



大数据分析概论

Introduction to Big Data Analysis

朱晓峰 主编

图书在版编目(CIP)数据

大数据分析概论 / 朱晓峰主编. — 南京: 南京大学出版社, 2018. 3

ISBN 978-7-305-19953-0

I. ①大… II. ①朱… III. ①数据处理—高等学校—教材 IV. ①TP274

中国版本图书馆 CIP 数据核字(2018)第 044959 号

大数据分析概论

Introduction to Big Data Analysis

朱晓峰 主编

出版发行 南京大学出版社
社 址 南京市汉口路 22 号 邮 编 210093
出 版 人 金鑫荣

书 名 大数据分析概论
主 编 朱晓峰
责任编辑 胥橙庭 王南雁 编辑热线 025-83593962

照 排 南京南琳图文制作有限公司
印 刷 南京人民印刷厂有限责任公司
开 本 787×1092 1/16 印张 25.25 字数 646 千
版 次 2018 年 3 月第 1 版 2018 年 3 月第 1 次印刷
ISBN 978-7-305-19953-0
定 价 55.00 元

网址: <http://www.njupco.com>
官方微博: <http://weibo.com/njupco>
官方微信号: njupress
销售咨询热线: (025) 83594756

- * 版权所有, 侵权必究
- * 凡购买南大版图书, 如有印装质量问题, 请与所购图书销售部门联系调换

南京大学出版社



前 言

“大数据分析”是当今科技行业最受欢迎的流行语之一,也是各领域人士极为关注的话题。飞速发展的中国,同样将大数据作为国家战略,企业实践不断涌现。

《大数据分析概论》是数据科学领域为数不多的理论与实践相结合的入门级教材,它通过详细剖析大数据分析基础理论和实例实训,全景展现了大数据分析各个阶段的基础知识、相关方法、关键技术和实用工具。

本书分为两个部分:第一部分,大数据分析的理论部分,包括“大数据分析概述”、“大数据分析的体系架构”、“大数据分析的关键技术”、“大数据分析的数据采集与存储”、“大数据分析的数据清洗”、“大数据分析的数据挖掘”和“数据可视化”;第二部分,大数据分析的实训部分,包括“体育行业 NBA 数据分析”、“金融行业贷款分析”、“服装行业库龄库存分析”、“公司财务数据分析”、“能源行业油井数据分析”、“政府行业财政收支分析”、“人力资源行业职位需求分析”和“互联网行业网站分析”等十二个不同行业的大数据分析。每个理论章节都以 1000 字左右的小案例引出本章内容,然后提供本章知识要点,最后提供案例思考题,并设计专门实验,方便学生认知和操作;每个行业实训,都包括背景分析、需求分析、大数据分析过程和分析结论。

本书由朱晓峰、赵柳榕讨论大纲,朱晓峰负责理论篇中的第 1—3 章;赵柳榕负责理论篇的第 4—6 章、实训篇的 9—10 章;张琳负责理论篇中的第 7 章,并和朱晓峰一起审核本书的理论部分;吴海东(福州大学)负责实训篇中的第 1—8 章;马小东(苏州国云)负责实训篇中的第 11 章,并和朱晓峰一起审核本书的实训部分;郑乐负责实训篇中的第 12 章。

本书在编写过程中,得到了南京工业大学校级教材重点项目的资助,得到了苏州国云数据科技有限公司、南京大学出版社的支持和帮助,尤其是经济与管理学院姚山季院长、苏州国云朱琼琼项目总监、南京大学出版社吴汀老师给予的项目申报、行业案例、分析工具、原始数据等方面的指导和帮助,在此表示衷心的感谢!

