大数据时代的统计与人工智能系列教材

数据科学导论

——R 与 Python 实现

主　编　吴喜之
副主编　刘　苗

高等教育出版社

大数据时代的统计与人工智能系列教材

回归分析——方法、数据与R的应用　刘　超　主编

▶ 数据科学导论——R与Python实现　吴喜之　主编

ISBN 978-7-04-052545-8

数据及其他资源

定价：48.00元

大数据时代的统计与人工智能系列教材

数据科学导论
——R与Python实现

主　编　吴喜之
副主编　刘　苗

高等教育出版社·北京

内容简介

本书的内容安排基本上按照数据科学的步骤及思维，所有讨论都基于实际数据案例并通过 R 和 Python 两种软件实现．我们尽可能地给出所有数据集的来源，读者可以通过相关网址及本书的二维码获取．

本书第一部分为读者介绍了两款软件的基本操作，帮助读者快速入门．第二部分内容包括数据及探索性数据分析、数据整理及清理，我们力求打破传统教科书中的“规整案例”模式，希望给读者呈现更“真实”的数据预处理的面貌，并通过图形可视化方法勾勒数据的基本“画像”．第三部分介绍有监督学习中十余种分类及回归学习方法．第四部分介绍降维和聚类两类无监督学习方法．这两部分是本书的核心内容所在．第五部分通过案例简要介绍网络爬虫技术，供读者参考．在课程内容的选择、课程进度及授课方式方面，我们认为必须给任课教师以最大的自由度和自主权，一切依学生的基础、条件及需要而定，介绍对学生最有帮助的内容，使学生的思维空间向无限的可能开放．

本书既可作为本科各专业二、三年级数据科学的入门课程教材，也可供广大实际工作者参考．

图书在版编目（CIP）数据

数据科学导论：R 与 Python 实现 / 吴喜之主编．--北京：高等教育出版社，2019. 11

ISBN 978-7-04-052545-8

Ⅰ. ①数… Ⅱ. ①吴… Ⅲ. ①数据处理 Ⅳ. ① TP274

中国版本图书馆 CIP 数据核字（2019）第 181807 号

Shuju Kexue Daolun —— R yu Python Shixian

策划编辑	吴淑丽	责任编辑	吴淑丽	封面设计	王 鹏	版式设计	徐艳妮
责任校对	吕红颖	责任印制	韩 刚				

出版发行	高等教育出版社	网　址	http://www.hep.edu.cn
社　址	北京市西城区德外大街 4 号		http://www.hep.com.cn
邮政编码	100120	网上订购	http://www.hepmall.com.cn
印　刷	河北新华第一印刷有限责任公司		http://www.hepmall.com
开　本	787mm × 1092mm　1/16		http://www.hepmall.cn
印　张	20.25		
字　数	480 千字	版　次	2019 年 11 月第 1 版
购书热线	010-58581118	印　次	2019 年 11 月第 1 次印刷
咨询电话	400-810-0598	定　价	48.00 元

本书如有缺页、倒页、脱页等质量问题，请到所购图书销售部门联系调换

版权所有　侵权必究

物 料 号　52545-00

前　言

1. 作为科学的数据科学

数据科学属于科学, 科学的方法或过程就是认识世界的方法或过程:

(1) 从感兴趣的问题得到观测或实验数据;

(2) 基于数据对产生数据的机制提出理论 (亦称假说或模型);

(3) 当理论与进一步的观测或实验数据有矛盾时, 则需要更新或替代旧的理论;

科学的过程就是上面 (1)–(3) 的不断重复.

由于所有的科学都以证据为基础, 因此很难区分数据科学和其他科学. 但人们通常认为, 其他科学主要处理有相对确定规律的自然现象, 而数据科学的对象则还包括社会科学这样具有很大不确定性的现象. 需要指出的是: 包括数据科学在内的所有科学理论在社会中的具体应用可能不属于科学, 而属于社会科学.

