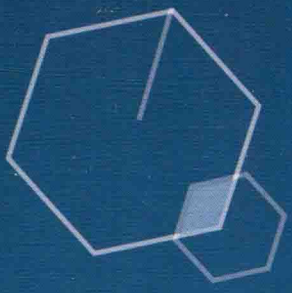
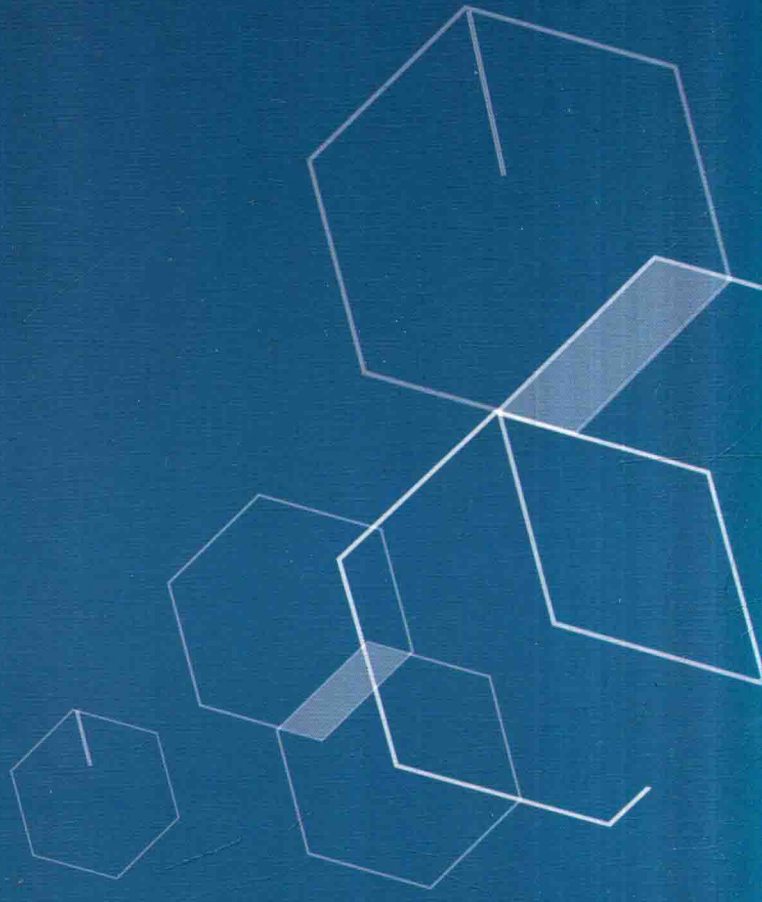
# 统计学习方法

（第2版）

李航 著





清华大学出版社

# 统计学习方法

（第2版）

李航 著

清华大学出版社
北京

## 内 容 简 介

统计学习方法即机器学习方法，是计算机及其应用领域的一门重要学科。本书分为监督学习和无监督学习两篇，全面系统地介绍了统计学习的主要方法。包括感知机、$k$近邻法、朴素贝叶斯法、决策树、逻辑斯谛回归与最大熵模型、支持向量机、提升方法、EM 算法、隐马尔可夫模型和条件随机场，以及聚类方法、奇异值分解、主成分分析、潜在语义分析、概率潜在语义分析、马尔可夫链蒙特卡罗法、潜在狄利克雷分配和 PageRank 算法等。

