《数据分析与挖掘》教学大纲

**课程名称：**数据分析与挖掘

**课程类别：**必修

**适用专业：**大数据技术类相关专业

**总学时：**64学时（其中理论48学时，实验16学时）

**总学分：**4.0学分

# 课程的性质

随着云时代的来临，大数据技术将具有越来越重要的战略意义。大数据已经渗透到每一个行业和业务职能领域，逐渐成为重要的生产要素，人们对于海量数据的运用将预示着新一轮生产率增长和消费者盈余浪潮的到来。大数据分析技术将帮助企业用户在合理时间内攫取、管理、处理、整理海量数据，为企业经营决策提供积极的帮助。大数据分析作为数据存储和挖掘分析的前沿技术，广泛应用于物联网、云计算、移动互联网等战略性新兴产业。虽然大数据目前在国内还处于初级阶段，但是其商业价值已经显现出来，特别是有实践经验的大数据分析人才更是各企业争夺的热门。为了满足日益增长的大数据分析人才需求，特开设Python数据分析与挖掘实战课程。

# 课程的任务

通过本课程的学习，使学生学会使用Python语言进行数据探索、数据预处理、分类与预测、聚类分析、时序预测、关联规则挖掘、智能推荐、偏差检测等操作，并完成大量数据挖掘工程案例，将理论与实践相结合，让学生熟练掌握使用Python语言对样本数据进行处理、挖掘建模，为将来从事数据分析研究、工作奠定基础。

# 课程学时分配

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 教学内容 | 理论学时 | 实验学时 | 其它 |
| 1 | 第1章 数据挖掘基础 | 1 | 0 |  |
| 2 | 第2章 Python数据分析简介 | 2 | 1 |  |
| 3 | 第3章 数据探索 | 3 | 3 |  |
| 4 | 第4章 数据预处理 | 3 | 3 |  |
| 5 | 第5章 挖掘建模 | 13 | 11 |  |
| 6 | 第6章 财政收入影响因素分析及预测模型 | 3 | 5 |  |
| 7 | 第7章 航空公司客户价值分析 | 3 | 4 |  |
| 8 | 第8章 商品零售购物篮分析 | 3 | 4 |  |
| 9 | 第9章 基于水色图像的水质评价 | 3 | 4 |  |
| 10 | 第10章 家用热水器用户行为分析与事件识别 | 5 | 9 |  |
| 11 | 第11章 电子商务网站用户行为分析及服务推荐 | 3 | 8 |  |
| 12 | 第12章 电商产品评论数据情感分析 | 5 | 9 |  |
| 13 | 第13章 基于Python引擎的开源数据挖掘建模平台（TipDM） | 2 | 2 |  |
|  | 总计 | 49 | 63 |  |