包括数据科学在内的科学理论有下面一些特点:

- 所有的理论 (模型或假说) 仅仅是对被研究对象规律的一种近似;
- 如果被研究的是自然规律, 那每一次对模型的更新就离真正的规律近了一步, 但到底距离多少, 永远也不知道;
- 如果被研究的是自然规律, 那么试验应该是可以重复的, 而且理论 (模型或假说) 应该可以证伪.

社会科学和科学是两个不同的概念. 科学也称为自然科学, 以区别于完全不同的社会科学. 社会科学的各种概念和准则依社会组成、种族、群体、地域、历史、国度、信仰、发展水平及时代的不同而不可能有完全一致的共识. 也就是说, 对于社会科学, 不可能有一个无争议的价值体系, 任何涉及社会科学的理论 (模型或假说) 都是不可能证伪的, 社会实验也是不可能重复的. 所有社会科学的理论 (模型或假说) 都只有局部或相对的意义.

2. 数据科学的要点

数据科学的步骤大体上应该是:

(1) 确定感兴趣的问题 (包括目标);

(2) 根据问题确定收集哪些变量的数据, 这通常需要大量艰苦的调查研究及定性分析;

(3) 对数据进行清理, 形成可以建模的形式, 这一步是必需的, 而且往往是耗费资源的;

(4) 对数据做探索性数据分析, 使用包括可视化在内的各种可能方法;

(5) 根据问题寻找或创造符合目标要求的模型, 其中包括对各种模型的对比和交叉验证;

(6) 对于得到的结果根据实际问题的性质 (最好由该领域的工作者) 做出判断; 在展示结果时, 尽量使用可视化方法;

(7) 应用于实际, 并根据需要回到 (1) 或 (2)。

上述相邻两步之间都可能要进行多次互动反复, 也就是说这些步骤不一定总是单方向的. 上述过程实际上反映了人类认识世界的过程. 简单地说就是数据和模型之间的反复互动过程. 比如, 在婴儿时期, 孩童们只有和父母或其他家庭成员打交道的经历 (数据), 形成了与人交往所遵循的最初模型; 这些模型随着和其他人打交道而不断变化, 比如和邻家儿童玩耍或幼儿园同学交往, 就必须学会和非家人或陌生人打交道 (从新数据得到新的模型); 如此下去, 逐渐成长; 这个过程伴随我们一生.

3. 数据科学和统计

统计可以而且应该定义为数据科学, 根据不列颠百科全书[①]: **Statistics: the science of collecting, analyzing, presenting and interpreting data.** 也就是说, **统计是收集、分析、展示及解释数据的科学.** 但是, 按照中国大百科全书[②]: 数理统计是 **"数学的一个分支学科".** 这种说法也出现在一些其他的百科全书之中. 为什么会这样呢? 这是由统计的历史造成的.

在 20 世纪, 气象学家、优生学家、社会达尔文学家、生物统计学家 Karl Pearson 以及遗传学家、生物学家 Ronald Aylmer Fisher 等创造的统计最基本的概念和方法, 属于典型的**数据驱动** (data driven) 思维.

在前计算机时代, 数据量相对较小, 信息量相对缺乏. 在没有计算机的情况下, 为了可计算性, 后继的以数学家为主的学者把统计逐渐发展成数学的一个分支, 推断基于对数据及模型的主观数学假定, 使用了大量数学演绎方法, 手段是基于汇总统计量对假想模型的参数进行假设检验和估计, 最终结论是关于总体性质的以 "显著性" 为代表的不确定结论. 前计算机时代发展的经典数理统计 (及其延伸) 的思维属于**模型驱动** (model driven).

而计算机时代现实世界的科技及经济的发展所需要的主要是关于个体特征的预测, 关注的是可预测性、可计算性及稳定性[③]. 其思维是**数据或问题驱动** (data/problem driven).