本书是统计机器学习及相关课程的教学参考书，适用于高等院校文本数据挖掘、信息检索及自然语言处理等专业的大学生、研究生，也可供计算机应用等专业的研发人员参考。



**图书在版编目(CIP)数据**

统计学习方法/李航著.—2 版.—北京：清华大学出版社，2019
ISBN 978-7-302-51727-6

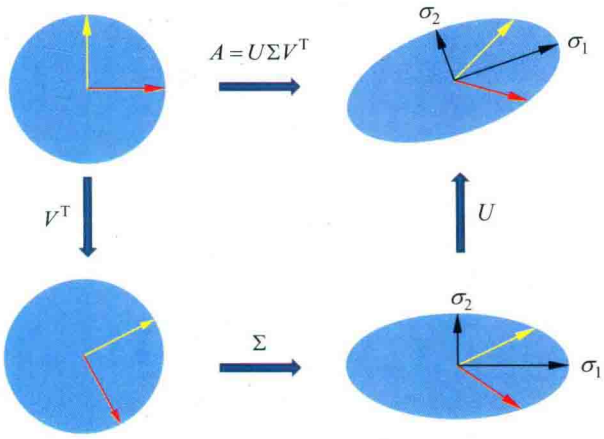
Ⅰ. ①统…　Ⅱ. ①李…　Ⅲ. ①统计学　Ⅳ. ①C8

中国版本图书馆 CIP 数据核字(2018)第 267477 号

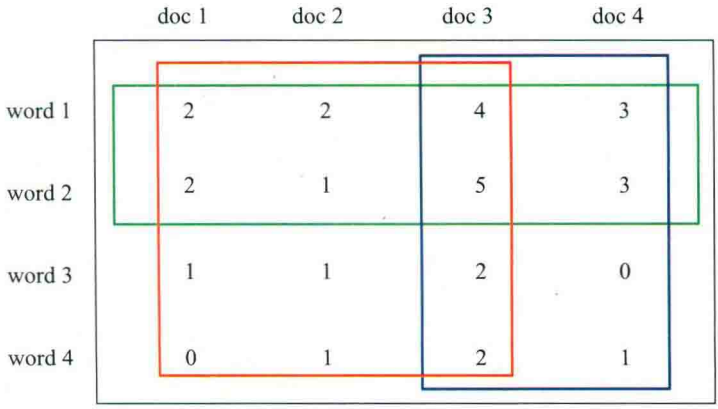
**责任编辑**：薛　慧
**封面设计**：李祥榕
**责任校对**：刘玉霞
**责任印制**：宋　林

**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
**印 装 者**：三河市龙大印装有限公司
**经　　销**：全国新华书店
**开　　本**：170mm×240mm　　**印　张**：30.25　　**插　页**：1　　**字　数**：593 千字
**版　　次**：2012 年 3 月第 1 版　2019 年 5 月第 2 版　　**印　次**：2019 年 5 月第 1 次印刷
**印　　数**：1～20000
**定　　价**：98.00 元