本书作者均为从事大数据类课程教学、实践的一线教师、实践者。因此,希望本书能够对大数据类课程的教学、学习和实践提供帮助。当然,本书难免有不妥之处,恳请广大读者对此教材提出宝贵意见,以期不断改进。

2.1.2 数据集中与标准化	32
2.1.3 数据报表与可视化	32
2.1.4 产品与运营分析	32
2.1.5 精细化运营	32

编 者

2018 年 2 月

目 录

理论篇

第一章 大数据分析概述	3
1.1 大数据分析的背景与基础	4
1.1.1 大数据分析的背景	4
1.1.2 大数据分析的基础	6
1.2 大数据分析的概念与原理	8
1.2.1 大数据分析的概念界定	8
1.2.2 大数据分析的基本原理	10
1.3 大数据分析的思维与误区	11
1.3.1 大数据分析的思维	11
1.3.2 大数据分析的误区	14
1.4 大数据分析的作用及影响	19
1.4.1 大数据分析对企业的作用和影响	19
1.4.2 大数据分析对社会的作用和影响	20
1.5 大数据分析的过程与对象	21
1.5.1 大数据分析的过程	21
1.5.2 大数据分析的对象	22
1.6 大数据分析的流程与基础模型	25
1.6.1 大数据分析的流程	25
1.6.2 大数据分析的基础模型	27
延伸阅读思考:大数据分析带来的改变	29
实验一:认知大数据分析的价值	29
第二章 大数据分析的体系架构	30
2.1 大数据分析的总体架构	31
2.1.1 基础 IT 系统	31
2.1.2 数据集中与标准化	32
2.1.3 数据报表与可视化	33
2.1.4 产品与运营分析	34
2.1.5 精细化运营	35

2.1.6 数据产品	36
2.2 大数据分析的技术体系	36
2.2.1 基于分析流程的大数据技术栈	36
2.2.2 基于主流软件的大数据技术栈	37
2.2.3 基于淘宝海量数据的大数据技术栈	39
2.3 大数据分析的产业架构	40
2.3.1 国外大数据分析的产业架构	40
2.3.2 国内大数据分析的产业架构	43
延伸阅读思考:携程大数据应用框架的重构	48
实验二:理解大数据分析的体系	49
第三章 大数据分析的关键技术	50
3.1 大数据分析的关键技术概述	51
3.1.1 基于大数据分析流程的关键技术	51
3.1.2 基于大数据生态的关键技术	53
3.1.3 大数据分析技术的发展趋势	55
3.2 大数据分析的基础架构 Hadoop	56
3.2.1 Hadoop 概述	56
3.2.2 Hadoop 的版本与选择	58
3.2.3 Hadoop 生态的四层架构	59
3.2.4 Hadoop 生态中的典型组件	61
3.2.5 Spark	65
3.3 大数据分析的云技术	68
3.3.1 云计算	68
3.3.2 云平台	71
3.4 大数据分析的存储技术	72
3.4.1 分布式文件系统	72
3.4.2 分布式数据库 HBase	73
3.4.3 NoSQL 数据库	74
延伸阅读思考:“大数据+人脸识别”助力众可贷	77
实验三:认知大数据分析工具——以“魔镜”为例	79
第四章 大数据分析的数据采集与存储	84
4.1 大数据采集概述	85
4.1.1 大数据采集的基本概念	85
4.1.2 大数据采集的数据源	86
4.1.3 大数据采集架构与场景	87
4.1.4 大数据采集的困境及对策	88
4.2 大数据采集工具	90