计算机的发展可以完全不利用数学假定而通过各种算法得到关于个体性质的较精确的结论, 这大大优于传统数理统计只能做涉及虚构总体性质的推断. 最早的大部分著名算法都是数学家和其他领域科学家创造的, 没有多少传统数理统计学家参与, 因此没有被广泛地认为是统计的一部分, 但这显然应该属于数据科学. 随着机器学习、数据挖掘及人工智能的发展, 这些基于算法模型的文化和基于数学模型的传统数理统计文化渐行渐远. 这里涉及的不是统计的字面定义, 而是人们心中潜在的定义及对自己贴什么标签的问题.

统计及其方向事实上是根据下面事实被潜在定义的:

(1) 研究生课程列表 (各个学校教授的意愿);

(2) 统计杂志发表的文章种类 (杂志的主编副主编们的偏好);

---

① Encyclopædia Britanica, `https://www.britannica.com/science/statistics`

② 《中国大百科全书》数学卷 593 页, 1988 年第一版.

③ 英文是: predictability, computability, stability.

(3) 博士生导师们的具体研究方向 (往往只是某方向的一个点);

(4) 任何人自己的知识范围.

但是, 作为服务于所有领域的数据科学, 能这样定义吗? 本书并不想把精力耗费在这些没有共识的定义上, 而只是想介绍数据科学的最基本及最主要的内容.

注重总体还是注重个体是传统统计及现代数据科学的重要区别. 在典型的传统数理统计中, 首先对数据及变量之间的关系做出数学假定; 然后通过数据对假定模型的参数做出推断: 假设检验及参数估计; 推断中所用的统计量大都是矩的函数 (都涉及 "$\sum$"); 得到的结果是以 P 值为代表的不确定结论. 有关参数都是涉及总体的, 比如总体均值, 总体方差, 或者线性回归中的系数等, 但这些个别参数是否能够代表总体, 比如收入的均值能够代表极端偏态收入总体的分布吗? **本书不会花费篇幅于前计算机时代基于假想数学模型的关于总体参数性质的不确定推断上.**

在现代数据科学中, 人们关注的是个体特征预测的准确性, 通常涉及关于个体的决策, 比如预测个人的信用程度、某患者的疾病种类、某公司的预期收益等. **本书将着重介绍基于数据本身做出对个体特征判断的数据科学方法.**

我们应该介绍对学生最有帮助的内容, 而不是某些陈旧教条所认为 "应该灌输" 的内容; 应该使学生的思维空间向无限的可能开放, 而不是约束在一个被主观认为 "正确" 的狭隘方向. 科学就是在这种开放无拘束的环境下不断发展的.

4. 本书使用的软件

本书准备使用 R 和 Python 两种编程语言来实现各种方法. 并且希望完全不会这两种软件的读者通过本书能够基本掌握这两种软件在数据科学中的使用. 当然, 对于每个读者来说, 学习软件应该根据自己的基础和需要进行. 软件必须服务于应用, 也只能从数据分析的能力上来体现使用软件的熟练程度.

2017 年, Python 已经超过了 R 在数据分析领域里占据了首位. 根据调查 (KDnuggets Poll 2017), 2017 年数据分析领域中使用 Python 的占 41%, 使用 R 的占 36%, 同时使用 R 及 Python 的占 12%, 而使用其他各种软件 (包括各种如 SAS、SPSS、Eviews、Stata、Matlab 等商业软件) 的总和仅有 11%. 在 2017 年通用软件领域中使用 Python 的人已经超过使用 JavaScript 和 Java 的人数而占据首位.

未来会不会编程将成为是否是 "文盲" 的标准. 我们这里讲的不是一两种计算机语言的编程, 而是下面意义上的**泛型编程能力**: 不同的程序语言有相似的思维方式和可比较的逻辑形式, 因此, 它们在某种意义上是相通的. 由于各种编程语言的相似性, 人们往往在掌握了一种编程语言之后就能很快地学会第二种、第三种, 等等. 编程语言之间的差别远远小于民族语言之间的差别.

