TURING 图灵数学·统计学丛书

统计机器学习界泰斗作品

Statistical Learning with Sparsity:
The Lasso and Generalizations

# 稀疏统计学习及其应用

[美] Trevor Hastie
Robert Tibshirani
Martin Wainwright 著

刘波 景鹏杰 译

非外借

全面介绍稀疏统计模型及其研究成果
用lasso模型解决大数据挖掘、机器学习等热点问题

CRC Press
Taylor & Francis Group

中国工信出版集团

人民邮电出版社
POSTS & TELECOM PRESS

TURING 图灵数学·统计学丛书

Statistical Learning with Sparsity:
The Lasso and Generalizations

# 稀疏统计学习及其应用

[美] Trevor Hastie
Robert Tibshirani
Martin Wainwright 著

刘波 景鹏杰 译

人民邮电出版社
北京

**图书在版编目(CIP)数据**

稀疏统计学习及其应用/（美）特里瓦·哈斯蒂（Trevor Hastie），（美）罗伯特·蒂伯沙拉尼（Robert Tibshirani），（美）马丁·韦恩怀特（Martin Wainwright）著；刘波，景鹏杰 译 —北京：人民邮电出版社，2018.1

（图灵数学·统计学丛书）

ISBN 978-7-115-47261-8

Ⅰ.① 稀… Ⅱ.①特… ②罗… ③马… ④刘… ⑤景… Ⅲ. ①统计学 Ⅳ.① C8

中国版本图书馆 CIP 数据核字（2017）第 282216 号

## 内 容 提 要

稀疏统计模型只具有少数非零参数或权重，经典地体现了化繁为简的理念，因而广泛应用于诸多领域。本书就稀疏性统计学习做出总结，以 lasso 方法为中心，层层推进，逐渐囊括其他方法，深入探讨诸多稀疏性问题的求解和应用；不仅包含大量的例子和清晰的图表，还附有文献注释和课后练习，是深入学习统计学知识的极佳参考。

本书适合算法、统计学和机器学习专业人士。

◆ 著 [美] Trevor Hastie, Robert Tibshirani, Martin Wainwright

译 刘 波 景鹏杰

责任编辑 朱 巍

责任印制 彭志环

◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路 11 号

邮编 100164 电子邮件 315@ptpress.com.cn

网址 http://www.ptpress.com.cn

北京隆昌伟业印刷有限公司印刷

◆ 开本：700×1000 1/16

印张：18.75 彩插：4

字数：368 千字 2018 年 1 月第 1 版

印数：1–3 000 册 2018 年 1 月北京第 1 次印刷

著作权合同登记号 图字：01-2015-8299 号

定价：89.00 元

读者服务热线：(010)51095186 转 600 印装质量热线：(010) 81055316

反盗版热线：(010)81055315

广告经营许可证：京东工商广登字 20170147 号

站在巨人的肩上

Standing on Shoulders of Giants


iTuring.cn

站在巨人的肩上

Standing on Shoulders of Giants


iTuring.cn

# 版 权 声 明

*Statistical Learning with Sparsity: The Lasso and Generalizations* / by Trevor Hastie, Robert Tibshirani, Martin Wainwright / ISBN: 978-1-4987-1216-3.