4.2.1	已有大数据采集工具的比较	90
4.2.2	大数据采集工具的设计	92
4.3	大数据存储	96
4.3.1	传统存储面临的挑战	96
4.3.2	大数据存储概述	98
4.3.3	大数据存储的技术路线	103
4.3.4	大数据存储和管理数据库系统	104
	延伸阅读思考:医疗大数据——数据收集或是最难点	108
	实验四:大数据分析的数据导入与编辑	109
第五章	大数据分析的数据清洗	117
5.1	大数据质量	118
5.1.1	大数据质量概述	118
5.1.2	大数据质量产生的根源	121
5.1.3	大数据质量问题的分类与实例	125
5.2	大数据清洗概述	127
5.2.1	大数据清洗定义	127
5.2.2	大数据清洗的对象	127
5.2.3	大数据清洗的总体架构	129
5.2.4	大数据清洗与数据质量的关系	130
5.3	大数据清洗的方法与工具	131
5.3.1	大数据清洗方法概述	131
5.3.2	可视化大数据清洗	132
5.3.3	大数据清洗的工具	135
5.4	大数据清洗的过程与具体内容	139
5.4.1	大数据清洗的过程	139
5.4.2	大数据清洗的具体内容	142
	延伸阅读思考:微软与谷歌的拼写检查	146
	实验五:大数据分析的数据清洗	147
第六章	大数据分析的数据挖掘	156
6.1	传统数据挖掘	157
6.1.1	数据挖掘的界定	157
6.1.2	数据挖掘的基本流程	160
6.1.3	数据挖掘面临的主要问题	164
6.2	大数据和数据挖掘	165
6.2.1	递进升级学说	165
6.2.2	一体两面学说	167
6.2.3	互相促进学说	168

00	6.2.4 其他学说	169
50	6.3 大数据挖掘的任务	170
80	6.3.1 分类	170
80	6.3.2 聚类	172
80	6.3.3 关联分析	173
201	6.3.4 估测和预测	174
101	6.4 大数据挖掘的流程	175
801	6.5 大数据挖掘的常用算法	177
001	6.5.1 决策树	177
111	6.5.2 遗传算法	183
301	6.5.3 神经网络	186
811	6.5.4 关联规则	189
151	6.5.5 粗糙集	191
251	延伸阅读思考:大数据预测——真的有那么神奇吗?	193
351	实验六:大数据挖掘实验	194
	第七章 大数据分析的数据展现	197
151	7.1 数据可视化概述	198
051	7.1.1 数据可视化的含义	198
081	7.1.2 数据可视化的应用价值和应用领域	200
161	7.1.3 数据可视化的工具	201
181	7.1.4 数据可视化步骤	205
281	7.2 数据可视化的基础要素	208
281	7.2.1 数据	208
081	7.2.2 图表	209
051	7.3 数据可视化的表现形式	214
511	7.3.1 数据可视化的常见方式	214
811	7.3.2 不同类型数据的展示	215
111	7.3.3 不同类型图形的展示	217
181	7.4 数据可视化的设计	220
121	7.4.1 设计的基本理念	220
111	7.4.2 图表设计技巧	222
091	7.4.3 配色方案设计	224
181	7.4.4 字体设计	227
181	7.4.5 应用场景设计	228
201	7.5 数据可视化的改进	231
581	7.5.1 总体思路	231
201	7.5.2 图表改进思路	233

延伸阅读思考:《卫报》的数据可视化与数据新闻	235
实验七:数据图表规范化和美化	238

实训篇

第一章 财务数据分析	247
1.1 实训背景知识	247
1.2 实训简介	247
1.2.1 原始数据情况	247
1.2.2 实训分析过程	248
1.3 实训过程	248
1.3.1 新建项目	248
1.3.2 数据导入	249
1.3.3 数据处理	249
1.3.4 数据分析	251
1.3.5 数据可视化	255
1.4 实训总结	259
1.5 实训思考题	260
第二章 库龄库存分析	261
2.1 实训背景知识	261
2.2 实训简介	261
2.2.1 原始数据情况	261
2.2.2 实训分析过程	261
2.3 实训过程	262
2.3.1 新建项目	262
2.3.2 数据导入	262
2.3.3 数据处理	263
2.3.4 数据分析	263
2.3.5 数据可视化	267
2.4 实训总结	268
2.5 实训思考题	268
第三章 销售数据分析	269
3.1 实训背景知识	269
3.2 实训简介	269
3.2.1 原始数据情况	269
3.2.2 实训分析过程	270
3.3 实训过程	270

