

Python数据分析实战



[Python数据分析实战 下载链接1](#)

著者:[印尼] 伊凡·伊德里斯 (Ivan Idris)

出版者:机械工业出版社

出版时间:2017-8-1

装帧:平装

isbn:9787111576402

数据分析是一个快速发展的领域，而Python已经演变成数据科学的主要语言，广泛应用于数据分析、可视化和机器学习等领域。本书从数据分析的基础内容入手，比如matplotlib、NumPy和Pandas库，介绍如何通过选择色彩图和调色板来创建可视化，之后深入统计数据分析，将帮助你掌握Spark和HDFS，为网络挖掘创建可迁移脚本。本书还详细阐述如何评价股票，检测市场有效性，使用指标和聚类等，并且还将使用多线程实现并行性，并加速你的代码。在本书结束时，读者将能够运用各种Python中的数据分析技术，并针对问题场景设计解决方案。

作者介绍:

目录: Contents?目录

译者序

前言

第1章为可重复的数据分析奠定基础1

1.1简介1

1.2安装Anaconda2

1.3安装数据科学工具包3

1.4用virtualenv和virtualenvwrapper创建Python虚拟环境5

1.5使用Docker镜像沙盒化Python应用6

1.6在IPythonNotebook中记录软件包的版本和历史8

1.7配置IPython11

1.8学习为鲁棒性错误校验记录日志13

1.9为你的代码写单元测试16

1.10配置pandas18

1.11配置matplotlib20

1.12为随机数生成器和NumPy打印选项设置种子23

1.13使报告、代码风格和数据访问标准化24

第2章创建美观的数据可视化28

2.1简介28

2.2图形化安斯库姆四重奏28

2.3选择Seaborn的调色板31

2.4选择matplotlib的颜色表33

2.5与IPythonNotebook部件交互35

2.6查看散点图矩阵38

2.7通过mpld3使用d3.js进行可视化40

2.8创建热图41

2.9把箱线图、核密度图和小提琴图组合44

2.10使用蜂巢图可视化网络图45

2.11显示地图47

2.12使用类ggplot2图49

2.13使用影响图高亮数据51

第3章统计数据分析和概率53

3.1简介53

3.2将数据拟合到指数分布53

3.3将聚合数据拟合到伽马分布55

3.4将聚合计数拟合到泊松分布57

3.5确定偏差59

3.6估计核密度61

3.7确定均值、方差和标准偏差的置信区间64

3.8使用概率权重采样66

3.9探索极值68

3.10使用皮尔逊相关系数测量变量之间的相关性71

3.11使用斯皮尔曼等级相关系数测量变量之间的相关性74

3.12使用点二列相关系数测量二值变量和连续变量的相关性77

3.13评估变量与方差分析之间的关系78

第4章处理数据和数值问题81

4.1简介81

4.2剪辑和过滤异常值81

4.3对数据进行缩尾处理84

4.4测量噪声数据的集中趋势85

4.5使用Box-Cox变换进行归一化88

- 4.6使用幂阶梯转换数据90
- 4.7使用对数转换数据91
- 4.8重组数据93
- 4.9应用logit()来变换比例95
- 4.10拟合鲁棒线性模型97
- 4.11使用加权最小二乘法考虑方差99
- 4.12使用任意精度进行优化101
- 4.13使用任意精度的线性代数103
- 第5章网络挖掘、数据库和大数据107
- 5.1简介107
- 5.2模拟网页浏览108
- 5.3网络数据挖掘110
- 5.4处理非ASCII文本和HTML实体112
- 5.5实现关联表114
- 5.6创建数据库迁移脚本117
- 5.7在已经存在的表中增加一列117
- 5.8在表创建之后添加索引118
- 5.9搭建一个测试Web服务器120
- 5.10实现具有事实表和维度表的星形模式121
- 5.11使用Hadoop分布式文件系统126
- 5.12安装配置Spark127
- 5.13使用Spark聚类数据128
- 第6章信号处理和时间序列132
- 6.1简介132
- 6.2使用周期图做频谱分析132
- 6.3使用Welch算法估计功率谱密度134
- 6.4分析峰值136
- 6.5测量相位同步138
- 6.6指数平滑法140
- 6.7评估平滑法142
- 6.8使用Lomb-Scargle周期图145
- 6.9分析音频的频谱146
- 6.10使用离散余弦变换分析信号149
- 6.11对时序数据进行块自举151
- 6.12对时序数据进行动态块自举153
- 6.13应用离散小波变换155
- 第7章利用金融数据分析选择股票159
- 7.1简介159
- 7.2计算简单收益率和对数收益率159
- 7.3使用夏普比率和流动性对股票进行排名161
- 7.4使用卡玛和索提诺比率对股票进行排名162
- 7.5分析收益统计164
- 7.6将个股与更广泛的市场相关联166
- 7.7探索风险与收益169
- 7.8使用非参数运行测试检验市场170
- 7.9测试随机游走173
- 7.10使用自回归模型确定市场效率175
- 7.11为股票价格数据库建表177
- 7.12填充股票价格数据库178
- 7.13优化等权重双资产组合183
- 第8章文本挖掘和社交网络分析186
- 8.1简介186
- 8.2创建分类的语料库186
- 8.3以句子和单词标记化新闻文章189

8.4词干提取、词形还原、过滤和TF-IDF得分	189
8.5识别命名实体	193
8.6提取带有非负矩阵分解的主题	194
8.7实现一个基本的术语数据库	196
8.8计算社交网络密度	200
8.9计算社交网络接近中心性	201
8.10确定中介中心性	202
8.11评估平均聚类系数	203
8.12计算图的分类系数	204
8.13获得一个图的团数	205
8.14使用余弦相似性创建文档图	206
第9章集成学习和降维	209
9.1简介	209
9.2递归特征消除	210
9.3应用主成分分析来降维	211
9.4应用线性判别分析来降维	213
9.5多模型堆叠和多数投票	214
9.6学习随机森林	217
9.7使用RANSAC算法拟合噪声数据	220
9.8使用Bagging来改善结果	222
9.9用于更好学习的Boosting算法	224
9.10嵌套交叉验证	227
9.11使用joblib重用模型	229
9.12层次聚类数据	231
9.13Theano之旅	232
第10章评估分类器、回归器和聚类	235
10.1简介	235
10.2直接使用混淆矩阵分类	235
10.3计算精度、召回率和F1分数	23
• • • • •	(收起)

[Python数据分析实战 下载链接1](#)

标签

数据分析

计算机

编程

Python

评论

数据分析体现在很多方面，比如说以做出决策或提出新的假设和问题为目的进行数据分析。数据科学以及大数据的热潮、高待遇以及经济回报让我想起了当数据存储和商业智能还是时髦词的年代。商业智能和数据存储的终极目标是构建应用于管理的可视化图表。这涉及很多政治和组织方面的利益，但是从技术的角度来看，这主要还是和数据库相关。数据科学则不是以数据库为中心，而是很大程度上依赖于机器学习。由于数据的量在不断地变多，机器学习变得越来越不可或缺。数据大量增长的背后是人口的快速增长以及新技术的层出不穷，比如说社交媒体和移动设备的出现。事实上，数据增长可能是我们唯一可以肯定的将一直持续的趋势。构建可视化图表和应用机器学习的区别就类似于搜索引擎的演进。

[Python数据分析实战_下载链接1](#)

书评

[Python数据分析实战_下载链接1](#)