

目 录

CONTENTS

算法/程序列表 / 1

第一部分 数据科学与工程概述

第 1 章

绪论

- 1.1 信息文明与数据简史 / 4
- 1.2 数据科学与工程的基本内涵 / 13
- 1.3 第四范式:数据密集型科学 / 18
- 1.4 数据科学与工程的应用 / 23
- 1.5 实践:以 Git 与 Python 为中心 / 32
- 1.6 本章小结 / 37
- 1.7 习题与实践 / 37

第 2 章

数据思维 与问题求解

- 2.1 问题求解与思维方式 / 40
- 2.2 计算思维与数据思维 / 43
- 2.3 计算思维与数据思维实例 / 51
- 2.4 实践:Python 问题求解 / 61
- 2.5 本章小结 / 66
- 2.6 习题与实践 / 67

第二部分 数据与计算的基础设施

第 3 章

数据的模型 与结构

-
- 3.1 比特与数据 / 72
 - 3.2 进制与数据表达 / 78
 - 3.3 数据的编码与存储 / 82
 - 3.4 数据的模型 / 86
 - 3.5 数据的结构 / 93
 - 3.6 实践:Python 数据结构 / 96
 - 3.7 本章小结 / 101
 - 3.8 习题与实践 / 101

第4章 数据的计算 与程序表达

- 4.1 数据的计算 / 104
- 4.2 算法分析 / 115
- 4.3 算法的实例 / 117
- 4.4 计算机编程语言 / 124
- 4.5 实践:Python 算法 / 129
- 4.6 本章小结 / 136
- 4.7 习题与实践 / 136

第5章 计算基础设施

- 5.1 数据处理的通用机器 / 139
- 5.2 程序执行过程 / 144
- 5.3 计算机系统结构 / 148
- 5.4 基础设施软件 / 154
- 5.5 云计算与数据中心 / 160
- 5.6 实践:基础设施数据采集与分析 / 164
- 5.7 本章小结 / 168
- 5.8 习题与实践 / 169

第6章 数据的全生命 周期管理

- 6.1 数据采集 / 173
- 6.2 数据存储 / 178
- 6.3 数据管理 / 181
- 6.4 数据计算 / 183
- 6.5 数据分析 / 185
- 6.6 数据展示 / 187
- 6.7 实践:Python 网络爬虫 / 196
- 6.8 本章小结 / 202
- 6.9 习题与实践 / 202

第7章 数据库系统

- 7.1 数据库的起源与发展 / 205
- 7.2 关系数据库 / 208
- 7.3 数据仓库与 OLAP / 217
- 7.4 SQL 语言 / 220
- 7.5 实践:SQL 数据处理与分析 / 226
- 7.6 本章小结 / 233
- 7.7 习题与实践 / 233

第8章 大数据系统

- 8.1 大数据的基本概念 / 237
- 8.2 Hadoop 和 Spark 生态 / 244
- 8.3 SQL 与 Hadoop 的组合 / 247
- 8.4 大数据系统的发展与未来 / 250
- 8.5 实践:Hadoop 与 Spark 大数据处理 / 253

8.6 本章小结 / 261

8.7 习题与实践 / 261

第三部分 数据分析的原理与方法

第9章

数据科学过程

9.1 数据科学过程基础 / 266

9.2 数据科学 workflow / 279

9.3 实践:KNIME 数据科学 workflow / 285

9.4 本章小结 / 294

9.5 习题与实践 / 295

第10章

统计分析的原理

10.1 数据科学的数学基础 / 297

10.2 概率与统计基础 / 307

10.3 统计建模:线性回归模型 / 314

10.4 数据分析的工具 / 319

10.5 实践:Python 统计分析 / 323

10.6 本章小结 / 327

10.7 习题与实践 / 327

第11章

机器学习方法

11.1 机器学习发展历史 / 331

11.2 机器学习方法 / 335

11.3 机器学习最新发展 / 340

11.4 经典机器学习算法 / 342

11.5 实践:Python 机器学习 / 350

11.6 本章小结 / 357

11.7 习题与实践 / 357

第12章 深度学习

- 12.1 深度学习介绍 / 360
- 12.2 深度学习价值 / 366
- 12.3 误差反向传播算法 / 368
- 12.4 卷积神经网络 / 371
- 12.5 深度学习工具 / 376
- 12.6 实践:Python 深度学习——手写汉字识别 / 379
- 12.7 本章小结 / 385
- 12.8 习题与实践 / 385

第13章 数据挖掘基础

- 13.1 初识数据挖掘 / 389
- 13.2 数据挖掘技术 / 394
- 13.3 典型数据挖掘算法 / 399
- 13.4 实践:Python 图像分类 / 411
- 13.5 本章小结 / 415
- 13.6 习题与实践 / 416

第14章 非结构化数据 挖掘

- 14.1 自然语言处理 / 418
- 14.2 语音信号处理 / 422
- 14.3 图像处理与理解 / 427
- 14.4 实践:Python 文本数据挖掘 / 432
- 14.5 本章小结 / 440
- 14.6 习题与实践 / 440