3.3.1	新建项目	270
3.3.2	数据导入	271
3.3.3	数据处理	271
3.3.4	数据分析	272
3.3.5	数据挖掘	279
3.3.6	数据可视化	280
3.4	实训总结	281
3.4.1	实训总结结论	281
3.4.2	实训总结建议	282
3.5	实训思考题	282
第四章	油井数据分析	283
4.1	实训背景知识	283
4.2	实训简介	283
4.2.1	原始数据情况	283
4.2.2	实训分析过程	284
4.3	实训过程	284
4.3.1	新建项目	284
4.3.2	数据导入	285
4.3.3	数据处理	285
4.3.4	数据分析	288
4.3.5	数据挖掘	291
4.3.6	数据可视化	292
4.4	实训总结	294
4.5	实训思考题	294
第五章	网站流量分析	295
5.1	实训背景知识	295
5.2	实训简介	296
5.2.1	原始数据情况	296
5.2.2	实训分析过程	296
5.3	实训过程	297
5.3.1	新建项目	297
5.3.2	数据导入	297
5.3.3	数据处理	298
5.3.4	数据分析	298
5.3.5	数据挖掘	304
5.4	实训总结	306
5.5	实训思考题	306

第六章 楼盘数据分析	307
6.1 实训背景知识	307
6.2 实训简介	307
6.2.1 原始数据情况	307
6.2.2 实训分析过程	308
6.3 实训过程	308
6.3.1 新建项目	308
6.3.2 数据导入	309
6.3.3 数据处理	309
6.3.4 数据分析	310
6.3.5 数据可视化	313
6.4 实训总结	314
6.5 实训思考题	314
第七章 贷款数据分析	315
7.1 实训背景知识	315
7.2 实训简介	315
7.2.1 原始数据情况	315
7.2.2 实训分析过程	316
7.3 实训过程	316
7.3.1 新建项目	316
7.3.2 数据导入	317
7.3.3 数据处理	317
7.3.4 数据分析	318
7.3.5 数据可视化	324
7.4 实训总结	325
7.5 实训思考题	325
第八章 NBA 数据分析	326
8.1 实训背景知识	326
8.2 实训简介	326
8.2.1 原始数据情况	326
8.2.2 实训分析过程	326
8.3 实训过程	327
8.3.1 新建项目	327
8.3.2 数据导入	327
8.3.3 数据处理	328
8.3.4 数据分析	329
8.3.5 数据可视化	333

8.4	实训总结	336
8.5	实训思考题	336
第九章 行业职位需求分析		337
9.1	实训背景知识	337
9.2	实训简介	337
9.2.1	原始数据情况	337
9.2.2	实训分析过程	338
9.3	实训过程	338
9.3.1	新建项目	338
9.3.2	数据导入	339
9.3.3	数据处理	339
9.3.4	数据分析	341
9.3.5	数据可视化	348
9.4	实训总结	349
9.5	实训思考题	350
第十章 水资源数据分析		351
10.1	实训背景知识	351
10.2	实训简介	351
10.2.1	原始数据情况	351
10.2.2	实训分析过程	352
10.3	实训过程	352
10.3.1	新建项目	352
10.3.2	数据导入	353
10.3.3	数据处理	353
10.3.4	数据分析	354
10.3.5	数据可视化	360
10.4	实训总结	362
10.5	实训思考题	362
第十一章 国民经济数据分析		363
11.1	实训背景知识	363
11.2	实训简介	363
11.2.1	原始数据情况	363
11.2.2	实训分析过程	364
11.3	实训过程	364
11.3.1	新建项目	364
11.3.2	数据导入	365
11.3.3	数据处理	365

11.3.4 数据分析·····	367
11.3.5 数据可视化·····	373
11.4 实训总结·····	374
11.5 实训思考题·····	374
第十二章 政府财政预算分析·····	375
12.1 实训背景知识·····	375
12.2 实训简介·····	375
12.2.1 原始数据情况·····	376
12.2.2 实训分析过程·····	376
12.3 实训过程·····	376
12.3.1 新建项目·····	376
12.3.2 数据导入·····	377
12.3.3 数据处理·····	377
12.3.4 数据分析·····	378
12.3.5 数据可视化·····	383
12.3.6 数据分析结果的分享·····	386
12.4 实训总结·····	386
12.5 实训思考题·····	386

