

《数据采集与数据预处理》课程教学大纲

一、课程简介

课程中文名	数据采集与数据预处理				
课程英文名	Data collection and data preprocessing			双语授课	<input type="checkbox"/> 是 <input checked="" type="checkbox"/> 否
课程代码	06122178	课程学分	2	总学时数	32
课程类别	<input type="checkbox"/> 通识教育课程 <input type="checkbox"/> 公共基础课程 <input checked="" type="checkbox"/> 专业教育课程 <input type="checkbox"/> 综合实践课程 <input type="checkbox"/> 教师教育课程	课程性质	<input type="checkbox"/> 必修 <input checked="" type="checkbox"/> 选修 <input type="checkbox"/> 其他	课程形态	<input type="checkbox"/> 线上 <input checked="" type="checkbox"/> 线下 <input type="checkbox"/> 线上线下混合式 <input type="checkbox"/> 社会实践 <input type="checkbox"/> 虚拟仿真实验教学
考核方式	<input type="checkbox"/> 闭卷 <input checked="" type="checkbox"/> 开卷 <input type="checkbox"/> 课程论文 <input type="checkbox"/> 课程作品 <input checked="" type="checkbox"/> 汇报展示 <input type="checkbox"/> 报告 <input checked="" type="checkbox"/> 课堂表现 <input type="checkbox"/> 阶段性测试 <input checked="" type="checkbox"/> 平时作业 <input type="checkbox"/> 其他（可多选）				
开课学院	财经学院		开课系(教研室)	金融系	
面向专业	金融工程		开课学期	第6学期	
课程负责人			审核人	余函	
先修课程	python程序设计				
后续课程	数据挖掘、数据挖掘实训等。				
选用教材	大数据采集、预处理与可视化. 葛继科、张晓琴、陈祖琴.人民邮电出版社, 2023年7月.				
参考书目	数据采集与预处理. 林子雨. 人民邮电出版社, 2022年2月.				
课程资源	www.python.org				
课程简介	本课程详细阐述了大数据领域数据采集与预处理的相关理论和技术。课程内容共8章, 包括概述、大数据实验环境搭建、网络数据采集、分布式消息系统Kafka、日志采集系统Flume、数据仓库中的数据集成、ETL工具Kettle、使用pandas进行数据清洗。				

二、课程目标

表 2-1 课程目标

序号	具体课程目标
课程目标 1	知识与理解：使学生掌握数据采集与数据预处理的基本概念、原理和方法，包括数据的来源、类型、结构等基础知识。理解数据采集过程中可能遇到的问题以及数据预处理的重要性，包括数据质量、数据完整性和数据一致性等方面的考虑。
课程目标 2	实际操作与应用：通过实验和项目实践，使学生能够在实际场景中运用所学知识和技能进行数据采集与预处理操作，解决实际问题。培养学生具备分析和评估数据采集与预处理结果的能力，能够根据实际需求调整和优化数据采集和预处理策略。
课程目标 3	创新思维与问题解决：鼓励学生运用创新思维解决数据采集与预处理过程中遇到的问题，探索新的方法和策略，提高数据处理效率和准确性。

表2-2 课程目标与毕业要求对应关系

毕业要求	指标点	课程目标
毕业要求3：专业知识【M】	3.1：具备扎实的数学、计算机、统计学基础知识。	课程目标1
毕业要求4：专业能力【M】	4.3：能运用所学的经济学、金融学、金融工程知识对经济环境进行分析、预测，并能将其应用于今后的业务决策中。	课程目标2
毕业要求8：终身学习【M】	8.3：掌握反思方法，运用批判性思维方法，对自身岗位实践、职业发展中的现实需要和问题进行探索实践。	课程目标3

