方法生成复杂分布(比如后验分布)随机数
	3.2 重要性抽样	
	3.3 分层抽样	
	3.4 EM 算法、数据扩充算法 3.4.1 EM 算法的收敛性 3.4.2 EM 算法的应用 3.4.3 EM 算法的改进变种	
	3.5 Bootstrap 方法 3.5.1 Bootstrap 的基本原则 3.5.2 非参数 Bootstrap 3.5.3 参数 Bootstrap 3.5.4 基于 Bootstrap 的回归分析 3.5.5 基于 Bootstrap 的纠偏分析 3.5.6 基于 Bootstrap 的统计推断	
	3.6 Metropolis-Hasting 抽样	
	3.7 逆跳 MCMC	
	3.8 Gibbs 抽样	

续表

内容模块	知识点	重点与难点
4. 优化方法	4.1 最速下降法 4.2 梯度下降法 4.3 Newton 法、拟牛顿法 4.4 ADMM 算法 4.5 内点法	掌握最优化理论的基本思想与算法并会软件实现
5. 云计算、并行计算	5.1 云计算的实施机制 5.1.1 基本概念 5.1.2 特殊云机制 5.1.3 云管理机制 5.1.4 云安全机制 5.2 分布式计算和云计算 5.2.1 分布式计算 5.2.2 云计算 5.2.3 二者区别与联系 5.3 集群技术 5.4 MPI、多线程 5.5 MapReduce 原理 5.5.1 MapReduce 简介 5.5.2 MapReduce 程序执行流程 5.5.3 MapReduce 工作原理	掌握大数据背景下云计算、并行计算的计算技能

七、考核要求

开卷考试、完成作业或论文。考核学生能否应用所学的统计计算方法解决实际相关的问题。

八、编写成员名单

张宝学(首都经济贸易大学)、刘扬(中央财经大学)、房祥忠(北京大学)、林明(厦门大学)

04 统计数据分析方法

一、课程概述

本课程是指用适当的统计分析方法对搜集的大量数据进行分析,提取有用信息和形成结论而对数据加以详细研究和概括总结的过程。统计数据分析方法是统计类课程体系中最重要的专业基础课之一,主要包括五大部分——回归分析、时间序列分析、多元统计分析、非参数统计分析和纵向数据分析。

二、先修课程

概率论、数理统计、随机过程。

三、课程目标

通过本课程的学习,使学生能够对统计数据分析的方法和思想有更进一步的理解,让学生能够熟练应用诸多的统计方法进行数据分析和建模,通过和应用领域的结合,对考虑的问题能够给出较为合理的解答。引导学生既重视理论又重视实际应用,将学生培养成复合型应用人才。

四、适用对象

本课程适用于应用统计硕士专业学位研究生。

五、授课方式

采用课堂讲授方法为主,案例教学和实验教学为辅,多种教学方式相结合。使学生掌握本课程的基本概念、基本理论和基本方法的理解,提高学生分析问题和解决问题的能力。

六、课程内容

内容模块	内容要求	重点与难点
1. 回归分析	1.1 回归分析的研究内容及建模过程;回归分析的应用及发展历史	建模过程
	1.2 简单线性回归模型:一元线性回归模型的建模;最小二乘估计方法及其估计量的性质;回归方程的有关检验、预测和控制的理论与应用	估计量的性质