谨以此书献给我们的父母亲：

Valerie Hastie、Patrick Hastie

Vera Tibshirani、Sami Tibshirani

Patricia Wainwright、John Wainwright

也献给我们的亲人：

Samantha、Timothy、Lynda

Charlie、Ryan、Jesss、Julie、Cheryl

Haruko、Hana

# 译 者 序

我们怀着无比敬仰的心情译完了这本书，因为本书的三位作者都是统计机器学习界的泰斗，他们在基于稀疏的统计学习理论方面的造诣有目共睹。能翻译此书是我们的荣幸。

随着大数据时代的到来，稀疏性成为研究大数据的重要手段。斯坦福大学的 Robert Tibshirani 于 1996 年首次提出将 $\ell_1$ 范数作为普通最小二乘的正则项，从而得到了著名的 lasso 模型。lasso 在拟和数据的同时，也利用 $\ell_1$ 范数让系数具有稀疏性来选择特征。这种稀疏性可增加模型的可解释性，并提高计算效率，这些特性对高维数据尤其重要。因此，$\ell_1$ 范数的这一特性很快就吸引了大批学者进行研究，并在各种机器学习模型中广泛使用，这些模型不仅仅包括广义线性模型，还包括图模型、信号处理中的压缩感知模型、矩阵填充模型等。$\ell_1$ 范数是一个凸函数，但不可微，因此，在求解以 $\ell_1$ 范数作为惩罚项的模型时通常比较麻烦。用来求解这类模型的常用算法是次梯度方法，但这种方法的效率通常不高，尤其在大规模的高维数据上更是如此。最近又出现了大量求解该问题的高效求解算法，比如坐标下降法等。

本书从最基本的 lasso 模型出发介绍了 lasso 模型的物理意义和相应的求解算法，然后介绍了 lasso 模型的推广和最新的研究成果。本书的特点是：(1) 采用深入浅出、图文并茂的方式介绍各种抽象理论；(2) 内容新颖，书中大量内容都是最近几年的研究成果，比如筛选规则等；(3) 涉及面广，书中的内容包含了与稀疏性相关的各个重要的研究领域，比如多元统计中的稀疏性问题等。因此，本书是广大机器学习研究人员、工程人员不可多得的参考书，也可用于研究生的机器学习课程的教材。

本书的第 1 章、第 2 章、第 5 章、第 7 章、第 8 章、第 10 章由重庆工商大学计算机科学与信息工程学院的刘波博士翻译，第 3 章、第 4 章、第 6 章、第 9 章、第 11 章由上海期货信息技术有限公司的景鹏杰翻译。

由于本书涉及的知识面广，内容也很新颖，因此许多术语尚无固定译法。我们虽然经过反复推敲和讨论，但仍然可能出现词不达意的情况。同时由于时间和能力有很，书中内容难免不出差错。若有问题或建议，读者可通过电子邮件 liubo7971@163.com 或 pjjing@foxmail.com 与我们联系，欢迎大家对本书的翻译进行指正或提出宝贵的建议。本书翻译的勘误信息会发布在 http://www.cnblogs.com/ml-cv/上，欢迎关注。

本书翻译过程得到如下项目资助：（1）重庆市教委研究项目“多核正则化机器学习理论研究”，项目号为 KJ130709；（2）重庆工商大学研究项目“基于多核学习的高维数据分析研究”，项目号为 2013-56-09；（3）大数据稀疏表示判别字典学习及其应用技术研究，项目号为 KJ1400612；（4）重庆工商大学研究生院教改项目“基于二维码的研究生互动教学改革”；（5）电子商务及供应链系统重庆市重点实验室项目“基于迹比率的特征选择及关键技术研究”，项目编号为 1456025。

感谢翻译过程中图灵公司的朱巍编辑给予的帮助和支持；感谢高敬雅老师，她给予了我们在统计学上的帮助；感谢重庆工商大学计信学院金融信息化专业的曾芳同学，她帮助我们录入了本书的公式。感谢刘波的妻子杨雪莉的支持，感谢刘波两个小女儿刘典、刘恩丫（此书完成时她才 7 个月大）对他的忍耐。

# 前　言

在这本专著中，我们将概述基于稀疏性的统计学习的最新研究。稀疏的统计模型仅具有少数非零参数或权值。它代表了“少即是多”的经典情形：与稠密模型相比，稀疏模型更容易估计和解释。在这个大数据的时代，对一个人或目标进行度量的特征数量可能更多，而且有可能比观测样本数更多。借助稀疏性假设，我们能够解决这些问题，并从大数据集中提取有用的、可重复性模式。

