

李群机器学习

Lie Group Machine Learning



李凡长 张 莉
杨季文 钱旭培 著
王邦军 何书萍

中国科学技术大学出版社

013734112

0152.5
10

当代科学技术基础理论与前沿问题研究

中国科学技术大学
校友文库

李群机器学习
Lie Group Machine Learning

李凡长 张 莉
杨季文 钱旭培 著
王邦军 何书萍



中国科学技术大学出版社



北航

C1641384

内 容 简 介

从历史经验看,研究机器学习应该“以认知科学为基础、数学方法为手段、可计算理论为标准、分析数据规律为目标、计算机技术为实现途径,并沿着这样的路径来构建机器学习的理论、技术、方法和应用体系”。因此,本书根据“模型+分析”的认知互补机制和李群理论,提出了李群机器学习框架。本书共分 11 章:引论,李群覆盖学习,李群深层结构学习,李群半监督学习,李群核学习,张量学习,标架丛上的联络学习,谱估计学习,Finsler 几何学习,同调边缘学习,基于范畴理论的学习方法。

本书可供计算机、自动化、认知科学、数学、管理科学、哲学等领域的大学高年级学生、硕士生、博士生和教师使用,也可供相关科技工作者参考。

图书在版编目(CIP)数据

李群机器学习/李凡长等著. —合肥:中国科学技术大学出版社,2013.4
(当代科学技术基础理论与前沿问题研究丛书:中国科学技术大学校友文库)
“十二五”国家重点图书出版规划项目

ISBN 978-7-312-03160-1

I. 李… II. 李… III. 李群—机器学习 IV. ①O152.5 ②TP181

中国版本图书馆 CIP 数据核字(2013)第 054388 号

出版发行 中国科学技术大学出版社

安徽省合肥市金寨路 96 号,230026

<http://press.ustc.edu.cn>

印 刷 合肥晓星印刷有限责任公司

经 销 全国新华书店

开 本 710 mm×1000 mm 1/16

印 张 23.25

字 数 390 千

版 次 2013 年 4 月第 1 版

印 次 2013 年 4 月第 1 次印刷

定 价 69.00 元

编 委 会

顾 问 吴文俊 王志珍 谷超豪 朱清时

主 编 侯建国

编 委 (以姓氏笔画为序)

王 水 史济怀 叶向东 朱长飞

伍小平 刘 兢 刘有成 何多慧

吴 奇 张家铝 张裕恒 李曙光

杜善义 杨培东 辛厚文 陈 颢

陈 霖 陈初升 陈国良 陈晓剑

郑永飞 周又元 林 间 范维澄

侯建国 俞书勤 俞昌旋 姚 新

施蕴渝 胡友秋 骆利群 徐克尊

徐冠水 徐善驾 翁征宇 郭光灿

钱逸泰 龚惠兴 童秉纲 舒其望

韩肇元 窦贤康 潘建伟

总 序

大学最重要的功能是向社会输送人才,培养高质量人才是高等教育发展的核心任务.大学对于一个国家、民族乃至世界的重要性和贡献度,很大程度上是通过毕业生在社会各领域所取得的成就来体现的.

中国科学技术大学建校只有短短的五十余年,之所以迅速成为享有较高国际声誉的著名大学,主要就是因为她培养出了一大批德才兼备的优秀毕业生.他们志向高远、基础扎实、综合素质高、创新能力强,在国内外科技、经济、教育等领域做出了杰出的贡献,为中国科大赢得了“科技英才的摇篮”的美誉.

2008年9月,胡锦涛总书记为中国科大建校五十周年发来贺信,对我校办学成绩赞誉有加,明确指出:半个世纪以来,中国科学技术大学依托中国科学院,按照全院办校、所系结合的方针,弘扬红专并进、理实交融的校风,努力推进教学和科研工作的改革创新,为党和国家培养了一大批科技人才,取得了一系列具有世界先进水平的原创性科技成果,为推动我国科教事业发展和社会主义现代化建设做出了重要贡献.

