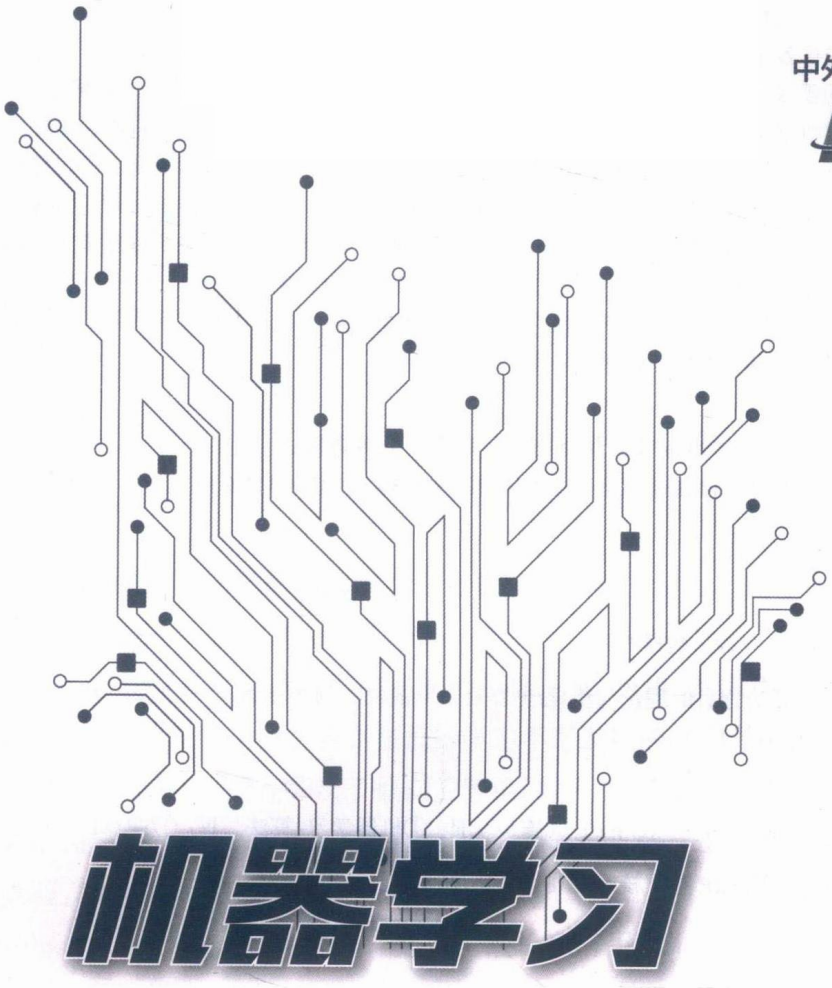

中外学者论
AI
机器学习
算法背后的理论与优化
史春奇 卜晶祎 施智平 著
清華大學出版社

中外学者论

AI



机器学习

算法背后的理论与优化

◎史春奇　卜晶祎　施智平　著

清華大學出版社
北京

内 容 简 介

以机器学习为核心的人工智能已经成为新一代生产力发展的主要驱动因素。新的技术正在向各行各业渗透，大有变革各个领域的趋势。传统产业向智慧产业的升级迫使原行业从业人员逐渐转型，市场上对相关学习材料的需求也日益高涨。帮助广大学习者更好地理解和掌握机器学习，是编写本书的目的。

本书针对机器学习领域中最常见的一类问题——有监督学习，从入门、进阶、深化三个层面由浅入深地进行了讲解。三个层面包括基础入门算法、核心理论及理论背后的数学优化。入门部分用以逻辑回归为代表的广义线性模型为出发点，引入书中所有涉及的知识点；进阶部分的核心理论涵盖了经验风险最小、结构风险最小、正则化及统一的分类边界理论；深化部分的数学优化则主要包括最大熵原理、拉格朗日对偶等理论在数学上的推导，以及对模型求解的主流最优化方法的探讨等。

