《数据分析理论与实践（双语）》教学大纲

课程编号：071403B

课程类型：□通识教育必修课 □通识教育选修课

☑学科基础课 □专业核心课

□专业提升课 □专业拓展课

总 学 时：48 讲课学时：32 实验（上机）学时：16

学　　分：3

考试类型：□考试 ☑考查

适用对象：信息管理与信息系统专业

□是 ☑否 适合作为其他专业学生的个性化选修课

先修课程：概率论与数理统计、数据结构

一、教学目标

本课程是信息管理和信息系统专业的学科基础课，是为适应信息时代快速增长的业务数据及基于数据的科学决策的要求而开设的，既重视学生对相关理论的系统学习，又强调培养学生发现问题、分析问题和解决问题的实践能力。通过本课程的学习，使学生对数据分析相关理论、方法和技术有较完整的认识，并熟练掌握一种常用数据分析工具如Python及相关数据分析模块，能独立有效地根据业务和数据选择合适的工具技术，对商业数据加工处理和科学建模，并为后续数据分析课程如机器学习与量化投资等打下坚实的理论和技术基础。

1、课程专业教学目标

目标1：了解数据分析方法和技术概貌，理解并掌握数据分析相关理论、知识和算法原理等；

目标2：了解Python语言特点和相关工具技术，熟练掌握Python语言基础编程和常用数据分析模块；

目标3：了解数据分析完整过程对应的方法和技术，能基于Python独立完成数据采集、数据预处理、探索性分析、构建模型及部署应用等数据分析过程。

2、课程思政目标

《数据分析理论与实践（双语）》课程思政教育的总体目标是把马克思主义立场观点方法的教育与科学精神的培养结合起来，提高学生正确认识问题、分析问题和解决问题的能力。具体目标包括：

（1）强化工程伦理教育：教育引导学生深刻理解并自觉实践数据科学伦理，增强职业责任感，培养学生精益求精的大国工匠精神，激发学生科技报国的家国情怀和使命担当。

（2）科学思维方法的训练：教育引导学生深刻理解数据科学分析方法，培养学生探索未知、追求真理、勇攀科学高峰的责任感和使命感。

（3）社会主义道德：教育引导学生深刻理解信息安全与数据滥用违法带来的道德和法律问题，培养遵纪守法、爱岗敬业、无私奉献、诚实守信、公道办事、开拓创新的职业品格和行为习惯。

二、教学内容及其与毕业要求的对应关系

1、教学内容

本课程的核心内容包括：（1）数据科学概念、数据类型、大数据、数据分析过程及相关实践应用；（2）Python数据科学工具，包括Jupyter Notebook、Numpy、Pandas、Scipy、Matplotlib、Statsmodels、Scikit-Learn等；（3）基于Numpy和Pandas等模块进行数据读写、预处理、聚合、连接合并等数据操作；（4）使用Python数据科学工具解决统计与计量问题，如描述性统计、方差分析、相关性分析、回归分析、主成分分析等；（5）数据可视化及高级应用。

2、教学方法和手段

根据教学目标，拟采用的教学方法有：课堂讲解基本概念和核心知识，上机作业训练解决问题的实践动手能力，提供大量的英文学习资料鼓励学生培养独立自学及解决问题能力，利用大作业项目的分组方式启发学生相互协作、互相共享以促进学生整体的能力提高。最后借助适度规模的项目作业，将所学知识进行融会贯通，促进学生掌握本门课程的知识和相关技术。

3、实践教学环节要求

使用上机实验方式提高实践动手能力，要求必须按时提交作业，每次作业评分计入平时成绩。

4、学习要求

为有效学习本课程，要求学生首先具备计算机系统、计算机网络、数据库、数据结构、数理统计等方面的基本知识。

自学时应该熟读大纲，提纲挈领地掌握数据科学的内容，随后，按照大纲和教师发放的讲义认真研读教材，同时还要阅读相关参考文献和学习资料，通过课后思考和练习题进行多角度和深层次的反复学习。

