



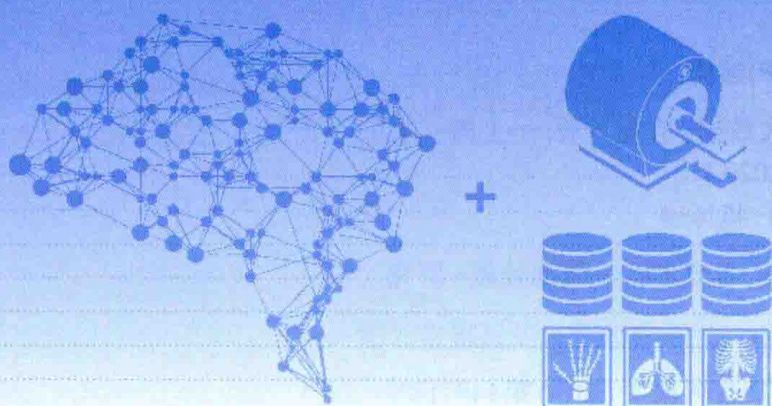
◎ 乔霓丹
—
著

深度学习 与 医学大数据

DEEP LEARNING AND MEDICAL BIG DATA



上海科学技术出版社



深度学习与医学大数据

Deep Learning and Medical Big Data

乔霓丹
著



上海科学技术出版社

图书在版编目(CIP)数据

深度学习与医学大数据 / 乔霓丹著. —上海: 上海科学技术出版社, 2020.1

ISBN 978-7-5478-4526-4

I. ①深… II. ①乔… III. ①医学—数据处理—研究
IV. ①R319

中国版本图书馆CIP数据核字 (2019) 第148717号

深度学习与医学大数据

乔霓丹 著

上海世纪出版(集团)有限公司 出版、发行
上海科学技术出版社
(上海钦州南路71号 邮政编码200235 www.sstp.cn)

上海盛通时代印刷有限公司印刷
开本 787×1092 1/16 印张 13.25 插页 4
字数 280千字
2020年1月第1版 2020年1月第1次印刷
ISBN 978-7-5478-4526-4 / R·1886
定价: 98.00元

本书如有缺页、错装或坏损等严重质量问题, 请向工厂联系调换

内容提要

随着数据量的几何级数增长，近 5 年，以数据为导向的预测和因果推断在学术界有巨大的进展。本书着重讨论这两方面的进展，以及如何应用这些成果进行医学科学研究。

本书首先以重症监护治疗病房患者的数据为例介绍如何进行医学数据的预处理、如何使用机器学习模型预测患者的死亡率。其重点在于深度学习，内容包括神经网络的基础知识、利用卷积神经网络分类诱发电位图像、利用递归神经网络预测疾病的复发，以及利用自编码算法去除图像噪声以生成新的模拟图像。本书还涉及基础医学中深度学习的应用及如何解释机器学习模型。

然而，医学实践中往往不仅需要预测某位患者的某项结局，而且需要知道哪些治疗能够改善结局。本书在神经网络模型之后以脓毒症患者的治疗为例，介绍强化学习的概念，从而为引入因果推断搭建桥梁。

本书最后两个专题着重讨论如何通过因果关系图直观地判断因果关系中的混杂因素、如何使用回归控制混杂因素、如何利用倾向得分控制混杂因素，以及如何利用逆概率加权控制混杂因素。结合强化学习的概念，最后讨论如何评估随时间变化的治疗，以及如何建立治疗策略。

本书可作为医学研究者、医学院学生和教师的参考用书。

作者简介

> **乔霓丹** 医学博士，复旦大学附属华山医院神经外科主治医师，上海市医师协会会员。2012年毕业于复旦大学上海医学院临床医学专业，获博士学位。2009—2010年在日本京都大学北野医院交流学习，方向为心血管与神经外科。2012年毕业后留院成为复旦大学附属华山医院神经外科医师，从事颅内鞍区肿瘤的微创手术工作。2017—2019年在职攻读美国哈佛大学医学院临床研究科学硕士，为美国麻省总医院神经内分泌科研究员。毕业后回国继续从事相关领域研究工作。