续表

内容模块	内容要求	重点与难点
1. 回归分析	1.3 多元线性回归模型及其基本假设；回归模型未知数的估计及其性质；回归方程及回归系数的显著性检验	估计的性质和显著性检验
	1.4 回归模型选择的评价标准：模型比较的标准，模型选择的交叉验证，变量选择方法	变量选择
	1.5 残差分析可以获取的信息，残差图，学生残差，异常值的检测	残差分析
	1.6 违背基本假设的情况：异方差、序列相关和多重共线性产生的原因，对最小二乘估计的性质和相关检验的影响，如何检验问题的存在和处理方法	异方差的处理方法
2. 时间序列分析	2.1 时间序列分析的基本内容，回顾时间序列的定义以及均值、方差和协方差求解，并举例说明	平稳性的定义及判断和纯随机性的检验
	2.2 一般线性过程、自回归过程、滑动平均过程和自回归滑动平均混合模型的基本性质以及适用场景	格林函数、逆函数、自相关和偏自相关函数
	2.3 平稳化的各种方法以及 ARIMA 模型得基本性质，并结合具体案例分析	平稳化的方法
	2.4 样本自相关函数和偏自相关函数的性质，非平稳性时间序列模型的处理方法，以及真实时间序列分析的举例	偏自相关函数的性质及非平稳性的识别
	2.5 矩估计、最小二乘估计和极大似然估计的性质，自助法的应用，以及残差分析及过度拟合的处理方法	估计的性质及自助法和残差分析
	2.6 最小均方误差、ARIMA 预测和预测极限和更新	条件期望，预测的极限和更新
	2.7 金融时间序列的特点、ARCH 模型、GARCH 模型的极大似然估计和模型诊断，以及 GARCH 模型的扩张	参数估计和模型诊断
	2.8 门限模型：非线性检验方法、一阶门限自回归模型、门限模型门限和非线性的检验门限模型的估计和模型诊断	门限模型的估计和模型诊断

续表

内容模块	内容要求	重点与难点
3. 非参数统计	3.1 掌握适应任意分布的统计量、计数统计量和秩统计量符号秩统计量,以及条件的适应任意分布的检验	秩统计量及相关理论
	3.2 掌握一样本 U 统计量,一样本 U 统计量的渐近分布和二样本 U 统计量的渐近分布	U 统计量的渐近分布
	3.3 线性秩统计量的定义、线性秩统计量分布的一些性质以及线性符号秩统计量	线性秩统计量的渐近性质
	3.4 次序统计量的分布、分位数的估计、分布函数的置信区间以及随机变量的容忍区间	分布函数的置信区间
	3.5 Spearman 秩相关系数、Kendall-tau 相关系数和 Kendall 协和系数	秩相关系数的检验和异常值的检测
4. 多元统计分析	4.1 多元统计分析的基本内容及应用领域,补充相关的矩阵代数的基本知识:如行列式、逆矩阵、矩阵的迹、二次型、正定阵以及矩阵微商等概念	正定矩阵和二次型
	4.2 统计距离,多元正态分布基本概念和定义及其有关的性质,多元正态随机变量的基本性质,均值向量和协方差阵的估计,Wishart 分布的定义和基本性质,Hotelling T2 和 Wilks 分布的定义及其基本性质	多元正态及其各种分布的性质
	4.3 多元正态分布均值向量和协方差阵的假设检验,含多个正态总体均值和协方差阵的假设检验、计算程序中有关假设检验的算法基础	总体均值和协方差阵的检验
	4.4 聚类分析的目的和意义、聚类分析中所使用的几种尺度的定义、8 种系统聚类方法的定义及其基本性质、模糊聚类方法及其基本性质,K-均值聚类和有序样品的聚类,有关聚类分析的算法基础	聚类的方法
	4.5 判别分析的目的和意义、判别分析中所使用的几种判别尺度的定义和基本性质	判别方法的算法基础
	4.6 主成分分析的目的和意义、主成分分析的数学模型及几何解释,主成分的推导及基本性质、计算程序中有关主成分分析的算法基础	主成分的算法
	4.7 因子分析的目的和基本思想、因子分析的数学模型,因子载荷阵的估计方法,因子旋转,因子得分、计算程序中有关因子分析的算法基础	因子旋转方法

续表

内容模块	内容要求	重点与难点
4. 多元统计分析	4.8 对应分析的目的和基本思想、对应分析方法的基本原理；简单介绍相关的计算程序	对应分析的基本原理
	4.9 典型相关分析的目的和基本思想、典型相关分析的数学模型、总体和样本的典型相关系数以及典型变量，典型相关系数的假设检验	典型相关系数的假设检验
	4.10 对数线性模型基本理论和方法，logistic 回归的基本理论和方法	logistic 回归的基本理论和方法
5. 纵向数据分析	5.1 纵向数据背景通过搜集到的一些具体纵向数据实例，介绍什么是纵向数据，对于这一特殊结构的数据，给出通常采用什么样的统计模型	纵向数据的理解
	5.2 纵向数据下线性模型：介绍普通线性模型如何应用在纵向数据这一特殊的数据形式下，介绍模型的意义、参数估计方法，随机效应模型提出的背景，参数估计方法以及经验似然方法	纵向数据线性模型统计推断
	5.3 广义线性模型对线性模型的推广，指数族分布、广义线性模型参数估计方法和推断方法，拟似然方法	广义线性模型统计推断
	5.4 纵向数据边际模型：包括均值参数估计，相关系数估计的矩估计方法，广义矩方法，拟加权最小二乘法，方差的参数估计方法	边际模型估计
	5.5 纵向数据下模型选择：介绍纵向数据下变量选择方法，相关变量选择准则有 QIC，推广的 QIC 相关信息准则，经验似然准则，伪高斯似然准则等	变量选择方法和准则
	5.6 拓展：相关前沿介绍	

