《R语言》课程教学大纲

**总主编 刘鹏 张燕**

**主编 程显毅 刘颖 朱倩**

**适合专业： 数据科学与大数据技术专业 课程编号：**

**先修课程：数理统计、数据库、大数据导论**

**学分: 4 总学时： 64**

# 一、课程性质、目的与要求

**课程性质：专业必修课。**

**课程目的：**本课程是各专业想了解大数据分析技术的学生必修的一门基础课程，具有很强的实践性和应用性。它以《数理统计》、《大数据导论》为基础，主要培养学生大数据分析和计算机编程基本思想和基本技能，为后续的数据挖掘应用性课程和数据产品开发课程的学习打好编程基础。

**课程要求：**本课程设置的目的是通过对R语言的语法规则、数据结构、数据清晰、数据变换、数据整合、数据可视化、数据建模构的学习，较好地训练学生利用计算机解决对数据进行分析和展现，使学生具有数据分析和数据展现的能力，为培养学生有较强数据开发能力打下良好基础。

# 二、教学内容

**理论总学时：32学时**

**第1章 绪论 1学时**

基本要求：

（1）了解R语言特点和优势。

（2）理解数学思维的基本原理。

（3）掌握统计思维的基本原理。

（4）理解逻辑思维的基本原理。

重点：树立正确的思维观。

难点：逻辑思维。

**第2章 R语言入门 1学时**

基本要求：

（1）理解R语言工作空间，环境变量的显示、保存和删除。

（2）了解R包的作用。

（3）能够安装R语言开发环境。

（4）能够编辑和运行R脚本。

（5）R语言基本语法。

重点：R语言脚本编辑和运行。

难点：R包。

**第3章 数据类型 4学时**

基本要求：

（1）掌握向量的产生、引用、合并等操作，包括x:y，seq()，c(),rnorm()。

（2）掌握矩阵的产生、引用、合并、转换等操作。

（3）理解数据的产生、引用、转换等操作。

（4）熟练掌握数据框的产生、引用、转换等操作。

（5）理解列表的产生、引用、转换等操作。

（6）掌握因子的作用、定义和转换。

（7）熟练掌握常量和变量。

（8）基本掌握字符串处理函数。

（9）掌握常用数据类型转换函数。

重点：数据框和向量的应用。

难点：列表和因子的应用。

**第4章 数据准备 2学时**

基本要求：

（1）能用R语言导入文本、Excel、数据库数据。

（2）能用R语言导出数据。

重点：freade和fwrite函数的使用。

难点：数据库数据导入。

**第5章 数据可视化 6学时**

基本要求：

（1）了解数据可视化概念和作用。

（2）理解可视化建议。

（3）了解图表的作用：成分关系图表（柱图、饼图）、对比关系图表（时序图、折线图、面积图）、相关关系图表（散点图、气泡图）、多维比较图表（雷达图、圆环图）。

（4）熟练掌握“散点点图”“折线线图”“直方图”“柱状图”“饼图”“箱线图”“密度图”“直线图”的绘制方法。

（5）理解ggplot2图层绘图原理，并熟练用ggplot2绘图。

重点：ggplot2绘图。

难点：图层原理。

**第6章 数据探索 2学时**

基本要求：

（1）熟练掌握缺失值判定，处理。

（2）熟练掌握异常值判定，处理。

（3）掌握数据统计特征分析方法。

重点：缺失值、异常值判定，分布分析，统计量分析。

难点：缺失值、异常值处理，对比分析，周期分析。

**第7章 数据变换 4学时**

基本要求：

（1）掌握数据规范化基本方法。

（2）能用R语言进行变量增减。

（3）掌握分组汇总,包括列联表table、排序order/sort/rank、聚集aggregate。

（4）掌握数据连接操作。

（5）掌握长表-宽表转换操作，包括stack 和 unstack。

重点：排序，分组汇总。

难点：分组汇总。

**第8章 高级编程 2学时**

基本要求：

（1）熟练掌握缺失值判定，处理。

（2）熟练掌握异常值判定，处理。

（3）掌握数据统计特征分析方法。

重点：缺失值、异常值判定，分布分析，统计量分析。

难点：缺失值、异常值处理，对比分析，周期分析。

**第9章 数据建模 8学时**

基本要求：

（1）理解数据建模概念和作用。

（2）了解数据建模常用方法。

（3）熟练使用rattle进行数据建模。

（4）掌握聚类rattle实现。

（5）掌握关联分析rattle实现。

（6）掌握决策树和随机森林的rattle实现。

（7）理解回归分析基本思想。

（8）能用回归分析模型解决实际应用问题。

（9）了解SVM和神经网络建模基本思想。

重点：rattle数据建模操作。

难点：回归分析原理、关联复习原理。

**第10章 模型评估 2学时**

基本要求：

（1）理解训练集、验证集和测试集作用。

（2）熟练用混淆矩阵进行模型评估。

（3）理解风险图和ROC曲线模型评估方法。

（4）了解模型评估的其它方法。

重点：数据集作用、混淆矩阵的使用。

难点：风险图和ROC方法。