工作期间的主要研究项目是基于神经解剖路径评估垂体腺瘤患者的视功能，在多个国内、国际会议上做相关进展的报告。近年来，研究工作集中在创新性地将新型数据科学方法应用于神经外科和神经内分泌疾病中，如利用机器学习预测垂体瘤的预后、利用深度学习优化视觉电生理的分析过程等。

至今发表SCI论文30余篇，在国内外多个专业会议上做大会演讲。负责上海市科学技术委员会青年科技英才扬帆计划1项，上海市科学技术委员会基础研究项目1项。参译《尤曼斯神经外科学》，参编《垂体瘤疑难病例汇编》。

前言

> 笔者有幸在 2017 年 7 月受复旦大学附属华山医院神经外科委派，前往美国哈佛大学医学院攻读医学研究硕士学位，同时在美国麻省总医院进行临床研究。在哈佛大学医学院的课程主要包括机器学习、因果推断、临床试验等。在美期间同时参与与麻省理工学院的合作项目，主要内容是神经网络及医学电子病历大数据的应用。

笔者深感利用医学大数据进行预测和因果推断是今后临床医学研究的方向，也很遗憾国内的医学教育还比较缺乏这方面的意识。因此，笔者自觉有义务将在哈佛大学和麻省理工学院学习的成果和感悟分享给国内的医学研究者和学生。

随着计算机计算能力的提升，数据的产生呈爆炸式几何级数增长，数据存储的成本也越来越低。在医学领域也是如此：从基础医学到临床医学，从基因数据到电子病历数据，从文本数据到图像数据，医学的各个方面都出现了“大数据”。利用医学大数据进行医学研究是近年来的热门话题，频频出现在各大主流媒体上。如何存储数据、利用数据进行科学研究的学科称为“数据科学”（data science），主要包括三大方面：数据的描述（description）、数据的预测（prediction）和因果推断（casual inference）。传统的统计学着重于数据的描述，如发病率、均值、误差等，其统计推断也基于简单的卡方检验、 t 检验、Logistic 回归分析等。随着数据量的几何级数增长，近年来以数据为导向的预测和因果推断在学术界发展迅捷。

利用数据进行预测需要解决的问题是，在某一人群中如何通过一系列观测值预判另一未观测值。传统的预测模型主要是线性回归模型，这也是最为广泛应用的模型。近年来，随着计算机计算速度的提升，以往难以完成的需要耗费大量计算能力的模型得



到了广泛应用。这些模型统称为机器学习 (machine learning)，它通过算法和统计模型的科学研究，利用计算机系统来有效地执行特定任务，不使用明确的指令而是依赖于模式和推理。这其中也包括近两年来最为热门的神经网络，也称为“深度学习” (deep learning)。其实，这些模型的诞生都要追溯至 50 年前，但是因为那时的计算能力无法与模型的复杂程度匹配，因而这些模型未被广泛应用。

本书前半部分以重症监护治疗病房患者的数据为例介绍如何进行医学数据的预处理，在一般机器学习中重点介绍利用“梯度提升决策树” (gradient boosting decision tree) 预测患者的死亡率，然后介绍聚类算法在基因数据中的应用。当然，本书前半部分的重点在于深度学习，包括神经网络的基础知识、利用卷积神经网络分类诱发电位图像、利用递归神经网络预测疾病的复发，以及利用自编码算法去除图像噪声并生成新的模拟图像。本书还涉及基础医学中深度学习的应用及如何解释机器学习模型。

然而，医学实践中往往不仅需要预测某位患者的某项结局，而且需要知道哪些治疗能够改善结局。本书在神经网络模型之后将以脓毒症患者的治疗为例，介绍强化学习的概念，从而为引入因果推断搭建桥梁。

