

**《大数据处理与分析》**

**指导书**

杨鲲鹏 编写

**信息工程学院计算机系**

**2022年 6 月**

目 录

[**1．课程教学大纲要求** 3](#_Toc106258604)

[**1.1 课程简介** 3](#_Toc106258605)

[**1.2 课程教学目标** 3](#_Toc106258606)

[**1.3 课程教学目标的考核方式** 4](#_Toc106258607)

[**2. 设计类型与选题任务要求** 4](#_Toc106258608)

[**2.1 选题类型** 4](#_Toc106258609)

[**2.2 选题指南** 4](#_Toc106258610)

[**3．课程教学目标的评价标准** 5](#_Toc106258611)

[**3.1 阶段成绩（25分）** 6](#_Toc106258612)

[**3.2 设计验收答辩（50分）** 6](#_Toc106258613)

[**3.3 设计报告（25分）** 6](#_Toc106258614)

# **1．课程教学大纲要求**

## **1.1 课程简介**

大数据处理与分析是数据科学与大数据技术专业的一门必修课，属于专业实践环节。该课程基于实践驱动的方式讲解大规模数据的清洗，转化，处理与分析的整个流程，教学目标在于使学生掌握大规模数据的处理与分析解决方案，主要包括大规模数据的清洗与转化技术，大规模数据的处理与方法,同时要求学生能够设计可行的试验方案，并有所创新。

## **1.2 课程教学目标**

### 1.2.1 课程教学目标

教学目标1：能够基于大数据平台，分析和设计特定需求下的大规模数据分析的解决方案，并有所创新；

教学目标2：能够基于大数据平台和方法，选择适用的大规模数据清洗、转换等处理技术，并设计可行的实验方案；

教学目标3：能够理解个人及负责人在团队中的角色划分，且胜任所担职责；

教学目标4：能够针对大规模数据处理与分析要点，合理撰写设计报告和呈现效果；

教学目标5：能够基于发展动态和行业需求，合理应用新颖的大数据处理与分析方法。

### 1.2.2 课程教学目标对毕业要求指标点的支撑关系

|  |  |  |  |
| --- | --- | --- | --- |
| **毕业要求** | **毕业要求2级指标点** | **课程教学目标** | **本课程对指标点的权重系数** |
| 3. 设计/开发解决方案 | 3.4 能够针对特定需求，对大数据复杂工程问题进行分解和细化，具有大规模数据分析、智能计算和可视化展现的设计和实现能力，并体现创新意识。 | 1 | 0.2 |
| 4. 研究 | 4.1 能够基于数据科学与大数据技术学科相关原理和方法，选择研究路线，设计可行的实验方案。 | 2 | 0.3 |
| 9. 个人和团队 | 9.2 能够理解个人及负责人在团队中的角色划分，且胜任相应的角色职责。 | 3 | 0.2 |
| 10. 沟通 | 10.2 能够就数据科学与大数据技术领域复杂工程问题与业界同行及社会公众进行有效沟通和交流，包括撰写报告和设计文稿、陈述发言、清晰表达或回应指令。 | 4 | 0.1 |
| 12. 终身学习 | 12.2 能够追踪数据科学与大数据技术领域发展动态和行业需求，有不断学习和适应发展的能力。 | 5 | 0.2 |

## **1.3 课程教学目标的考核方式**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 课程 | | 大数据处理与分析 | | | | | | |
| 教学目标 | | 1 | 2 | 3 | 4 | 5 | ∑ | |
| 课程对指标点的权重系数 | | 0.2 | 0.3 | 0.2 | 0.1 | 0.2 | 1.0 | |
| 教学及考核环节 | | 教学及考核环节对指标点的权重系数 | | | | | 权重∑ | 成绩 |
| 方式 | 权重 |
| 阶段考核 | 0.25 |  | 0.6 | 0.4 |  |  | 1 |  |
| 设计验收 | 0.50 | 0.2 | 0.5 | 0.2 | 0.1 |  | 1 |  |
| 设计报告 | 0.25 | 0.6 |  |  | 0.2 | 0.2 | 1 |  |
| ∑ | 1.0 | 0.8 | 1.1 | 0.6 | 0.3 | 0.2 | 3 |  |
| 达成度 | |  |  |  |  |  |  |  |

# **2. 设计类型与选题任务要求**

## **2.1 选题类型**

本课程设计属于大数据处理与分析主题，其选题内容及数据直接来自大数据竞赛平台，主题涉及电子商务、语音识别、计算机视觉、舆情分析等诸多领域，学生小组可根据兴趣自行选择选题。

## **2.2 选题指南**

如下表所示，共有8 道题。学生小组需要选择其中一个大数据处理与分析题目。选定选题后需要提交确认并经过审核。

所有大数据分析预处理的项目选题，均来自于实际的大数据竞赛平台（如Kaggle、阿里云天池等）的实际赛题或相关基础数据。学生小组需要完成的内容包含但不限于对项目的理解、数据分析及可视化、算法模型的分析以及数据分析思路等。对于利用数据分析的基础结果进行进一步的数据挖掘或应用模型进行AI预测等项目需求，仅需要参照课程中提供的若干标准方式或训练模型进行初步尝试即可，完成程度和效果仅作为加分项，不作为项目的基本要求。