# 三、实践教学要求

**实践总学时：32学时**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 序号 | 实验项目名称 | 实验  类型 | 实验  要求 | 实验内容简介 | 基本要求 | 学时  分配 |
| 1 | 数据可视化 | 验证 | 必做 | 1、成分关系图表（柱图、饼图）、对比关系图表（时序图、折线图、面积图）、相关关系图表（散点图、气泡图）、多维比较图表（雷达图、圆环图）。  2、ggplot2绘图。  3、shiny绘图。 | 熟练掌握 | 4 |
| 2 | 数据探索 | 验证 | 必做 | 1、利用md.pattern函数识别缺失值。2、  2、箱线图检验离群点。  3、LOF法检测异常值。  4、通过散点图矩阵显示异常值。  5、分布分析、周期分析。 | 熟练掌握 | 2 |
| 3 | 数据变换 | 验证 | 必做 | 1、LASSO方法数据降维。  2、transform函数使用。  3、stack和unstack函数使用。  4、melt 函数使用。  5、aggregate函数使用。 | 熟练掌握 | 2 |
| 4 | 高级编程 | 验证 | 必做 | 1、if-else、ifelse和switch。  2、repeat循环。  3、while循环。  4、for循环。 | 熟练掌握 | 2 |
| 5 | 数据建模 | 应用 | 必做 | 1、聚类。  2、关联分析。  3、随机森林。  4、回归分析。 | 熟练掌握 | 4 |
| 6 | 模型评估 | 应用 | 必做 | 1、混淆矩阵。  2、ROC曲线。 | 理解 | 2 |
| 7 | 影响大学平均录取分数线因素分析 | 应用 | 必做 | 根据教材第11章的实验背景、数据说明和实验目标完成实验项目的操作。 | 理解 | 4 |
| 8 | 收视率分析 | 应用 | 必做 | 根据教材第12章的实验背景、数据说明和实验目标完成实验项目的操作。 | 理解 | 4 |
| 9 | Rhadoop | 验证 | 选做 | 根据教材第13章的实验背景、数据说明和实验目标完成实验项目的操作。 | 理解 | 4 |
| 10 | SparkR | 验证 | 选做 | 根据教材第14章的实验背景、数据说明和实验目标完成实验项目的操作。 | 理解 | 4 |

注：1、实验类型：演示、验证、应用、设计、研究。

2、实验要求：必做、选做。

# 四、课时分配

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 章节 | 内容 | 理论环节时数 | 实验时数 |
| 1 | 第1章 | 绪论 | 1 |  |
| 2 | 第2章 | R语言入门 | 1 |  |
| 3 | 第3章 | 数据类型 | 4 |  |
| 4 | 第4章 | 数据准备 | 2 |  |
| 5 | 第5章 | 数据可视化 | 4 | 4 |
| 6 | 第6章 | 数据探索 | 2 | 2 |
| 7 | 第7章 | 数据变换 | 4 | 2 |
| 8 | 第8章 | 高级编程 | 2 | 2 |
| 9 | 第9章 | 数据建模 | 6 | 4 |
| 10 | 第10章 | 模型评估 | 2 | 2 |
| 11 | 第11章 | 影响大学平均录取分数线因素分析 |  | 4 |
| 12 | 第12章 | 收视率分析 |  | 4 |
| 13 | 第13章 | Rhadoop |  | 4 |
| 14 | 第14章 | SparkR |  | 4 |
| 合计： | | | 32 | 32 |
| 总学分 | 4 | 总学时 | 64 | |

# 五、建议教材与教学参考书

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 书名 | 编者 | 出版社 | 版本 |
| 1 | 云计算导论 | 丛书总主编：刘鹏  主编：王成喜 | 清华大学出版社 | 1 |
| 2 | 大数据导论 | 丛书总主编：刘鹏  主编：付雯 | 清华大学出版社 | 1 |
| 3 | 大数据实践 | 丛书总主编：刘鹏  主编：袁晓东 | 清华大学出版社 | 1 |
| 4 | 数据清洗 | 丛书总主编：刘鹏  主编：李法平 | 清华大学出版社 | 1 |
| 5 | 数据挖掘基础 | 丛书总主编：刘鹏  主编：陶建辉 | 清华大学出版社 | 1 |
| 6 | 大数据系统运维 | 丛书总主编：刘鹏  主编：姜才康 | 清华大学出版社 | 1 |
| 7 | 数据标注工程 | 主编：刘鹏 | 清华大学出版社 | 1 |
| 8 | Python语言 | 丛书总主编：刘鹏  主编：李肖俊 | 清华大学出版社 |  |

# 六、教学形式与考核方式

**教学形式：**本课程采用课堂讲授、分析和实践教学等手段开展教学活动。

**考核方式：**本课程采用平时考察与期末闭卷考试相结合的考核方法。平时考察主要考察课堂表现、作业完成情况，平时成绩占30%（课堂表现及考勤占5%，作业占25%）。期末闭卷考试占70%，期末考试主要采用闭卷考试的方式。命题要求覆盖大纲重点内容，题型不少于四种，难易程度适中。