

Python数据科学与机器学习



[Python数据科学与机器学习 下载链接1](#)

著者:[美]弗兰克·凯恩

出版者:人民邮电出版社

出版时间:2019-7-1

装帧:平装

isbn:9787115512413

本书介绍了使用Python进行数据分析和高效的机器学习，首先从一节Python速成课开始，然后回顾统计学和概率论的基础知识，接着深入讨论与数据挖掘和机器学习相关的60多个主题，包括贝叶斯定理、聚类、决策树、回归分析、实验设计等。

作者介绍:

Frank Kane, Sundog
Software公司创始人，曾在Amazon和IMDb工作近9年，在分布式计算、数据挖掘和机

器学习等领域持有17项专利。

目录: 第1章 入门 1

1.1 安装Enthought Canopy 1

1.2 使用并理解IPython/Jupyter Notebook 6

1.3 Python基础——第一部分 9

1.4 理解Python代码 11

1.5 导入模块 13

1.5.1 数据结构 13

1.5.2 使用列表 14

1.5.3 元组 17

1.5.4 字典 18

1.6 Python基础——第二部分 20

1.6.1 Python中的函数 20

1.6.2 循环 23

1.6.3 探索活动 24

1.7 运行Python脚本 24

1.7.1 运行Python代码的其他方式 25

1.7.2 在命令行中运行Python脚本 25

1.7.3 使用Canopy IDE 26

1.8 小结 28

第2章 统计与概率复习以及Python实现 29

2.1 数据类型 29

2.1.1 数值型数据 30

2.1.2 分类数据 30

2.1.3 定序数据 31

2.2 均值、中位数和众数 32

2.2.1 均值 32

2.2.2 中位数 33

2.2.3 众数 34

2.3 在Python中使用均值、中位数和众数 35

2.3.1 使用NumPy包计算均值 35

2.3.2 使用NumPy包计算中位数 36

2.3.3 使用SciPy包计算众数 37

2.4 标准差和方差 40

2.4.1 方差 40

2.4.2 标准差 42

2.4.3 总体方差与样本方差 42

2.4.4 在直方图上分析标准差和方差 44

2.4.5 使用Python计算标准差和方差 44

2.4.6 自己动手 45

2.5 概率密度函数和概率质量函数 45

2.5.1 概率密度函数 45

2.5.2 概率质量函数 46

2.6 各种类型的数据分布 47

2.6.1 均匀分布 47

2.6.2 正态分布或高斯分布 48

2.6.3 指数概率分布与指数定律 50

2.6.4 二项式概率质量函数 50

2.6.5 泊松概率质量函数 51

2.7 百分位数和矩 52

2.7.1 百分位数 53

2.7.2 矩 56

2.8 小结	60
第3章 Matplotlib与概率高级概念	61
3.1 Matplotlib快速学习	61
3.1.1 在一张图形上进行多次绘图	62
3.1.2 将图形保存为文件	63
3.1.3 调整坐标轴	64
3.1.4 添加网格	65
3.1.5 修改线型和颜色	65
3.1.6 标记坐标轴并添加图例	68
3.1.7 一个有趣的例子	69
3.1.8 生成饼图	70
3.1.9 生成条形图	71
3.1.10 生成散点图	72
3.1.11 生成直方图	72
3.1.12 生成箱线图	73
3.1.13 自己动手	74
3.2 协方差与相关系数	74
3.2.1 概念定义	75
3.2.2 相关系数	76
3.2.3 在Python中计算协方差和相关系数	76
3.2.4 相关系数练习	80
3.3 条件概率	80
3.3.1 Python中的条件概率练习	81
3.3.2 条件概率作业	84
3.3.3 作业答案	85
3.4 贝叶斯定理	86
3.5 小结	88
第4章 预测模型	89
4.1 线性回归	89
4.1.1 普通最小二乘法	90
4.1.2 梯度下降法	91
4.1.3 判定系数或 r^2	91
4.1.4 使用Python进行线性回归并计算 r^2	92
4.1.5 线性回归练习	94
4.2 多项式回归	95
4.2.1 使用NumPy实现多项式回归	96
4.2.2 计算 r^2 误差	98
4.2.3 多项式回归练习	98
4.3 多元回归和汽车价格预测	99
4.3.1 使用Python进行多元回归	100
4.3.2 多元回归练习	102
4.4 多水平模型	102
4.5 小结	104
第5章 使用Python进行机器学习	105
5.1 机器学习及训练/测试法	105
5.1.1 非监督式学习	106
5.1.2 监督式学习	107
5.2 使用训练/测试法防止多项式回归中的过拟合	109
5.3 贝叶斯方法——概念	113
5.4 使用朴素贝叶斯实现垃圾邮件分类器	115
5.5 k均值聚类	118
5.6 基于收入与年龄进行人群聚类	121
5.7 熵的度量	123
5.8 决策树——概念	124