5、与毕业要求的关系

数据分析理论与实践（双语）是信管专业培养方案的重要组成部分，也是未来IT领域内学生职业发展的必要知识组成部分。学习和了解数据科学的理论和具体的开发技术是量化投资专业和商务智能方向培养的基本要求。

三、各教学环节学时分配

以表格方式表现各章节的学时分配，表格如下：

**教学课时分配**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **序号** | **章节内容** | **讲课** | **实验** | **其他** | **合计** |
| 1 | 数据科学概述 | 2 |  |  | 2 |
| 2 | Python基础 | 4 | 2 |  | 6 |
| 3 | Numpy基础 | 2 | 2 |  | 4 |
| 4 | Pandas基础 | 2 | 2 |  | 4 |
| 5 | 数据获取与存储 | 4 | 2 |  | 6 |
| 6 | 数据操纵与预处理 | 2 |  |  | 2 |
| 7 | 数据可视化 | 4 | 2 |  | 6 |
| 8 | 概率与统计基础 | 2 |  |  | 2 |
| 9 | 假设检验与统计推断 | 2 | 2 |  | 4 |
| 10 | 回归分析 | 4 | 2 |  | 6 |
| 11 | 主成分分析与数据规约 | 2 | 2 |  | 4 |
| 12 | 机器学习与Python工具 | 2 |  |  | 2 |
| 合计 |  | 32 | 16 |  | 48 |

四、教学内容

1. 第一章 数据科学概述
2. 什么是数据科学
3. 数据科学相关技术
4. 数据科学应用

教学内容：介绍数据科学概念及与其他学科的关系，数据类型与大数据，相关工具、技术、应用案例，理解数据科学的应用领域，思考生活、工作与学习中如何应用数据科学。

重点和难点：数据分析、数据科学相关概念、工具和技术。

考核要求：了解数据科学的相关概念与范畴，理解数据科学的应用范围和前景，如欺诈检测、推荐系统、图像识别、量化投资等，初步了解相关数据类型与大数据相关技术。

复习思考题：数据科学在生活、学习中的应用及伦理问题。

课程思政切入点：提出大数据和数据科学所带来的伦理问题，培养学生树立职业道德和职业伦理理念。

1. 第二章 Python基础
2. Python语言简介
3. Python语法基础
4. Python数据类型
5. Python函数与类
6. 字符串、正则表达式与文件操作

教学内容：Python语言开发环境和相关工具，Python基础语法、数据类型、数据表示、常用函数、函数与类封装、字符串与处理、基本文件操作等，Python 2与Python 3比较，编程常见问题与处理方法。

重点和难点：Python内置数据类型和数据结构，函数与类封装。

考核要求：了解Python的基本语法、基本数据类型、模块和类创建方法，理解不同数据类型的表示方法、区别与用途，掌握基本文本数据的读取、变量存取与变换处理的相关编程方法，能够独立编写Python基本程序并正确运行。

复习思考题：Python内置数据结构列表、元组、集合与字典的特点和区别；Python语言支持哪类编程模式，与Java语言有何区别。

实验作业一：基于随机数构建列表和字典对象，并对其进行存取操作。

1. 第三章 Numpy基础
2. Numpy简介
3. 多维数组与基本操作
4. 数组运算与对齐广播机制
5. Numpy文件读写

教学内容：Numpy模块介绍，多维数组与基本操作，数组的索引、切片与遍历，条件与布尔数组，多维数组的形状操纵、连接与拆分、聚合运算，数组数学运算与广播，基于文件读写数据等。