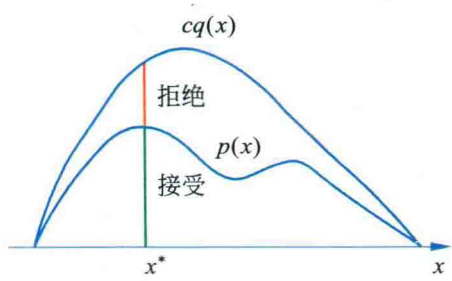
产品编号：081329-01



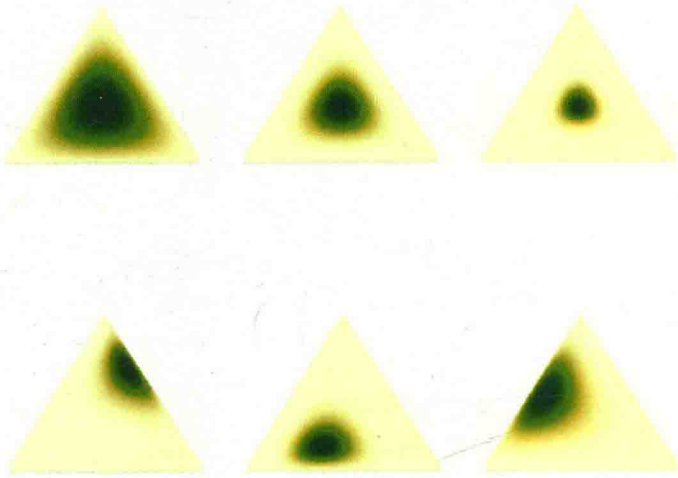

彩图 15.1　奇异值分解的几何解释



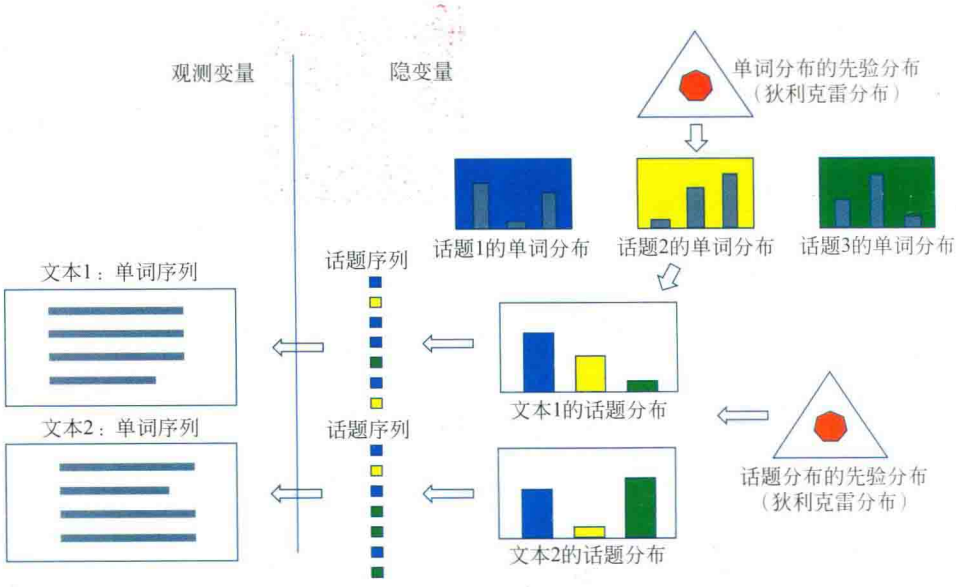

彩图 18.1　概率潜在语义分析的直观解释




彩图 19.1　接受-拒绝抽样法



彩图 20.1　狄利克雷分布例




彩图 20.3　LDA 的文本生成过程

献给我的母亲

# 第 2 版序言

《统计学习方法》第 1 版于 2012 年出版，讲述了统计机器学习方法，主要是一些常用的监督学习方法。第 2 版增加了一些常用的无监督学习方法，由此本书涵盖了传统统计机器学习方法的主要内容。

在撰写《统计学习方法》伊始，对全书内容做了初步规划。第 1 版出版之后，即着手无监督学习方法的写作。由于写作是在业余时间进行，常常被主要工作打断，历经六年多时间才使这部分工作得以完成。犹未能加入深度学习和强化学习等重要内容，希望今后能够增补，完成整本书的写作计划。

《统计学习方法》第 1 版的出版正值大数据和人工智能的热潮，生逢其时，截至 2019 年 4 月本书共印刷 25 次，152000 册，得到了广大读者的欢迎和支持。有许多读者指出本书对学习和掌握机器学习技术有极大的帮助，也有许多读者通过电子邮件、微博等方式指出书中的错误，提出改进的建议和意见。一些高校将本书作为机器学习课程的教材或参考书。有的同学在网上发表了读书笔记，有的同学将本书介绍的方法在计算机上实现。清华大学深圳研究生院袁春老师精心制作了第 1 版十二章的课件，在网上公布，为大家提供教学之便。众多老师、同学、读者的支持和鼓励，让作者深受感动和鼓舞。在这里向所有的老师、同学、读者致以诚挚的谢意！

能为中国的计算机科学、人工智能领域做出一点微薄的贡献，感到由衷的欣慰，同时也感受到作为知识传播者的重大责任，让作者决意把本书写好。也希望大家今后不吝指教，多提宝贵意见，以帮助继续提高本书的质量。在写作中作者也深切体会到教学相长的道理，经常发现自己对基础知识的掌握不够扎实，通过写作得以对相关知识进行了深入的学习，受益匪浅。

