

数据挖掘与机器学习

基础及应用

许桂秋 吴丽镛 张文明 主编
龙法宁 王子琦 王珂 朱琳玲 徐强 副主编

大数据技术与应用丛书

《大数据导论（通识版）》

《数据采集及预处理基础与应用》

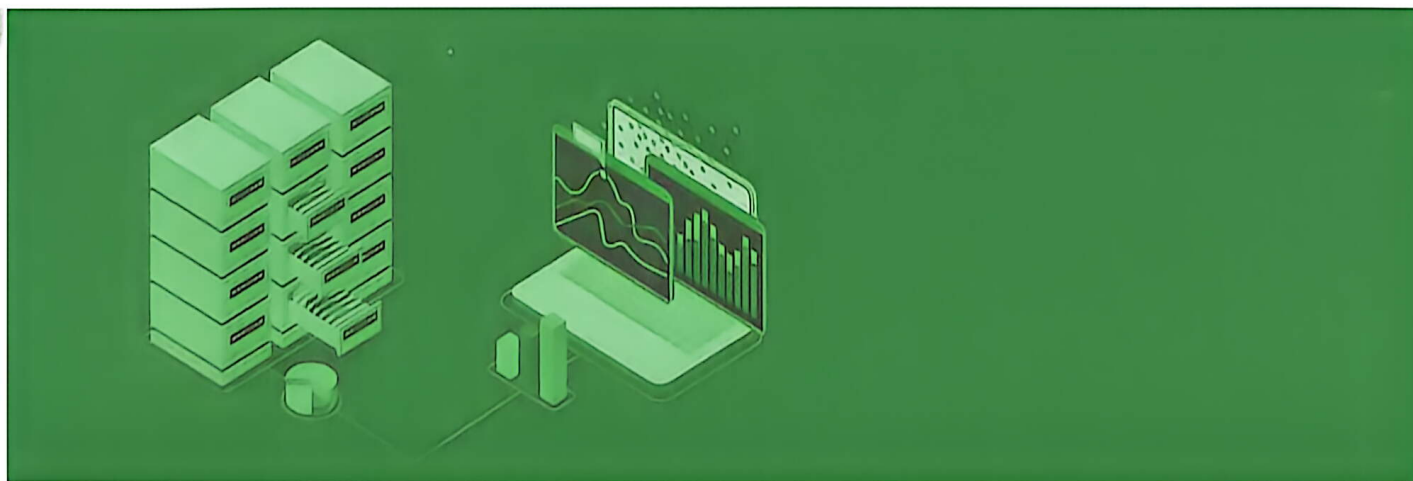
《分布式数据库基础与应用》

《数据挖掘与机器学习基础及应用》

《数据可视化基础与应用》

《大数据处理技术基础与应用（Hadoop+Spark）》

《Python编程基础与应用——任务式案例教程》



分类建议：计算机 / 大数据

人民邮电出版社网址：www.ptpress.com.cn

ISBN 978-7-115-64576-0



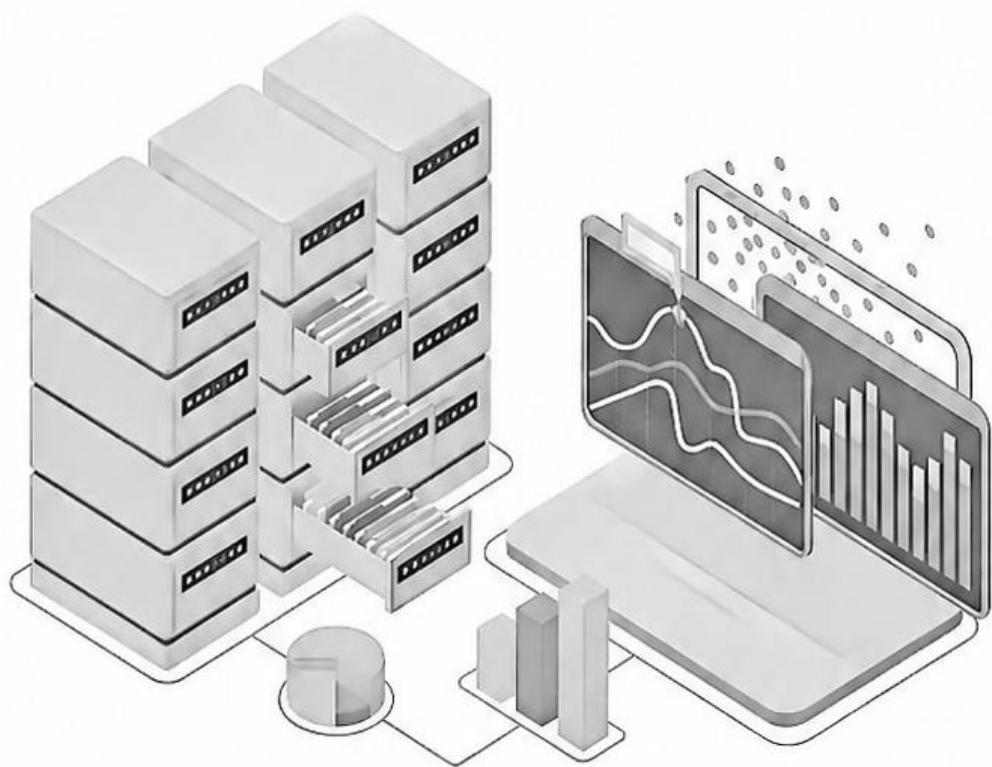
9 787115 645760 >

定价：69.80 元

数据挖掘与机器学习

基础及应用

许桂秋 吴丽楠 张文明◎主 编
龙法宇 王子茹 王珂 朱琳玲 徐强◎副主编



人民邮电出版社
北京

图书在版编目 (CIP) 数据

数据挖掘与机器学习基础及应用 / 许桂秋, 吴丽镐,
张文明主编. — 北京: 人民邮电出版社, 2024. 8.
(大数据技术与应用丛书). — ISBN 978-7-115-64576

-0

I. TP274; TP181

中国国家版本馆 CIP 数据核字第 2024CN3518 号

内 容 提 要

这是一本全面介绍数据挖掘与机器学习的大数据专业类图书, 阅读本书可以提升读者对大数据分析
与挖掘的认知及动手能力。本书共 10 章, 由浅入深地讲解数据挖掘与机器学习的基本概念与流程、相关
算法与实现工具。全书理论与实践相结合, 既有新技术的深度, 也有行业应用的广度, 使读者可以全面
了解数据挖掘与机器学习相关技术。

本书可以作为高等学校计算机、数据科学与大数据技术等相关专业“机器学习”或者“数据挖掘”
课程的教材, 也可作为从事机器学习与数据挖掘、数据分析相关工作的技术人员的参考书。

