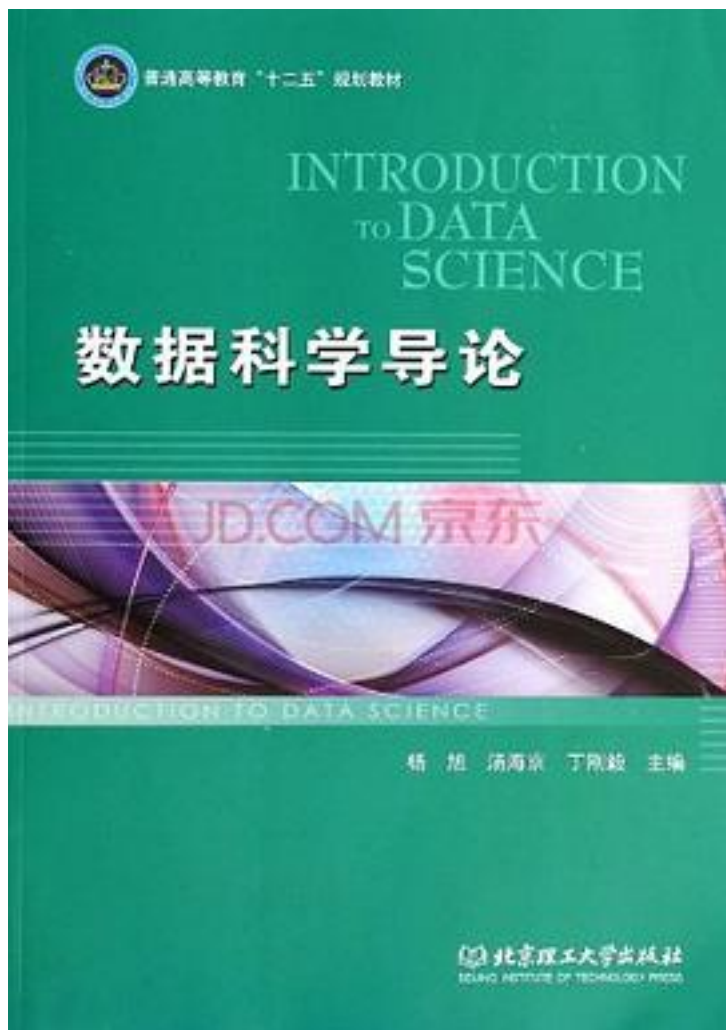


# 数据科学导论



[数据科学导论\\_下载链接1](#)

著者:阿尔贝托·博斯凯蒂

出版者:机械工业出版社

出版时间:2016-8-10

装帧:平装

isbn:9787111544340

本书由两位资深数据科学家撰写，融合其多年从事数据科学相关的教学和科研工作经验

，借助现有的Python语法和结构知识，全面而系统地讲解进行数据科学分析和开发的相关工具、技术和\*佳实践，包含清晰的代码和简化的示例。通过阅读本书，你将深入理解Python核心概念，成为高效数据科学实践者。

本书共6章，系统介绍了进行数据科学分析和开发所涉及的关键要素。书中首先介绍Python软件及相关工具包的安装和使用；接着不仅讲解数据加载、运算和改写等基本数据准备过程，还详细介绍特征选择、维数约简等高级数据操作方法；并且建立了由训练、验证、测试等过程组成的数据科学流程，结合具体示例深入浅出地讲解了多种机器学习算法；然后介绍了基于图模型的社会网络创建、分析和处理方法；最后讲解数据分析结果的可视化及相关工具的使用方法。

作者介绍:

Alberto Boschetti

数据科学家、信号处理和统计学方面的专家。他拥有通信工程专业博士学位，现在伦敦居住和工作。基于所从事的项目，他每天都要面对包括自然语言处理、机器学习和概率图模型等方面的挑战。他对工作充满激情，经常参加学术聚会、研讨会等学术活动，紧跟数据科学技术发展的前沿。

Luca Massaron

数据科学家、市场研究总监，是多元统计分析、机器学习和客户洞察方面的专家，有十年以上解决实际问题的经验，使用推理、统计、数据挖掘和算法为利益相关者创造了巨大的价值。他是意大利网络受众分析的先锋，并在Kaggler上获得排名前十的佳绩，随后一直热心参与一切与数据分析相关的活动，积极给新手和专业人员讲解数据驱动知识发现的潜力。他崇尚大道至简，坚信理解数据科学的本质能带来巨大收获。

