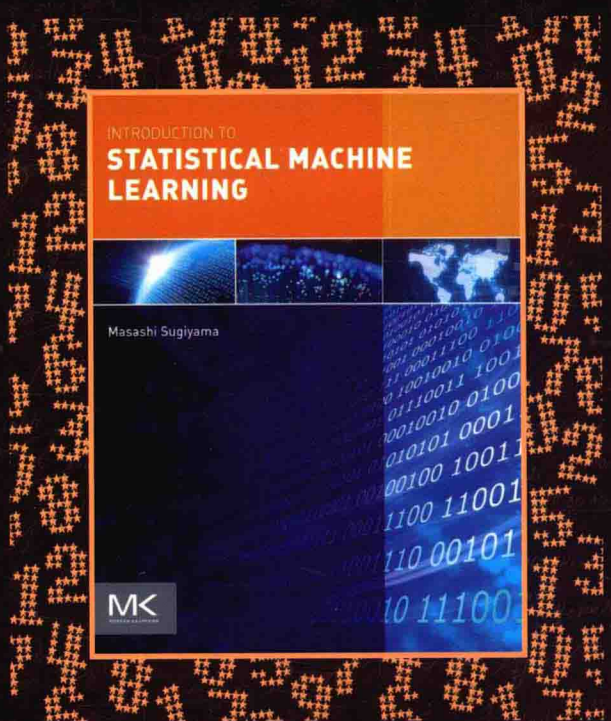
ELSEVIER

HZ BOOKS
华章IT

数据科学与工程技术丛书

统计机器学习导论

[日] 杉山将（Masashi Sugiyama）著
谢宁 李柏杨 肖竹 罗宇轩 等译



INTRODUCTION TO
STATISTICAL MACHINE LEARNING

机械工业出版社
China Machine Press

数据科学与工程技术丛书

INTRODUCTION TO STATISTICAL MACHINE LEARNING

统计机器学习导论

[日] 杉山将（Masashi Sugiyama）著

谢宁 李柏杨 肖竹 罗宇轩 等译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

统计机器学习导论 /（日）杉山将（Masashi Sugiyama）著；谢宁等译 . —北京：机械工业出版社，2018.4
（数据科学与工程技术丛书）
书名原文：Introduction to Statistical Machine Learning

ISBN 978-7-111-59679-0

I. 统… II. ① 杉… ② 谢… III. 机器学习 IV. TP181

中国版本图书馆 CIP 数据核字（2018）第 067578 号

本书版权登记号：图字 01-2016-4752

ELSEVIER
Elsevier (Singapore) Pte Ltd.
3 Killiney Road, #08-01 Winsland House I, Singapore 239519
Tel: (65) 6349-0200; Fax: (65) 6733-1817

Introduction to Statistical Machine Learning
Masashi Sugiyama

ISBN-13: 9780128021217

《统计机器学习导论》（谢宁 李柏杨 肖竹 罗宇轩 等译）
ISBN: 978-7-111-59679-0

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：唐晓琳　　责任校对：殷 虹
印 刷：三河市宏图印务有限公司　　版 次：2018 年 5 月第 1 版第 1 次印刷
开 本：185mm × 260mm 1/16　　印 张：22
书 号：ISBN 978-7-111-59679-0　　定 价：89.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991 88361066　　投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259　　读者信箱：hzjsj@hzbook.com

译 者 序

机器学习是计算机科学的重要分支之一，旨在研究原理、算法以及能够像人类一样学习的系统的应用。同时，其亦是一门交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。机器学习作为人工智能的核心部分，是计算机获得智能的根本途径，其应用遍及人工智能的各个领域，发挥着不可替代的重要作用。

本书是日本人工智能和机器学习领域的新一代领军人物杉山将(Masashi Sugiyama)的统计机器学习力作。本书致力于讲解数学背景及多种机器学习技术的实用化算法。其结构清晰，内容丰富，案例详实，系统地介绍了统计机器学习的概念、技术及应用。通过对本书的学习，读者可以了解统计机器学习的基本概念和知识，同时培养统计机器学习的基本技能。阅读本书需要了解计算机科学、概率论与统计学等相关基础知识。本书适用于计算机及相关专业的本科生、研究生以及相关领域的研究人员和专业技术人员。

本书翻译工作得到了课题组成员的鼎力支持和大力协作。

谢宁作为本次翻译活动的倡议者和联络人，负责和参与翻译了前言、作者简介及第1章，并帮助分析、修改各章中的疑难点。

第一部分(第1章)，由谢宁和李柏杨共同完成。

第二部分(第2～10章)，由李煜玮、周飞宇、苏秋霖和文洋负责翻译。

第三部分(第11～20章)，由王磊和徐颖负责翻译。

第四部分(第21～29章)，由罗宇轩负责翻译。

第五部分(第30～39章)，由李柏杨负责翻译。

此外，肖竹负责翻译第一到五部分的引言内容。在翻译组内部审校阶段，谢宁担任内部审校总负责人，李柏杨、肖竹和张帅担任内部审校主要负责人。

本书中文版能够出版发行，首先要感谢本书的作者杉山将教授，是他为我们著作了一本好书。其次要感谢机械工业出版社华章公司引进了本书的中文版权，使得我们能够获得为博士导师杉山将教授翻译此书的机会，并实现将其介绍给国内广大读者的良好愿望。此外，特别感谢本书的编辑曲熠以及所有为此书的出版做出贡献