三、课程学习内容与方法

(一) 理论学习内容及要求

表3-1 课程目标、学习内容和教学方法对应关系

序号	课程模块	学习内容	课程目标	学习重点难点	教学方法	学时
1	大数据概述	1. 了解大数据时代的特点	课程目标 1、2、3	1、重点：大数据相关概念及特征、 大数据思维 2、难点：大数据系统	讲授法	4
		2. 理解大数据的概念、特征及作用				
		3. 熟悉大数据系统				
		4. 具备大数据思维能力，了解大数据伦理及安全				
2	Python程序设计	1. 使学生掌握Python的安装，熟练掌握数据类型与运算符，熟悉程序控制结构，熟练掌握函数与模块的应用	课程目标 1、2、3	1、重点：数据类型与运算符、程序控制结构 2、难点：函数与模块、文件	讲授法	4
		2. 掌握文件读写操作方法，为应用Python程序设计语言进行数据采集、预处理及可视化奠定基础。				
3	大数据采集技术	1、大数据采集概述	课程目标 1、2、3	1、重点：网络爬虫技术 2、难点：大数据采集平台	讲授法	4
		2、网络爬虫技术				
		3、数据抽取技术				
		4、数据采集实例				

4	大数据预处理技术	1. 数据预处理概述	课程目标 1、2、3	1、重点：数据清洗、数据集成 2、难点：数据归约、数据脱敏	讲授法 案例法	4
		2. 数据清洗				
		3. 数据集成				
		4. 数据变换				
		5. 数据归约				
		6. 数据脱敏				
5	Excel数据获取与预处理	1. Excel数据获取	课程目标 1、2、3	1、重点：数据获取、数据清洗与转换 2、难点：数据抽取与合并	讲授法	4
		2. Excel数据清洗与转换				
		3. 数据抽取与合并				
		4. Excel数据获取与预处理综合实例				
6	Python数据预处理	1. Python数据预处理基础	课程目标 1、2、3	1、重点：数据的分组、分割、合并和变形，缺失值、异常值和重复值处理 2、难点：文本数据分析	讲授法	6
		2. 数据的分组、分割、合并和变形				
		3. 缺失值、异常值和重复值处理				
		4. 文本数据分析				
7	数据可视化技术	1. 数据可视化概述	课程目标 1、2、3	1、重点：Python数据可视化方法 2、难点：Pyecharts数据可视化方法	讲授法 案例法	6
		2. 数据可视化的理论基础				
		3. Python数据可视化方法				
		4. Pyecharts数据可视化方法				
		2.apriori算法				

		3.协同过滤（基于item）				
		3.使用NLTK进行文本分析案例				

四、课程考核

（一）考核内容与考核方式

表4-1 课程目标、考核内容与考核方式对应关系

课程目标	考核内容	所属学习模块/项目	考核占比	考核方式
课程目标 1	大数据概述	1	20%	开卷考试
	Python程序设计	2		
课程目标 2	Excel数据获取与预处理	5	50%	开卷考试
	Python数据预处理	6		
	数据可视化技术	7		
课程目标 3	大数据采集技术	3	30%	开卷考试
	大数据预处理技术	4		

表4-2 课程目标与考核方式矩阵关系

课程目标	考核方式				考核占比（此处需与上表基本一致）
	开卷考试60%	课堂表现10%	汇报展示10%	作业完成20%	
课程目标1	20%	20%	20%	20%	$20\%=60\%*20\%+10\%*20\%+10\%*20\%+20\%*20\%$
课程目标2	50%	50%	50%	50%	$50\%=60\%*50\%+10\%*50\%+10\%*50\%+20\%*50\%$
课程目标3	30%	30%	30%	30%	$30\%=60\%*30\%+10\%*30\%+10\%*30\%+20\%*30\%$

评分依据：

（二）成绩评定

1.平时成绩评定

（1）课堂表现（25分）：通过学生在课堂上回答问题情况，来评价学生对相关的知识的掌握水平。

（2）作业完成（50分）：围绕课程的学习目标进行作业的设计，通过课后练习的形式给予一定的评价，训练学生的数据采集与预处理思维方法和思维能力。

（3）汇报展示（25分）：通过学生对自行设计的程序进行展示，来评价学生对相关知识的应用能力。

2.期末成绩评定

期末考核主要考察学生对机器学习基本概念、基本原理和基本方法的理解与运用等，考核内容为教学大纲所要求内容，重点考察学生的数据采集与预处理基础知识的掌握。要求学生掌握数据采集与预处理基本概念、基本原理和方法，运用python等工具解决实际应用问题。方式为开卷考试。