本书由浅入深，从个别到普便，从自然算法到优化算法，从各个角度深入剖析了机器学习，力求帮助读者循序渐进地掌握机器学习的概念、算法和优化理论。

图书在版编目（CIP）数据

机器学习：算法背后的理论与优化/史春奇，卜晶祎，施智平著. —北京：清华大学出版社，2019
(中外学者论 AI)
ISBN 978-7-302-51718-4

Ⅰ. ①机… Ⅱ. ①史… ②卜… ③施… Ⅲ. ①机器学习-算法 Ⅳ. ①TP181

中国版本图书馆 CIP 数据核字(2018)第 267470 号

责任编辑：王　芳　王冰飞
封面设计：台禹微
责任校对：焦丽丽
责任印制：丛怀宇
出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课 件 下 载：http://www.tup.com.cn，010-62770175 转 4506
印 装 者：北京嘉实印刷有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　张：**12.75　　**字　　数：**277 千字
版　　次：2019 年 7 月第 1 版　　**印　　次：**2019 年 7 月第 1 次印刷
定　　价：69.00 元

产品编号：078279-01

史春奇博士 毕业于日本京都大学，美国 Brandeis University 博士后，现为港辉金融信息 Vice President。曾任通用电气（中国）有限公司资深数据科学家。

卜晶祎 毕业于上海交通大学，现为友邦保险集团人工智能主管。曾就职于通用电气（中国）研究开发中心有限公司，任资深数据科学家；曾任飞利浦亚洲研究院高级研究员。

施智平博士 首都师范大学信息工程学院教授，院长，毕业于中科院计算技术研究所。于 2012 年和 2017 年获北京市科学技术奖三等奖两次，中国计算机学会高级会员，形式化方法专委会委员，人工智能学会会员，智能服务专委会委员，IEEE/ACM 会员。

前言

Preface

在当今的人工智能领域中，最热门的技术毫无疑问当属深度学习。深度学习在 Geoffrey Hinton、Yoshua Bengio、Yann LeCun 和 Juergen Schmidhuber 等巨擘们持续不断的贡献下，在文本、图像、自然语言等领域均取得了革命性的进展。当然，深度学习只是机器学习的一个分支，能取得当前的成就也是建立在机器学习不断发展的基础之上。在机器学习领域，很多著名科学家 (如图 1 所示) 提出了他们的理论，做出了他们的贡献。Leslie Valiant 提出的概率近似正确学习 (Probably Approximately Correct Learning, PAC) 理论打下了计算学习理论的基石，并在此后提出了自举 (Bootstrapping) 思想。Vladimir Vapnik 提出的支持向量机 (Support Vector Machine, SVM) 是一个理论和应用都十分强大的算法。与此同时他所提出的经验风险最小与结构风险最小理论，以及背后更深层次的 VC 维 (Vapnik-Chervonenkis dimension) 理论，为部分统一分类问题提供了理论基础。Judea Pearl 提出

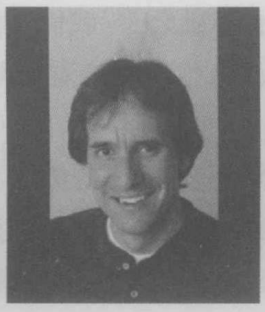
(a) Leslie Valiant

(b) Vladimir Vapnik

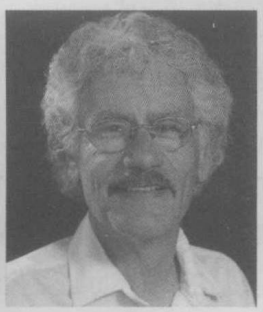
(c) Judea Pearl

(d) Michael I.Jordan

(e) Leo Breiman



(f) Robert Schapire



(g) Jerome H.Friedman

图 1　机器学习领域 (支持向量机、集成学习、概率图模型) 的著名科学家

了贝叶斯网络，而 Michael I. Jordan 则在此基础上发展了概率图模型。Leo Breiman 在集成 (Ensemble) 学习的思想下设计了随机森林 (Random Forest) 算法，Robert Schapire 和 Jerome H. Friedman 则基于 Boosting 分别发明了 AdaBoost 和 Gradient Boosting 算法。至此，机器学习中最耀眼的算法 —— 支持向量机、集成学习和概率图模型交相辉映，为整个机器学习理论的发展奠定了深厚的基础。

本书首先尝试把机器学习的经典算法，包括逻辑回归 (Logistic Regression)、支持向量机和 AdaBoost 等，在经验风险最小和结构风险最小的框架下进行统一，并且借助 Softmax 模型和概率图模型中的 Log-Linear 模型阐述它们的内在联系；其次从熵的角度解读概率分布、最大似然估计、指数分布族、广义线性模型等概念；最后深入剖析用于求解的最优化算法及其背后的数学理论。

