新时代 • 技术新未来

移动通信大数据分析
——数据挖掘与机器学习实战

[中] 欧阳晔（Ye Ouyang）
[中] 胡曼恬（Mantian Hu）
[法] 亚历克西斯 • 休特（Alexis Huet）
[中] 李中源（Zhongyuan Li）
著

徐俊杰　译

清華大學出版社
北　京

北京市版权局著作权合同登记号　图字：01-2019-5926

图书在版编目（CIP）数据

移动通信大数据分析：数据挖掘与机器学习实战 / 欧阳晔等著；徐俊杰译 . —北京：清华大学出版社，2020.12
（新时代 • 技术新未来）
书名原文：Mining Over Air: Wireless Communication Network Analytics
ISBN 978-7-302-54124-0

Ⅰ . ①移…　Ⅱ . ①欧… ②徐…　Ⅲ . ①移动网－数据采集②机器学习　Ⅳ . ① TP274 ② TP181

中国版本图书馆 CIP 数据核字 (2019) 第 247853 号

责任编辑：刘　洋
封面设计：徐　超
版式设计：方加青
责任校对：王凤芝
责任印制：杨　艳

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：小森印刷（北京）有限公司
经　　销：全国新华书店
开　　本：187mm×235mm　　**印　　张：**13.25　　**字　　数：**239 千字
版　　次：2020 年 12 月第 1 版　　**印　　次：**2020 年 12 月第 1 次印刷
定　　价：99.00 元

产品编号：085561-01

内 容 简 介

本书以 4G / 5G 无线技术、机器学习和数据挖掘的新研究和新应用为基础，对分析方法和案例进行研究；从工程和社会科学的角度，提高读者对行业的洞察力，提升运营商的运营效益。本书利用机器学习和数据挖掘技术，研究移动网络中传统方法无法解决的问题，包括将数据科学与移动网络技术进行完美结合的方法、解决方案和算法。

本书可以作为研究生、本科生、科研人员、移动网络工程师、业务分析师、算法分析师、软件开发工程师等的参考书，具有很强的实践指导意义，是不可多得的专业著作。

致 谢

我谨将最诚挚的爱献给我的女儿欧阳琳琅和我的妻子徐蓉蓉，感谢你们陪伴我并给我无尽的支持。

——欧阳晔博士

我想借此机会感谢我的父母黄兴和胡学峰，感谢你们对我一直以来的宝贵支持！

——胡曼恬博士

谢谢我的妻子敏珠和我的儿子丹尼尔，你们给了我力量，让我脚踏实地、阔步向前！

——李中源

主要作者简介

欧阳晔　博士

■ 亚信科技首席技术官、高级副总裁

欧阳晔博士目前全面负责亚信科技的技术与产品的研究、开发与创新工作。加入亚信科技之前，欧阳晔博士曾任职于美国第一大移动通信运营商威瑞森电信（Verizon）集团，担任通信人工智能系统部经理，是威瑞森电信的 Fellow。欧阳晔博士在移动通信领域拥有丰富的研发与大型团队管理经验，工作中承担过科学家、研究员、研发经理、大型研发团队负责人等多个角色。欧阳晔博士专注于移动通信、数据科学与人工智能领域跨学科研究，致力于 5G 网络智能化、BSS/OSS 融合、通信人工智能、网络切片、MEC、网络体验感知、网络智能优化、5G 行业赋能、云网融合等领域的研发创新与商业化。

欧阳晔博士在多个国际标准、技术、工业和学术组织中担任职务，包括 3GPP 和欧洲通信标准组织（ETSI）公司代表、IEEE 5G 峰会工业界主席、IEEE Sarnoff 行业主席、IEEE 工业互联网（ICII）工业界主席、IEEE GLOBECOM 高层管理论坛主席、IEEE 大数据委员会执行委员、IEEE 计算、网络及通信国际会议（ICNC）研讨会主席、IEEE 无线通信研讨会（WTS）及 IEEE 无线与光通信会议（WOCC）大数据委员会主席、机械工业出版社专家咨询委员会委员等，并在多种期刊担任编委和审稿人，以及在多个学术会议担任审稿人。

欧阳晔博士在工业界与学术界获得多项荣誉与奖励，包括 2018—2019 年度 TMForum 电信业未来数字领袖大奖、2017 年美国杰出亚裔工程师奖、2017 年 IEEE 国际大数据会议最佳论文奖、2017 年美国电信业创新大奖和最佳 OSS/BSS 产品奖、2017 年北美最佳运营商大数据系统奖、2016 年美国电信业创新大奖、2015 年 IEEE 无线通信年会“无线

通信跨领域贡献奖”、2012 年美国总统科学技术与政策办公室电信大数据研究基金等。

欧阳晔博士发表了 30 余篇学术论文，拥有 40 余项专利，提出 10 余项国际标准，著有 5 本学术书籍。欧阳晔博士拥有中国东南大学无线电工程系学士学位、美国哥伦比亚大学硕士学位、美国塔夫茨大学硕士学位和美国斯蒂文斯理工学院博士学位。

胡曼恬

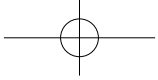
现任香港中文大学工商管理学院市场学系副教授、营销工程中心主任。她曾获美国 Society for Marketing Advance 学会博士论文竞赛最佳论文奖。她的主要研究方向是运用前沿的实证方法进行数据分析和挖掘，探索和解释 TMT、汽车、电商和 FinTech 等行业中的消费者行为，特别是社交网络、口碑效应及人际互动在营销活动中的作用及影响。其研究成果发表于 *Marketing Science*、*Management Science*、*The International Journal of Research in Marketing* 等国际顶尖营销类学术期刊。胡教授担任香港数码分析协会荣誉顾问，并为国内外市场研究公司、电信企业及手机制造商提供营销策略咨询。 胡教授本科毕业于复旦大学，于纽约大学 Stern 商学院取得博士学位。

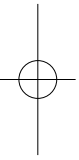
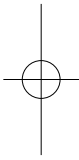
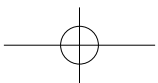
译 者 简 介

徐俊杰

TM Forum 高级技术协作总监，2002 年获英国赫瑞瓦特大学分布式多媒体硕士学位。先后在中国石油大学、Comverse、亚信、华为等单位担任讲师、全球培训师、咨询顾问、解决方案专家等职务，为全球超过三十个国家的运营商、厂商和咨询公司等提供过培训和咨询服务。

翻译出版了 4 本专业图书：《数字经济大趋势：正在到来的商业机遇》《跨界与融合：互联网时代企业合作模式与商业新机遇》《架构即服务：企业数字化运营架构设计与演进》《数字经济生存之道：电信运营商转型》。

推荐序一

第五代移动通信（The Fifth-Generation，5G）与人工智能（Artificial Intelligence，AI）作为 21 世纪最新的一组通用目的技术（General Purpose Technology，GPT），与 19 世纪、20 世纪以电力、内燃机、计算机和互联网为主的 GPT 一样，将极大地促进人类社会从工业化、信息化到数字化的变革发展。全球通信运营商们，从 3G 时代开始逐渐探索自动化与智能化的技术在通信网络与业务生产系统中的应用。结合大数据的发展，通信生态系统中网络与业务的特征数据得以细粒度地被记录、存留在数据仓库或者数据湖中。那么对这些数据进行有效、准确的分析，形成主动性与预测性的决策，促进通信网络与业务运营效率的提升，成为全球通信运营商们数字化转型中一个重要的课题。