由于仅仅利用模型进行数据预测终究无法指导医学实践，因此需要数据科学的第三大方面：因果推断。因果推断需要解决的问题是，评价一种观测值的变化是否带来另一种观测值的变化；如果有变化，变化的幅度又是多大。医学科学中进行因果推断最好的方法当属随机对照临床研究。如果需要研究某种药物对某种疾病是否有效，通常的方法是选取一个该疾病的患者队列，然后进行随机分组：一组使用该药物，另外一组不使用或使用安慰剂，比较两组患者的结局（缓解率、治愈率或死亡率等）是否不同，从而得出该药物是否有效的结论。然而，并不是所有的因果问题都可以用随机对照研究来解决，比如想要研究吸烟是否引起肺癌，就不能使用随机分组。不仅是干预措施有害时不能使用随机对照研究，结局出现时间太长的研究（比如喝咖啡是否引起死亡率上升）也不能使用随机对照研究。而且随机对照研究得出的治疗效果往往只是倾向性治疗的效果 (intention-to-treat)，即研究组与对照组对比的效果。实际研究过程中，部分患者可能存在不依从甚至组间交叉。医学实践现实中需要的往往是实际治疗 (per-protocol) 的效果，因此很多情况下需要利用观察性数据进行正确的因果推断。



观察性数据与随机对照研究的最大差别在于观察性数据中有很多混杂因素，因此控制混杂因素是利用观察性数据进行因果推断的关键。本书最后两个专题着重讨论如何通过因果关系图直观地判断因果关系中的混杂因素，如何使用回归控制混杂因素，如何利用“倾向得分”（propensity score）控制混杂因素，以及如何利用逆概率加权控制混杂因素。结合强化学习的概念，最后将讨论如何评估随时间变化的治疗，以及如何建立治疗策略。

本书所用软件为 R 和 Python，因为传统的 SPSS 以及 Stata 对机器学习及进阶的因果推断支持性不佳。希望读者在阅读本书之后，能对数据科学有较全面的认识，也能运用本书中的模型与方法进行实际的医学研究；希望读者在今后的研究工作中区分所涉及的问题是预测问题还是因果推断问题（但同时这两者又通过相同的模型紧密地联系在一起）。

本书文字描述部分皆为笔者在美期间的学习所获及研究感悟。数据案例部分大部分为临床中收集的去除隐私信息的案例，部分来源于网络公开数据，小部分为模拟数据。代码部分大部分为作者自行编写，部分代码改编自一些引用文献，笔者也列出了这些引用文献。

本书文字中概念性的名词以中文表达并在括号中加注英文。如无特殊说明，直接用英文的表示数据表，用“[列变量名]”表示列变量名，用引号表示具体变量的值。

本书的写作得到了哈佛大学公共卫生学院 Miguel Hernan 教授、哈佛大学医学院 David Sontag 教授、麻省理工学院 Leo Anthony Celi 教授及谷歌研究员 Ian Goodfellow 等资深学者的指导，在此深表感谢。

本书受上海市科学技术委员会青年科技英才扬帆计划（17YF426700）、上海市科学技术委员会基础研究项目（17JC1402100），以及 2018 年度米尔斯坦美亚医学基金会转化医学研究基金（2018 Milstein Medical Asian American Partnership Foundation Fellowship Award in Translational Medicine）支持。

由于笔者经验有限，书中缺点和不足之处还望读者不吝指正。

乔覓丹

2019 年 7 月

致 谢

> 出版这本书要感谢的单位、领导、专家和朋友实在太多了，无论用多么华丽的词语都难以表述到位。在这里，我对在本书写作过程中提供材料的单位、朋友，给予支持的单位及领导、专家和朋友致以真挚的感谢！

感谢复旦大学附属华山医院神经外科主任周良辅教授、毛颖教授，以及复旦大学附属华山医院林建华主任，是他们的支持，让我有了去哈佛医学院学习的机会。感谢他们让我在华山医院有一个良好的环境实践和推进数据科学的建设工作。

感谢复旦大学附属华山医院神经外科赵曜教授，他是我的导师，他积极支持本书的编写工作。感谢复旦大学附属华山医院神经外科王镛斐教授，他为本书的编写提供了宝贵的思路。

感谢哈佛大学医学院院长 Ajay Singh 教授、哈佛大学医学院继续教育学院 Finnian McCausland 教授，他们为我在哈佛大学的学习提供了支持，并协调哈佛医学院院内资源。感谢美国麻省总医院 Tritos Nicholas 教授、Brooke Swearingen 教授支持本书的编写工作。