重点和难点：Numpy多维数组构建、检索、连接与拆分，数组和矩阵运算的对齐与广播机制。

考核要求：理解Numpy的功能特点、常用函数和操作、使用方法等，掌握基于Numpy的多维数组创建和常用操纵方法，能够基于Numpy读写数据并进行矩阵运算。

复习思考题：Numpy数组之间如何进行算数运算，对应的对齐和广播机制是什么；多维数组如何进行连接和拆分。

实验作业二：基于Numpy函数生成随机数、整数序列，构建多个多维数组并对其进行检索、切片、连接和算数运算。

课程思政切入点：基于Python语法和Numpy数据处理所要注意的细节和规范，引导学生养成良好的职业品格和行为习惯，培养学生的社会主义工匠精神。

1. 第四章 Pandas基础
2. Pandas简介
3. Pandas数据结构与操作
4. Pandas数学运算与对齐广播机制

教学内容：Pandas模块介绍与安装，Series与DataFrame数据结构的定义、元素选择、赋值、检索切片、函数运算、过滤操作等，排序、重建索引、删除元素、常用函数等多维数组，Series与DataFrame之间的连接、合并、算数运算等操作。

重点和难点：Series与DataFrame的结构特征和常用操作，Series与DataFrame之间算数运算的对齐与广播机制。

考核要求：了解Pandas常用的数据结构和支持功能，理解Series与DataFrame的结构特征、数据组织形式及与多维数组的区别，掌握Series与DataFrame数据结构的定义和常用方法，能够基于Pandas处理多维数组及二维平面数据。

复习思考题：Series与DataFrame对象之间如何进行数学运算，其对齐与广播机制与Numpy的多维数组对象有何异同；相比多维数组，DataFrame对象的检索切片和转换形状操作有何异同；

实验作业三：基于Numpy函数生成随机数、整数序列，使用Pandas模块构建多个Series和DataFrame对象并对其进行检索、切片、连接和数学运算。

课程思政切入点：基于Python语法、Numpy和Pandas数据处理所要注意的细节和规范，引导学生养成良好的职业品格和行为习惯，培养学生的社会主义工匠精神。

1. 第五章 数据获取与存储
2. 文本文件CSV、JSON等读写操作
3. XML、HTML文件读写
4. 网络爬虫技术
5. 数据库访问

教学内容：CSV与文本文件的读写，文本文件的正则表达式解析，HTML文件的读写，XML文件的读写，微软Excel文件的读写，JSON数据、HDF5格式数据的读写，Pickle的对象序列化与反序列化，关系数据库和非关系数据库的交互操作。

重点和难点：如何保存和恢复对象即序列化与反序列化，字符串与文本对象的正则表达式解析，关系数据库和非关系数据库的交互操作。

考核要求：掌握统计学中的基本方法，如描述性统计、分布、协方差、方差分析、相关性分析等，并能够结合实际业务问题和所提供的案例数据，独立使用Python中的基本工具解决统计学中的上述基本问题。

复习思考题：关系数据库与非关系数据库的区别及Python交互接口差异；保存和恢复对象都有哪些方法（Pickle和对象自带方法）。

实验作业四：读取指定网页数据（带有Table结构的静态网页）和CSV文件数据，将其转换成DataFrame对象，进行基本的清洗和转换操作后存入MySQL关系数据库。

1. 第六章 数据操纵与预处理
2. 数据预处理概述
3. 数据连接、合并与分组聚合运算
4. 数据清洗与数据探索

教学内容：数据合并、缺失值处理、删除重复值、格式转换等数据预处理，数据连接、数据透视表、数据替换、函数映射等，数据离散化与分箱、异常值检测，数据转换、随机抽样、字符串处理、正则表达式，数据聚合、分组与遍历等。

重点和难点：基于Pandas的多数据源连接、合并、聚合，数据的离散化、分箱与随机抽样操作。

考核要求：了解DataFrame对象间数据连接、合并与分组聚合的操作类型，掌握基于Pandas的数据连接、转换、离散化和分箱、重置排列、数据聚合、分组遍历等方法，能够对获取的多类数据源独立进行多种数据预处理操作。

