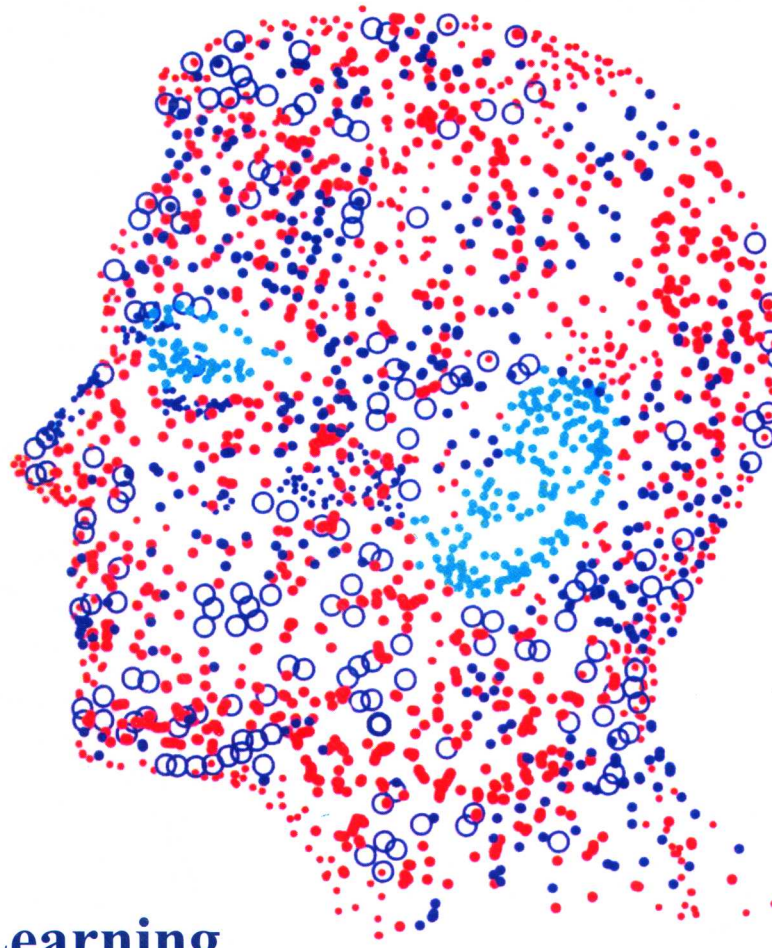
TURING 图灵程序设计丛书

CAMBRIDGE

机器学习

[英] Peter Flach ◎著　　段菲 ◎译



Machine Learning

The Art and Science of Algorithms that Make Sense of Data

中国工信出版集团

人民邮电出版社
POSTS & TELECOM PRESS

TURING 图灵程序设计丛书

机器学习

[英] Peter Flach ◎著　段菲 ◎译

Machine Learning

The Art and Science of Algorithms that Make Sense of Data

人民邮电出版社
北京

图书在版编目（CIP）数据

机器学习 / (英) 弗拉赫 (Flach, P.) 著 ; 段菲译.
-- 北京 : 人民邮电出版社, 2016.1
（图灵程序设计丛书）
ISBN 978-7-115-40577-7

Ⅰ. ①机… Ⅱ. ①弗… ②段… Ⅲ. ①机器学习
Ⅳ. ①TP181

中国版本图书馆CIP数据核字(2015)第233670号

内 容 提 要

本书是最全面的机器学习教材之一。书中首先介绍了机器学习的构成要素(任务、模型、特征)和机器学习任务，接着详细分析了逻辑模型(树模型、规则模型)、几何模型(线性模型和基于距离的模型)和概率模型，然后讨论了特征、模型的集成，以及被机器学习研究者称为“实验”的方法。作者不仅使用了已有术语，还引入了一些新的概念，同时提供了大量精选的示例和插图解说。

本书适合具备机器学习理论基础的理工科、信息技术类学生，以及数学、自动化、计算机科学等专业人员阅读。

◆ 著　[英] Peter Flach
译　段　菲
责任编辑　岳新欣
执行编辑　戴　玮　王春青
责任印制　杨林杰
◆ 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
北京天宇星印刷厂印刷
◆ 开本：787×1092　1/16
印张：18.25　彩插10
字数：515千字　2016年1月第1版
印数：1-4 000册　2016年1月北京第1次印刷
著作权合同登记号　图字：01-2013-3888号

定价：79.00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：京崇工商广字第0021号

版权声明

Machine Learning: The Art and Science of Algorithms that Make Sense of Data first edition (978-1-107-42222-3) by Peter Flach first published by Cambridge University Press 2012.

All rights reserved.

This simplified Chinese edition for the People's Republic of China is published by arrangement with the Press Syndicate of the University of Cambridge, Cambridge, United Kingdom.

© Cambridge University Press & Posts & Telecom Press 2016.