本书的主要内容

全书分为 9 个章节，从单一算法到统一框架，再到一致最优化求解，各章节的设置如下。

第 1 章，首先提出并探讨几个基本问题，包括回归思想、最优模型评价标准、数理统计与机器学习的关系等。然后介绍两个最简单、最常见的有监督学习算法 —— 线性回归和逻辑回归，并从计算的角度分析两种模型内在的关联，从而为学习“广义线性模型”打下基础。在本章的最后部分初步讲解两个模型的求解方法 —— 最小二乘法和最大似然估计。

第 2 章，主要内容是线性回归的泛化形式 —— 广义线性模型。本章详细介绍广义线性模型，并在第 1 章的基础上从 Fisher 信息、KL 散度、Bregman 距离的角度深入讲解最大似然估计。本章可以看作是第 3 章的基础引入。

第 3 章，在前两章的基础上提出泛化误差和经验风险最小等概念，并且将最小二乘和最大似然并入损失函数的范畴。在此基础之上，我们便将逻辑回归、支持向量机和 Ada Boost 算法统一到分类界面的框架下。至此，我们会看到不同的算法只是分别对应了不同的损失函数。

第 4 章，介绍经验风险最小的不足与过拟合的概念，之后引出正则化。紧接着介绍有监督学习算法中的常见正则化方法，包括 L_1 和 L_2 正则化 XG Boost 和树。本章从两个角度对 L_1 和 L_2 正则化进行深入讲解 —— 贝叶斯和距离空间。这两个观点分别对应本书后续的两大部分 —— 熵和最优化。

第 5 章，介绍贝叶斯统计和熵之间的关系，并且基于熵重新解读了最大似然估计、指

数分布族等概念。本章可以看作是前四章中出现的内容在熵概念下的再定义。同时也为下一章的 Log-Linear 模型作出铺垫。

第 6 章，介绍 Softmax 和 Log-Linear 的变化，并且将第 3 章的二分类界面泛化到多分类界面，把分类问题的思路扩展到了多分类和结构分类。在本章中通过 Log-Linear 关联了概率图模型，通过 Softmax 关联了深度学习。

第 7 章，承接第 4 章中 L_1 和 L_2 正则化在最优化角度的解释，从凸共轭开始递进地推导出拉格朗日对偶、Fenchel 对偶、增广拉格朗日乘子法、交替方向乘子法。

第 8 章，介绍有监督学习模型在机器学习场景下的统一求解方法 —— 随机梯度下降法及其改进算法。本章对随机梯度下降法进行了收敛性分析，并根据分析结果针对其缺点着重介绍了两类改进策略 —— 方差缩减和加速与适应。

第 9 章，主要对数学意义上的最优化方法进行探讨，可以看作是连接第 7 章和第 8 章的桥梁。第 7 章的内容是本章的理论部分，而第 8 章的内容则是本章介绍的算法应用在机器学习场景中的特例，主要内容包括一阶、二阶最优化算法及其收敛性分析。