-
- ◆ 主 编 许桂秋 吴丽镐 张文明
副 主 编 龙法宁 王子琦 王 珂 朱琳玲 徐 强
责任编辑 张晓芬
责任印制 马振武
 - ◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路 11 号
邮编 100164 电子邮件 315@ptpress.com.cn
网址 <https://www.ptpress.com.cn>
三河市祥达印刷包装有限公司印刷
 - ◆ 开本: 787×1092 1/16
印张: 16.5 2024 年 8 月第 1 版
字数: 381 千字 2024 年 8 月河北第 1 次印刷

定价: 69.80 元

读者服务热线: (010)81055493 印装质量热线: (010)81055316
反盗版热线: (010)81055315

前言

数据挖掘与机器学习是计算机科学和人工智能学科中一个非常重要的研究领域，也是一些交叉学科的重要支撑技术。在过去的数十年中，互联网以及各种信息系统产生了大量数据。数据的爆炸性增长激起了人们对新技术和自动化工具的需求，以便于帮助我们将海量数据转换成信息和知识。数据挖掘与机器学习作为一种前沿的分析工具，引起了产业界和学术界的广泛关注，并快速成为计算机研究领域的一个热点。

本书是数据挖掘与机器学习的入门级图书，主要描述数据挖掘与机器学习的相关定义与算法，包含数据挖掘与机器学习中各种模式的概念和实现方式、算法的具体使用技巧等。本书可帮助读者了解数据挖掘与机器学习的内容，并让读者能根据书中提供的案例完成数据挖掘的各种模式操作。

全书共 10 章，可分为 4 个部分。

第一部分是基础概述，包括第 1 章和第 2 章。第 1 章阐述数据挖掘与机器学习的发展历史、基本概念、算法分类、一般流程，以及主要的应用领域；第 2 章详细介绍实现数据挖掘与机器学习的基本工具，例如 Numpy、pandas、Matplotlib 和 scikit-learn 数据科学分析库，以及它们的使用方法。