为反映中国科大五十年来的人才培养成果,展示我校毕业生在科技前沿的研究中所取得的最新进展,学校在建校五十周年之际,决定编辑出版《中国科学技术大学校友文库》50种.选题及书稿经过多轮严格的评审和论证,入选书稿学术水平高,被列入“十一五”国家重点图书出版规划.

入选作者中,有北京初创时期的第一代学生,也有意气风发的少年班毕业生;有“两院”院士,也有中组部“千人计划”引进人才;有海内外科研院所、大专院校的教授,也有金融、IT行业的英才;有默默奉献、矢志报国的科技将军,也有在国际前沿奋力拼搏的科研将才;有“文革”后留美学者中第一位

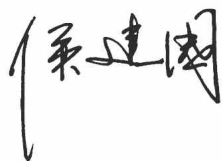
担任美国大学系主任的青年教授,也有首批获得新中国博士学位的中年学者……在母校五十周年华诞之际,他们通过著书立说的独特方式,向母校献礼,其深情厚谊,令人感佩!

《文库》于2008年9月纪念建校五十周年之际陆续出版,现已出书53部,在学术界产生了很好的反响.其中,《北京谱仪II:正负电子物理》获得中国出版政府奖;中国物理学会每年面向海内外遴选10部“值得推荐的物理学新书”,2009年和2010年,《文库》先后有3部专著入选;新闻出版总署总结“‘十一五’国家重点图书出版规划”科技类出版成果时,重点表彰了《文库》的2部著作;新华书店总店《新华书目报》也以一本书一个整版的篇幅,多期访谈《文库》作者.此外,尚有十数种图书分别获得中国大学出版社协会、安徽省人民政府、华东地区大学出版社研究会等政府和行业协会的奖励.

这套发端于五十周年校庆之际的文库,能在两年的时间内形成现在的规模,并取得这样的成绩,凝聚了广大校友的智慧和母校的感情.学校决定,将《中国科学技术大学校友文库》作为广大校友集中发表创新成果的平台,长期出版.此外,国家新闻出版总署已将该选题继续列为“十二五”国家重点图书出版规划,希望出版社认真做好编辑出版工作,打造我国高水平科技著作的品牌.

成绩属于过去,辉煌仍待新创.中国科大的创办与发展,首要目标就是围绕国家战略需求,培养造就世界一流科学家和科技领军人才.五十年来,我们一直遵循这一目标定位,积极探索科教紧密结合、培养创新拔尖人才的成功之路,取得了令人瞩目的成就,也受到社会各界的肯定.在未来的发展中,我们依然要牢牢把握“育人是大学第一要务”的宗旨,在坚守优良传统的基础上,不断改革创新,进一步提高教育教学质量,努力践行严济慈老校长提出的“创寰宇学府,育天下英才”的使命.

是为序.



中国科学技术大学校长
中国科学院院士
第三世界科学院院士
2010年12月

序

机器学习作为人工智能的一个分支,近年来在科学研究中越来越扮演着重要角色.产生这种影响的主要原因,是众多学科的发展需要智能化方法来丰富相关学科的科学内涵.而在智能化方法中,机器学习是最为基础和关键的核心技术,如机器学习算法在生物信息学、物理学、化学、材料分析学等学科中的应用.但是,我们应该清醒地认识到,尽管机器学习在现代科学研究中有很重要的作用,但关于机器学习问题的实质究竟是什么却尚无定论.从1943年的神经网络模型及函数逼近论方法的研究、以符号演算为基础的规则学习和决策树学习的产生,以及之后的认知心理学中归纳、解释、类比等概念的引入,至2010年图灵奖获得者哈佛大学教授 Leslie Valiant 提出的计算学习理论和2011年图灵奖获得者 Judea Pearl 提出的概率和因果性推理演算法,不同的学习方法各有优缺点.