的排校人员，是他们的辛勤劳动才使本书能够付诸印刷和出版，在此表示深深的感谢和崇高的敬意！

本书对原著的错误之处做了一些修正，在原著难懂或需要提醒的地方添加了一些译者说明。尽管我们在翻译过程中力图做得更好，但因个人的业务水平、英文水平乃至中文文学水平的限制，以及翻译过程中的粗心和不够严谨，可能使得本书中文版中存在错误、不足和不当之处。热切期望读者对本书提出宝贵意见、建议和勘误，并欢迎与我们联络(seanxiening@gmail.com)。

2018 年 3 月

前　言

机器学习是计算机领域的一个学科，旨在研究原理、算法以及能够像人类一样学习的系统的应用。近年来，计算机和传感器的发展使得我们能够访问不同领域的海量数据(如文本、音频、图片、电影、电子商务、电气、医学和生物学等)。在此类大数据的分析和利用方面，机器学习起到了核心的作用。

本书致力于讨论机器学习的数学背景及多种机器学习技术的实用化算法。目标读者定位于计算机和相关专业的本科生和研究生。在工作中应用机器学习技术的工程师和分析数据的科学家也会从本书中获益。

本书特色在于每章的主题简明扼要，给出具体机器学习技术的数学推导并附以简洁的 MATLAB 程序。由此，读者在学习数学概念的同时，可掌握多种机器学习技术的实用价值。全部 MATLAB 程序可以从如下网址获得：

http://www.ms.k.u-tokyo.ac.jp/software/SMLbook.zip

本书第一部分给出机器学习领域的简要概述。紧接着，第二部分介绍了概率和统计的基本概念，它们构成了统计机器学习的数学基础。第二部分的成文基于：

Sugiyama, M.
Probability and Statistics for Machine Learning,
Kodansha, Tokyo, Japan, 2015. (in Japanese).

第三部分和第四部分分别在生成和判别框架下，介绍了一系列实用机器学习算法。随后，第五部分介绍高级论题，进而处理更具挑战的机器学习任务。第三部分的成文基于：

Sugiyama, M.
Statistical Pattern Recognition: Pattern Recognition Based on Generative Models,
Ohmsha, Tokyo, Japan, 2009. (in Japanese),

第四部分和第五部分的成文基于：

Sugiyama, M.
An Illustrated Guide to Machine Learning,
Kodansha, Tokyo, Japan, 2013. (in Japanese).

在此感谢东京大学和东京工业大学相关研究组的研究员和学生针对本书早期手稿给出的有价值的反馈。

杉山将
东京大学

作者简介

杉山将 分别于1997、1999和2001年获得日本东京工业大学颁发的计算机专业工学学士、硕士和博士学位。2001年，他被东京工业大学聘为助理教授，随后于2003年晋升为副教授。2014年，他转到东京大学任教授。2003至2004年，他获得亚历山大·冯洪堡特基金会研究奖学金并前往位于德国柏林的弗朗霍夫研究所从事研究工作。2006年，他获得欧盟委员会项目Erasmus Mundus奖学金并前往英国爱丁堡大学从事研究工作。他曾被授予2007年度IBM学者奖(表彰他在机器学习领域中非平稳性方面的贡献)、2011年度日本信息处理学会颁发的Nagao特别研究员奖，以及由文部科学省颁发的科学技术领域的青年科学家奖(表彰他对机器学习领域密度比范式的贡献)。他的研究兴趣包括机器学习与数据挖掘的理论与算法，及其广泛的应用(如信号处理、图像处理和机器人控制等)。

目　录

第四部分 统计机器学习的判别式方法

第五部分　高级主题

第一部分

绪　论

第 1 章　统计机器学习

第 1 章

统计机器学习

计算机与网络的近期发展使得我们能够即刻访问大量信息，如文字、声音、图像与影像。此外，日志、消费记录和病历等广泛的个人数据也日复一日地累积起来。如此大量的数据被称为*大数据*(big data)。并且，存在一股通过从数据中抽取有用的知识来创造新价值与商机的增长趋势。这一过程通常被称为*数据挖掘*(data mining)。*机器学习*是用于抽取有用知识的关键技术。

1.1 学习的类型

依据可用数据的类型，机器学习能够分为*监督学习*(supervised learning)、*非监督学习*(unsupervised learning)和*强化学习*(reinforcement learning)。