选题直接来自相关大数据平台，相关数据需要在相关平台报名参赛获取，也可在选题确定后直接提供。

**表1 大数据处理与分析课程选题指南**

|  |  |  |
| --- | --- | --- |
| 序号号 | 题目名称 | 设计基本要求 |
| 1 | 用户情感可视化分析 | 以网络舆情分析为背景，要求根据用户的评论来对品牌的议题进行数据分析与可视化，并应当通过数据可视化图表和数据分析方法，对感兴趣的内容进行探索性数据分析。 |
| 2 | 产品关联分析 | 以购物车分析为背景，要求对品牌的历史订单数据，挖掘频繁项集与关联规则；并应当利用订单数据进行数据分析，为企业提供销售策略和产品关联组合推荐的初步分析结果，以辅助企业提升销量，为消费者提供更适合的商品推荐 |
| 3 | 汽车产品聚类分析 | 以汽车产品竞品分析为背景，通过数据的聚类，为汽车提供聚类分类。对于指定的车型，可以通过聚类分析找到其竞品车型。进一步地，可以利用车型数据，进行车型画像的分析，为产品的定位、竞品分析提供数据决策支持。 |
| 4 | 新闻推荐场景下的用户行为预测 | 以新闻APP中的新闻推荐为背景，要求根据用户历史浏览点击新闻文章的数据信息预测用户未来点击行为，即用户的最后一次点击的新闻文章。 |
| 5 | 语义分割之地表建筑物识别 | 以计算机视觉为背景，要求对航拍图像进行数据处理与分析，并使用给定的航拍图像训练模型并完成地表建筑物识别任务。 |
| 6 | 语音识别之食物声音识别 | 以语音识别为背景，要求对提供的语音数据进行处理和分析，并使用提供的语音数据训练模型完成语音分类的任务 |
| 7 | 二手车交易价格预测 | 以二手车市场为背景，要求针对交易平台的二手车交易记录数据进行处理和分析，并预测二手汽车的交易价格。 |
| 8 | 自选主题 | 在天池、Kaggle等大数据竞赛平台自选主题，完成数据分析和处理，并针对相应数据处理的数据挖掘或者数据训练任务进行必要的拓展。要求：自选主题需要包括基础的数据分析和处理任务，且提交选题资料进行审核。  除大数据竞赛相关平台外，也可以自选其他主题和数据完成项目。 |

# **3．课程教学目标的评价标准**

采用评价量表法，通过平时考核、演示答辩和设计报告，分别评价课程目标的达成值。

## **3.1 阶段成绩（25分）**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 课程目标 | 2 | | 3 | |
| 满分值 | 15 | | 10 | |
| 评价描述 | 7-15 | 有明确合理的项目大数据处理与分析的研究设计方案，且已开始程序设计 | 5-10 | 有明确合理的项目大数据处理与分析相关软件工具的应用以及应用理由 |
| 0-6 | 缺乏项目大数据处理与分析研究设计方案，还没有进入编程阶段 | 0-4 | 缺乏项目大数据处理与分析相关软件工具的应用思路 |

## **3.2 设计验收答辩（50分）**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 课程目标 | 1 | | 2 | | 3 | | 4 | |
| 满分值 | 10 | | 25 | | 10 | | 5 | |
| 评价描述 | 7-10 | 合理选用大数据处理与分析的算法，理解到位，应用得当 | 15-25 | 有明确合理的大数据处理与分析框架设计和实现方案，程序能够正确运行 | 7-10 | 有明确合理的大数据处理与分析相关软件工具的应用以及应用理由 | 4-5 | 有明确的个人任务和协作要求，且完成效果好 |
| 4-6 | 能够选用大数据处理与分析算法，但不懂原理，应用较差 | 7-14 | 有一定的大数据处理与分析系统设计方案，但部分模块没有实现 | 4-6 | 有一定的大数据处理与分析相关软件工具的应用以及应用理由 | 2-3 | 个人任务和协作要求不够明确，完成效果一般 |
| 0-3 | 不能理解大数据处理与分析算法，应用效果差 | 0-6 | 缺乏大数据处理与分析设计方案，大部分模块没有实现 | 0-3 | 缺乏大数据处理与分析相关软件工具的应用以及应用理由 | 0-1 | 不明确个人任务和协作要求，完成效果不好 |

## **3.3 设计报告（25分）**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 课程目标 | 1 | | 4 | | 5 | | 6 | |
| 满分值 | 10 | | 5 | | 5 | | 5 | |
| 评价描述 | 7-10 | 合理选用大数据处理与分析算法，理解到位，应用得当 | 4-5 | 有明确合理的大数据处理与分析框架设计和实现方案，程序能够正确运行 | 4-5 | 可以有效和别人沟通并完成个人任务和协作要求，且完成效果好 | 4-5 | 完全可以通过自我学习，查阅相关文献创新项目中的亮点 |
| 4-6 | 能够选用大数据处理与分析算法，但不懂原理，应用较差 | 2-3 | 有一定的大数据处理与分析设计方案，但部分模块没有实现 | 2-3 | 基本可以完成和别人沟通并完成个人任务和协作要求，且完成效果好 | 2-3 | 基本可以通过自我学习，查阅相关文献创新项目中的亮点 |
| 0-3 | 不能理解大数据处理与分析算法，应用效果差 | 0-1 | 缺乏大数据处理与分析设计方案，大部分模块没有实现 | 0-1 | 无法完成和别人沟通并完成个人任务和协作要求，且完成效果好 | 0-1 | 无法通过自我学习，查阅相关文献创新项目中的亮点 |