还要感谢麻省理工学院 Leo Anthony Celi 教授，他积极关注和支持本书的编写工作，为本书提供了大量最新的相关资料。感谢哈佛公共卫生学院 Miguel Hernan 教授，他为本书提供因果推断的最新资料，他的观点、思想在本书相关专题中得到了体现。

感谢山东大学经济学院陈强教授为本书做的审校工作。感谢上海市科学技术委员会及米尔斯坦美亚医学基金会对我的支持，以及对本书出版的支持。

感谢上海科学技术出版社的各级领导和本书的责任编辑等相关人员，是他们对笔者的充分信任和大力支持，是他们富有效率



的协调、严谨的作风和辛勤工作，使得本书能很快与读者见面。

感谢复旦大学附属华山医院心内科罗心平教授、华山医院防保科周萍科长。感谢我的妻子和一对女儿对我的支持与理解。

最后，谨向帮助、支持和鼓励我完成本书工作的家人和所有朋友致以深深的敬意和诚挚的感谢！

乔竞丹

2019年5月于波士顿

目 录



1 机器学习基础 / 1

- 1.1 数据概况 / 2
- 1.2 数据的预处理和特征选取 / 3
- 1.3 缺失值的处理与插补 / 8
- 1.4 交叉验证 / 12
- 1.5 模型建立 / 13
- 1.6 模型比较 / 20

2 梯度提升决策树 / 23

- 2.1 超参数 / 26
- 2.2 特征重要性 / 32
- 2.3 模型的临床应用 / 33
- 2.4 模型集成 / 35
- 2.5 机器学习的报告要点 / 38

3 聚类算法 / 41

- 3.1 各种聚类算法 / 42
- 3.2 主成分分析 / 46
- 3.3 聚类算法的直观显示 / 48



4 神经网络 / 51

- 4.1 感知器 / 52
- 4.2 全连接神经网络的训练 / 53
- 4.3 控制过拟合 / 58
- 4.4 公开数据来源 / 61

5 卷积神经网络 / 67

- 5.1 卷积运算 / 68
- 5.2 池化运算 / 71
- 5.3 简单卷积神经网络的构建和训练 / 71
- 5.4 图像样本量扩大 / 78
- 5.5 迁移学习 / 81
- 5.6 可解释的卷积神经网络 / 86
- 5.7 开放图像数据库 / 88
- 5.8 卷积神经网络的意义与不足 / 89

6 自编码和对抗生成神经网络 / 91

- 6.1 自编码算法基础 / 92
- 6.2 自编码算法降噪 / 97
- 6.3 变分自编码算法 / 100
- 6.4 变分自编码算法生成虚拟图像 / 104

6.5 对抗生成神经网络生成虚拟图像 / 105

7 递归神经网络 / 107

7.1 递归神经网络原理 / 108

7.2 递归神经网络构建 / 109

7.3 长短期记忆网络 / 111

7.4 门控递归神经网络 / 113

7.5 LSTM和GRU的构建 / 113

7.6 卷积神经网络和递归神经网络的叠加 / 115

8 自然语言处理和电子病历 / 119

8.1 从单词到向量 / 120

8.2 利用传统自然语言处理寻找脑外伤患者 / 122

8.3 利用神经网络寻找脑外伤患者 / 125

8.4 电子病历系统中神经网络的应用 / 131

9 可解释的机器学习 / 133

9.1 预测蛋白-蛋白间结合 / 134

9.2 预测基因-蛋白间结合 / 139

9.3 机器学习的解释 / 145



10 深度强化学习 / 147

- 10.1 强化学习 / 148
- 10.2 利用Q学习预测脓毒症的治疗策略 / 149
- 10.3 利用深度强化学习预测治疗策略 / 154
- 10.4 强化学习的不足之处 / 161