七、考核要求

采用闭卷考试和课程项目结合，对学生知识的掌握及应用所学知识解决实际问题能力进行考核。

八、编写成员名单

张虎（中南财经政法大学）、刘禄勤（武汉大学）、史代敏（西南财经大学）、杨仲山（东北财经大学）

05 机器学习与数据挖掘

一、课程概述

本课程是面向应用统计及相关专业硕士研究生开设的专业基础课,其教学目的是使学生掌握常用机器学习与数据挖掘方法,理解其基本思想和算法步骤。通过计算机实验和在经济学、金融学、生物信息学、计算机科学等学科领域中的应用实例,熟悉机器学习与数据挖掘的科学方法和具体运用,增强学生对机器学习与数据挖掘的学习兴趣。

二、先修课程

要求学生事先受过基本编程训练(熟悉 Matlab/R/Python 软件),并具有线性代数、微积分和概率统计方面的基础知识。

三、课程目标

通过本课程的学习,使学生掌握常用机器学习与数据挖掘方法的基本思想、算法及其具体应用,在通过计算机分析和解决实际问题的能力方面得到进一步的培养和训练。同时使学生了解该领域的研究趋势,具备初步的科研创新能力。

四、适用对象

本课程适用于应用统计硕士专业学位研究生。

五、授课方式

教学方式主要由教师运用多媒体讲授方法的基本思想和算法步骤,并进行实例分析与学生上机实验运用方法解决实际问题构成,适合双语教学。

六、课程内容

(标“*”号的节为可选教学内容或者教学程度可根据专业教学实施情况灵活把握)

内容模块	知识点	重点与难点
1. 概论	1.1 机器学习与数据挖掘的基本概念和功能	1. 机器学习与数据挖掘的基本概念和功能。
	1.2 机器学习与数据挖掘的基本应用概述	
	1.3 学习问题类型基本划分	

续表

内容模块	知识点	重点与难点
1. 概论	1.4 模型评价	2. 模型评价的重要概念：训练误差与测试误差、过拟合与欠拟合、偏差方差平衡、模型可解释性和预测准确性的平衡
	实验要求：掌握 Matlab/R/Python 软件的基本操作和命令（包括矩阵计算、绘图等相关命令）	
2. 线性回归、模型选择与正则化	*2.1 简单线性回归模型、多重线性回归模型：参数估计、模型评价	*简单线性回归模型、多重线性回归模型的基本思想，参数估计及其准确性评价、模型准确性评价，分析应用多重线性回归的四个基本问题：因变量和自变量之间是否存在关系、重要变量选择、模型的拟合优度、模型的预测准确性
	*2.2 线性回归的运用，包括定性自变量的处理、线性模型假定条件不成立时各种情形的判定和处理	*掌握定性自变量的处理方法、线性模型假定条件不成立时的各种情形的判定与处理，包括可加性假定、因变量与自变量之间的非线性、模型残差相关性、非常数模型残差、异常值、高杠杆点、多重共线性等
	2.3 变量选择方法：最佳子集选择	最佳子集选择的基本思想和运用
	2.4 收缩方法：岭回归、Lasso 及调节参数选取	收缩方法的基本思想，岭回归和 Lasso 在理论和应用方面的区别和联系
	2.5 线性回归和 K 最近邻回归（非线性回归）的区别和联系	理解线性回归和 K 最近邻回归（非线性回归）的区别和联系
	实验要求：灵活运用软件进行线性回归分析、最佳子集选择、岭回归、Lasso	