复习思考题：DataFrame的连接、合并、分组聚合等操作与关系数据库的SQL操作有何异同，连接与合并都有哪些方式。

课程思政切入点：提出稍复杂的数据分析和管理决策场景，训练学生学思维方法，提高学生正确认识问题、分析问题和解决问题的能力。

1. 第七章 数据可视化
2. 数据可视化概述
3. 绘图常用模块与使用
4. 数据可视化实操

教学内容：介绍常用绘图模块Matplotlib、Seaborn和Pyecharts等，绘图模块的安装与结构，直方图、散点图、饼图、条形图、线图、3D曲面图等绘制方法、参数设置及保存输出。

重点和难点：Matplotlib的不同接口与提供的各种可视化技术，不同绘图类型选择、绘图分区与图形叠加。

考核要求：了解数据可视化的概念、工具技术、不同绘图类型特点，熟练掌握Matplotlib包的常用可视化工具与调用方法，能够根据提供的数据和案例分析选择合适的技术，并能正确分析处理及可视化展示。

复习思考题：常见绘图类型都有哪些，各有什么特点；Python常用绘图模块都有哪些（除了Matplotlib、Seaborn和Pyecharts等独立绘图模块，某些数据分析模块如DataFrame、Scikit-learn等也自带绘图功能）。

实验作业五：给定数据源文件（带时间列属性），使用Pandas读取数据，使用绘图模块如Matplotlib绘制时间趋势线图、直方图、饼图等，将绘图结果保存到文件，并使用Pandas自身绘制简单图形进行比较。

1. 第八章 概率与统计基础
2. 概率与统计概述
3. 描述性统计相关指标与技术
4. 数据特征与分布
5. 相关工具与技术

教学内容：介绍描述数据特征的描述性统计指标、方差与标准差、协方差与相关性，离散变量与连续变量的分布特征、伯努利分布、二项式分布与正态分布及可视化表示，概率质量函数、概率密度函数、累积分布函数计算与Scipy模块。

重点和难点：离散变量与连续变量的分布特征、函数表示与可视化表示，协方差、相关性与因果关系。

考核要求：掌握统计学中的基本方法，如描述性统计、分布、协方差、方差分析、相关性分析等，并能够结合实际业务问题和所提供的案例数据，独立使用Python中的基本工具解决统计学中的上述基本问题。

复习思考题：伯努利分布与二项式分布有何关系，协方差与相关性有何关系。

课程思政切入点：基于不同维度的视图和统计分析方法，引导学生基于马克思辩证唯物主义思想，辩证看待不同分析方法的优缺点，提高学生正确认识问题的能力。

1. 第九章 假设检验与统计推断
2. 假设检验与统计推断相关概念
3. 零假设与相关检验
4. 相关工具与技术

教学内容：介绍假设检验基本概念和相关指标，零假设、p-value、中心极限定理、z检验、学生t检验，假设检验中的错误类型，A/B测试，相关scipy模块功能。

重点和难点：假设检验的显著性指标与零假设，假设检验错误类型与混淆矩阵。

考核要求：理解假设检验的基本概念和相关指标，掌握scipy模块相关功能及混淆矩阵。

复习思考题：什么是混淆矩阵，在药物测试及司法判定中两类错误的影响如何；零假设与显著性指标的关系。

实验作业六：读取给定实验数据文件，分别给出数值列和字符列的描述性统计，使用numpy或Pandas计算输出协方差和相关性系数矩阵，对相关性显著的列变量提出假设并基于scipy模块进行检验，最后绘出某列数据的直方图和线图（观察是否为正态分布）。

1. 第十章 回归分析
2. 线性回归方法
3. 逻辑回归方法
4. 其他回归方法
5. Python统计分析与机器学习模块

教学内容：介绍简单线性回归、多元线性回归、非线性回归相关概念、理论和方法，Python相关模块statsmodels、sklearn的导入、使用和分析，分析结果指标的解释和输出，过度拟合、拟合不足、模型泛化与偏差-方差权衡，样本集拆分与交叉验证等。

