

Scala机器学习



[Scala机器学习 下载链接1](#)

著者:Alexander Kozlov

出版者:机械工业出版社

出版时间:2017-7

装帧:平装

isbn:9787111572152

《Scala机器学习》全面而系统地讲解怎么使用Scala在Spark平台上实现机器学习算法，其中Scala的版本Jj为2.11.7，Spark采用基于Hadoop 2.6的版本，都是比较新的版本，并且书中还提供大量有针对性的编程实例，可以帮助你快速提高自己的工程实战能力。全书共10章，第1章介绍数据分析师如何开始数据分析；第2章介绍数据驱动过程；第3章介绍Spark体系结构以及MLlib所支持几个算法；第4章介绍机器学习的基本原理，讨论两种不同的机器学习方法——监督学习和无监督学习；第5章通过具体的算法实例介绍回归和分类；第6章详细介绍显示、存储以及改进非

结构化数据的方法；第7章深入介绍Scala的图（graph）库以及算法的实现；第8章探讨Scala与R和Python的集成；第9章介绍自然语言处理（NLP）的一些常用算法，同时介绍一些特别适合Scala编程的算法；第10章介绍现有Scala监控解决方案。

作者介绍:

亚历克斯·科兹洛夫（Alex Kozlov），是一名多学科的大数据科学家。自1991年来到硅谷起就创办了几家计算机和数据管理公司。期间，他师从Daphne Koller和John Hennessy两位教授，于1998年获得斯坦福大学博士学位。他目前是企业安全初创公司E8 Security的首席解决方案架构师，曾在Cloudera、HP公司的HPLabs工作。

罗棻，重庆工商大学计算机科学与信息工程学院教师，主要从事计算机视觉、计算机算法的研究。同时对Scala编程感兴趣。

刘波，重庆工商大学计算机科学与信息工程学院教师，主要从事机器学习理论、计算机视觉和最优化技术研究，同时爱好Hadoop和Spark平台上的大数据分析，也对Linux平台的编程和Oracle数据库感兴趣。

目录: 译者序

前言

第1章探索数据分析1

1.1 Scala入门2

1.2 去除分类字段的重复值2

1.3 数值字段概述4

1.4 基本抽样、分层抽样和一致抽样5

1.5 使用Scala和Spark的Note—book工作8

1.6 相关性的基础12

1.7 总结14

第2章数据管道和建模15

2.1 影响图16

2.2 序贯试验和风险处理17

2.3 探索与利用问题21

2.4 不知之不知23

2.5 数据驱动系统的基本组件23

2.5.1 数据收集24

2.5.2 数据转换层25

2.5.3 数据分析与机器学习26

2.5.4 UI组件26

2.5.5 动作引擎28

2.5.6 关联引擎28

2.5.7 监控28

2.6 优化和交互28

2.7 总结29

第3章使用Spark和MLlib30

3.1 安装Spark31

3.2 理解Spark的架构32

3.2.1 任务调度32

3.2.2 Spark的组件35

3.2.3 MQTT、ZeroMQ、Flume和Kafka36

3.2.4 HDFS、Cassandra、S3和Tachyon37

3.2.5 Mesos、YARN和Standalone38

3.3应用	38
3.3.1单词计数	38
3.3.2基于流的单词计数	41
3.3.3SparkSQL和数据框	45
3.4机器学习库	46
3.4.1SparkR	47
3.4.2图算法：Graphx和Graph—Frames	48
3.5Spark的性能调整	48
3.6运行Hadoop的HDFS	49
3.7总结	54
第4章监督学习和无监督学习	55
4.1记录和监督学习	55
4.1.1lirs数据集	56
4.1.2类标签点	57
4.1.3SVMWithSGD	58
4.1.4logistic回归	60
4.1.5决策树	62
4.1.6bagging和boosting：集成学习方法	66
4.2无监督学习	66
4.3数据维度	71
4.4总结	73
第5章回归和分类	74
5.1回归是什么	74
5.2连续空间和度量	75
5.3线性回归	77
5.4logistic回归	81
5.5正则化	83
5.6多元回归	84
5.7异方差	84
5.8回归树	85
5.9分类的度量	87
5.10多分类问题	87
5.11感知机	87
5.12泛化误差和过拟合	90
5.13总结	90
第6章使用非结构化数据	91
6.1嵌套数据	92
6.2其他序列化格式	100
6.3Hive和Impala	102
6.4会话化	104
6.5使用特质	109
6.6使用模式匹配	110
6.7非结构化数据的其他用途	113
6.8概率结构	113
6.9投影	113
6.10总结	113
第7章使用图算法	115
7.1图简介	115
7.2SBT	116
7.3Scala的图项目	119
7.3.1增加节点和边	121
7.3.2图约束	123
7.3.3JSON	124
7.4GraphX	126

- 7.4.1谁收到电子邮件130
- 7.4.2连通分量131
- 7.4.3三角形计数132
- 7.4.4强连通分量132
- 7.4.5PageRank133
- 7.4.6SVD++134
- 7.5总结138
- 第8章Scala与R和Python的集成139
 - 8.1R的集成140
 - 8.1.1R和SparkR的相关配置140
 - 8.1.2数据框144
 - 8.1.3线性模型150
 - 8.1.4广义线性模型152
 - 8.1.5在SparkR中读取JSON文件156
 - 8.1.6在SparkR中写入Parquet文件157
 - 8.1.7从R调用Scala158
 - 8.2Python的集成161
 - 8.2.1安装Python161
 - 8.2.2PySpark162
 - 8.2.3从Java／Scala调用Python163
 - 8.3总结167
- 第9章Scala中的NLP169
 - 9.1文本分析流程170
 - 9.2Spark的MLlib库177
 - 9.2.1TF—IDF177
 - 9.2.2LDA178
 - 9.3分词、标注和分块185
 - 9.4POS标记186
 - 9.5使用word2vec寻找词关系189
 - 9.6总结192
- 第10章高级模型监控193
 - 10.1系统监控194
 - 10.2进程监控195
 - 10.3模型监控201
 - 10.3.1随时间变化的性能202
 - 10.3.2模型停用标准202
 - 10.3.3A／B测试202
 - 10.4总结202
- • • • • (收起)

[Scala机器学习_下载链接1](#)

标签

统计学与机器学习

Scala

而知也无涯-2019

评论

也不知道是理解不了国外写书人的思维，是翻的太烂，还是本来就很烂？

干货不多，讲解不深入，大段简单代码的粘贴。适合完全没有Scala/spark基础的小白读者了解机器学的基础。

[Scala机器学习_下载链接1](#)

书评

[Scala机器学习_下载链接1](#)