# 教学内容及学时安排

### 理论教学

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **序号** | **章节名称** | **主要内容** | **教学目标** | **学时** |
| 1 | 数据挖掘基础 | 1. 掌握数据挖掘的流程
2. 了解常用的数据挖掘建模工具
 | 掌握数据挖掘的流程 | 1 |
| 2 | Python数据分析简介 | 1. 掌握Python开发平台的搭建方法
2. 了解Python的运行方式、基本命令
3. 了解Python的数据结构
4. 了解Python数据分析工具
5. 了解配套附件使用设置
 | 1. 掌握Python开发平台的搭建
2. 了解Python的运行方式、基本命令、数据结构以及库的导入与添加
3. 熟悉Python数据挖掘相关库的安装和使用方法
4. 了解书本配套附件的使用
 | 2 |
| 3 | 数据探索 | 1. 掌握缺失值分析的方法
2. 掌握异常值分析的方法
3. 掌握一致性分析的方法
4. 掌握分布分析的方法
5. 掌握对比分析的方法
6. 掌握统计量分析的方法
7. 掌握周期性分析的方法
8. 掌握贡献度分析的方法
9. 掌握相关性分析的方法
10. 了解基本的统计特征函数
11. 了解统计绘图函数
 | 1. 掌握Python中数据质量分析的方法
2. 掌握Python中数据特征分析的方法
3. 熟悉Python主要数据探索函数
 | 3 |
| 4 | 数据预处理 | 1. 掌握缺失值、异常值处理方法
2. 了解常见的数据集成的方法
3. 掌握数据规范化和离散化的方法
4. 了解属性构造的方法
5. 了解小波变换的方法
6. 掌握属性归约和数值归约的方法
7. 了解Python主要数据预处理函数
 | 1. 掌握Python中数据清洗的方法
2. 掌握Python中数据集成的方法
3. 掌握Python中数据归约的方法
4. 了解Python主要数据预处理函数
 | 3 |
| 5 | 挖掘建模 | 1. 了解常用的分类与预测算法
2. 了解回归分析
3. 掌握决策树的原理
4. 掌握人工神经网络的原理
5. 掌握分类与预测算法的评价方法
6. 了解Python分类预测模型特点
7. 了解常用聚类分析算法
8. 掌握K-Means聚类算法的原理
9. 掌握聚类分析算法的评价方法
10. 了解Python主要聚类分析算法
11. 掌握Apriori算法的原理
12. 掌握时间序列预处理的方法
13. 了解平稳时间序列分析
14. 掌握非平稳时间序列分析
15. 了解Python主要时序模式算法
16. 了解离群点检测方法
17. 了解基于模型的离群点检测方法
18. 了解基于聚类的离群点检测方法
 | 1. 掌握Python中分类与预测的方法
2. 掌握Python中聚类分析的方法
3. 掌握Python中关联规则分析的方法
4. 掌握Python中时序模式的分析方法
5. 掌握Python中离群点的检测方法
 | 13 |
| 6 | 财政收入影响因素分析及预测模型 | 1. 分析财政收入预测背景
2. 了解财政收入预测的方法
3. 熟悉财政收入预测的步骤与流程
4. 了解描述性统计分析
5. 分析计算结果
6. 了解相关性分析
7. 分析计算结果
8. 了解Lasso回归方法
9. 分析Lasso回归结果
10. 了解灰色预测算法
11. 了解SVR算法
12. 分析预测结果
 | 1. 了解财政收入预测的背景知识，分析步骤和流程
2. 掌握相关性分析方法与应用
3. 掌握用Lasso模型特征选取方法
4. 掌握灰色预测和支持向量回归算法的原理与应用
 | 3 |
| 7 | 航空公司客户价值分析 | 1. 分析航空公司现状
2. 认识客户价值分析
3. 熟悉航空客户价值分析的步骤与流程
4. 描述性统计分析
5. 相关性分析
6. 处理缺失值与异常值
7. 标准化L、R、F、M、C这5个特征
8. 了解K-Means聚类算法
9. 分析聚类结果
10. 模型应用
 | 1. 了解RFM模型的基本原理