在通信运营商生态系统中利用海量数据做自动化与智能化分析，有两条主线在平行发展。在网络领域，我们称之为网络智能化（Network Intelligence），即在网络基础设施或应用管理系统中利用统计学、数据科学、人工智能等技术，在网络的规划、建设、优化、运维的全生命周期中构建敏捷、自动化与智能化的决策与运行机制。网络智能化的决策与运行机制通常由智能化的信息系统来承载实现。这一智能化新系统既可以作为网络基础设施的一部分与网络设施融合存在，也可以作为独立的智能化网络信息系统存在，与网络基础设施通过一套标准化的互联互通规则对网络设施本身进行智能化管理和运行。在业务领域，我们称之为商业智能（Business Intelligence），即在业务支撑系统（Business Supporting System，BSS）中利用统计学、数据科学、人工智能等技术，在业务的运维与运营的全生命周期中构建敏捷、自动化与智能化的决策与运行机制。智能化的决策机制被注入和融入业务支撑体系的各种生产与运行系统中，例如客户关系管理（Customer Relationship Management，CRM）、计费系统（Billing System）、经营分析系统等。

本书作者在移动通信领域拥有丰富的技术管理经验，亲身经历、领导并实践了过去10年中通信领域的数据科学在美国通信运营商蓬勃发展的历程。本书的内容以数据科学和移动通信网络理论为基础，应用于运营商真实的业务场景，将通信大数据与机器学习算法技术深入地应用于通信运营商网络领域与业务领域的各种实际案例中。书中的每一个通信场景案例都用实证分析和量化数据分析的形式呈现，作者将通信网络与业务领域的知识与机器学习算法相结合，演绎并推导出量化可执行的决策，为运营商探索数字化时代以数据驱动网络与业务运营提供了很多宝贵的经验总结。

作为一本在通信大数据领域中技术结合案例分析，并立足于实践的图书，它既适合广大通信、信息、计算机领域的研究生和运营商与通信业软硬件企业的研发人员学习参考，也适合对移动通信、数据科学、人工智能技术感兴趣的读者阅读。

田溯宁　博士

亚信科技董事长

2020 年 11 月于北京

推荐序二

在过去的数十年中，电信行业在大数据的使用及数据分析技术领域始终是领先者。正因如此，电信行业可以更好地了解自己的网络、业务、市场和客户。随着更新兴和强大的网络技术不断演进，电信行业也在持续发展，从 IP 到各代的蜂窝网络，其在可提供的数据与数据问题分析方面都做出了突破性的贡献。电信网络生态系统中产生的数据，包括从物理层到应用层的数据，各种业务的数据，以及用户画像数据等，使得电信行业的业务专家和数据科学家可以探索一个全新的范式——数据驱动的运营，从而更好地运营电信业务与网络。不同于传统的方法，例如仿真与统计分析，电信数据分析是利用通信原理与数据科学领域的知识结合，对通信生态系统中的业务与网络做基于数据驱动的洞察决策。数据驱动的洞察决策需要通信运营商知道为什么（Know Why），即为什么网络与业务的表现有所改变，也需要知道如何（Know How），即如何改进结果，例如业务质量与体验在某一颗粒度上的迅速改进，而不是依赖于传统自动化数据工具的人工分析。

这本书将数据挖掘（尤其是机器学习）与网络相融合。数据挖掘与分析扮演着一辆汽车的角色，通过决策智能的隧道开往洞察决策的终点。本书致力于缩小网络与业务的商业问题与运营商形成可执行决策之间的鸿沟。本书的作者融合了通信与数据科学领域的知识，用实证研究来分析电信运营中的各种典型问题。

在网络领域，数据驱动的网络分析已经开始针对网络全生命周期赋能，包括网络规划、部署、优化和维护。本书介绍了基于统计学、数据挖掘和机器学习等技术的数据分析，以量化分析的形式阐述如何更好地规划、优化和运营现代移动通信网络。相比传统的方法，一套数据分析方法集体现了更好的准确性、稳定性、健壮性，从而保证电信运营商的网络运维的服务质量可以维持在优秀的水平，并体现了更高的精细度水准。由此，

运营商可以给用户带来更好的体验，并使自己的运营、管理和维护工作得到显著的效率提升。

在业务领域，本书详细地介绍了商业智能，商业智能主要用来解决电信市场、客户关系管理、客户服务等领域的商业问题。本书介绍了一套数据分析方法集，用于解决不同的电信业务问题，例如推测用户的离网流失，评估终端质量，用户行为画像，分析用户体验感知等。

作为通信数据科学领域的一名老兵，我见证并经历了全世界通信领域的大数据分析在过去 20 年的发展。本书是一本很及时且关键的里程碑式的著作，系统总结了先进的数据分析技术如何赋能通信业的网络和业务两个领域的成果。欧阳晔博士不仅是我学术上紧密的合作者，也是美国威瑞森电信的 Fellow 和通信人工智能系统部经理。我相信他在威瑞森电信的通信数据科学的经验，会对通信业同人们运用数据科学对移动通信技术演进持续赋能提供很大的帮助。

本书适合通信行业的数据科学家、数据工程师、业务专家和管理者以及电信管理、数据科学、电子工程、计算机工程专业的研究生阅读和学习。

大卫·贝兰格　博士

AT&T（美国电话电报）公司首席科学家，AT&T 香农研究院副总裁

美国斯蒂文斯理工学院教授

目 录

第 1 章 概 述

第 2 章 电信分析方法论

第 3 章　LTE 网络性能趋势分析

第 4 章　热门设备就绪和返修率分析

第 5 章　VoLTE 语音质量评估

第 6 章 移动 APP 无线资源使用分析

第 7 章 电信数据的异常检测

第 8 章 基于大数据分析的 LTE 网络自优化

第 9 章 电信数据和市场营销

第 10 章 传染式客户流失

第 11 章 基于社交网络的精准营销

第 12 章 社交影响和动态社交网络结构

第 1 章
概　述

1.1　电信业大数据分析

电信生态系统是一个天然的大数据仓库，对于那些知道如何挖掘它的人来说，它就是一个智慧宝库。然而，不能简单认为“大数据”就是集合大量的数据，这是因为电信大数据分析不单纯是数据库问题，而是一个如何理解电信数据的问题。得益于网络的演进和智能手机数量的暴增，电信运营商（CSPs）可以获取海量的用户、网络和应用数据，这些数据是极具价值的信息资产。同样，获益于大数据分析的强大能力，电信运营商才能对网络模式和消费者行为有更深刻的洞察。

大数据最初为电信业而生。当谈到大数据时，电信业由于在日常业务过程中所采集数据的绝对广度和深度而具有其独特的优势。电信运营商每天都处于大数据世界中，大数据已经成为电信业无处不在的一部分，每秒都有大量的数据通过互联互通产生，如用户发起语音、视频或数据呼叫，发送短信，上网等。