大数据分析的世界与价值

2009年2月19日, Nature 上面有一篇文章, "Coastal influenza search engine query data", 描述了 Google 基于用户的搜索引擎(其中包括搜索频率以及用户 IP 地址等信息)的汇总信息, 成功"预测"了流感病人的数量。

那么, Google 为什么要做这件事情呢? 在美国, 由疾控中心(CDC)统计各个地区的疾病就诊人数, 然后汇总并公布。但是, 这个公平的数据一出来, 也就是说真实的流感的全国就诊人数, 要在两周之后才加进。Google 建立了一个预测平台, 把这个数据提前公布出来。因此, Google 做这件事情的预测什么时候流感来, 而是将 CDC 已经获得但是延迟时公布的数据提前公布出来。"越及时的数据, 价值越高", 数据是有价值意义的, 所以, 它在公共管理领域还是商业领域都具有重大的意义。

Google 成功"预测"流感病人的例子成为经典案例的原因在于它成功了, Google 就真正证明了大数据是"万能的", 因为 Google 在数据的处理, 只用了很简单的 Logistic 回归关系, 却成功地预测了复杂的。Google 用了简单的方法, 预测复杂的问题, 充分证明了 Google 的大数据一切!

大数据的观点之一认为, 海量的数据可以弥补模型的不完善, 如果模型甚至根本就不需要。这种观点目前仍然处于争论中, 很多理论和方法的专家们对此既困惑又试图批判。但无论如何, Google 的例子证明了支持大数据的一方, 如果 Google 的案例是成功的, 那么支持, 拥有海量数据可以解决任意复杂的问题, 大数据解决大问题!

就在 Nature 发表论文的时候, Google 的预测还是准确的, 不过存在一定的偏差, 偏差最大甚至高出了标准值(CDC 公布的结果)将近一倍。

Google 预测的失败也确实是由于过度地依赖于数据, 导致很多被忽略的结果产生了很大的影响。对客观世界进行预测需要模型, 模型首先来自于理论, 数据对模型进行训练, 对模型进行优化完善。大数据观点强调模型对数据可能地忽略理论构造这一部分的意义, 这就有可能带来隐患。

因此, Google 的案例就是一个很好的大数据的应用, 同时也为大数据道路起到了很好的指示灯作用。



案例导读

大数据分析的光环与陷阱

2009年2月19日, Nature 上面有一篇文章, “Detecting influenza epidemics using search engine query data”, 论述了 Google 基于用户的搜索日志(其中包括搜索关键词、用户搜索频率以及用户 IP 地址等信息)的汇总信息, 成功“预测”了流感病人的就诊人数。

那么, Google 为什么要做这件事情呢? 在美国, 由疾控中心 CDC 专门负责统计美国本土各个地区的疾病就诊人数, 然后汇总并公布。但是, 这个公布的数据一般要延迟两周左右, 也就是说当天的流感的全国就诊人数, 要在两周之后才知道, Google 就利用搜索引擎搭建了一个预测平台, 把这个数据提前公布出来。因此, Google 做的工作并不是实际意义上的预测什么时候流感来, 而是将 CDC 已经获得但是没及时公布的数据提前给“猜”出来, 然后公布出来。“越及时的数据, 价值越高”, 数据是有价值属性的。所以, Google 的工作无论在公共管理领域还是商业领域都具有重大的意义。

Google 成功“预测”流感病人的例子成为经典案例的深层次原因在于, 如果在这个案例上成功了, Google 就真正证明了大数据是“万能的”。因为 Google 在这项研究中对于数据的处理, 只用了很简单的 Logistic 回归关系, 却成功地预测了复杂的流感规模的问题。Google 用了简单的方法, 预测复杂的问题, 充分证明了 Google 的大数据价值观——大就是一切!