*监督学习*可谓机器学习最基本的类型。它被视为一种学生学习的过程，即向导师提问并回答。在机器学习情境中，学生对应于计算机，导师对应于计算机的用户；计算机从问与答的成对样本中学习一种从问题到其答案的映射。监督学习的目标在于获得*泛化能力*(generalization ability)。其指的是一种能够为从未被学习过的问题猜出恰当答案的能力。因此，用户不必再将每件事情都教给计算机，而计算机仅通过学习一小部分知识就能够自动地应对未知情况。监督学习已经被成功地应用到广泛的真实问题中，如：手写字母识别、语音识别、图像识别、垃圾邮件过滤、信息检索、在线广告、推荐系统、脑电波分析、基因分析、股票价格预测、天气预报和天文数据分析。监督学习的问题也会被具体称为*回归*(regression)，当答案是一个实数值(如：温度)；*分类*(classification)，如果答案是一个分类值(如："是"或"否")以及*排序*(ranking)，如果答案是一个数列值(如："好"、"中"或"劣")。

*非监督学习*被认为是，导师不存在并且学生自学。在机器学习情境中，计算机通过互联网自动地收集数据并且尝试在没有用户任何指导下抽取有用的知识。因此，非监督学习比监督学习更加自动化，尽管其目标不一定指定清楚。非监督学习的典型任务包括*数据聚类*(data clustering)和*异常点检测*(outlier detection)。同时，这些非监督学习技术已经就广泛的真实问题取得巨大成功，如系统诊断、安全、事件检测和社交网络分析。非监督学习也通常被用作监督学习的预处理过程。

*强化学习*与监督学习类似，也是以使计算机获得对没有学习过的问题做出正确解答的泛化能力为目标，但是在学习过程中，不设置导师提示对错、告知最终答案的环节。相

反，导师评价(evaluate)学生的行为并给予其反馈。强化学习的目标是基于来自导师的反馈，使得学生提高其行为，从而最大化导师的评价。强化学习是一个人与机器人的行为的重要的模型。它已经被广泛地应用于诸多领域，如：自主机器人控制、电脑游戏和营销策略优化。在强化学习背后，监督学习与非监督学习的方法，如：回归、分类与聚类，都常被使用。

本书重点在于监督学习与非监督学习。对于强化学习，请参考文献[99，105]。

1.2　机器学习任务举例

在本节，会详尽介绍不同的监督学习与非监督学习任务。

1.2.1　监督学习

回归(regression)的目标是为了从其样本近似估计一个实数值函数(如图 1.1 所示)。让我们不妨定义输入为 d 维实数向量 $\boldsymbol{x}$，输出为实数标量 y，学习目标函数为 $y=f(\boldsymbol{x})$。这个学习目标函数 f 被认为是未知的，但是其输入输出样本对$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$是可以被观测的。在实践中，被观测输出值 y_i 可能被某些误差ϵ_i 所污染，即

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i$$

在这个设定中，x_i 对应学生向导师提出的一个问题；y_i 对应导师给学生的回答。噪声ϵ_i 或许对应导师的口误或者学生的误解。学习目标函数 f 对应导师的知识，这些知识使得他(她)可以回答任何问题。回归的目标在于让学生学习这个函数，学生由此也可以回答任何问题。泛化的水平能够通过真实函数 f 与其近似 $\hat{f}$ 的接近程度来测量。

另一个方面，以监督学习方式，分类则是一个模式识别(pattern recognition)问题(如图 1.2 所示)。不妨用 d 维向量 $\hat{\boldsymbol{x}}$ 来定义输入模式和由标量 $y\in\{1, \cdots, c\}$定义其类别，其中 c 定义类别数目。为了训练分类器，输入-输出对样本$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$以与回归一样的方式提供。倘若真实分类规则被定义为 $y=f(\boldsymbol{x})$，那么分类也能够被视为函数逼近问题。然而，在回归情境下，一个本质的不同在于在 y 中不存在接近程度这一概念 y：$y=2$ 比 $y=1$更接近 $y=3$，但是“y 和 y‘是否相同’”是分类唯一关心的。

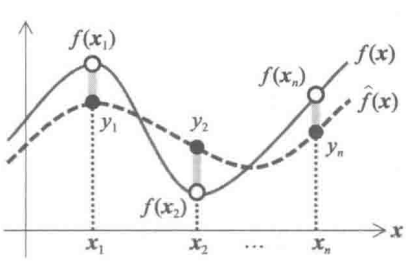


图 1.1　回归

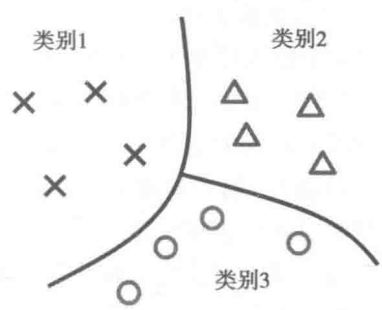


图 1.2　分类

在监督学习中，排序(ranking)问题旨在学习样本 $\boldsymbol{x}$ 的次序 y。因为次序有顺序，如：$1<2<3$，排序更像是回归而非分类。为此，排序问题也可以被视为有序回归(ordinal regression)。然而，不同于回归，提出输出值 y 是没有必要被预测的，但也仅仅是其相对