第二部分是数据挖掘与机器学习的算法使用，包括第 3~7 章。这 5 章详细介绍数据挖掘与机器学习中回归、分类、聚类等算法与应用，以及关联规则与协同过滤，最后介绍特征工程、降维与超参数调优的内容。

第三部分是进阶部分，包括第 8 章和第 9 章。这两章通过图像分类之猫狗识别和基于 NLTK 实现文本数据处理这两个案例介绍文本与图像的相关分析方法。

第四部分主要介绍深度学习的相关内容，只包括第 10 章。这章以 Fashion MNIST 数据集为例，介绍基于深度学习的图像处理方法以及模型的搭建、优化、保存等处理过程。

由于编者水平有限，书中难免存在不足之处，恳请广大读者批评指正。

编者

2024 年 6 月

目录

| | |
|-------------------------|----|
| 第 1 章 数据挖掘与机器学习概述 | 1 |
| 1.1 数据挖掘与机器学习的发展历史 | 1 |
| 1.1.1 数据时代 | 1 |
| 1.1.2 数据挖掘的技术发展 | 2 |
| 1.1.3 机器学习的技术发展 | 5 |
| 1.1.4 人工智能、数据挖掘与机器学习的关系 | 7 |
| 1.2 数据挖掘与机器学习的相关概念 | 8 |
| 1.2.1 数据挖掘的定义 | 8 |
| 1.2.2 机器学习的定义 | 8 |
| 1.2.3 数据库与数据仓库 | 9 |
| 1.2.4 知识发现 | 9 |
| 1.3 数据挖掘与机器学习的算法分类 | 12 |
| 1.3.1 类/概念描述：特征和区分 | 12 |
| 1.3.2 回归分析 | 13 |
| 1.3.3 分类 | 14 |
| 1.3.4 预测 | 14 |
| 1.3.5 关联分析 | 15 |
| 1.3.6 聚类分析 | 15 |
| 1.3.7 异常检测 | 16 |
| 1.3.8 迁移学习 | 17 |
| 1.3.9 强化学习 | 17 |
| 1.4 数据挖掘与机器学习的一般流程 | 19 |
| 1.4.1 确定分析目标 | 19 |
| 1.4.2 收集数据 | 19 |
| 1.4.3 数据预处理 | 19 |

| | | |
|--------------|------------------------|-----------|
| 1.4.4 | 数据建模 | 20 |
| 1.4.5 | 模型训练 | 21 |
| 1.4.6 | 模型评估 | 21 |
| 1.4.7 | 模型应用 | 21 |
| 1.5 | 数据挖掘与机器学习的应用领域及面临的问题 | 21 |
| 1.5.1 | 电商领域 | 22 |
| 1.5.2 | 金融领域 | 22 |
| 1.5.3 | 医疗领域 | 22 |
| 1.5.4 | 电信领域 | 23 |
| 1.5.5 | 自然语言处理领域 | 23 |
| 1.5.6 | 工业领域 | 25 |
| 1.5.7 | 艺术创作领域 | 26 |
| 1.5.8 | 数据挖掘与机器学习应用面临的问题 | 27 |
| 第 2 章 | 数据科学分析入门 | 28 |
| 2.1 | 数据科学分析库 | 28 |
| 2.1.1 | NumPy | 31 |
| 2.1.2 | pandas | 33 |
| 2.1.3 | Matplotlib | 37 |
| 2.1.4 | scikit-learn | 40 |
| 2.2 | 数据科学分析库的使用方法及应用示例 | 43 |
| 2.2.1 | NumPy 基本使用方法 | 43 |
| 2.2.2 | pandas 基本使用方法 | 47 |
| 2.2.3 | Matplotlib 基本使用方法 | 49 |
| 2.2.4 | scikit-learn 基本使用方法 | 57 |
| 2.2.5 | 金融贷款数据可视化 | 60 |
| 第 3 章 | 回归算法与应用 | 71 |
| 3.1 | 回归预测问题 | 71 |
| 3.1.1 | 回归预测简介 | 71 |
| 3.1.2 | 常见的回归数据集 | 72 |
| 3.2 | 女性身高与体重的线性回归预测 | 73 |
| 3.2.1 | 线性回归原理与应用场景 | 73 |
| 3.2.2 | 数据导入与查看 | 74 |
| 3.2.3 | 绘制女性身高和体重的散点图 | 75 |
| 3.2.4 | 使用 Statsmodels 进行建模与评估 | 77 |
| 3.2.5 | 模型的建立与优化 | 79 |
| 3.2.6 | 线性回归算法优缺点 | 82 |