这里所陈述的观点代表了统计学和机器学习方面全体研究人员的工作，我们感谢大家对这个令人激动的领域所作出的贡献。我们要特别感谢斯坦福大学和加州大学伯克利分校的同事，感谢我们的合作者，以及现在及过去在这个领域工作的同学们。他们分别是：Alekh Agarwal, Arash Amini, Francis Bach, Jacob Bien, Stephen Boyd, Andreas Buja, Emmanuel Candes, Alexandra Chouldechova, David Donoho, John Duchi, Brad Efron, Will Fithian, Jerome Friedman, Max G'Sell, Iain Johnstone, Michael Jordan, Ping Li, Po-Ling Loh, Michael Lim, Jason Lee, Richard Lockhart, Rahul Mazumder, Balasubramanian Narashimhan, Sahand Negahban, Guillaume Obozinski, Mee-Young Park, Junyang Qian, Garvesh Raskutti, Pradeep Ravikumar, Saharon Rosset, Prasad Santhanam, Noah Simon, Dennis Sun, Yukai Sun, Jonathan Taylor, Ryan Tibshirani[①], Stefan Wager, Daniela Witten, Bin Yu, Yuchen Zhang, Ji Zhou, and Hui Zou。感谢我们的编辑 John Kimmel 对本书的建议和支持。

Trevor Hastie, Robert Tibshirani, 斯坦福大学
Martin Wainwright, 加州大学伯克利分校

①本书文献中提到的 $\text{Tibshirani}_2$ 是 Ryan Tibshirani，而 Tibshirani 则是其父 Robert Tibshirani。

# 目　　录

# 第1章 引 言

“我从来都不记记分卡或击球率。我讨厌统计。我会把必须知道的东西记在脑海里。”

这段话是棒球投手 Dizzy Dean 说的，他曾在 1930 ~1947 年参加美国职业棒球大联盟的比赛。

一晃 75 年过去了，世界发生了很大的变化！如今，人们在科学、娱乐、商业和工业各领域收集和挖掘大量数据，并对其进行研究和应用。医学家们通过研究患者的基因组选择最佳的治疗方法，并由此了解这些疾病产生的根本原因。在线电影和网上书店会研究客户的评价，以便向他们推荐新的电影或书籍。社交网络会研究其会员及好友的资料，优化在线体验。而且，现在多数大联盟棒球队都有统计员收集和分析击球手和投手的详细信息，帮助球队经理和队员做出更好的决策。

由此可知，这个世界淹没在了数据中。而 Rutherford D. Roger 等人则说：

“我们淹没在了信息的海洋里，却渴求着知识。”

海量信息亟待整理，取其精华去其糟粕。为了成功完成这项工作，人们期望真实情况得以简化：也许人体内大约 30 000 个基因并非都与癌症的发展过程直接相关；也许只需要客户对 50 或 100 部电影做出评价就足以揭示他们的爱好；也许左撇子投手对付左撇子击球手会比较轻松。

这些情形背后都有简单性假设。**稀疏性**（sparsity）是简单性的一种形式，这也是本书的中心主题。简而言之，在一个稀疏统计模型中，仅有较少参数（也称预测子，predictor）在发挥重要作用。本书将介绍如何利用稀疏性来恢复一组数据中的基础信号。

最典型的例子是线性回归，即有 $N$ 组观测值，每组观测值由一个输出变量 $y_i$ 和 $p$ 个相关预测子变量（也称特征）$x_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ 所组成。线性回归的目标是通过预测子来预测输出值，既要正确预测将来的数据，又要找出哪些预测子在起重要作用。一个线性回归模型可设为：

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + e_i \tag{1.1}$$

其中，$\beta_0$ 和 $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ 是未知参数，$e_i$ 为误差项。这些参数可用最小二乘法来估计，即最小化最小二乘目标函数：

$$\underset{\beta_0,\beta}{\text{minimize}} \sum_{i=1}^{N} \left( y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \tag{1.2}$$