重点和难点：多元线性回归和逻辑回归与Python模块实现，回归分析结果解释。

考核要求：了解线性回归与非线性回归的概念、方法、相关指标及拟合常见问题，掌握Python相关模块如statsmodels、sklearn、scipy的使用和编程，能够独立对商业数据选择合适的Python模块进行回归分析与结果指标解读。

复习思考题：statsmodels和sklearn进行回归分析的使用方法有何不同；为何会有过度拟合和拟合不足问题。

实验作业七：读取给定实验数据文件，分别给出数值列和字符列的描述性统计和相关性系数矩阵，选定因变量进行多元线性回归或逻辑回归，分别使用statsmodels和sklearn模块进行分析并比较分析结果。

课程思政切入点：指出华人在统计和计量领域的学术贡献，如邹至庄检验等，培养学生的爱国主义情怀，激发学生科技报国的家国情怀和使命担当。

1. 第十一章 主成分分析与数据规约
2. 数据规范化概念与方法技术
3. 数据规约与降维
4. Python相关工具与技术

教学内容：介绍数据规范化、标准化、约减（降维）、压缩、特征提取等概念和方法，数据降维方法分类、图像压缩方法，因子分析法与主成分分析PCA方法，奇异值分解SVD等。

重点和难点：主成分分析PCA方法的原理和实现，SVD图像压缩方法。

考核要求：了解数据规约的常用方法、原理和工具技术，掌握基于Python的数据标准化、数据降维及图像压缩方法，能够独立对提供的商业数据选择合适的数据规约方法并编程实现。

复习思考题：如何基于PCA方法提取数据的主要特征；什么是验证性因子分析与探索性因子分析，Python中的PCA方法属于哪类。

实验作业八：使用sklearn.datasets模块自动生成指定特征数（大于3）的大量数据集，进行主成分分析提取特征维度，并将数据进行投影绘制散点图显示。

课程思政切入点：基于主成分分析抽取主要特征和降维约简的方法效果，引导学生基于马克思辩证唯物主义思想，辩证分析问题，化繁为简，抓住主要矛盾，提高学生分析问题和解决问题的能力。

1. 第十二章 机器学习与Python工具
2. 机器学习简介
3. 机器学习相关模块和技术
4. 分类、聚类和神经网络常用算法

教学内容：介绍机器学习的概念、分类、方法和Python相关模块，scikit-learn、TensorFlow、keras、pytorch等模块定位、功能与安装使用，分类和聚类相关算法、神经网络、深度学习等相关知识。

重点和难点：机器学习分类和常用算法，Python工具对应的解决方案。

考核要求：通过本部分的学习，使学生了解机器学习的概念、分类与应用，理解不同机器学习算法类别的差异、用途、相应问题及评价方法（如过度拟合与拟合不足、正确性等）

复习思考题：机器学习与传统的统计计量有何关系，应用有何差异；有监督学习和无监督学习常见的算法有哪些。

五、考核方式、成绩评定

本课程采用平时作业、期末开卷考试相结合的考核方式，尤其注重实践环节的考核。各部分所占比重为平时成绩占50%，期末考试成绩占50%。

平时成绩包括考勤成绩（10%）和作业成绩（40%）。

六、主要参考书及其他内容

[1] 韦斯·麦金尼（Wes McKinney） 著，徐敬一 译．利用Python进行数据分析（原书第2版）．北京：机械工业出版社．2018.08

[2] 赫苏斯·罗格尔-萨拉查（Jesús Rogel-Salazar） 著，白皓 刘江一 上官明乔 刁娟译 译．数据科学与分析：Python语言实现．北京：机械工业出版社．2019.04

[3] 杰克·万托布拉斯（Jake VanderPlas），陶俊杰 陈小莉译.Python数据科学手册.北京：人民邮电出版社.2018.02

[4] Fabio Nelli.Python Data Analytics(2nd Edition).美国加州：Apress.2018.10

执笔人：付东普 教研室主任：胡磊　　　　　系教学主任审核签名：