5. 学习及研究都需要问题驱动

对于念了十几年书的人, "如何学习" 似乎不是问题. 但为什么不同的人学习的效果差别那么大呢? 其实, **每个人在其生命的任何阶段都可以改变学习和思维的方式, 并因此而重新塑造以后的人生.**

- **应该根据所面对问题的需要来学习，也就是拉动式学习．这种学习可以保持很强的持续学习动力，**相信每个已经工作的人都对此有程度不同的感受．当你遇到非解决不可的实际问题时，你就会千方百计寻找解决的途径，而且在此过程中会不断产生疑问并自己寻求答案．人们就是在解决实际问题中成长的．**这种学习过程皆以自学为主．**学校的书本教学天生缺乏这种动力，做习题及做实验都是对这种缺陷的弥补．教师在讲授任何知识时都应让学生充分理解其实际意义．
- **要想解决实际问题，必须首先抓住问题的本质，即主要框架，必要时才关注某些非本质的细节．**对于任何学科，如果理解了其主要的框架或理念，就可以很快掌握其基本思想，而且很容易做出推广和创新．在数据科学中，有很多模型，关键是要知道模型的主要原理，只有在必要时才关注那些非本质的细节或小技巧．人们爬山时只有关注周围的总体地形及特征，才能了解环境并欣赏其美丽，也不会迷路；但如果拿着放大镜爬山，仅注意细小的砂石草木，则难免会摔跟头和迷路．**死记硬背得到的现成知识肯定不会超过一个硬盘或图书馆，而硬盘和图书馆不会产生任何新的思想和价值．**
- **真正有意义而又需要我们解决的实际问题通常没有现成的答案，也没有人可以指导，因此必须坚持独立思考，坚持批判性思维，通过实践获得能力．**很多人可能都有这样的经验，有人带领（或者跟着导航）去各种地方，即使你去了很多次，还是不认识路，但如果自己寻找路径，可能很快就可以找到路．**跟着别人的思路走，不会产生自己的思路，只有经过自己的摸索，才能找得到方向．**家长或老师管教得越多的人，往往越不会自己思考或生存．课堂教学应该是启发式的而非大量细节的灌输．**科学发展的基础是批判性思维．科学是在否定中发展的．**任何科学理论的最终结局是被新的理论所替代或更新．
- **一个不需要解决问题的人生还算是人生吗？**

吴喜之

2019 年 5 月

目　录

第一部分　本书的安排及软件 …… 1

第 1 章　本书的内容安排及教学建议 …… 3

1.1　适用对象 …… 3

1.2　内容及教学建议 …… 3

1.3　根据需要和疑问调整教学内容 …… 5

1.3.1　拉动式学习方法 …… 5

1.3.2　根据学生的需要教学 …… 6

1.3.3　老师应该让学生产生质疑 …… 6

第 2 章　软件准备 …… 7

2.1　软件准备 …… 7

2.2　R——为领悟而运行 …… 7

2.2.1　R 简介 …… 7

2.2.2　安装和运行 …… 8

2.2.3　试试 R 编程 …… 10

2.3　Python——为领悟而运行 …… 20

2.3.1　Python 简介 …… 20

2.3.2　安装和运行 …… 21

2.3.3　试试 Python 编程 …… 22

2.3.4　Numpy 模块 …… 26

2.3.5　Pandas 模块 …… 31

2.3.6　Matplotlib 模块 …… 32

2.4　习题 …… 34

第二部分　数据及探索性数据分析 …… 35

第 3 章　数据的形式、获得及简单描述 …… 37

3.1　数据的意义及获取 …… 37

3.1.1　数据 …… 37

3.1.2　数据的获取 …… 38

3.2 标准格式数据 …… 39
3.2.1 变量和观测值 …… 39
3.2.2 数据的输入和简单展示 …… 40
3.2.3 数据阵中的元素 …… 41
3.2.4 变量的种类 …… 41
3.2.5 统计与计算机术语的区别 …… 43
3.3 简单的数据汇总 …… 44
3.3.1 一些汇总统计量 …… 44
3.3.2 汇总统计量的计算 …… 45
3.3.3 从 “被平均” 谈均值 …… 46
3.3.4 例 3.1 数据的简单汇总 …… 47
3.4 数据的图形描述 …… 48
3.4.1 程序包的安装和使用 …… 48
3.4.2 散点图 …… 49
3.4.3 直方图 …… 53
3.4.4 盒形图 …… 54
3.4.5 条形图 …… 55
3.4.6 三维曲面图及等高线图 …… 56
3.4.7 更多的图形 * …… 57
3.5 本章的 Python 代码 …… 63
3.6 习题 …… 73