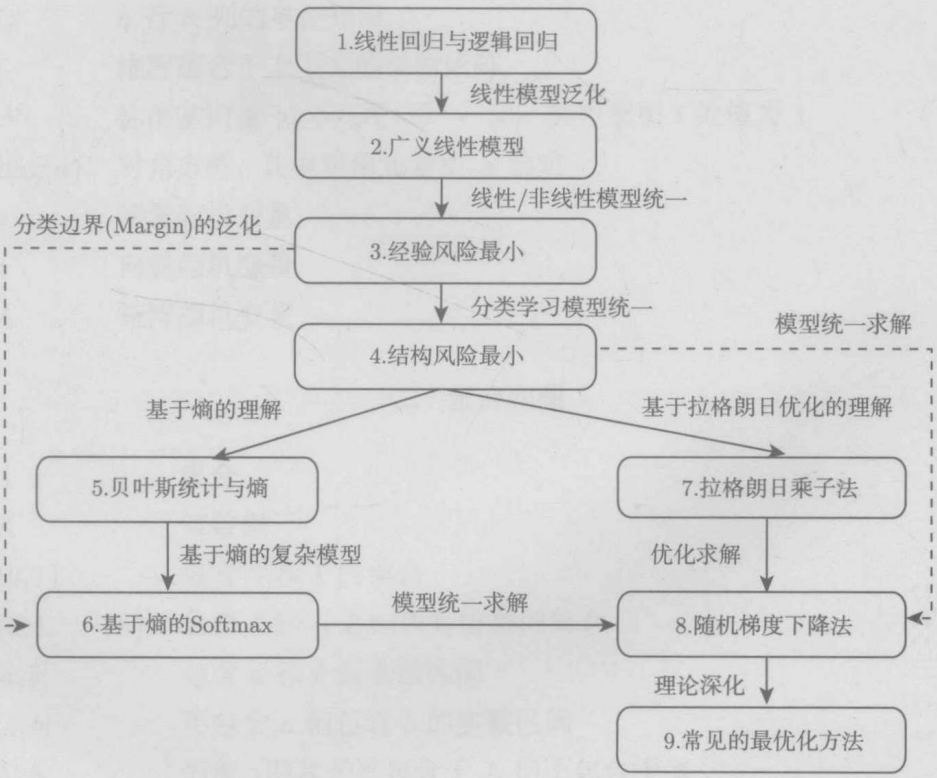


图 2　章节结构关系图

史春奇与卜晶祎共同为本书的第一作者。书中第 3~6 章主要由史春奇博士撰写，第 1、2、7~9 章主要由卜晶祎撰写，施智平教授参与了本书的组织结构设计并提出了很多宝贵意见。由于作者的能力与水平有限，本书对机器学习的探讨难免会有不全面、不深刻等不足之处，敬请各位读者批评指正，如蒙赐教将不胜感激。

各个章节结构之间的关系如图 2 所示。对于基础稍浅的读者，可以按照图示循序渐进地阅读；对于有一定基础的读者，可以跳过部分章节直接阅读感兴趣的章节。

作　者

2019 年 1 月

M 数学符号

Mathematical symbol

本部分介绍本书所使用的数学符号。

一、数和数组

a	标量 (整数或实数)
$\boldsymbol{a}$	向量
$\boldsymbol{A}$	矩阵
$\boldsymbol{A}$	张量
$\boldsymbol{I}_n$	n 行 n 列的单位矩阵
$\boldsymbol{I}$	维度蕴含于上下文的单位矩阵
$\boldsymbol{e}^{(i)}$	标准基向量 $[0,\cdots,0,1,0,\cdots,0]$，其中索引 i 处值为 1
$\text{diag}(\boldsymbol{a})$	对角方阵，其中对角元素由 $\boldsymbol{a}$ 给定
$\boldsymbol{a}$	标量随机变量
$\boldsymbol{a}$	向量随机变量
$\boldsymbol{A}$	矩阵随机变量

二、集合和图

A	集合
$\mathbf{R}$	实数集
$\{0,1\}$	包含 0 和 1 的集合
$\{0,1,\cdots,n\}$	包含 0 和 n 之间所有整数的集合
$[a,b]$	包含 a 和 b 的实数区间
$(a,b]$	不包含 a 但包含 b 的实数区间
$\mathbb{A}\backslash\mathbb{B}$	差集，即其元素包含于 $\mathbb{A}$ 但不包含于 $\mathbb{B}$
$\mathcal{G}$	图
$\text{Pa}_{\mathcal{G}}(x_i)$	图 $\mathcal{G}$ 中 x_i 的父节点

三、索引

a_i	向量 $\boldsymbol{a}$ 的第 i 个元素，其中索引从 1 开始
a_{-i}	除了第 i 个元素，$\boldsymbol{a}$ 的所有元素
$\boldsymbol{A}_{i,j}$	矩阵 $\boldsymbol{A}$ 的 i,j 元素
$\boldsymbol{A}_{i,:}$	矩阵 $\boldsymbol{A}$ 的第 i 行
$\boldsymbol{A}_{:,i}$	矩阵 $\boldsymbol{A}$ 的第 i 列
$\boldsymbol{A}_{i,j,k}$	三维张量 $\boldsymbol{A}$ 的 (i,j,k) 元素
$\boldsymbol{A}_{:,:,i}$	三维张量的二维切片
a_i	随机向量 $\boldsymbol{a}$ 的第 i 个元素

四、线性代数中的操作

$\boldsymbol{A}^\top$	矩阵 $\boldsymbol{A}$ 的转置
$\boldsymbol{A}^+$	$\boldsymbol{A}$ 的 Moore-Penrose 伪逆
$\boldsymbol{A}\odot\boldsymbol{B}$	$\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的逐元素乘积 (Hadamard 乘积)
$\det(\boldsymbol{A})$	$\boldsymbol{A}$ 的行列式