本书是一部机器学习的基本读物，要求读者拥有高等数学、线性代数和概率统计的基础知识。书中主要讲述统计机器学习的方法，力求系统全面又简明扼要地阐述这些方法的理论、算法和应用，使读者能对这些机器学习的基本技术有很好的掌握。针对每个方法，详细介绍其基本原理、基础理论、实际算法，给出细致的数学推导和具体实例，既帮助读者理解，也便于日后复习。

第 2 版增加的无监督学习方法，王泉、陈嘉怡、柴琛林、赵程绮等帮助做了认真细致的校阅，提出了许多宝贵意见，在此谨对他们表示衷心的感谢。清华大学出版社的薛慧编辑一直对本书的写作给予非常专业的指导和帮助，在此对她表示衷心的感谢!

由于本人水平有限，本书一定存在不少错误，恳请各位专家、老师和同学批评指正。

李 航

2019 年 4 月

# 第 1 版序言

计算机与网络已经融入人们的日常学习、工作和生活之中，成为人们不可或缺的助手和伙伴。计算机与网络的飞速发展完全改变了人们的学习、工作和生活方式。智能化是计算机研究与开发的一个主要目标。近几十年来的实践表明，统计机器学习方法是实现这一目标的最有效手段，尽管它还存在着一定的局限性。

本人一直从事利用统计学习方法对文本数据进行各种智能性处理的研究，包括自然语言处理、信息检索、文本数据挖掘。近 20 年来，这些领域发展之快，应用之广，实在令人惊叹！可以说，统计机器学习是这些领域的核心技术，在这些领域的发展及应用中起着决定性的作用。

本人在日常的研究工作中经常指导学生，并在国内外一些大学及讲习班上多次做过关于统计学习的报告和演讲。在这一过程中，同学们学习热情很高，希望得到指导，这使作者产生了撰写本书的想法。

国内外已出版了多本关于统计机器学习的书籍，比如，Hastie 等人的《统计学习基础》，该书对统计学习的诸多问题有非常精辟的论述，但对初学者来说显得有些深奥。统计学习范围甚广，一两本书很难覆盖所有问题。本书主要是面向将统计学习方法作为工具的科研人员与学生，特别是从事信息检索、自然语言处理、文本数据挖掘及相关领域的研究与开发的科研人员与学生。

本书力求系统而详细地介绍统计学习的方法。在内容选取上，侧重介绍那些最重要、最常用的方法，特别是关于分类与标注问题的方法。对其他问题及方法，如聚类等，计划在今后的写作中再加以介绍。在叙述方式上，每一章讲述一种方法，各章内容相对独立、完整；同时力图用统一框架来论述所有方法，使全书整体不失系统性，读者可以从头到尾通读，也可以选择单个章节细读。对每一种方法的讲述力求深入浅出，给出必要的推导证明，提供简单的实例，使初学者易于掌握该方法的基本内容，领会方法的本质，并准确地使用方法。对相关的深层理论，则予以简述。在每章后面，给出一些习题，介绍一些相关的研究动向和阅读材料，列出参考文献，以满足读者进一步学习的需求。本书第 1 章简要叙述统计学习方法的基本概念，最后一章对统计学习方

法进行比较与总结。此外，在附录中简要介绍一些共用的最优化理论与方法。

本书可以作为统计机器学习及相关课程的教学参考书，适用于信息检索及自然语言处理等专业的大学生、研究生。

本书初稿完成后，田飞、王佳磊、武威、陈凯、伍浩铖、曹正、陶宇等人分别审阅了全部或部分章节，提出了许多宝贵意见，对本书质量的提高有很大帮助，在此向他们表示衷心的感谢。在本书写作和出版过程中，清华大学出版社的责任编辑薛慧给予了很多帮助，在此特向她致谢。

由于本人水平所限，书中难免有错误和不当之处，欢迎各位专家和读者给予批评指正。

李 航

2011 年 4 月 23 日

# 目　录

## 第1篇　监督学习