2. 掌握K-Means算法的基本原理与使用方法
3. 比较不同类别客户的客户价值，制定相应的营销策略
 | 3 |
| 8 | 商品零售购物篮分析 | 1. 分析零售企业商品销售现状
2. 了解某商品零售企业的基本数据情况
3. 熟悉购物篮分析的基本流程与步骤
4. 分析热销商品
5. 分析售出商品的结构
6. 了解Apriori算法的基本原理与使用方法
7. 构建零售商品的Apriori模型
8. 根据模型结果提出商品销售策略
 | 1. 熟悉购物篮分析的实现流程与步骤
2. 掌握Apriori算法的基本原理与使用方法
3. 分析商品销售状况与商品结构合理性
4. 分析零售商品间的关联关系
 | 3 |
| 9 | 基于水色图像的水质评价 | 1. 了解水质评价背景与挖掘目标
2. 熟悉水质评价的分析步骤与流程
3. 图像切割
4. 特征提取
5. 构建决策树水质评价分类模型
6. 水质评价
 | 1. 掌握图像处理算法中的颜色矩提取
2. 掌握决策树算法的使用
3. 了解决策树算法在水质评价实例中的应用
 | 3 |
| 10 | 家用热水器用户行为分析与事件识别 | 1. 分析家用热水器行业现状
2. 熟悉家用热水器用户行为分析的步骤与流程
3. 分析热水器的水流量状况
4. 删除冗余特征
5. 划分用水事件
6. 确定单次用水事件时长阈值
7. 构建用水时长与频率特征
8. 构建用水量与波动特征
9. 筛选候选洗浴事件
10. 模型构建
11. 评估模型
 | 1. 掌握使用Python进行数据预处理的方法
2. 掌握数据转换，属性提取过程
3. 掌握BP神经网络算法的使用
4. 了解BP神经网络算法在用户行为分析与事件识别案例中的应用
 | 5 |
| 11 | 电子商务网站用户行为分析及服务推荐 | 1. 了解智能推荐服务应用场景
2. 了解某法律网站的基本情况
3. 分析原始数据用户点击的网页类型，得到统计分析结果
4. 根据原始数据用户浏览网页次数的情况进行统计分析
5. 通过原始数据用户在浏览页面时的情况得到网页排名的统计分析
6. 清除数据探索分析过程中发现与目标无关的数据
7. 识别翻页的网址，并对其进行还原，然后对用户访问的页面进行去重操作
8. 对数据进行手动网址分类，对处理后的数据进行特征选取
9. 基于物品的协同过滤算法，计算出物品之间的相似度
10. 根据物品的相似度和用户的历史行为给用户生成推荐列表
11. 对模型进行评价，判断推荐系统的好坏
 | 1. 熟悉网站智能推荐的步骤与流程
2. 掌握简单的统计分析方法，运用于网页类型与点击次数的统计
3. 掌握对某网站数据进行预处理的方法，包括数据去重、数据变换和特征选取
4. 掌握使用协同过滤算法对某网站进行智能推荐的方法
 | 3 |
| 12 | 电商产品评论数据情感分析 | 1. 了解电商企业现状
2. 了解网络上发布内容的技术和Web文档中提取信息的技术
3. 对电商产品评论数据进行预处理，包括行分词处理，去除停用词，查看分词效果
4. 基于情感词表进行情感词匹配，并对情感词的倾向进行修正
5. 对情感分析结果进行检验
6. 了解主题模型，以及LDA模型原理与参数估计方法
7. 掌握寻找最优主题数的方法
8. 建立相应的LDA模型
9. 进行LDA主题分析
 | 1. 熟悉电商产品评论数据实现情感分析的步骤与流程
2. 了解如何使用Python语言对互联网信息进行爬取
3. 掌握文本分析的预处理方法
4. 对预处理后的评论数据进行情感分析
5. 使用LDA模型对正、负面评论数据进行主题分析
 | 5 |
| 13 | 基于Python引擎的开源数据挖掘建模平台（TipDM） | 1. 平台简介
2. 导入数据
3. 配置输入源组件
4. 配置缺失值处理组件
5. 配置纪录选择组件
6. 配置数据标准化组件
7. 配置K-Means组件
 | 1. 了解TipDM数据挖掘建模平台的模板、数据源、工程及系统组件
2. 了解TipDM数据挖掘建模平台的本地化部署的步骤
3. 掌握快速构建数据挖掘工程的方法
 | 2 |
| 学 时 合 计 | 49 |