近年来，电信业的数据呈指数级增长。智能手机、移动宽带、物联网和 5G 网络等带来了海量数据，同时也给电信网络生态系统带来了众多且不可预测的变化，如更多的信令流量、新应用的并发连接及每个数据应用连接所消耗的数据流量的变化，这些变化带来的结果是数据使用量的大幅增加和带宽消耗的爆炸式增长 [1]。

集成电路技术的发展速度最先被英特尔的戈登 • E. 摩尔注意到，被称为摩尔定律，摩尔定律似乎在分析用户在使用电信网络中产生和传输的数据时更加适用。近年来，全

球 4G 用户呈指数级增长。截至 2017 年第一季度，全球 LTE 用户总量达到 21 亿 [2]，这种惊人的增长推动了网络流量的快速增长。2016 年全球移动数据流量增长了 63%。全球移动数据流量从 2015 年年底的每月 4.4EB 上升到 2016 年年底每月 7.2EB。全球移动数据流量在过去 5 年中增长了 18 倍 [3]。在 5G 时代，物联网通信产生的数据量预计将超过人类产生的数据，预计 2020 年将有 320 亿台设备产生 44 万亿 GB 的数据 [4]。

随着终端设备、网络、应用和服务产生的数据呈滚雪球效应式增长，电信数据分析对于电信运营商真正了解网络、客户、业务和行业本身变得至关重要。大多数运营商积极利用数据分析来提高网络效率、客户分群和提高盈利能力，并取得了一定的成功。多年来，电信运营商实际上已经使用了各种技术手段来处理这些数据，包括统计分析、数据挖掘、知识管理和商业智能。随着海量数据被准确地捕获并加以专业分析，这些数据将有助于我们洞悉事物的本质，从而提高内部效率。展望未来，电信业大数据分析面临着更严峻的挑战：如何获得更深入的理解、洞察事物内在本质、模式和关联关系、从大量数据中提炼出有意义的信息，最终采取富有洞察力的行动来增加整个电信价值链的收入和利润（从网络运营到产品开发再到营销、销售和客户服务，甚至数据变现）[5]。电信运营商甚至可以利用这些深入洞察的数据来帮助诸如农业、电力公共事业和医疗保健等其他行业。

电信运营商处于电信大数据领域的核心，拥有数据“金矿”，这些数据使它们能够以较高的水平来理解其网络、服务和用户。近年来，电信运营商受到非传统竞争对手的挤压、降低成本的压力，以及客户忠诚度变化和动态技术环境的冲击，面临诸如 Google 和 Facebook 等 OTT 互联网公司蚕食收入的竞争局面，而大数据则为电信运营商提供了一个独特的利器，运营商需要更好地利用大数据技术使自身更具竞争力，扭转近年来收入和利润下降的不利局面 [6]。

1.2 电信大数据分析的驱动力

在 CT 与 IT 不断融合的背景下，整个电信业的外部环境日益严峻，体现在运营商之间竞争惨烈、网络中立法规的不利影响、OTT“玩家”直接威胁和技术变革等各个方面。电信运营商需要走出“舒适区”，避免被非传统“玩家”击败。因此，在电信业中充分

利用大数据来增加自身竞争力的驱动力是显而易见的。

第一，电信业处于竞争日益激烈的严峻环境中，激烈的市场竞争导致利润和每用户平均收入（ARPU）的下降。是否可以通过电信数据分析建立一个新的商业模式来扭转这种下降趋势？通过分析，实现电信数据变现是一种选择。

第二，像 Facebook、Google、Snapchat、Netflix 等这样的 OTT“玩家”不花一分钱就可以使用运营商的网络。在运营商的免费网络之上，OTT“玩家”向客户提供语音、数据和内容服务。很明显，这直接影响了电信运营商的收入。网络中立政策使得 OTT 公司可以在电信运营商投巨资（CAPEX/OPEX）打造的信息高速公路上免费“行驶”，导致运营商在传统语音和数据服务上的 APRU 值持续下降。

第三，电信公司需要在信息技术和电信（ICT）融合的过程中跟上技术变革不断前进的步伐。与仅仅拥有应用层数据的 OTT“玩家”相比，电信运营商的独特优势是拥有从物理层到应用层的全栈数据。有了正确的分析解决方案和产品，电信运营商将对整个 ICT 行业有更全面的认知，这将帮助运营商扭转在与 OTT“玩家”的竞争中所处的不利局面。

第四，受上次全球经济衰退的影响，电信业也不例外地跌入低谷。电信运营商都面临着提高运营效率和降低成本的巨大压力，同时还要将服务质量保持在最佳水平。电信分析的目标集中在两个方面：提高内部业务效率和利用新的商业模式实现数据变现。大多数电信运营商已经开始利用对其内部数据的分析结果来提升其网络、设备、服务、应用、客户和运营效率，并取得了一些成功。然而，大数据潜力比想象中的还要诱人：需要将数据分析变现扩展到整个电信价值链，从网络运营到新产品开发、市场营销和销售、客户关怀，甚至将电信数据变现扩展到其他行业。

1.3　大数据分析对电信产业价值链的益处

电信运营商大数据分析是洞察的载体。电信业数据分析同时涉及数据科学和电信领域的分析方法、分析工具和分析技术。所有这些需要在数据之间建立关联、识别趋势和模式并预测结果。大数据分析为电信业发掘数据宝藏奠定了基础。

例如，如果电信运营商希望改善其零售商店的客户服务，它可以通过大量客户触点

收集客户的感受数据。对这些数据的后续分析将帮助电信运营商深入洞悉这个客户，如客户更喜欢的服务、服务的使用频率、总体品牌感受等。这些分析结果极具价值，但在深入探讨电信运营商如何优化其服务以满足传统市场和新兴市场的需求时，这些数据分析结果仅仅是一个出发点。因此，尽管数据分析显然有助于更深入地了解客户的统计资料和感受，但对这些发现进行更加仔细的洞察，以得到重要的深刻见解，是电信运营商的重要工作。

那么，电信业能从大数据中获得哪些收益？理解这点非常关键。

第一，必须更好地了解网络。网络分析帮助移动运营商（MNOs）更好地利用内部的网络信息，使网络运行更可靠、更健壮、更具可扩展性。网络分析帮助电信运营商在网络的整个生命周期中受益：网络规划、网络部署和网络维护（优化）。在网络规划阶段，分析首先为未来的网络需求做好准备。网络规划分析有助于电信运营商在网络估算时了解网络流量的未来需求，可以提前精准规划新的网络基础设施或网络扩容的投资性支出（CAPEX）。在网络部署和优化阶段，电信运营商可以通过分析诊断方法，充分利用网络分析来优化其网络性能和质量。

第二，更好地了解客户。大数据的能力使得通过网络、设备、应用、社交媒体等方面获取的数据信息更容易地了解客户的概况、行为和模式，有助于进一步建立以用户为中心的指标体系，了解用户体验质量。

第三，更好地理解应用程序。在运营商网络之上运行的各种互联网应用程序给无线网络带来了许多不可预测的变化，如更大的信令流量、新应用程序的并发连接及每个数据应用程序连接所消耗的数据流量的变化。电信运营商可以利用分析来更好地了解应用程序如何影响自身的网络和服务，并相应地深入了解应用程序模式和消费者行为。