第 4 章 数据整理和清理案例 …… 75
4.1 GDP 数据案例 …… 75
4.1.1 形成某年各国的多指标数据 …… 77
4.1.2 形成某国各个时期不同指标的多元时间序列 …… 78
4.1.3 形成某指标的各个时期不同国家的多元时间序列 …… 79
4.2 世界卫生组织案例之一 …… 80
4.2.1 数据的提取 …… 81
4.2.2 删除多余的行和列 …… 82
4.2.3 把多个数据合并成一个数据 …… 83
4.3 世界卫生组织案例之二 …… 84
4.3.1 输入并识别数据 …… 85
4.3.2 删除多余的行和列 …… 85
4.3.3 合并数据 …… 86
4.3.4 清理数据中的其他问题 …… 88

4.4 数据中的缺失值 …… 89
4.4.1 缺失值的审视与标记 …… 89
4.4.2 删除缺失值的方法及问题 …… 93
4.4.3 用认定值或某些准则来填补缺失值 …… 95
4.4.4 利用可预测模型来填补缺失值 …… 96
4.5 本章的 Python 代码 …… 101
4.5.1 读入例 4.1 数据形成 2016 年的各国数据 …… 101
4.5.2 形成某国各个时期不同指标的多元时间序列 …… 102
4.5.3 某指标的各个时期不同国家的多元时间序列 …… 102
4.5.4 世界卫生组织案例一 …… 103
4.5.5 每个数据集只要一年的数据来合并 …… 104
4.5.6 世界卫生组织案例二 …… 104
4.5.7 数据的缺失值 …… 107
4.6 习题 …… 113

第三部分　有监督学习 …… 115

第 5 章　有监督学习概论 …… 117
5.1 “学习” 的概念 …… 117
5.1.1 有监督学习 …… 118
5.1.2 无监督学习 …… 119
5.2 模型和拟合 …… 119
5.2.1 训练决策树模型分类的例子 …… 119
5.2.2 训练简单最小二乘线性回归模型的例子 …… 129
5.2.3 模型的普遍意义和过拟合 …… 132
5.2.4 警告: 训练集的 MSE 和测试集的 MSE 的计算不一样! …… 133
5.3 模型评价 …… 134
5.3.1 交叉验证 …… 134
5.3.2 分类问题交叉验证的预测精度标准 …… 135
5.3.3 回归问题交叉验证预测精度标准 …… 135
5.3.4 Z 折交叉验证时提取各折下标集的 R 函数 …… 136
5.4 本章的 Python 代码 …… 138
5.4.1 例 5.1 数据的决策树分类 …… 138
5.4.2 例 5.2 数据的决策树分类 …… 139
5.4.3 例 5.3 的简单线性回归 …… 140
5.4.4 Z 折交叉验证时提取各折下标集的 Python 函数 …… 141
5.5 习题 …… 143

第 6 章 最小二乘线性回归 146
6.1 基本概念 146
6.2 一个回归实例 147
6.3 多自变量线性回归系数大小有意义吗? 149
6.4 本章的 Python 代码 150

第 7 章 Logistic 回归 151
7.1 基本概念 151
7.2 Logistic 回归及 ROC 曲线 154
7.3 本章的 Python 代码 157