总之,从历史经验看,研究机器学习应该“以认知科学为基础、数学方法为手段、可计算理论为标准、分析数据规律为目标、计算机技术为实现途径,并沿着这样的路径来构建机器学习的理论、技术、方法和应用体系”.这样对已知的机器学习方法就有了检验依据,从而有可能形成以机器学习技术为主体的平台和产业,真正实现机器学习技术引领未来科技的目的.因此,以“模型+分析”的认知互补机制研究机器学习的理论和算法可能是最合适的途径之一.其根本原因如下:

第一,机器学习是机器通过算法对输入数据的加工过程.

这种解释可以把机器和人进行对应.正如 Neisser 在《认知心理学》一书中指出的:“认知是指信息经过感知输入的变化、简化、加工、存储、恢复、使

用的全过程。”这种观点强调信息在人脑中的流动过程,从感知信息开始直到最终作出对策结束。

第二,机器学习是机器对符号的处理过程。

这种解释可以把机器和人脑对符号的操作与处理过程进行对应,符号代表着不同于自身的一种表示,如语言、标记、数学符号等。

第三,机器学习是机器进行问题求解的过程。

这种解释可以理解成类似于人脑利用已有的知识和经验对当前所面临问题进行求解的过程。

第四,机器学习是机器对数据进行优化的过程。

人类学习的目的是认知自然规律、社会规律等,在其学习过程中优化是非常重要的任务;机器学习的目标是分析数据规律,因此,在机器学习过程中优化也是非常重要的任务。

上述四种对机器学习不同解释的共同特征是:

第一,机器学习系统是一个符号系统;

第二,机器学习系统是一个多层次、多方法,可以融合处理不同来源数据的系统;

第三,机器学习系统处理数据容量的能力是有限的;

第四,机器学习系统具有处理“定性和定量”数据的能力;

第五,机器学习系统具有“学习、推理和分析”相融合的能力。

本书正是以这种途径来介绍李群机器学习(Lie Group Machine Learning, LML)的,并致力于回答如何用模型的连续理论解决现实的离散数据、如何用最少数据构造的模型处理更多的数据、如何用非结构化方法解决结构化数据、如何用线性方法解决非线性数据等问题。

本书的姊妹篇《机器学习理论及应用》自2009年出版后深受读者欢迎,尤其是“李群机器学习”篇在学术界产生了重要影响。为了进一步满足广大读者的要求,我们将近几年的最新成果进行整理,形成了本书。本书的主要目标是探索“李群学习模型+分析”的机器学习新的理论体系,这不仅体现了数学美的观点,更重要的是为实现“用简单模型去解决复杂问题”的认知观提供了范例;既有别于逻辑方法,也不同于概论方法;尤其是面对大数据、物联网、云计算、认知科学、空间科学、物质科学等快速发展的时代,打破了过去人们认为群理论只能存在脑海里而不能很好地和现实问题关联的模

式,真正体现了“尽管数学不是万能的,但没有数学万万不能”的科学理念.例如晶体数据分析、天文数据分析、蜜蜂家族成员数据分析、人类社会中家庭成员数据分析等,为了快速揭示这些数据规律,李群学习算法毫无疑问是比较好的方法之一.

本书共有 11 章内容.除第 1 章外,其余各章具有相对的独立性,读者可以任意选择一章进行学习.

本书的总体设计及主要工作由李凡长完成,其余参与本书编写的人员包括张莉、杨季文、钱旭培、王邦军、何书萍、管文文、李军、高聪、陈明、李祥亮、周丽丽、董梦璇、徐寒香、鲜敏、何文慧.

本书所有研究内容得到了国家自然科学基金项目(61033013;60775045;60970067)、江苏省自然科学基金项目(SBK201222725;BK2011284)、江苏省青蓝工程、苏州大学国家自然科学基金预研项目(SDY2011B09)、苏州大学东吴学者计划(14317360)、苏州大学科技创新团队(SDT2012B02)、江苏省“计算机科学与技术”和“软件工程”重点学科资助,在此一并表示衷心的感谢!

愿本书能够给读者带来收获和快乐,同时由于作者水平有限,书中难免存在不足和错误,如读者发现问题或持不同观点,请随时与李凡长联系 (Email:lfzh@suda.edu.cn).