1.4 电信大数据的实现范围

与电信业传统的数据仓库和数据库技术相比，大数据分析可以为未来的网络需求做好准备，也可以了解客户体验的质量[7]。特别地，大数据可帮助移动运营商利用其网络中的潜在信息和数据，使其网络更健壮、更优化和更具可扩展性。凭借实时计算能力，大数据通过实时分析网络流量或模式来帮助优化路径和服务质量。大数据使得从网络数

据或社交媒体信息中详细地了解客户变得更容易，有助于建立以用户为中心的 KPI 体系来更好地理解用户体验。在本章中，我们将介绍电信业中所使用的大数据技术、用例、最新的研究成果和面临的挑战、在网客户和市场的分析以及商业模型。

1.4.1　网络分析

移动运营商需要通过网络可视化来了解网络如何服务其内部管理和外部客户。移动通信网络中的基站（eNB）数据采集故障可能导致服务降级或服务中断，更换设备通常比维修更昂贵，所以维护工作既不能太早，也不能太晚。当下，电信网络正从传统的硬件和以设备为中心的部署向基于云的部署过渡。网络功能虚拟化（Network Function Visualization，NFV）或软件定义网络（Software Defined Networking，SDN）[8,9] 为所有网络功能组件中较重要的组件，这两者的目的都是虚拟化网络应用程序和网络连接。大数据分析工具保存网络中的非结构化、流式和传感器数据。在当前的大数据工具中，Hadoop 或 Spark 平台存储和处理来自网络的非结构化、流式和传感器数据。移动运营商通过将实时信息与历史数据进行比较，得出最优的维护计划。通过 Spark 或其他机器学习库提供的 MLlib 或 ML（高级 API），算法可以帮助移动运营商分析它们的网络，在设备损坏之前进行修复以减少维护成本和防止服务中断。

1．通话中断分析

移动运营商在扩展其宽带服务的同时需要聚焦与提高其网络性能 [10]，因为网络故障或网络中断会导致通话中断和语音质量下降。此类事件会损害电信运营商的声誉，也会增加其客户流失。因此，移动运营商应持续监控其网络以防止此类故障，要尽早从根本上解决问题。不满意的客户可能不会频繁投诉通话中断，但这些客户流失的可能性会加大，他们可能会转寻其他能提供更好的服务 / 信号覆盖的运营商。

为了解决这些问题，移动运营商可以分析用户产生的通话详单（CDR）数据，并与相应的时间段的网络设备日志进行关联，然后对通话中断原因进行分类。在 Hadoop 大数据平台中，Flume 是处理数据导入的工具，能够将数百万条通话详单注入 Hadoop 中。在实时机制中，Apache Storm 通过模式识别算法来发现这些数据中的各种故障模式。

2. 异常检测

在无线网络中，异常是偏离正常网络行为的异常流量模式[11]。数据挖掘和机器学习中，异常被称为不正常、偏差或极端值。无线网络中的异常可以由各种因素引起，如新特性的实现、网络入侵或灾难事件。在许多情况下，例如入侵事件的异常值仅仅是多个数据点的序列，而不是单个数据点。近年来，网络监控设备的容量越来越大，能够以较高的采样率采集数据。

利用大数据平台，精心设计的异常检测系统可以帮助我们从大量噪声数据中提取有用信息。在大多数应用程序中，收集的数据由多个进程生成，即共现数据。共现数据通常是两组基本观测联合出现：一组的流量数据（观测值— W）与另一组中的生成实体（时间戳或节点 ID-D）相关联。对共现数据建模（具有生成实体的流量数据）是异常检测中的基本问题。当一般分布随生成实体（时隙或节点 ID）变化时，有效的异常检测会识别出这种变化。

利用像 Spark 中的 MLlib 等机器学习库，可以有效地检测和识别网络行为模式或网络异常值。

3. 网络性能健康度

传统网络优化的工作流程遵循一些常规步骤。网络系统性能工程师通常先从运营支持系统（Operation Support System，OSS）工具中提取 KPI 统计数据、观察原始数据、可视化 KPI 趋势，利用该领域知识或一些人为制定的规则（如 KPI 阈值）来查找异常模式、异常及致命的 KPI。当锁定问题后，工程师需要从服务降级、覆盖 / 容量黑洞、巨额流量用户、容量瓶颈等方面确定该问题的根本原因。同时，工程师还要检查网络修复工单，以验证性能方面存在的问题。完成上述步骤后，工程师根据收集和分析的所有信息，通过一些网络优化工具，结合自己的领域知识、经验和一些半自动化的解决方案，最终形成解决方案。

显然，工程师们每天都盯着成千上万的 KPI 来评估网络性能，这不是一个明智的办法。应该采用一种类似于人工智能的科学方法，如树或类似神经元的模型，自上而下以分治的方式过滤掉噪声数据和不太重要的信息，使工程师能专注于关键的 KPI。通过这种方式，将工程师们从烦琐的目视任务中解脱出来，更专注于性能诊断和优化，这是网络优化中最有价值的一步。网络性能健康可以在不同级别定义和计算网络性能健康度

（Network Performance Healthiness，NPH）以评估和可视化网络性能，如蜂窝小区、eNodeB 或在大数据平台上预定义的 Geo-bin。

4．智能网络规划

移动运营商需要基于高级分析所得到的网络规划解决方案来联合和关联来自不同网络数据库的信息帮助运营商进行网络投资规划、投资预测和投资优化。网络规划系统必须采用先进的分析手段，并且与 OSS 系统密切配合。两个系统结合可以促进容量优化，并且为网络规划工程师提供“假设”场景的方案预演能力。

5．基站优化

4G 和未来的 5G 网络旨在实现在自组织网络（Self-Organizing Network，SON）中定义的功能。SON 最重要的功能之一是自我优化，包括蜂窝小区自动管理它们之间的交互方式、管理它们的功耗，以及它们如何均衡流量负载和切换小区间的流量。这些功能的实现取决于移动运营商是否可以利用上下文信息来增强网络性能。这些功能包含用户信息，如特定领域的用户体验，以及对应的不同类型的服务和用户行为模式所产生的体验差异。

6．以用户为中心的无线分流

应用程序可以从远程基站监控系统、DPI 系统、话单系统、回传网络管理系统等采集大量的数据，大数据和机器学习技术用来处理和分析这些大批量的数据。根据用户的订购级别、正在使用的应用程序及不同类型的蜂窝基站的流量负载，这些数据被实时地分流至不同的蜂窝小区。

在 4G 网络中，Wi-Fi 分流被普遍使用。语义分析工具可以将用户信息与他们的在网价值相关联以智能地决定哪个用户应该被分流到 Wi-Fi。

7．拥塞控制

无线接入网络（Radio Access Network，RAN）拥塞是移动运营商面临的主要问题之一。将用户信息、订购服务、位置信息结合起来，可以实现单个子蜂窝级别的可视化。由于拥塞事件持续时间较短，利用大数据分析来发现问题并提前做好预案对运营商来说非常关键。

1.4.2 用户与市场分析

1. 客户流失预测

在移动通信业，留住客户是最重要的挑战之一。流失预测是指对预测处于离网风险中的客户进行预测。获得新客户比留住老客户需要更大的成本。

借助预测模型和机器学习算法，我们能够精确地识别可能会流失的客户。基于所采集的关于用户使用、投诉、交易、社交媒体等数据，算法会创建权重因素来识别客户是否正在离网。

2. 用户画像

用户画像是将市场或客户基于他们之间行为的相似性划分成不同组的过程。这种方法在运营商客户数量日益增长时颇为流行，是运营商做出战略决策的关键组成部分。例如，运营商可以基于客户分组为客户量身定做产品、识别高价值和长期客户、发掘潜在的客户。

