

《R 语言基础》 教学大纲

大数据分析基础系列

大纲目录

1	课程目的、任务	2
2	课程简介	2
3	教学方式和实践环节的特色	2
4	教材及参考书目	2
5	考核方式与评价结构比例	3
6	讲授大纲	4

课程名称: R 语言基础

适合专业: 数据科学与大数据技术、大数据技术与应用、统计学、应用统计学、统计统计、应用数学等本科专业

1 课程目的、任务

本课程是为数据科学与大数据技术及相关专业学生开设的一门课程,《R 语言基础》是一门基于 R 进行数据分析的基础课程,本课程的目的是帮助学生从零开始对 R 语言有全方位了解,掌握 R 语言数据处理、统计分析和可视化最为基础知识、技术和基本应用,为使用 R 语言进行深入的数据分析与行业应用打下扎实的基础。本课程是后期统计分析、机器学习、可视化等相关课程的预修课程。

2 课程简介

- R 语言是一个自由、免费、源代码开放的语言环境。它功能完备,语言灵活且不依赖操作系统,是一个用于统计计算和统计制图的优秀工具。在 2010 年,美国统计协会(ASA)授予 R 语言为(第一届)统计计算及图形奖。截止目前为止,R 是数据科学领域使用用户最多的数据挖掘与编程语言,且其用户数量在不断增加。R 已经成为几乎所有国内外高校统计类课程标配的教学软件,R 的书集遍及经济、金融、生物、医学、生态、电子商务、航空、旅游、心理、法律等各个领域,并成大数据分析的宠儿:覆盖机器学习、人工智能和可视化等。R 语言在国内业界也逐步成为主流的分析工具,每年全国性的 R 语言会议推进 R 的迅猛发展。
- 本课程共 5 大模块,从 R 入门、R 数据集创建与管理、R 绘图初步、R 数据探索与比较分析和 R 统计建模等方面,让零基础学员从各个角度对 R 的使用进行学习。
- 本课程为基础课程,学员可以通过学习对 R 语言进行全方位的了解与掌握,并为 R 的进阶学习与实际数据分析做好基础铺垫。

3 教学方式和实践环节的特色

- 教学方式: 课堂讲授,配合大数据平台演示。
- 实践环节:
 1. 课堂讲授中提供一定的思考题供学生练习或讨论;
 2. 课后布置作业,作业量 2—3 小时,作业中以基本练习为主。

4 教材及参考书目

- 建议教材:

1. R.I. Kabacoff 著, 高涛, 肖南, 陈钢译, R 语言实战 (R in Action: Data Analysis and Graphics with R), 人民邮电出版社, 2013.
2. 徐珉久 [韩], 武传海 (译). R 语言与数据分析实战. 中国工信出版集团, 人民邮电出版社, 2017.1

- 参考书目:

1. 汤银才, R 语言与统计分析, 高等教育出版社. 2008
2. Joseph Adler, R in a Nutshell, California: O'Reilly Meda, Inc., 2nd Ed. 2012.
3. Jared P. Lander, R for Everyone: Advanced Analytics and Graphics. New York: Addison-Wesley. 2014.
4. Hadley Wickham, Garrett Golemund, R for Data Science - Import, Tidy, Transform, Visualize, and Model data, O'Reilly. 2017.
5. Jim Albert, M. Rizzo, R by example, Springer, 2012.
6. Prabhanjan N. Tattar, Suresh Ramaiah, B.G. Manjunath. A Course in Statistics with R(ACSWR), for master students. Wiley. 2016. [*]
7. Norman Matloff. The art of R programming, No Starch Press, inc., 2011.
8. Paul Murrell, R Graphics, Chapman & Hall/CRC. 2006.
9. Deepayan Sarkar, Lattice: Multivariate Data Visualization with R, Springer. 2008.
10. Hadley Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer. 2009. (中译本: 殷腾飞, 统计之都, 2014)
11. Donato Teutonico, ggplot2 Essentials, PACKT, 2015.
12. Alboukadel Kassambara, ggplot2 - Guide to Create Beautiful Graphics in R (2nd Ed). 2013.
13. Mark P.J. van der Loo, Edwin de Jonge, Learning Rstudio for R Statistical Computing. PACKT, 2012.
14. John Verzani, Getting Started with Rstudio, O'Reilly, 2011.
15. Nicholas J. Horton and Ken Kleinman. Using R and RStudio for Data Management Statistical Analysis and Graphics(2nd Ed.), 2015.

5 考核方式与评价结构比例

- 考核内容为: 平时表现、作业完成情况、平时测验、期终闭卷考试。
- 评价结构比例:

1. 总评成绩由平时成绩和考试成绩两部分组成,一般平时成绩占 40%,考试成绩占 60%。
2. 平时成绩由平时表现、作业、平时测验成绩组成,一般平时表现占 10%,作业占 30%,平时测验成绩占 60%。

6 讲授大纲

M1: R 入门

1. 大数据与数据科学
 - 数据科学与分析工具
 - R 及其优势
 - R 安装与配置
 - R 包安装与使用
 - R 资源与帮助
2. R 快速入门教程
 - R 中的基本语法
 - R 中的数据对象及其属性
 - R 的工作空间与管理
 - R 编程基础
 - R 程序调试
3. R 编辑器与 RStudio
 - R 常用编辑器
 - Rstudio 功能与使用技巧
 - Rstudio 进阶
 - 项目管理
4. Rmarkdown 介绍
 - Rmarkdown
 - 文学化统计编程

M2: R 数据集创建与管理

1. 数据集的创建与处理
 - 常用数据对象与创建

- 数据的读取
- 数据的存储
- 数据集的处理
- 数据集的合并与子集提取
- 缺失值的处理
- 数据表数据的切片、切块与组合

2. 向量的操作

- 合并数据框的行 (向量) 与列 (向量)
- apply 系列函数
- 数据分组并调用函数
- 数据拆分与合并
- 数据排序
- 访问数据中的列
- 查找符合条件的数据索引
- 分组运算

3. 数据对象的其他操作

- 赋值与常用运算
- 基本的数学运算
- 用于矩阵的运算
- 正则表达式
- 与统计分布相关的函数

M3: R 绘图初步

1. 基本的绘图命令

- 大趋势: 信息可视化
- R 绘图基础: 低级与高级绘图命令
- 基本绘图函数: plot, points, lines, curve, par
- 绘图三要素设置详解 (颜色, 点型, 线型)
- 绘图信息补充 (title, text, legend, axis)

- R 窗口和图形设备 (X11, pdf, png)

2. 一维数据的可视化

- 常用统计分布与 4 类函数
- 一维离散变量的分布图示
- 一维连续变量的分布图示
- 一维连续分布诊断图
- 非参数密度估计与展示

3. 多维数据的可视化

- 二个离散变量的分布图示
- 二个混合变量的分布图示
- 二个连续变量的分布图示
- 多变量的可视化

M4: R 统计分析初步

1. 数据分析基础

2. R 中常用的统计函数

3. 描述性统计分析

- 常用描述性统计量及其计
- 单个连续型变量描述性统计量的获取
- 分组计算描述性统计量

4. 相关性分析

5. 相关性检验

M5: R 统计建模

1. 回归模型

- `lm()` 函数中的公式表示
- 一元线性回归
- 多元线性回归
- 回归预测

2. 分类数据的统计推断

- 广义线性模型概述
- glm() 函数介绍
- Logistic 回归
- Poisson 回归

3. 模型的检验与比较

- 回归模型诊断
- 变量的选择
- 模型比较
- 异常值判断
- 预测与交叉验证