第 8 章 决策树及其组合方法 159
8.1 决策树 159
8.1.1 决策树分类 159
8.1.2 决策树回归 168
8.2 有放回再抽样简介 170
8.2.1 有放回再抽样 171
8.2.2 经验分布 171
8.2.3 OOB 数据 172
8.2.4 非等权放回再抽样 173
8.3 Bagging 173
8.3.1 Bagging 的原理 173
8.3.2 Bagging 分类 174
8.3.3 Bagging 回归 175
8.4 随机森林 176
8.4.1 随机森林分类 177
8.4.2 随机森林回归 181
8.4.3 一个解释随机森林的程序包 * 184
8.5 AdaBoost 分类 189
8.6 本章的 Python 代码 191
8.6.1 决策树分类 191
8.6.2 决策树回归 194
8.6.3 Bagging 分类 195
8.6.4 Bagging 回归 196
8.6.5 随机森林分类 196
8.6.6 随机森林回归 197
8.6.7 AdaBoost 分类 197

第 9 章　支持向量机 199
9.1　支持向量机分类 199
9.2　支持向量机回归 201
9.3　本章的 Python 代码 202
9.3.1　支持向量机分类 202
9.3.2　支持向量机回归 202

第 10 章　人工神经网络 204
10.1　基本概念 204
10.2　神经网络分类 207
10.3　神经网络回归 209
10.4　本章的 Python 代码 212
10.4.1　人工神经网络分类 212
10.4.2　人工神经网络回归 212

第 11 章　朴素贝叶斯 214
11.1　朴素贝叶斯原理 214
11.2　朴素贝叶斯方法分类 215
11.3　本章的 Python 代码 216

第 12 章　K 最近邻方法 217
12.1　K 最近邻方法概要 217
12.1.1　K 最近邻分类 218
12.1.2　K 最近邻回归 220
12.2　本章的 Python 代码 221
12.2.1　K 最近邻方法分类 221
12.2.2　K 最近邻方法回归 221

第 13 章　有监督学习模型比较案例及第三部分习题 223
13.1　多分类问题例子 223
13.1.1　二分类问题 224
13.1.2　回归问题 226
13.2　本章的 Python 代码 228
13.3　习题 232

第四部分 无监督学习 ····· 235

第 14 章 降维: 主成分分析 ····· 237

14.1 变量之间的关系与降维的可能性 ····· 237
14.2 两维空间的降维 ····· 238
14.3 通过特征值问题求各个主成分 ····· 242
14.4 各个观测值在新坐标中的记分 ····· 245
14.5 主成分分析在图像处理中的应用 ····· 247
14.6 本章的 Python 代码 ····· 249
14.7 习题 ····· 255

第 15 章 聚类分析 ····· 256

15.1 聚类分析概论 ····· 256
15.1.1 点间距离 ····· 257
15.1.2 类间距离 ····· 258
15.1.3 分层聚类 ····· 259
15.1.4 K 均值聚类 ····· 261
15.1.5 聚类数目的选择 ····· 264
15.1.6 图像色彩的聚类 ····· 266
15.2 本章的 Python 代码 ····· 271
15.3 习题 ····· 274

第五部分 网络爬虫 ····· 277

第 16 章 网络抓取/网络爬虫入门 ····· 279

16.1 引言 ····· 279
16.1.1 目的 ····· 279
16.1.2 手段 ····· 280
16.2 HTML 网页结构 ····· 280
16.2.1 简单的网页 ····· 281
16.2.2 元素 (element) 和标签 (tag) 的语法 ····· 282
16.2.3 HTML 的属性 ····· 287
16.3 网络抓取的步骤 ····· 288
16.3.1 使用 Python 软件 ····· 288
16.3.2 使用 R 软件 ····· 292
16.4 网络爬虫的若干案例 ····· 294
16.4.1 抓取新京报动新闻标题 ····· 294
16.4.2 马克思恩格斯全集文章的抓取 ····· 296

16.4.3　抓取狄更斯小说《双城记》全部原文内容 …… 297
16.4.4　抓取豆瓣《月亮与六便士》书评 …… 299
16.4.5　抓取微博账号 “公安部儿童失踪信息紧急发布平台” 儿童失踪信息 …… 302
16.5　表格数据爬取 …… 303
16.6　习题 …… 305

参考文献 …… **306**

第一部分

本书的安排及软件