通过用户画像，运营商可以识别高价值的忠诚客户，实现有针对性的营销和客户维系活动，以降低客户流失率。更广泛的客户细分根据客户需求为每个细分市场提供适合的产品，从而提升客户满意度。通过大数据技术，运营商能够根据汇集到的客户数据和使用历史来进行更有效的客户细分，以开展更有针对性的营销活动。

3. 预测式营销和抢先式客户关怀

在高度竞争的环境中，移动运营商面临的主要挑战是客户维系和从客户获得收入。实时分析技术可帮助运营商主动分析、关联和洞察数据，破解客户流失和收入损失难题。对消费者数据进行实时分析可以洞悉客户的购买模式。这些模式是高度个性化的，需要对他们的购买行为迅速做出响应，让客户明白他们究竟想要购买什么、应该在什么时候购买。与此同时，企业能够获得实时数据并加以分析，以便未来产品销售更具针对性。

通过大数据分析技术，运营商能够获取大量的营销活动工具。这些工具具备数据管理、营销活动管理和性能监控等功能，可用于处理需要筛选的海量数据。

4. 位置服务

根据位置信息，运营商可以更深入地了解用户，这些基于地图的可视化信息可以用

于许多分析服务中。除了位置服务，定位技术也可以替代 Wi-Fi 位置信息，从而为用户提供更好的服务。

1.4.3　创新的商业模式

1．数据开放和 API 使能

对于当下绝大多数的应用，应用程序接口（Application Programming Interface，API）让运营商能够更好地实现内外部数据的交换。通过精心设计的 API，应用开发工程师可以将客户链接到各种新的应用上。

2．使用支付数据增加销售

移动运营商可采集实地交易数据，并提供给商业顾客。通过此功能，可采集并分析客户的支付和交易数据，基于客户的喜好向客户推送个性化的电子优惠券和促销信息。

3．场景化的供需匹配

运营商可以根据场景满足用户需求并推荐相关产品，比如用户通过手机在商场、购物中心和超市周围寻找他们感兴趣的产品。运营商可以与其商业合作伙伴共享这些信息，并提供潜在的用户群。该功能可以帮助合作伙伴销售新产品，将特定的营销活动推送给特定的用户群，为买家和商家双方创造一个更稳定、更高效的供需市场。

1.5　本书概要

第 1 章：概述，全面阐述电信业大数据分析技术。

第 2 章：电信分析方法论，涵盖可用于电信业分析的机器学习算法，介绍回归方法、分类方法、聚类方法、预测方法、ARIMA 模型和强化学习。

第 3 章：LTE 网络性能趋势分析，介绍网络性能分析的过程，如网络性能预测策略、网络资源与性能指标之间的关系、网络资源预测及评估 RRC 连接设置建立的应用。

第 4 章：热门设备就绪和返修率分析，介绍设备退修率和设备就绪的预测策略、模型和实现结果。

第 5 章：VoLTE 语音质量评估，介绍电信网络 VoLTE 语音质量的定义、方法和试验结果。

第 6 章：移动 APP 无线资源使用分析，展示移动资源管理和使用的工具、算法和试验结果。

第 7 章：电信数据的异常检测，介绍异常值识别模型及其在电信业的比较。

第 8 章：LTE 网络自优化，重点介绍 SON（自组织网络）及其实现，APP-SON 是大数据平台上 4G 和未来 5G 网络的自我优化解决方案。

第 9 章：电信数据和市场营销，介绍电信营销、社交网络和网络测量。

第 10 章：传染式客户流失，主要研究电信业的客户流失问题及社交学习和网络效应的动态模型。

第 11 章：基于社交网络的精准营销，介绍网络效应的渠道、建模策略问题及它们的发现与应用。

第 12 章：社交影响和动态社交网络结构，涵盖网络结构对社交影响多元分析的模型。

参 考 文 献

[1] Big Data & Advanced Analytics in Telecom: A Multi-Billion-Dollar Revenue Opportunity. https://www.huawei.com/ilink/en/download/HW_323807.

[2] Ericsson Mobility Report June 2017.

[3] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021 White Paper.

[4] Digital Cosmos to Include 32 Billion Devices Generating 44 Trillion GB of Data by 2020. http://www.industrytap.com/digital-cosmos-include-32-billion-devices-generating-44-trilliongb-data-2020/28791.

[5] Benefiting from big data: A new approach for the telecom industry. https://www.strategyand.pwc.com/reports/benefiting-big-data.

[6] Analytics: Real-world use of big data in telecommunications. https://www-935.ibm.com/services/us/gbs/thoughtleadership/big-data-telecom/.

[7] Hilbert, Martin. “Big data for development: A review of promises and challenges.” Development Policy Review 34.1(2016): 135–174.

[8] Virtualization, Network Functions. “NETWORk FUNCTION VIRTUALIzATION.”

[9] Han, Bo, et al. “Network function virtualization: Challenges and opportunities for innovations.” IEEE Communications Magazine 53.2(2015): 90–97.

[10] Lee, Jonghun. “Method and system for preventing call drop by restricting overhead message updated in 1X system during 1xEV-DO traffic state.” U.S. Patent No. 7, 394, 787. 1 Jul. 2008.

[11] Garcia-Teodoro, Pedro, et al. “Anomaly-based network intrusion detection: Techniques, systems and challenges.” computers & security 28.1–2(2009): 18–28.

第 2 章
电信分析方法论

过去的几年里，我们见证了移动运营商数据的迅猛增长，网络侧的数据流量也大幅增加。除了存储和管理这些流量数据外，另一个主要的挑战是如何选择和使用这些大量数据来更好地认识网络。实际上，由于不可能对这些网络流量数据进行手动处理和分析，简单的统计汇总不足以呈现数据蕴含的全部信息，这就需要有新的策略来管理和理解这些数据。在过去几十年中，人们在机器学习算法的基础上创建了一系列工具。开发这些工具的目的是提供高级分析，以发现和解释数据中的复杂模式。这些算法通常分为两类：①监督学习，指的是从已标记的训练集中预测输出或分类对象的技术；②无监督学习，指的是描述或分割对象以推断数据的隐藏结构的技术。本章介绍了一系列常用的机器学习算法。选择这些算法，是因为它们在电信领域有十分重要的作用。在接下来几章的方法论部分会介绍这些算法，其专门用于网络分析、评价或检测目的。

本章包括 4 个主要部分：回归方法、分类方法、聚类方法和预测方法。对于每个部分，介绍了分析算法，并将重点放在指导思想的理解上。详细内容可以在文献 [1]（回归、分类、聚类）和文献 [2]（预测）中找到。

2.1 回归方法

回归方法是用于估计预测变量与连续目标变量之间关系的监督学习方法。最常见的用法是，回归分析在给定自变量的前提下，估计因变量的条件期望。不太常见的用法是，在给定自变量前提下，关注因变量条件分布的分位数或其他位置参数。在所有情况下，