- 5.8.1 决策树实例 126
- 5.8.2 生成决策树 127
- 5.8.3 随机森林 127
- 5.9 决策树——使用Python预测录用决策 128
- 5.9.1 集成学习——使用随机森林 132
- 5.9.2 练习 133
- 5.10 集成学习 133
- 5.11 支持向量机简介 135
- 5.12 使用scikit-learn通过SVM进行人员聚集 137
- 5.13 小结 140
- 第6章 推荐系统 141
- 6.1 什么是推荐系统 141
- 6.2 基于项目的协同过滤 145
- 6.3 基于项目的协同过滤是如何工作的 146
- 6.4 找出电影相似度 149
- 6.5 改善电影相似度结果 155
- 6.6 向人们推荐电影 159
- 6.7 改善推荐结果 165
- 6.8 小结 167
- 第7章 更多数据挖掘和机器学习技术 168
- 7.1 k最近邻的概念 168
- 7.2 使用KNN预测电影评分 170
- 7.3 数据降维与主成分分析 176
- 7.3.1 数据降维 176
- 7.3.2 主成分分析 177
- 7.4 对鸢尾花数据集的PCA示例 178
- 7.5 数据仓库简介 182
- 7.6 强化学习 184
- 7.6.1 Q-learning 185
- 7.6.2 探索问题 186
- 7.6.3 时髦名词 186
- 7.7 小结 188
- 第8章 处理真实数据 189
- 8.1 偏差-方差权衡 189
- 8.2 使用k折交叉验证避免过拟合 192
- 8.3 数据清理和标准化 196
- 8.4 清理Web日志数据 198
- 8.4.1 对Web日志应用正则表达式 198
- 8.4.2 修改1——筛选请求字段 200
- 8.4.3 修改2——筛选post请求 201
- 8.4.4 修改3——检查用户代理 203
- 8.4.5 筛选爬虫与机器人 204
- 8.4.6 修改4——使用网站专用筛选器 205
- 8.4.7 Web日志数据练习 206
- 8.5 数值型数据的标准化 207
- 8.6 检测异常值 208
- 8.6.1 处理异常值 209
- 8.6.2 异常值练习 211
- 8.7 小结 211
- 第9章 Apache Spark——大数据上的机器学习 212
- 9.1 安装Spark 212
- 9.1.1 在Windows系统中安装Spark 213
- 9.1.2 在其他操作系统上安装Spark 214
- 9.1.3 安装Java Development Kit 214

9.1.4 安装Spark	217
9.2 Spark简介	227
9.2.1 可伸缩	227
9.2.2 速度快	228
9.2.3 充满活力	229
9.2.4 易于使用	229
9.2.5 Spark组件	229
9.2.6 在Spark中使用Python还是Scala	230
9.3 Spark和弹性分布式数据集	231
9.3.1 SparkContext对象	231
9.3.2 创建RDD	232
9.3.3 更多创建RDD的方法	233
9.3.4 RDD操作	233
9.4 MLlib简介	235
9.4.1 MLlib功能	235
9.4.2 MLlib特殊数据类型	236
9.5 在Spark中使用MLlib实现决策树	236
9.6 在Spark中实现k均值聚类	245
9.7 TF-IDF	250
9.7.1 TF-IDF实战	250
9.7.2 使用TF-IDF	251
9.8 使用Spark MLlib搜索维基百科	251
9.8.1 导入语句	252
9.8.2 创建初始RDD	252
9.8.3 创建并转换HashingTF对象	253
9.8.4 计算TF-IDF得分	254
9.8.5 使用维基百科搜索引擎算法	254
9.8.6 运行算法	255
9.9 使用Spark 2.0中的MLlib数据框API	255
9.10 小结	259
第 10 章 测试与实验设计	260
10.1 A/B测试的概念	260
10.1.1 A/B测试	260
10.1.2 A/B测试的转化效果测量	262
10.1.3 小心方差	263
10.2 t检验与p值	263
10.2.1 t统计量或t检验	264
10.2.2 p值	264
10.3 使用Python计算t统计量和p值	265
10.3.1 使用实验数据进行A/B测试	265
10.3.2 样本量有关系吗	267
10.4 确定实验持续时间	268
10.5 A/B测试中的陷阱	269
10.5.1 新奇性效应	270
10.5.2 季节性效应	271
10.5.3 选择性偏差	271
10.5.4 数据污染	272
10.5.5 归因错误	272
10.6 小结	273
• • • • •	(收起)

标签

Python

python

实践者解答

数据科学

人工智能

豆瓣

计算机科学

计算机

评论

等Python入门后就看Python数据科学//来了来了//总体来说阅读体验很好，就是python代码解释得偏少

非常不错的一本入门书 可以让你很快提起兴趣 没有复杂的数学推导
看理论还是去看吴恩达或者花书 学用Python调包战士，这本书做启蒙非常不错

[Python数据科学与机器学习_下载链接1](#)

[Python数据科学与机器学习_下载链接1](#)