目录: 译者序

前言

第1章 新手上路1

1.1 数据科学与Python简介1

1.2 Python的安装2

1.2.1 Python 2还是Python 33

1.2.2 分步安装3

1.2.3 Python核心工具包一瞥4

1.2.4 工具包的安装7

1.2.5 工具包升级9

1.3 科学计算发行版9

1.3.1 Anaconda10

1.3.2 Enthought Canopy10

1.3.3 PythonXY10

1.3.4 WinPython10

1.4 IPython简介10

1.4.1 IPython Notebook12

1.4.2 本书使用的数据集和代码18

1.5 小结25

第2章 数据改写26

2.1 数据科学过程26

2.2 使用pandas进行数据加载与预处理27

2.2.1 数据快捷加载27

2.2.2 处理问题数据30

2.2.3 处理大数据集32

2.2.4 访问其他数据格式36

2.2.5 数据预处理	37
2.2.6 数据选择	39
2.3 使用分类数据和文本数据	41
2.4 使用NumPy进行数据处理	49
2.4.1 NumPy中的N维数组	49
2.4.2 NumPy ndarray对象基础	50
2.5 创建NumPy数组	50
2.5.1 从列表到一维数组	50
2.5.2 控制内存大小	51
2.5.3 异构列表	52
2.5.4 从列表到多维数组	53
2.5.5 改变数组大小	54
2.5.6 利用NumPy函数生成数组	56
2.5.7 直接从文件中获得数组	57
2.5.8 从pandas提取数据	57
2.6 NumPy快速操作和计算	58
2.6.1 矩阵运算	60
2.6.2 NumPy数组切片和索引	61
2.6.3 NumPy数组堆叠	63
2.7 小结	65
第3章 数据科学流程	66
3.1 EDA简介	66
3.2 特征创建	70
3.3 维数约简	72
3.3.1 协方差矩阵	72
3.3.2 主成分分析	73
3.3.3 一种用于大数据的PCA变型—Randomized PCA	76
3.3.4 潜在因素分析	77
3.3.5 线性判别分析	77
3.3.6 潜在语义分析	78
3.3.7 独立成分分析	78
3.3.8 核主成分分析	78
3.3.9 受限玻耳兹曼机	80
3.4 异常检测和处理	81
3.4.1 单变量异常检测	82
3.4.2 EllipticEnvelope	83
3.4.3 OneClassSVM	87
3.5 评分函数	90
3.5.1 多标号分类	90
3.5.2 二值分类	92
3.5.3 回归	93
3.6 测试和验证	93
3.7 交叉验证	97
3.7.1 使用交叉验证迭代器	99
3.7.2 采样和自举方法	100
3.8 超参数优化	102
3.8.1 建立自定义评分函数	104
3.8.2 减少网格搜索时间	106
3.9 特征选择	108
3.9.1 单变量选择	108
3.9.2 递归消除	110
3.9.3 稳定性选择与基于L1的选择	111
3.10 小结	112
第4章 机器学习	113

4.1 线性和逻辑回归	113
4.2 朴素贝叶斯	116
4.3 K近邻	118
4.4 高级非线性算法	119
4.4.1 基于SVM的分类算法	120
4.4.2 基于SVM的回归算法	122
4.4.3 调整SVM	123
4.5 组合策略	124
4.5.1 基于随机样本的粘合策略	125
4.5.2 基于弱组合的分袋策略	125
4.5.3 随机子空间和随机分片	126
4.5.4 模型序列—AdaBoost	127
4.5.5 梯度树提升	128
4.5.6 处理大数据	129
4.6 自然语言处理—瞥	136
4.6.1 词语分词	136
4.6.2 词干提取	137
4.6.3 词性标注	137
4.6.4 命名实体识别	138
4.6.5 停止词	139
4.6.6 一个完整的数据科学示例—文本分类	140
4.7 无监督学习概述	141
4.8 小结	146
第5章 社会网络分析	147
5.1 图论简介	147
5.2 图的算法	152
5.3 图的加载、输出和采样	157
5.4 小结	160
第6章 可视化	161
6.1 matplotlib基础介绍	161
6.1.1 曲线绘图	162
6.1.2 绘制分块图	163
6.1.3 散点图	164
6.1.4 直方图	165
6.1.5 柱状图	166
6.1.6 图像可视化	167
6.2 pandas的几个图形示例	169
6.2.1 箱线图与直方图	170
6.2.2 散点图	171
6.2.3 平行坐标	173
6.3 高级数据学习表示	174
6.3.1 学习曲线	174
6.3.2 验证曲线	176
6.3.3 特征重要性	177
6.3.4 GBT部分依赖关系图	179
6.4 小结	180
• • • • •	(收起)

[数据科学导论\\_下载链接1](#)

## 标签

Python

数据科学

编程

机器学习

数据挖掘

数据分析

python

科学

## 评论

只是教你如何函数，但没有告诉使用的原因，并没有从数学上给出合适的定义。写得不如文档详细，内容还不到文档的水平

-----  
入门，部分指导启发作用 条理没有想象中好

-----  
还行

-----  
!

---

介绍概念、工具包

---

数据科学属于相对较新的知识领域，它需要成功融合线性代数、统计建模、可视化、计算语言学、图形分析、机器学习、商业智能、数据存储和检索等众多学科。本书将从介绍建立基本的数据科学工具箱开始。接着，它将引导你进入完整的数据改写和预处理阶段。我们还需要花一定量时间来解释数据类型的转换、修复、探索和处理等核心活动。然后，我们将演示高级数据科学操作，建立变量和假设选择的实验流程，优化超参数，有效地使用交叉验证和测试。最后，我们将完成数据科学精要的概述，介绍主要的机器学习算法、图的分析技术和所有用于呈现结果的可视化方法。

---

此书大致介绍了在数据科学中常用的python包的中的常用函数，正如题所言是一本入门级应用tutorial。一本薄薄180页的小书，和大牛们理论扎实的专著自然没法比。各开源包的文档固然细致全面，但毕竟对入门级研究人员来说，很多功能都用不到。况且，很可能他们连自己需要的是什么功能自己都不甚清楚。而此书就是解决这个问题，大致带着读者走进数据分析中的python应用世界，由简及难，完整地走过一个数据分析的流程。

---

[数据科学导论\\_下载链接1](#)

## 书评

本书由两位资深数据科学家撰写，融合其多年从事数据科学相关的教学和科研工作经验，借助现有的Python语法和结构知识，全面而系统地讲解进行数据科学分析和开发的相关工具、技术和\*佳实践，包含清晰的代码和简化的示例。通过阅读本书，你将深入理解Python核心概念，成为高效数据...

---

[数据科学导论\\_下载链接1](#)