需要估计被称作回归函数的自变量函数。在回归分析中，针对回归函数的预测，使用概率分布描述因变量变化特征吸引了大家的注意。回归分析广泛用于预测和预报，在这些领域它的使用与机器学习领域有很大的重叠。回归分析还用于理解自变量中哪些与因变量相关，并探索这些关系的形式。在受限制的情况下，回归分析可用于推断自变量和因变量之间的因果关系。

在电信领域，有些变量可能会比其他变量更容易收集，因为未知或不受控制的内部或外部因素，使得收集的变量和关键性能指标（KPI）之间的关系可能仍然不明确。在这种情况下，我们可以通过概率模型来具体化预测变量和目标变量之间的关系，从而理解和控制其中的不确定性。这些方法有助于为一组新的预测变量预测目标变量的值，并建立预测的置信度区间。在本节中，我们将简要介绍常用的回归方法，给出关键点以便于理解后续章节。最简单的回归方法是线性回归，第 2.1.1 节给出了定义（线性回归的简要历史，请参阅文献 [3]），2.1.2 节介绍 LOESS[4]（代表局部回归）和广义可加模型 [5]（GAM），给出了一种灵活的方法来推导预测变量与目标变量之间的非线性关系。最后，通过 Lasso 回归 [6]，2.1.3 节引入了通过自动选取特征实现正则化的概念。

如果目标变量是分类的，则我们改用监督分类方法（如 2.2 节所示）研究预测变量与目标变量之间的关系。

2.1.1　线性回归

线性回归是一种线性方法，用于建模标量因变量 y 和一个或多个表示为 X 的解释变量（或自变量）之间的关系。只有一个解释变量的情况称为简单线性回归。对于多个解释变量的情况，该过程称为多元线性回归。线性回归是第一种被严谨研究的回归分析类型，并在实际应用中广泛使用。这是因为与未知参数非线性相关的模型相比，未知参数线性相关的模型更容易拟合，结果估计量的统计特性也更容易确定。

在线性回归中，使用线性函数对关系进行建模，其中未知模型参数通过数据估计得到这种模型称为线性模型 [3]。最常见的情况是，假设给定 X 值的 y 的条件均值是 X 的仿射函数；不太常见的情况是，给定 X 值的 y 的条件分布的中值或其他分位数表示为 X 的线性函数。与所有形式的回归分析一样，线性回归侧重于给定 X 时 y 的条件概率分布，而不是多元分析领域中 y 和 X 间的联合概率分布。在经典线性回归中，假设目标变量 y

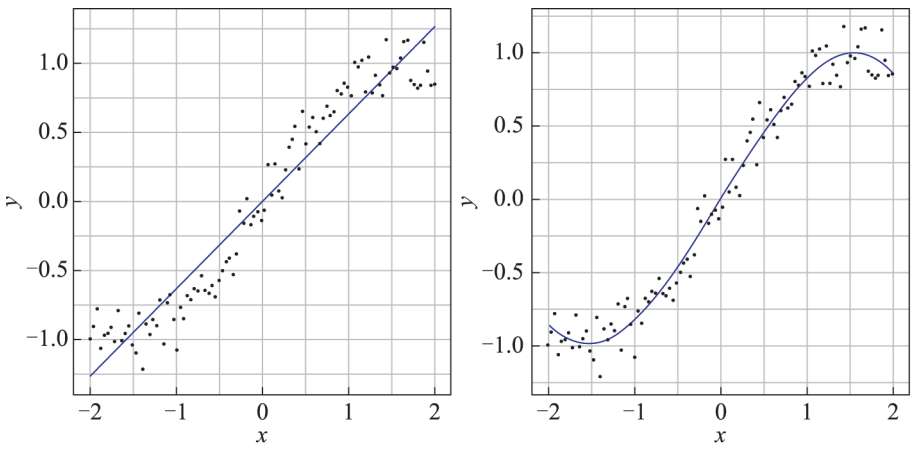
与预测量 x_1，…，x_l 之间的关系是线性的。在最经典的形式中，线性回归采用下述等式：

$$y=\beta_0+\beta_1 x_1+\cdots+\beta_l x_l+\varepsilon$$

其中预测量是固定的已知值，β_0，…，β_l 是要拟合的固定未知参数，ε 是服从高斯分布的随机变量，其均值为 0，方差为 σ^2。此外，假设对于所有的数据实例对 i 和 j，相关随机变量 ε_i 和 ε_j 是独立的，响应变量 y 因此服从高斯分布。

计算拟合参数的常用算法是最小二乘法，可以直接通过矩阵计算得到这些系数（详见文献 [7]）。在这种回归方法中，所有的预测量都对最终的预测起着积极或消极的作用。图 2.1（a）是用一个预测量进行线性拟合的例子。

如图 2.1（a）所示，直接线性回归不足以精确拟合预测量和目标变量之间的复杂关系。多项式回归是线性回归的一种直接扩展，它能在固定阶数 D 下拟合多项式。我们不直接取 x_1，…，x_l，而是首先计算 d 从 1 到 D 的幂值 x_1^d，…，x_l^d，从而得到 D_1 特征。该模型具有 D_1+1 个拟合系数（包含 D_1 个特征值加上常数系数），而不是经典线性模型中的 l+1 个拟合系数。使用前述最小二乘法进行拟合。图 2.1（b）给出了一个阶数为 3 的多项式拟合例子。



（a）使用线性回归。我们观察到，拟合可以获得x和y间关系的主要趋势，然而，这种算法无法获得非线性变化

（b）使用3阶多项式回归。对于这种拟合，虽然我们使用多项式拟合正弦函数，但仍会获取其整体趋势。我们注意到，用这种拟合进行推断最终会导致非常差的结果

图 2.1　用两种回归算法推导 x 和 y 之间的关系

在区间 [−2，2] 上采样 100 个点 x。x 和 y 之间的关系由 y=sin（x）+ε 定义，其中每个 ε 服从方差 σ^2=1/100 的中心高斯分布。

2.1.2　非线性回归

非线性回归是回归分析的一类，其中观测数据由模型参数的非线性组合的函数建模，该函数依赖一个或多个自变量。通过连续近似的方法拟合数据。

非线性回归方法在拟合中有更大的灵活性，其代价是缺乏对基本模型的理解。这些算法性能非常高，但如果没有选择正确的超参数，就可能导致过拟合（有关性能和过拟合的详细介绍，请参阅文献 [8] 第 5 章）。

最常见的非线性回归是局部加权回归 LOESS[4]，这也是一种非参数回归。该方法的思想是对每个兴趣点进行局部回归，可以描述如下：为了拟合给定预测量的新点，选取邻近该点的数据子集。然后，对子集进行低阶多项式拟合（通过赋予邻近兴趣点的观测值更大的权重）。对兴趣点的拟合就定义为该点的 LOESS 拟合。我们可以看到，必须选择两个超参数：多项式拟合的阶数和所选总数据集的百分比。通常情况下，如果将拟合的阶数设置为 1 或 2，一个更高的数字将会导致过拟合和结果不稳定的问题；可以改变选择的总数据集的百分比，使拟合更平滑或更不平滑（越高越平滑）。