11 因果推断简介 / 163

- 11.1 反事实模型 / 164
- 11.2 随机对照研究 / 164
- 11.3 非随机对照研究 / 165
- 11.4 因果推断还是预测 / 165
- 11.5 因果关系图 / 166
- 11.6 分层分析 / 167
- 11.7 回归 / 170
- 11.8 交互作用 / 171

12 控制混杂因素的新方法 / 173

- 12.1 匹配分析 / 174
- 12.2 倾向得分 / 175
- 12.3 逆概率和稳定的逆概率加权 / 179
- 12.4 失访偏差的校正 / 182

12.5 随机对照研究中的偏差校正 / 183

12.6 工具变量 / 184

12.7 断点回归 / 185

12.8 随时间变动的治疗和混杂 / 188

12.9 动态治疗方案 / 189

附录：软件安装 / 193



1

机器学习基础

本章要点

- 从原始数据中提取预测指标
- 预处理预测指标
- 处理缺失值
- 构建简单的机器学习模型
- 比较不同的模型

预测重症监护治疗病房（intensive care unit, ICU）患者死亡率对比较药物的疗效、比较护理的有效性、比较手术和其他干预措施的有效性有着重要的意义。以往的研究利用 APACHE、SOFA 和 SAPS-II 等评分标准来预测 ICU 患者的死亡，这些评分是利用患者的一般情况、生命体征以及实验室检查等计算出的。其预测的准确程度一般使用准确率或受试者工作特征（receiver operating characteristic, ROC）下面积（area under curve, AUC）来表示，但以往的研究中这两项指标都不高。因此，利用机器学习来构建预测模型、辅助临床预测有着一定的意义，能够帮助临床医生更好地判断病情。本专题选取 PhysioNet 网站中的公开数据作为案例教学，预测 ICU 患者的死亡率。读者可以至 <https://physionet.org/challenge/2012/> 中，点击 Training set A 后的 “[tarball]” 下载预测变量数据（下载完成后解压缩），点击 “[Outcome-related descriptors]” 下载结局变量数据（下载完成后更改后缀为 .csv，csv 文件可以用 Microsoft Excel 程序读取写入）。将上述文件存入此专题中设置好的工作文件夹中。

1.1 数据概况

该数据集是在 MIMIC III ICU 数据库中选择的，本书第 4 个专题会详细介绍该数据库。数据中包含 4 000 名患者，每名患者的数据存于一个 txt 文本中，文本中包含 3 列数据：时间（Time）、指标（Parameter）、值（Value）。指标中前 6 个是一般变量 [住院号（RecordID）、年龄（Age）、性别（Gender）、身高（Height）、ICU 类别（ICUType）和体重（Weight）]，其余指标（表 1.1）都是患者的实验室检查值，可能有多个观测值，也可能没有，这些指标都是在入住 ICU 后的最初 48 小时内记录的。每个指标都有一个相关的时间标记，指示自 ICU 入院以来的观察时间，以小时和分钟为单位。值为 -1 表示丢失或未知数据（如没有记录患者的身高）。

表 1.1 36 个指标含义及单位

Albumin (g/dL) 白蛋白	MAP (mmHg) 有创平均动脉压
ALP (U/L) 碱性磷酸酶	MechVent (0: 无, 1: 有) 机械通气
ALT (U/L) 丙氨酸转氨酶	Na (mmol/L) 钠
AST (U/L) 天冬氨酸转氨酶	NIDiasABP (mmHg) 无创舒张动脉压
BUN (mmol/L) 尿素氮	NIMAP (mmHg) 无创平均动脉压
Bilirubin (mmol/L) 胆红素	NISysABP (mmHg) 无创收缩动脉压
Cholesterol (mmol/L) 胆固醇	PaCO ₂ (mmHg) 动脉 CO ₂ 分压
Creatinine (mmol/L) 肌酐	PaO ₂ (mmHg) 动脉血氧分压
DiasABP (mmHg) 有创舒张动脉压	pH (0~14) 动脉 pH
FiO ₂ (0~1) 吸入氧分压	Platelets (10 ⁹ /L) 血小板