五、微积分

$\dfrac{\mathrm{d}y}{\mathrm{d}x}$	y 关于 x 的导数
$\dfrac{\partial y}{\partial x}$	y 关于 x 的偏导
$\nabla_{\boldsymbol{x}} y$	y 关于 $\boldsymbol{x}$ 的梯度
$\nabla_{\boldsymbol{X}} y$	y 关于 $\boldsymbol{X}$ 的矩阵导数
$\nabla_{\boldsymbol{X}} y$	y 关于 $\boldsymbol{X}$ 求导后的张量
$\dfrac{\partial f}{\partial \boldsymbol{x}}$	$f:\mathbf{R}^n\to\mathbf{R}^m$ 的 Jacobian 矩阵 $\boldsymbol{J}\in\mathbf{R}^{m\times n}$
$\nabla_{\boldsymbol{x}}^2 f(\boldsymbol{x})$ or $\boldsymbol{H}(f)(\boldsymbol{x})$	f 在点 $\boldsymbol{x}$ 处的 Hessian 矩阵
$\int f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$	$\boldsymbol{x}$ 整个域上的定积分
$\int_S f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$	集合 S 上关于 $\boldsymbol{x}$ 的定积分

六、概率和信息论

$a \perp b$	a 和 b 相互独立的随机变量
$a \perp b \mid c$	给定 c 后条件独立
$P(a)$	离散变量上的概率分布
$p(a)$	连续变量 (或变量类型未指定时) 上的概率分布
$a \sim P$	具有分布 P 的随机变量 a
$\mathbb{E}_{x\sim P}[f(x)]$ 或 $\mathbb{E}f(x)$	$f(x)$ 关于 $P(\mathrm{x})$ 的期望
$\mathrm{Var}(f(x))$	$f(x)$ 在分布 $P(x)$ 下的方差
$\mathrm{Cov}(f(x), g(x))$	$f(x)$ 和 $g(x)$ 在分布 $P(x)$ 下的协方差
$H(x)$	随机变量 x 的香农熵
$D_{\mathrm{KL}}(P\|Q)$	P 和 Q 的 KL 散度
$\mathcal{N}(x; \mu, \Sigma)$	均值为 μ，协方差为 Σ，x 上的高斯分布

七、函数

$f : \mathbb{A} \to \mathbb{B}$	定义域为 $\mathbb{A}$、值域为 $\mathbb{B}$ 的函数 f
$f \circ g$	f 和 g 的组合
$f(\boldsymbol{x}; \boldsymbol{\theta})$	由 $\boldsymbol{\theta}$ 参数化，关于 $\boldsymbol{x}$ 的函数 (有时为简化表示，忽略 $\boldsymbol{\theta}$ 记为 $f(\boldsymbol{x})$)
$\ln x$	x 的自然对数
$\sigma(x)$	Logistic sigmoid, $\dfrac{1}{1+\exp(-x)}$
$\zeta(x)$	Softplus, $\ln(1+\exp(x))$
$\|\|\boldsymbol{x}\|\|_p$	$\boldsymbol{x}$ 的 L^p 范数
$\|\|\boldsymbol{x}\|\|$	$\boldsymbol{x}$ 的 L^2 范数
x^+	x 的正数部分, 即 $\max(0, x)$
$\mathbf{1}_{\mathrm{condition}}$	如果条件为真则为 1，否则为 0

有时候使用函数 f，它的参数是一个标量，但应用到一个向量、矩阵或张量：$f(\boldsymbol{x})$、$f(\boldsymbol{X})$、$f(X)$。这表示逐元素地将 f 应用于数组。例如，$C = \sigma(X)$，则对于所有合法的 i、j 和 k，$C_{i,j,k} = \sigma(X_{i,j,k})$。

八、数据集和分布

p_{data}	数据生成分布
$\hat{p}_{\text{train}}$	由训练集定义的经验分布
$\mathbb{X}$	训练样本的集合
$\boldsymbol{x}^{(i)}$、$\boldsymbol{x}_i$	数据集的第 i 个样本 (输入)
$y^{(i)}$、$\boldsymbol{y}^{(i)}$、y_i 或 $\boldsymbol{y}_i$	监督学习中与 $\boldsymbol{x}^{(i)}$ 关联的目标
$\boldsymbol{X}$	$m \times n$ 的矩阵，其中行 $\boldsymbol{X}_{i,:}$ 为输入样本 $\boldsymbol{x}^{(i)}$

本套丛书从人工智能科普、人工智能相关计算、人工智能前沿技术以及人工智能相关创新创业等方面出发，内容涵盖人工智能的各个方面，可以使读者全面了解人工智能。

《人工智能：来路与前程》

《人工智能世代的创新与创业》

《计算人生》

《强化学习》

……

目 录

Contents