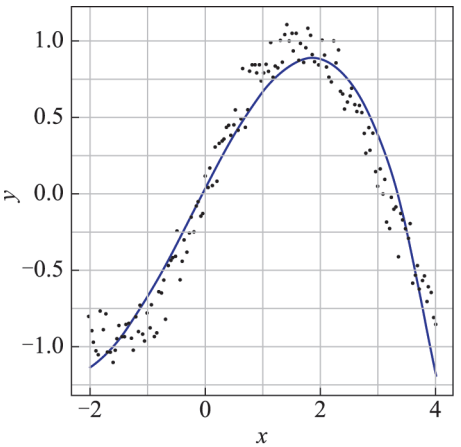
图 2.2 给出了 LOESS 与固定参数化模型相比较的有趣例子。在图 2.2（a）中，进行了 3 阶多项式回归，但回归不能灵活地提取曲线的整体行为（这可以在 x=1.5 左右显示，其中拟合曲线位于散点图下方）；在图 2.2（b）中，LOESS 回归可以正确拟合散点图，并可以连续预测区间 [−2，4] 内的新点。

另一种实现灵活非线性回归的更为复杂的方法是 GAM[5, 9]。与 LEOSS 相比，只要在数据集中有足够的观测量（如 1000 个元素），GAM 的拟合效果就会更好。在 GAM 中，目标变量 y 与预测量 x_1，…，x_l 的关系与以下等式相关联：

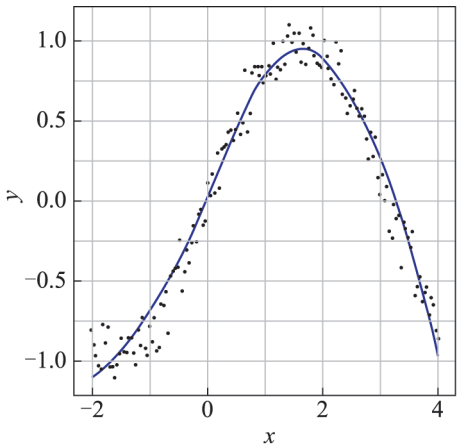
$$y=\beta_0+f_1(x_1)+\cdots+f_l(x_l)+\varepsilon$$

其中，f_1，…，f_l 表示输入变量之间的非线性关系，β_0 是常数项，ε 服从高斯分布。可以用非参数反向拟合算法来估计函数 f_j。该算法在每一步中迭代，并用三次样条来近似函数 f_j。

非线性函数的其他例子包括指数函数、对数函数、三角函数、幂函数、高斯函数和洛伦兹曲线等。



（a）使用3阶多项式回归，在x=1.5左右，拟合曲线位于散点图下方



（b）使用LOESS回归（在*R*中用ggplot2包自动拟合）。我们可以观察到拟合线正确停留在散点图的中间，并且可以连续地预测[−2,4]中的新点

图 2.2　用两种回归算法推导 x 和 y 间的关系

在区间 [−2，4] 中采样 150 个点。x 和 y 之间的关系与图 2.1 中描述的关系相似。

2.1.3　特征选择

特征选择也称为特征工程，它是使用数据的领域知识，生成用于机器学习算法的特征值的过程。特征工程是机器学习应用的基础，既有难度且成本很高。可采用自动特征学习来消除人工特征工程。在图 2.1 和图 2.2 中，只选择了一个特征来解释目标变量。在实际应用中，电信行业收集到的特征数量很大（数十个或数百个变量），在这种情况下会出现一个称为维度灾难的新问题：回归算法可以连续地拟合一个固定集合，但不能推广到新的未知数据。我们称在这种情况下使用这种算法会导致过拟合。

通常，精确拟合目标变量所需的特征数量要低于可用的特征数量，因为某些变量可能与目标变量无关，或者某些预测量之间可能是相关的。

与人工选择特征不同，一种称为正则化的方法可以帮助加权，甚至选择感兴趣的特征，从而有效地解释目标变量。更简单的正则化方法实现是约束某些系数（称为收缩），如果系数过大就加以惩罚约束。

在线性回归的情况下，约束涉及 $\beta_0, \cdots, \beta_l$ 系数。最小绝对收缩和选择算子[6]（LASSO）回归对系数的绝对和（常数系数除外）施加了一个条件，如下式所示：

$$\sum_{k=1}^{1}|\beta_k|\leqslant t$$

其中，t 是要选择的参数。理论分析显示，这是对系数的一个硬约束，这意味着它们中的某些值可能完全等于零。因此，该方法能够对预测量进行特征选择。岭回归 [10] 与 LASSO 回归有一些相似之处，但具有常规约束函数（通过约束平方和而不是绝对值）：

$$\sum_{k=1}^{1}|\beta_k^2|\leqslant t$$

岭回归能够收缩系数，并且广泛用于正则化拟合。最后，弹性网正则回归 [11] 是岭回归和 LASSO 回归之间的折中方法，给出了另一种收缩系数的方法。约束函数是前面两个正则化的线性组合：

$$\alpha\sum_{k=1}^{1}\beta_k^2+(1-\alpha)\sum_{k=1}^{1}|\beta_k|\leqslant t$$

还存在选择特征的其他方法，例如，通过迭代选择能够解释目标变量的特征（参考文献 [12] 中的子集选择，参考文献 [13] 中的单变量滤波方法）。

除了上述特征工程方法，另一种有效的方法是主成分分析（Principal Component Analysis，PCA）。PCA 是为降低高维数据集中相互关联特征的维数而设计。PCA 通过计算特征间的相关性，将数据从高维特征空间映射到低维特征空间。映射后，所有数据点都可以用低维特征空间中的主正交分量表示。在所有有序分量中，第一分量被认为是保留原始特征最大信息的分量。在低维特征空间中，第一个分量位于第一个坐标轴上，第二个分量位于第二个坐标轴上，以此类推 [14]。如矩阵 $\boldsymbol{X}$ 中的每一行代表一个数据点，矩阵中的每一列代表不同的特征。PCA 变换由一组带权重的 p 维向量定义：

$$\boldsymbol{W}_{(k)}=\left(\overline{\boldsymbol{w}}_{(1)},\overline{\boldsymbol{w}}_{(2)},\cdots,\overline{\boldsymbol{w}}_{(p)}\right)_{(k)}$$

$\boldsymbol{X}$ 中的 p 维向量可以映射到具有主成分分值的新向量上：

$$\boldsymbol{t}_{(i)}=\left(\boldsymbol{t}_{(1)},\boldsymbol{t}_{(2)},\cdots,\boldsymbol{t}_{(m)}\right)_{(i)}$$

其中，$\boldsymbol{t}_{(k,i)}=x_{(i)}\cdot\boldsymbol{w}_{(k)}$，其中 i=1,⋯,n，k=1,⋯,m。t 为 x 的最大可能方差。

利用 PCA，我们提取出的分量可以提高训练过程的效率，降低计算复杂度。在该算法中，特征变换的过程并不是简单地丢弃特征，每个主分量都是根据原始特征计算出的组合结果。

2.2 分类方法

在某些应用中，与 2.1 节不同的是，我们可能想研究预测量 $x_1,\cdots,x_l$ 和某个分类目标变量 y。由于目标变量是分类的，回归方法不能正确地拟合和预测该目标变量，因此为这类数据开发了新工具，称为分类方法。在这种分类方法中，我们先重点讨论 0-1 分类：我们假设目标变量只能取值 1（如对应一次成功）或 0（如对应一次失败）。然后，分类任务必须理解预测量是如何与目标变量有关联的，并且通常输出成功的概率。