李凡长

2012 年 8 月

目 次

总序	i
序	iii
第 1 章 引论	1
1.1 引言	1
1.2 李群机器学习的基本概念	3
1.3 李群机器学习的公理假设	4
1.4 李群机器学习的学习模型	6
1.5 李群机器学习中的 Dynkin 图几何学习算法	8
1.6 李群机器学习的分类器设计	12
1.7 本书内容安排	16
参考文献	18
第 2 章 李群覆盖学习	22
2.1 李群机器学习覆盖算法的相关理论	22
2.2 李群机器学习的单连通覆盖算法	29
2.3 李群机器学习的多连通覆盖算法	38
2.4 覆盖算法在分子对接中的应用	45
2.5 本章小结	61

参考文献	61
第 3 章 李群深层结构学习	65
3.1 引言	65
3.2 李群深层结构学习	67
3.3 李群逐层学习算法	78
3.4 李群深层结构启发式学习算法	84
3.5 本章小结	88
参考文献	89
第 4 章 李群半监督学习	93
4.1 引言	93
4.2 基于李群的半监督学习模型	102
4.3 基于线性李群的半监督学习算法	107
4.4 基于参数李群的半监督学习算法	117
4.5 本章小结	126
参考文献	127
第 5 章 李群核学习	132
5.1 矩阵群学习算法	132
5.2 李群上的 Gauss 分布	141
5.3 李群内均值的计算	145
5.4 李群均值学习算法	149
5.5 李群核学习算法	156
5.6 实例应用	164
5.7 本章小结	172
参考文献	173
第 6 章 张量学习	179
6.1 基于张量方法的数据约简	179

6.2 基于张量场的数据约简模型	184
6.3 基于张量场的学习模型及算法设计	194
6.4 本章小结	209
参考文献	210
第 7 章 标架丛上的联络学习	214
7.1 标架丛上的纵空间学习模型	214
7.2 标架丛上的纵空间联络学习算法	219
7.3 标架丛上的横空间学习模型及相关算法过程	221
7.4 标架丛上的横空间与纵空间学习算法的应用	223
7.5 本章小结	225
参考文献	225
第 8 章 谱估计学习	230
8.1 谱估计概念及定义	230
8.2 相关理论基础	232
8.3 同步谱估计学习算法	238
8.4 图像特征流形的比较原则	245
8.5 图像特征流形拓扑不变性的谱估计学习算法	250
8.6 基于图像特征流形拓扑不变性谱估计的聚类算法	253
8.7 本章小结	254
参考文献	255
第 9 章 Finsler 几何学习	261
9.1 基本概念	261
9.2 基于 Finsler 度量的 KNN 算法	263
9.3 基于 Finsler 度量的几何学习算法	269
9.4 本章小结	277
参考文献	277

第 10 章	同调边缘学习	281
10.1	边缘学习算法	281
10.2	基于同调代数的边缘划分方法	284
10.3	同调边缘学习算法设计与分析	294
10.4	本章小结	296
	参考文献	296
第 11 章	基于范畴理论的学习方法	300
11.1	引言	300
11.2	学习表达式的范畴表示	310
11.3	学习表达式的映射机制	318
11.4	学习表达式映射机制的分类器设计	327
11.5	实例分析	336
11.6	本章小结	347
	参考文献	347
汉英名词对照		351

第 1 章 引 论

1.1 引 言

李群机器学习作为机器学习领域中的一种新的学习方法,一方面继承了流形学习的优点,另一方面则借用了李群的思想,形成了具有创新特色的学习范式;自 2004 年提出至今^[1-47],已引起了如加拿大、爱尔兰、芬兰、意大利、美国等国内外同行的广泛关注.该方法从李群的概念出发,充分利用李群的微分流形和群的优势,既提供了描述数据的几何表示方法,又给出了具体的代数求解方案.例如,群保持了系统的完备性,微分提供了具体的代数计算方法,流形给出了几何表示方法等,这正符合认知理论中的定性和定量表示相结合的认知模式.另外,从认知过程来看,人脑在认识客观世界中的对象时,首先关注的是表征问题的一个稳定点,然后依次分析图像的结构.对这种认知模式,在李群结构中,最小生成元就是这个稳定点,只要找到这个最小生成元,就可以利用李群方法对认知图像进行了分析.因此,李群机器学习既符合学习认知规律,又满足计算机解决现实问题的条件^[53].