3.总成绩评定

总成绩（100%）=平时成绩（40%）+期末成绩（60%）

（三）评分标准

表5 评分标准（非试卷考核项目）

考核项目	评分标准				
	优秀 (100>x≥90)	良好 (90>x≥80)	中等 (80>x≥70)	及格 (70>x≥60)	不及格 (x<60)
作业完成	(1) 按时全部完成很好 (40%) (2) 书写工整 (20%) (3) 答案正确 (40%)	(1) 按时完成较好 (40%) (2) 书写工整 (20%) (3) 答案大部分正确 (40%)	(1) 按时全部完成一般 (40%) (2) 书写一般 (20%) (3) 答案基本正确 (40%)	(1) 按时完成大部分作业差 (40%) (2) 书写一般 (20%) (3) 答案基本正确 (40%)	(1) 不能按时完成 (40%) (2) 书写潦草 (20%) (3) 错误较多 (40%)
课堂表现	(1) 课堂主动回答问题、回答正确，且能进行解释 (50%) (2) 提问、讨论发言观点正确，问题有深度、有创新 (50%)	(1) 课堂主动回答问题，回答正确，但解释欠清楚 (50%) (2) 提问、讨论发言观点正确，但问题无深度或无创新 (50%)	(1) 课堂回答问题大部分正确，且不能解释 (50%) (2) 提问、讨论发言观点基本正确，但问题无深度、无创新 (50%)	(1) 课堂测验、回答问题错误率在30~50%之间，且不能解释 (50%) (2) 提问、讨论发言观点有部分错误 (50%)	(1) 课堂测验、回答问题错误率超过50%，且不能解释 (50%) (2) 提问、讨论发言观点错误，思路不清晰，逻辑不严密 (50%)
汇报展示	(1) 汇报学生亲切大方，语言有感染力，语言简洁易懂，无口头禅；抑扬顿挫，富有节奏；肢体语言运用适当。(20%)	(1) 汇报学生亲切大方，语言较有感染力，语言较简洁易懂，语言节奏感较强；肢体语言运用较适当。	(1) 汇报语言感染力及语言简洁度一般，语言节奏一般；肢体语言运用一般。(20%) (2) 课件制作	(1) 汇报学生语言有感染力较差，语言不够简洁，肢体语言运用不够适当。(20%) (2) 课件制作	(1) 汇报学生没有语言有感染力，语言不够简洁易懂，肢体语言运用不适当。(20%) (2) 课件制作

	<p>(2) 课件制作精美, 汇报逻辑性强, 过渡自然, 注重语言运用能力的培养。所呈现的容量和难度恰当。讲授的知识观点正确, 无知识性错误, 份量适中。(50%)</p> <p>(3) 学生参与度高, 活动积极, 思维活跃, 情绪饱满, 注意力集中。按时完成汇报任务(30%)</p>	<p>(20%)</p> <p>(2) 课件制作较好, 汇报逻辑性较强, 语言运用能力较强。所呈现的容量和难度适当。讲授的知识观点正确, 无知识性错误, 份量适中。(50%)</p> <p>(3) 学生参与度高, 情绪较饱满, 注意力比较集中。按时完成汇报任务(30%)</p>	<p>一般, 汇报逻辑性不强, 所呈现的容量和难度恰当。讲授的知识观点基本正确, 全面掌握知识能力不够, 份量适中。(50%)</p> <p>(3) 学生参与度一般, 思维较活跃, 注意力集中度一般。按时完成汇报任务(30%)</p>	<p>粗糙, 汇报逻辑性较差, 所呈现的容量和难度不够恰当。讲授的知识观点不完全正确, 部分有知识性错误。(50%)</p> <p>(3) 学生参与度较差, 活跃度差, 情绪不够饱满, 注意力不够集中。基本能完成汇报任务(30%)</p>	<p>差, 汇报逻辑性差, 所呈现的容量和难度不恰当。讲授的知识观点不正确, 有知识性错误,(50%)</p> <p>(3) 学生参与度差, 未按时完成汇报任务(30%)</p>
--	---	---	---	---	---

五、其它说明

本课程大纲依据2023版金融工程专业人才培养方案, 由财经学院金融系讨论制定, 财经学院教学工作委员会审定, 教务处审核批准, 自2023级开始执行。