本节主要介绍逻辑回归 [15, 16]（2.2.1 节）。这种基本方法可以看作线性回归模型的延伸，尽管很简单但十分常用。还有许多其他的分类算法，用于处理非线性关系。在 2.2.2 节中将给出主要算法的简单描述。

2.2.1 逻辑回归

逻辑回归在某种程度上命名不是很精确，因为它是一种分类方法而不是回归方法。该方法通常专用于解决二元分类问题。作为一个线性分类器，它不能捕捉复杂的非线性模式。此外，它对预测变量内的相关性很敏感。因此，使用时必须检查相关性，以避免某些变量过拟合和过度置信。逻辑回归的主要优点是运行速度快，线性模型识别可靠性高。此外，作为白盒模型，每个特征对目标变量的影响都易于理解。

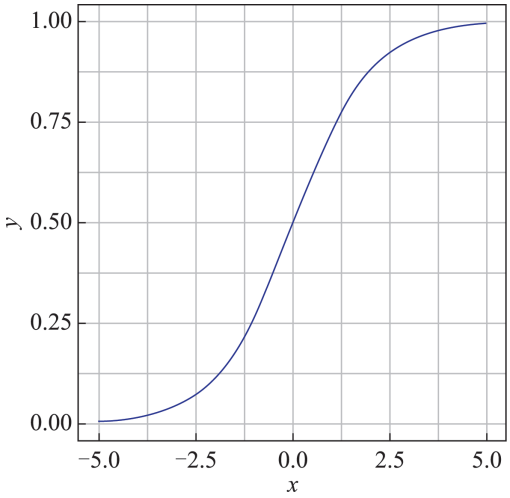


图 2.3 逻辑模型 $x \mapsto \frac{1}{1+\exp(-x)}$

与线性回归输出在实线上取值不同的是，逻辑回归约束输出值在 0 到 1。这样，拟合的结果可以看作目标值等于 1 的概率。

将输出值从实线约束到（0,1）的方法是使用 Sigmoid 函数做映射，定义如左（见图 2.3）：

$$\sigma(x)=\frac{1}{1+\exp(-x)}$$

因此，逻辑模型可以写成：

$$p(y=1\,|\,x,\beta_0,\cdots,\ \beta_d)=\sigma(\beta_0+\beta_1x_1+\cdots+\beta_lx_l)$$

在这个模型中，$\beta_0,\cdots,\beta_l$ 是需要优化的实

数参数。给定带标记的训练数据集，当目标变量为 1 时，我们希望 p（y=1|x,β_0,⋯,β_d）逼近 1；当目标变量为 0，p（y=1|x,β_0,⋯,β_d）逼近 0。

解决这一优化问题最有效的方法是寻找参数使训练数据集的似然最大化，由于这个问题没有解析解（与线性回归不同），我们使用迭代算法来近似参数，通常采用梯度下降算法（详见参考文献 [17]）。

在图 2.4 中，我们来观察一个对样本数据进行单预测量逻辑回归的例子。样本数据（图中黑色部分）显示，预测量的低值（高值）与 y=1 的低概率（高概率）有关。在这种情况下，可以使用逻辑回归将预测量的溢出空间线性分离。在拟合线上（图中蓝色部分），我们观察到 y=1 的概率随着预测值的增加而增加，分离线（图中以橙色显示）以低错误率对样本集进行分类。

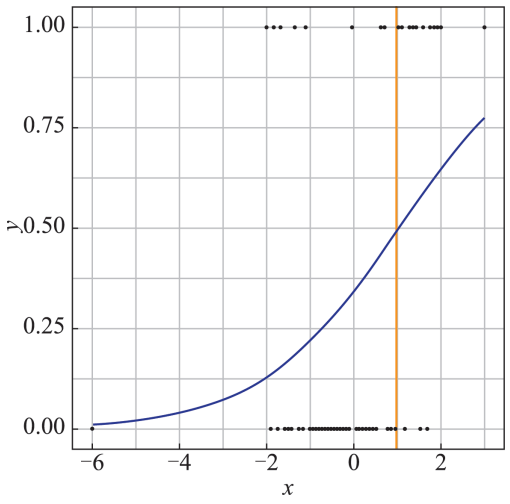


图 2.4　逻辑回归用于样本数据

样本数据以黑色显示，拟合线以蓝色显示，p（$y|x,\beta$）>1/2 和 p（$y|x,\beta$）<1/2 之间预测分割线以橙黄色显示。

2.2.2　其他分类方法

本节将扼要介绍电信领域中用到的其他分类方法：k 近邻、支持向量机（Support Vector Machine，SVM）和决策树。

k 最近邻算法 [18] 是非参数分类技术。为了确定一个新点属于哪个类，我们选择训练数据集中距离该点最近（欧氏距离）的 k 个数据点。每个参考点在训练数据集中都有类别标签，对 k 个标记的参考点进行多数投票来预测新点。在 KNN 算法中，只有参数 k 必须固定，它负责调整学习算法的能力。

支持向量机（SVM）[19] 是一种几何分类方法。主要思想是通过最小化经验分类误差和最大化几何边界，将特征空间分成两个半空间（在其简单形式中）。正则化参数是一种常见的、允许软间隔（soft margin）并减少支持向量机过拟合的方法。SVM 的一个优点是，使用所谓的核技巧（kernel trick），将数据点映射到更高维空间，然后在这个新空间上进行线性分类，从而实现非线性分类。支持向量机的主要问题是当训练集规模变大时，其计算开销很大，因此并不总是适合于解决电信领域的任务。

决策树是一种可以根据事先训练好的树进行分类的方法（参阅参考文献 [20]）。树的每个节点根据特征空间的一个条件，依次将特征空间的区域划分为两个子区域。例如，根节点条件可以为 $x_1 \leqslant 2$，在这种情况下，左边的子树与半空间 $x_1 \leqslant 2$ 相关，右边的子树与 $x_1>2$ 相关。然后，每个节点与输入空间的一个区域相关，该区域根据目标变量标记为 0 或 1。当需要拟合一个新的点时，我们通过显示这个点属于哪个区域来预测目标变量。随机森林 [21] 是基于计算大量决策树的延伸，在不考虑前期树的情况下对每个树进行计算，然后将这些树组合推导出单个预测。这种方法有许多优点：与决策树相比，它不易受过拟合的影响、可监控每个特征的重要性、树的训练速度相对较快。所以，随机森林从 21 世纪开始成为颇受欢迎的分类方法。

随机森林和梯度提升树 [22] 是两种基于树的集成分类学习算法。它们都是在训练时通过构建多个决策树来运行的，并利用了多个弱分类器的功能。随机森林和梯度提升树的区别在于随机森林基于套袋（bagging），而梯度提升树是基于提升（boosting）建立的。

2.3 聚类方法

在电信领域应用的一类重要的机器学习算法是无监督学习 [23, 25]。在这种学习算法中，数据没有被标记，算法试图学习数据的结构来理解数据。最常见的无监督算法是聚类算法，它将异构数据集划分为组，每个组都有一定的相似性。这种算法的一个关注点是理