| | | |
|--------------|----------------------|------------|
| 3.3 | 泰坦尼克号数据集的逻辑回归预测 | 82 |
| 3.3.1 | 逻辑回归原理与应用场景 | 82 |
| 3.3.2 | 任务描述 | 83 |
| 3.3.3 | 数据探索与可视化 | 84 |
| 3.3.4 | 数据预处理 | 87 |
| 3.3.5 | 模型的建立与预测 | 90 |
| 第 4 章 | 分类算法与应用 | 92 |
| 4.1 | 数据挖掘 | 92 |
| 4.1.1 | 数据挖掘分类 | 92 |
| 4.1.2 | 常见的分类数据集 | 93 |
| 4.2 | 使用 KNN 算法实现电影分类 | 97 |
| 4.2.1 | KNN 算法的基本原理 | 97 |
| 4.2.2 | 使用 Python 实现 KNN 算法 | 98 |
| 4.2.3 | KNN 算法在电影分类中的应用 | 99 |
| 4.2.4 | 电影分类的可视化 | 102 |
| 4.3 | 基于文档的空间向量模型应用 | 104 |
| 4.3.1 | 空间向量原理与应用场景 | 104 |
| 4.3.2 | 文档的空间向量模型应用实现 | 105 |
| 4.4 | 使用支持向量机进行数据分类 | 108 |
| 4.4.1 | 支持向量机原理与应用 | 108 |
| 4.4.2 | 线性可分与线性不可分 | 111 |
| 4.4.3 | 基于支持向量机的乳腺癌数据分类实现 | 115 |
| 4.4.4 | 确定超参数 | 116 |
| 4.4.5 | 过拟合问题的解决 | 118 |
| 4.5 | 使用决策树进行数据分类 | 121 |
| 4.5.1 | 决策树概述 | 121 |
| 4.5.2 | ID3 算法 | 123 |
| 4.5.3 | 基于决策树的乳腺癌数据分类实现 | 124 |
| 4.6 | 使用随机森林进行波士顿房价回归预测 | 127 |
| 4.6.1 | 集成学习概述 | 127 |
| 4.6.2 | 随机森林概述 | 134 |
| 4.6.3 | 使用随机森林进行波士顿房价回归预测的实现 | 136 |
| 4.7 | 模型的评判和保存 | 137 |
| 第 5 章 | 聚类算法与应用 | 142 |
| 5.1 | 无监督学习问题 | 142 |

| | | |
|--------------|---------------------------------|------------|
| 5.1.1 | 无监督学习 | 142 |
| 5.1.2 | 聚类的基本概念与原理 | 143 |
| 5.1.3 | 常见的聚类数据集 | 143 |
| 5.2 | 使用划分聚类对航空客户群进行分析 | 144 |
| 5.2.1 | 划分聚类基本原理 | 144 |
| 5.2.2 | 任务描述 | 148 |
| 5.2.3 | 航空客户群数据预处理 | 148 |
| 5.2.4 | 模型的建立 | 150 |
| 5.3 | 使用层次聚类挖掘运营商基站信息 | 151 |
| 5.3.1 | 层次聚类算法原理 | 151 |
| 5.3.2 | 任务描述 | 152 |
| 5.3.3 | 导入数据 | 153 |
| 5.3.4 | 数据的特征压缩 | 154 |
| 5.3.5 | 层次聚类实现数据挖掘 | 154 |
| 5.4 | 聚类效果评测 | 155 |
| 第 6 章 | 关联规则与协同过滤 | 157 |
| 6.1 | 推荐算法简介 | 157 |
| 6.2 | 关联规则 | 158 |
| 6.2.1 | 关联规则基本原理 | 158 |
| 6.2.2 | 关联规则的挖掘过程 | 159 |
| 6.2.3 | Apriori 算法 | 159 |
| 6.3 | 使用协同过滤进行电影推荐 | 169 |
| 6.3.1 | 协同过滤算法的概念 | 169 |
| 6.3.2 | 任务描述 | 170 |
| 6.3.3 | 电影推荐协同过滤的实现 | 170 |
| 6.3.4 | 推荐算法库 Surprise 简介与应用示例 | 173 |
| 第 7 章 | 特征工程、降维与超参数调优 | 179 |
| 7.1 | 特征工程 | 179 |
| 7.1.1 | 数据总体分析 | 179 |
| 7.1.2 | 数据预处理 | 180 |
| 7.1.3 | 数据预处理案例分析——美国高中生的社交数据案例分析 | 190 |
| 7.2 | 降维与超参数调优 | 193 |
| 7.2.1 | 降维 | 193 |
| 7.2.2 | 实现降维 | 193 |
| 7.2.3 | 超参数调优 | 195 |