文献[31]提出了用基于李群理论的一个 Bayes 方法来学习视觉感知,运用基于指数矩阵的图形生成模型,从包含极小变换的输入数据来得到一个学习李群算子的非监督学习算法.文献[32]使用运动表示的内在李群结构来求平均,用特殊正交群 $SO(3)$ 和特殊 Euclid 群 $SE(3)$ 的李代数来定义李群上的平均值,从而提出了全局相容运动估计的李代数求均值法.文献[33]提出了一种新的二维轮廓的不变特征、不变信号.它是对在从李群变换群中得到的许多变换下的

轮廓不变程度的度量,其中李群理论提供了在一个变换动作下,点位置的局部变换和此变换的全局描述的一种联系,并提供了自然的起始点,说明了不变信号在转移、旋转和轮廓缩放后本身保持不变.文献[34]用李群方法来寻找特征空间中的健壮和稳定特性,提出了基于李群理论的支配子空间不变量的概念和拟不变量的一种特殊类型,并给出了支配子空间不变量(DSI)算法.文献[35]给出了一种由有噪声的二维图像流对三维刚体运动进行神经计算的新方法;将三维图像流解释为一个线性信号变换,由神经网络来执行,其基本信号是三维 Euclid 群的六个极小生成子的二维向量场.神经网络模拟结果表明,在随机噪声导致传统的代数方法失败时,此方法仍能保证其可靠性.文献[36]给神经系统的权空间赋予了一个特殊李群结构,用李群的微分几何结构来学习这些属性并揭露不同学习方法间内在属性的联系,从而提出了基于李群的非监督神经学习.文献[37]提出了在 Stiefel 流形上基于测地线和接近于测地线的曲线簇(拟测地线)的新的神经网络学习算法,详细地解释了李群本身的梯度和测地线如何导出其在齐性空间上的对应,强调了李群而非正规化约束在这个流形上的作用.文献[38]提出了图像分析中的约束主轮廓进化的李群方法.在此方法中,使用平面作用的李群及其极小生成子的李代数的对应关系,适当地调整 Enler-Lagrange 下降方法,以此保证在保持原函数的下降方程不变时,曲线进化发生在被选变换群的轨道上.此方法的优点在于无需任何修改就能直接实现曲线进化,在许多跟踪和分割的应用中起到了重要作用.文献[39]引入了一个能够在分布式控制网络上执行的线性算子类,称为回归正交变换(ROTS).用回归正交变换可以表示特殊正交群 $SO(n)$ ($n \geq 4$),通过对其对应的内在的李群结构中提取梯度流来找到一个适合工作的回归正交变换.文献[40-54]提出了一种新的度量学习问题的框架结构,将候选度量集限制在一个参数族中,此度量是在一个李群变换下模糊信息量的拖回度量.文献[41]提出了机器学习中的群论和非交换调和分析相结合的一种学习问题研究框架,给出了群作用下核不变量的对称理论、在 Fourier 谱下群定向核的转换不变量的特性及置换学习的 Per-Mceptron 算法等相关内容.下面给出李群机器学习的基本概念.

1.2 李群机器学习的基本概念

定义 1.1(李群机器学习) 一般用 G 表示输入空间, M 表示输出空间. 令 $G \subseteq \mathbb{R}^D, M \subseteq \mathbb{R}^d (d < D)$. 借用李群的定义, G 对 M 的左作用可用如下映射 φ 表示:

$$\varphi: G \times M \rightarrow M$$

满足:

- (1) 对于所有的 $x \in M, \varphi(e, x) = x$;
- (2) 对于所有的 $g, h \in G, x \in M, \varphi(g, \varphi(h, x)) = \varphi(gh, x)$.