续表

内容模块	知识点	重点与难点
3. 分类	3.1 分类问题与方法概述	1. 理解线性回归不能用于解决分类问题的原因(两类问题例外); 2. 简单 logistic 回归模型和多重 logistic 回归模型的基本思想、参数估计及其准确性评价、模型准确性评价; 3. 线性判别分析、二次判别分析的基本思想、原理和运用; 4. 理解 logistic 回归、判别分析、K 最近邻分类等分类方法的联系与区别
	3.2 logistic 回归模型、多重 logistic 回归模型:参数估计、模型评价	
	3.3 线性判别分析、二次判别分析	
	3.4 K 最近邻分类	
	3.5 logistic 回归、判别分析、K 最近邻分类等分类方法的联系与区别	
	实验要求:灵活运用软件使用 logistic 回归、线性判别分析、二次判别分析、K 最近邻分类解决分类问题	
4. 重复抽样方法	4.1 交叉验证方法,包括验证集方法、留一法、K-折交叉验证	1. 交叉验证法的基本思想和运用。 2. K-折交叉验证的偏差方差平衡,交叉验证方法在回归问题和分类问题中的使用。 3. 自助法的基本思想和运用
	4.2 自助法(Bootstrap)	
	实验要求:灵活运用软件,使用交叉验证方法选取最优学习方法解决回归或分类问题	
5. 非监督学习	5.1 非监督学习概述	1. *主成分分析的基本思想和运用,高维情形下主成分分析的计算。 2. *K-均值聚类:基本思想、算法和运用。 3. *分层聚类:基本思想、算法和运用、树状图的解释。 4. 高斯混合模型:基本思想、EM 算法和模型运用
	*5.2 维数约简方法:主成分分析	
	5.3 聚类分析:*K-均值聚类、高斯混合模型;*分层聚类(距离度量包括最小距离、最大距离、平均距离、重心距离、离差平方和等)	
	实验要求:运用软件进行主成分分析、K-均值聚类、分层聚类、高斯混合模型聚类	

续表

内容模块	知识点	重点与难点
6. 基于树的方法	6.1 决策树：回归树、分类树；树与线性回归的联系和区别	1. 决策树的基本思想和运用。 2. 树与线性回归的联系和区别、优劣势比较。 3. 集成学习 Bagging、Random Forests、Boosting 的基本思想和运用。
	6.2 集成学习：Bagging、Random Forests、Boosting	
	实验要求：灵活运用软件进行决策树分类；Bagging、Random Forests、Boosting 分类	
7. 支持向量机	7.1 最大间隔分类器	1. 最大间隔分类器。 2. 支持向量分类器。 3. 核函数与支持向量机 SVM。 4. 多类问题 SVM。 5. SVM 与 logistic 回归的联系与区别。
	7.2 支持向量分类器	
	7.3 核函数与支持向量机 SVM	
	7.4 多类问题 SVM	
	7.5 SVM 与 logistic 回归的联系与区别	
	实验要求：灵活使用 SVM 进行分类	
8. 神经网络	8.1 神经元模型	1. 神经元模型的基本思想和激活函数的概念。 2. 感知机的基本思想、多层前馈神经网络结构。 3. 误差逆传播算法的基本思想。 4. 训练神经网络时的常见问题。 5. *了解深度学习的发展历程，掌握其基本思想，了解常用的几种深度学习方法（CNN、RNN、LSTM、GAN 等）
	8.2 感知机与多层网络	
	8.3 误差逆传播算法	
	8.4 训练神经网络时的常见问题：初值问题、过拟合问题、数据的标准化、隐含层数与神经元数的确定、局部极小问题	
	*8.5 深度学习简介	
	实验要求：灵活使用神经网络分析数据	

七、考核要求

考核方式主要包括考勤、平时作业、期末考试三个部分。平时作业包括概念、推理等练习题以及上机作业；期末考试建议采用上机编程的方式进行考核，以学生掌握和应用机器学习与数据挖掘方法的情况为考核标准，建议采用能将文字和代码融为一体的数据分析报告生成利器 R Markdown 或 Matlab Publish 来完成，利于教师阅卷以及培养学生撰写数据分析报告的能力。

八、编写成员名单

石磊(云南财经大学)、周勇(华东师范大学)、徐寅峰(西安交通大学)、赵彦云(中国人民大学)