第四部分 数据应用与社会问题

第15章

数据综合应用

15.1 搜索引擎 / 446

15.2 智能运维 / 460

15.3 开源数字年报 / 470

15.4 本章小结 / 474

15.5 习题与实践 / 475

第16章

数据道德

与职业行为准则

16.1 开放的世界 / 478

16.2 数据科学与工程职业规划 / 483

16.3 数据隐私与社会问题 / 488

16.4 数据与人工智能伦理 / 495

16.5 本章小结 / 498

16.6 习题与实践 / 499

文献阅读 / 500

参考文献 / 503

附录 / 505

算法 / 程序列表

第1章 绪论 / 3

程序 1.1 第一个 Python 数据科学程序 / 36

第 2 章 数据思维与问题求解 / 39

程序 2.1 递归加法 / 52

程序 2.2 最小值_循环 / 52

程序 2.3 最小值_递归 / 53

程序 2.4 最小值_分治 / 54

程序 2.5 验证帕斯卡的分析 / 56

程序 2.6 估计 π 值 / 58

程序 2.7 开平方 1 “笨办法” / 62

程序 2.8 开平方 2 二分法 / 63

程序 2.9 开平方 3 牛顿法 / 64

程序 2.10 开平方 4 蒙特卡罗法 / 66

第 3 章 数据的模型与结构 / 71

程序 3.1 变量的赋值 / 97

程序 3.2 栈的实现 / 97

程序 3.3 简单树的实现 / 99

程序 3.4 用列表创建简单树 / 99

程序 3.5 二叉树类的定义 / 99

程序 3.6 二叉树中插入左子节点 / 100

程序 3.7 二叉树中插入右子节点 / 100

程序 3.8 获取和设置根值以及获得左右子树 / 100

第 4 章 数据的计算与程序表达 / 103

算法 4.1 函数 search for X / 112

程序 4.2 交换变量 a 和 b 的值 / 117

算法 4.3 冒泡排序 / 118

算法 4.4 汉诺塔问题的解 / 120

算法 4.5 树排序 / 124

程序 4.6 冒泡排序 / 130

程序 4.7 选择排序 / 132

程序 4.8 插入排序 / 133

程序 4.9 快速排序 / 135

程序 4.10 希尔排序 / 137

第 5 章 计算基础设施 / 138

程序 5.1 替换函数 1 / 167

程序 5.2 替换函数 2 / 167

程序 5.3 替换函数 3 / 167

程序 5.4 替换函数 4 / 167

程序 5.5 程序性能测试 / 168

第 6 章 数据的全生命周期管理 / 171

程序 6.1 散点图 / 191

程序 6.2 网络爬虫 / 198

程序 6.3 绘制散点图 / 200

程序 6.4 绘制正弦、余弦曲线 / 200

程序 6.5 绘制等高线图 / 201

第 7 章 数据库系统 / 204

程序 7.1 查询客户总消费额 / 212

程序 7.2 数据库事务 / 213

程序 7.3 创建表 / 228

程序 7.4 SQL 查询 1 / 228

程序 7.5 SQL 查询 2 / 229

程序 7.6 SQL 查询 3 / 229

程序 7.7 SQL 查询 4 / 230

程序 7.8 SQL 查询 5 / 230

程序 7.9 SQL 查询 6 / 231

程序 7.10 SQL 分析 1 / 231

程序 7.11 SQL 分析 2 / 231

程序 7.12 SQL 分析 3 / 232

程序 7.13 SQL 分析 4 / 232

程序 7.14 SQL 分析 5 / 232

第 8 章 大数据系统 / 235

程序 8.1 map 代码 / 258

程序 8.2 reduce 代码 / 258

程序 8.3 用 Spark 进行 WordCount / 261

第 10 章 统计分析的原理 / 296

程序 10.1 文本词频统计 / 323

程序 10.2 线性回归模型 / 325

第 11 章 机器学习方法 / 329

程序 11.1 损失函数 / 350

程序 11.2 梯度计算函数 / 350

程序 11.3 梯度下降算法 / 351

第 12 章 深度学习 / 359

程序 12.1 基于 VGG 模型的手写汉字识别模型 / 384

第 13 章 数据挖掘基础 / 387

程序 13.1 KNN 算法模型 / 413

程序 13.2 训练 KNN / 414

第 14 章 非结构化数据挖掘 / 417

程序 14.1 词云制作 / 433

程序 14.2 文本分类实践 / 434

第 15 章 数据综合应用 / 445

程序 15.1 使用倒排索引的检索处理 / 455

程序 15.2 基于文档和查询关联度的检索 / 456

程序 15.3 基于查询单词的文档和查询关联度的检索 / 457

程序 15.4 基于排序的索引构建 / 458

程序 15.5 基于合并的索引构建 / 459