### 实验教学

|  |  |  |  |
| --- | --- | --- | --- |
| **序号** | **实验项目名称** | **实验要求** | **学时** |
| 1 | Python数据分析简介 | 1. 在Windows系统上安装Anaconda
2. 运行Python基本命令的程序
3. 导入库
4. 添加第三方库
 | 1 |
| 2 | 数据探索 | 1. 使用Python进行数据质量分析
2. 使用Python进行数据特征分析
3. 使用Python进行特征统计
4. 使用Python进行绘图
 | 3 |
| 3 | 数据预处理 | 1. 使用Python进行数据清洗
2. 使用Python进行数据集成
3. 使用Python进行数据归约
 | 3 |
| 4 | 挖掘建模 | 1. 使用逻辑回归算法进行分类预测
2. 使用决策树算法进行分类预测
3. 使用人工神经网络算法进行分类预测
4. 使用K-Means进行聚类分析
5. 使用TSNE进行数据降维并展示聚类结果
6. 使用Apriori算法进行关联分析
7. 使用ARIMA模型进行时间序列分析
8. 使用Python离散点检测
 | 11 |
| 5 | 财政收入影响因素分析及预测 | 1. 分析财政收入数据特征的相关性
2. 使用Lasso回归选取财政收入预测的关键特征
3. 使用灰色预测和SVR构建财政收入预测模型
4. 评价SVR模型
 | 5 |
| 6 | 航空公司客户价值分析 | 1. 处理数据缺失值与异常值
2. 构建航空客户价值分析的关键特征
3. 标准化LRFMC 5个特征
4. 构建K-Means聚类模型
5. 评价K-Means聚类模型
 | 4 |
| 7 | 商品零售购物篮分析 | 1. 使用统计学知识分析热销商品
2. 使用商品结构图分析售出商品的结构
3. 使用apriori函数构建零售商品的Apriori模型
 | 4 |
| 8 | 基于水色图像的水质评价 | 1. 图像切割和特征提取
2. 数据划分
3. 使用决策树算法构建水质评价分类模型
4. 使用混淆矩阵和准确率进行水质评价
 | 4 |
| 9 | 家用热水器用户行为分析与事件识别 | 1. 探索分析热水器的水流量状况
2. 删除冗余属性
3. 划分用水事件
4. 确定单次用水事件时长阈值
5. 构建用水行为特征
6. 筛选候选洗浴事件
7. 构建BP神经网络模型
8. 评价BP神经网络模型
 | 9 |
| 10 | 电子商务网站用户行为分析及服务推荐 | 1. 使用pymysql库中的pm连接数据库
2. 统计101、107和199等网页类型
3. 统计用户在网页的点击次数
4. 分析网页排名
5. 删除不符合规则的网页
6. 还原翻页网址
7. 划分正确的网页类型
8. 将处理后的数据转换成0-1二元型数据
9. 构建智能推荐模型
10. 评价智能推荐模型
 | 8 |
| 11 | 电商产品评论数据情感分析 | 1. 去除评论数据的数字、字母
2. 对评论数据进行去重
3. 对评论数据进行分词处理
4. 根据停用词库去除评论文本中的停用词
5. 绘制词云图，查看分词效果
6. 将情感词表与分词结果进行匹配
7. 对情感值的方向进行修正，并计算情感分析的准确率
8. 分别对正面评论和负面评论绘制词云，查看情感分析效果
9. 对正面情感词与负面情感词构建语料库，并建立文档-词条矩阵
10. 使用LDA主题模型，找出不同主题数下的主题词，寻找最优主题数
11. 输入正面情感与负面情感评论求解LDA模型，并分析结果
 | 9 |
| 12 | 基于Python引擎的开源数据挖掘建模平台（TipDM） | 1. TipDM数据挖掘建模平台的本地化部署
2. 导入数据
3. 配置输入源组件
4. 配置缺失值处理组件
5. 配置记录选择组件
6. 配置数据标准化组件
7. 配置K-Means组件
 | 2 |
| 学 时 合 计 | 63 |

# 考核方式

突出学生解决实际问题的能力，加强过程性考核。课程考核的成绩构成 = 平时作业（10%）+ 课堂参与（20%）+ 期末考核（70%），期末考试建议采用开卷形式，试题应包括基本概念、数据探索、数据预处理、分类与预测、聚类分析、时序预测、关联规则挖掘、智能推荐、偏差检测等部分，题型可采用判断题、选择、简答、应用题等方式。

# 教材

### 教材

张良均，谭立云，刘名军，江建明．《Python数据分析与挖掘实战（第2版）》[M]．北京：机械工业出版社．2019．