| | | |
|---------------|-----------------------------------|------------|
| 7.2.4 | 交叉验证案例分析 | 196 |
| 第 8 章 | 图像分类之猫狗识别 | 198 |
| 8.1 | 图像分类的基础知识 | 198 |
| 8.1.1 | 图像分类的定义与应用场景 | 198 |
| 8.1.2 | 图像分类的实现方法 | 199 |
| 8.1.3 | 图像分类数据集 | 202 |
| 8.2 | 任务内容 | 204 |
| 8.3 | 环境准备 | 205 |
| 8.3.1 | 安装相关库 | 205 |
| 8.3.2 | 新建一个项目 | 205 |
| 8.4 | 数据准备 | 206 |
| 8.4.1 | 获取 Dags vs. Cats 数据集 | 206 |
| 8.4.2 | 添加训练集和测试集到项目 | 207 |
| 8.4.3 | 获取训练集及测试集 | 207 |
| 8.5 | 构建 CNN 模型 | 210 |
| 8.6 | 训练模型 | 214 |
| 8.7 | 测试模型 | 217 |
| 第 9 章 | 基于 NLTK 实现文本数据处理 | 221 |
| 9.1 | 文本数据处理的相关概念 | 221 |
| 9.1.1 | 常用的文本数据处理技术 | 221 |
| 9.1.2 | 中英文的文本数据处理方法对比 | 222 |
| 9.2 | 文本数据处理关键技术应用 | 224 |
| 9.2.1 | 文本分词技术 | 224 |
| 9.2.2 | 文本向量化技术 | 226 |
| 9.2.3 | 关键词提取 | 229 |
| 9.3 | NLTK 环境搭建 | 230 |
| 9.3.1 | NLTK 简介 | 230 |
| 9.3.2 | NLTK 的安装与使用 | 231 |
| 9.4 | NLTK 实现词条化 | 232 |
| 第 10 章 | 基于深度学习识别 Fashion MNIST 数据集 | 234 |
| 10.1 | CNN 简介 | 234 |
| 10.1.1 | 多层感知机和 CNN | 234 |
| 10.1.2 | CNN | 235 |
| 10.2 | LeNet-5 网络模型 | 235 |

| | |
|---|-----|
| 10.3 Fashion MNIST 数据集..... | 236 |
| 10.3.1 Fashion MNIST 数据集简介..... | 236 |
| 10.3.2 数据集的下载与使用..... | 237 |
| 10.3.3 查看 Fashion_MNIST 数据集..... | 237 |
| 10.4 搭建模型识别 Fashion MNIST 数据集..... | 238 |
| 10.4.1 数据初始化处理..... | 238 |
| 10.4.2 搭建 LeNet-5 模型..... | 239 |
| 10.4.3 训练与评估模型..... | 240 |
| 10.4.4 卷积输出可视化..... | 242 |
| 10.5 改进 LeNet-5 模型实现 Fashion MNIST 数据集识别..... | 244 |
| 10.5.1 数据初始化处理..... | 244 |
| 10.5.2 搭建与训练模型..... | 245 |
| 10.5.3 训练与评估模型..... | 247 |
| 10.5.4 测试集预测..... | 248 |
| 10.5.5 保存模型与网络结构..... | 250 |
| 10.6 使用自然测试集进行预测..... | 251 |
| 10.6.1 图像预处理..... | 251 |
| 10.6.2 预测结果..... | 252 |