大数据的观点之一认为, 海量的数据可以弥补模型的不足, 如果数据足够大, 理论模型甚至根本就不需要。这种观点目前仍然处于争论中, 偏重理论和实证(强调数据和统计方法)的专家们对此既惶恐又试图批判。但无论如何, Google 对于流感预测的研究无疑站在了支持大数据的一方, 如果 Google 的案例是成功的, 那么或许, 拥有海量数据就真的意味着可以解决任意复杂的问题, 大数据解决大问题!

截至 Nature 发表论文的时候, Google 的预测还是准确的, 不过到后来就发生了很大的偏差, 偏差最大甚至高出了标准值(CDC 公布的结果)将近一倍^①。

Google 预测的失败也确实是过度地依赖于数据, 导致很多被忽略了的因素对预测的结果产生了很大的影响。对客观世界进行预测需要模型, 模型首先来自于理论构造, 其次需要数据对模型进行训练、对模型进行优化完善。大数据观点强调模型对数据训练的依赖, 而尽可能地忽略理论构造这一部分的意义, 这就有可能带来隐患。

因此, Google 的案例既是一个很好的大数据的应用, 同时也为大数据在未来的发展道路起到了很好的指示灯作用。

^① 数据大湿的博客. 从 Google 预测流感引发的大数据反思[EB/OL]. (2015-07-12). http://blog.sina.com.cn/s/blog_1464091000102vmeb.html.

学习目标

- 理解大数据分析产生的背景和现状
- 理解和掌握大数据的真正含义与价值
- 理解和掌握大数据分析的基本概念
- 熟悉正确的大数据分析思维
- 了解大数据分析的影响与价值
- 熟悉大数据分析的过程与对象
- 熟悉大数据分析的流程

1.1 大数据分析的背景与基础

大数据分析是数学与计算机科学相结合的产物,在 20 世纪早期就已确立,但直到计算机的出现才使得实际操作成为可能,并使得大数据分析得以推广。在学习大数据分析时,应该首先了解它的产生背景和基础。

1.1.1 大数据分析的背景

大数据分析的产生有其深刻的时代背景和历史的必然性,是 IT 技术的发展变革以及商务应用需求驱动的必然结果。

1. 数据的价值,已有时日

数据,已经渗透到当今每一个行业和业务职能领域,成为重要的生产因素。社会对于海量数据的挖掘和运用,预示着新一波生产率增长和消费者盈余浪潮的到来。数据在物理学、生物学、环境生态学等领域以及军事、金融、通信等行业存在已有时日,却因为近年来互联网和信息行业的发展而引起社会关注。数据正在迅速膨胀并变大,它决定着企业的未来发展,虽然很多企业可能并没有意识到由数据爆炸性增长带来的问题隐患,但是随着时间的推移,社会将越来越多地意识到数据对企业的重要性。

三分技术,七分数据,得数据者得天下。维克托·迈尔-舍恩伯格在《大数据时代》一书中举了诸多例证,都是为了说明一个道理:在大数据时代已经到来时要用大数据思维去发掘大数据的潜在价值。

大数据就是核心竞争力。全世界都在高呼大数据时代来临的优势:一家超市如何从一个 17 岁女孩的购物清单中,发现了她已怀孕的事实;或者将啤酒与尿不湿放在一起销售,神奇地提高了双方的销售额。实际上,数据已经无处不在,衣食住行、喜怒哀乐、吃喝玩乐都以数据的形式存在。通过数据来记录这个世界,再通过研究数据去发现这个世界。正如 IBM 所言:大数据时代——用智慧的分析洞察、构建智慧的地球。

2. 数据的数量,与日俱增

数据的数量,到底有多大? 一组名为“互联网上一天”的数据显露无疑:一天之中,互联网产生的全部内容可以刻满 1.68 亿张 DVD;发出的邮件有 2 940 亿封之多(相当于美国两年的纸质信件数量);发出的社区帖子达 200 万个(相当于《时代》杂志 770 年的文字量);卖出的手