右作用是满足条件 $\psi(e, x) = x$ 和 $\psi(\psi(x, g), h) = \psi(x, gh)$ 的一个映射 $\psi: M \times G \rightarrow M$. 记 $\varphi(g, \cdot) = \varphi_g(\cdot)$.

定义 1.2(轨道) 若 φ 是 G 在 M 上的作用, 且 $x \in M$, 则 x 的轨道定义如下:

$$\text{Orb}(x) = \{\varphi_g(x) \mid g \in G\} \subset M$$

在有限维情形下, $\text{Orb}(x)$ 是 M 的浸入子流形. 对于 $x \in M$, φ 在 x 处的稳定(或对称)群由

$$G_x = \{g \in G \mid \varphi_g(x) = x\} \subset G$$

给出, 并且 G_x 是一个李子群. $\text{Orb}(x)$ 的流形结构由微分同胚 $[g] \in G/G_x \mapsto g \cdot x \in \text{Orb}(x)$ 定义, 由此可以判定 G/G_x 是一个光滑流形.

定义 1.3(作用的可迁性、有效性和自由性) (1) 作用的可迁性: 存在唯一的一个轨道, 或等价地, 对于每一个 $x, y \in M$, 有一个 $g \in G$, 使得 $g \cdot x = y$;

(2) 作用的有效性(或一一对应): $\varphi_g = \text{id}_M$ 蕴涵 $g = e$, 即 $g \mapsto \varphi_g$ 是一对的;

(3) 作用的自由性: 它没有不动点, 即 $\varphi_g(x) = x$ 蕴涵 $g = e$, 或者等价地, 对于每一个 $x \in M, \varphi_g = \{e\}$, 并且每一个自由的作用是一对一的.

定义 1.4(轨道空间) 等价类的集合 M/G 即称为轨道空间.

定义 1.5(学习表达式之间的等价性判定) 设 M 和 N 表示两类学习表达

式形成的流形, G 是李群, 并设 $\varphi_g: M \rightarrow M, \psi_g: N \rightarrow N$. 构造一个映射 $f: M \rightarrow N$, 使之关于这两个作用是等价的, 如果对于所有的 $g \in G$, 有 $f \circ \varphi_g = \psi_g \circ f$. 如果 M/G 和 N/G 都有正规投影光滑浸没的光滑流形, 则等变映射 $f: M \rightarrow N$ 可以诱导出一个光滑映射 $f_G: M/G \rightarrow N/G$.

1.3 李群机器学习的公理假设

文献[13-15]给出了李群机器学习的四个公理假设, 现归纳如下.

1.3.1 假设 1: 李群机器学习的泛化能力假设公理

该公理包括以下三条.

(1) 利用李群、李代数、李子群、李子代数之间的关系来描述泛化能力. 可如下描述: 设 G 是一个李群, 对于 G 的李代数 $\mathfrak{g}$ 的任何一个李子代数 $\mathfrak{h}$, 存在唯一的一个 G 的连通李子群 H , 使得图 1.1 是可交换的.



图 1.1 李群、李代数、李子群、李子代数之间的泛化关系图

(2) 利用学习空间的内积不变性来描述泛化能力. 可如下描述: 设 V 是 n 维线性空间, G 是 $GL(V)$ 的李子群, 如果有 $(g(x), g(y)) = (x, y)$ ($\forall x, y \in V, g \in G$), 则称 V 中的内积 (x, y) 在 G 下是不变的.

(3) 利用中心函数来描述泛化能力. 可如下描述: 设在紧致李群 G 中取定总体积为 1 的左右不变的黎曼结构, f 为其上任何一个中心函数, 则有

下面的 Weyl 积分公式:

$$\int_G f(g) dg = \frac{1}{|W|} \int f(t) |Q(t)|^2 dt$$

其中 $|W|$ 表示 W 中元素的个数. 当 $t = \exp(H)$ ($H \in \mathfrak{h}$) 时, $Q(t)$ 可以表示为

$$Q(\exp(H)) = \sum_{\sigma \in W} \text{sign}(\sigma) e^{2\text{xi}(\sigma(\delta), H)}$$