This book is in copyright. No reproduction of any part may take place without the written permission of Cambridge University Press and Posts & Telecom Press.

This edition is for sale in the People's Republic of China (excluding Hong Kong SAR, Macao SAR and Taiwan Province) only.

此版本仅限在中华人民共和国境内(不包括香港、澳门特别行政区及台湾地区)销售。

谨以此书献给Hessel Flach（1923—2006）。

试读结束：需要全本请在线购买： www.ertongbook.com

推荐序

人工智能、大数据分析、机器人等领域在近年来日益引人瞩目，而机器学习则是其中一类非常重要的理论和工具。Peter Flach教授的这部著作可作为机器学习的入门图书，帮助广大迫切希望了解和掌握机器学习的同学和工程师奠定良好的基础。

本书各章节的选题恰到好处，不但对经典机器学习框架中的模型做了非常系统的梳理和分类，涵盖了机器学习基础知识的主要部分，如不同的学习模型、特征、集成学习，而且还对机器学习实验，尤其是机器学习算法的评价(ROC分析)给予了特别的关注，这是十分难能可贵的(在一定程度上弥补了同类教科书的空白)。只要理解了上述内容，相信读者便掌握了机器学习的基本要素，同时有能力进一步就一些更专门、更前沿的主题，如在线学习、主动学习、强化学习、深度学习等，进行更为深入的学习和探索。从这个意义上讲，作者对本书的导论性著作的定位已充分地达成了。

对于如何帮助读者充分理解书中的知识点，作者也予以了充分考虑。书中配有相当数量的图解和实例，旨在突出机器学习理论的直观性。这些内容的数学门槛不高，对工程师和工科学生无疑会有很大的帮助。在各章节中，作者还对每种方法的历史影响进行了介绍，相信会十分有助于提升读者的学习兴趣。作者作为在该领域耕耘多年的研究人员，拥有非常丰富的实践经验，在不少章节中都详尽地分享了实践经验，比如特征和实验这两章。相信这些宝贵的经验会为读者朋友们具体实践机器学习理论带来更好的可操作性。

最后要提到译者段菲的翻译，通篇用语规范、表达准确，总体说来是非常不错的翻译版本。作为计算机视觉领域的优秀研究人员，他本人也在使用机器学习方面有着丰富的研究和实践经验，相信这些都为这本书的翻译质量带来不少保证。

张益民
英特尔(中国)研究院首席研究员
2015年11月10日于北京

序

本书的写作始于2008年夏，当时我就职的布里斯托大学授予我为期一年的研究经费。出于两点考虑，我决定着手编写一部著作，对机器学习进行一般性介绍：一是这样一部著作所涵盖的知识点会对市面上的许多专业书籍起到补充作用；二是通过编写这部著作，我能够获取一些新知识，所谓教学相长嘛。

任何试图编写一部机器学习导论性著作的人都会面临这样一个挑战：如何在公平对待机器学习领域无与伦比的丰富内容时，还能保证其内在原理的统一性。如果过于强调这门学科的多样性，则可能使该书最终成为一本“菜谱”式图书而丧失统一性；若过于专注自己感兴趣的领域，则可能会错失许多其他有趣的方向和应用。经过反复斟酌，我确定了本书的基本宗旨，即统一性和多样性“两手都要抓，两手都要硬”。具体表现是通过将任务和特征分开处理(每个机器学习方法都有这个东西，但大家往往不会过多关注它们)，以实现统一性；通过覆盖大量逻辑模型、几何模型和概率模型，以实现内容的多样性。

显然，指望在区区三百页的篇幅内深入介绍所有的机器学习内容是不现实的。后记中列出了我最终决定舍弃的，但值得进一步研究的重要领域和方向。在我看来，机器学习是统计学和知识表示“联姻”的产物，因而我在书中主题的取舍上也有意识地强化了这种观点。比如，在介绍与统计学联系更紧密的内容之前，我会用大量篇幅来介绍决策树和规则学习。纵贯全书，我都会特别关注“直观”，希望通过大量例子和图解来帮助读者培养和加强直观理解，其中许多例子都源于我对机器学习中ROC应用相关的研究工作。

如何阅读本书

本书内容是以“线性”方式呈现的，也就是说读者可按章节顺序逐页阅读。然而，这并不意味着你不能随便挑一章进行阅读，因为我在写作时已力图使内容模块化。

例如，对于那些希望尽快了解其第一个学习算法的读者，可直接从介绍两类分类的2.1节开始，然后直接跳转至第5章学习决策树的相关算法，这样在知识连贯性方面不会有什么问题。阅读完5.1节之后，还可以直接跳转到第6章的前两节去学习基于规则的分类器。

或者，对线性模型感兴趣的读者可在学习完2.1节之后，转而阅读3.2节关于回归任务的相

关内容，之后再跳转到第7章学习线性回归。第4～9章关于逻辑模型、几何模型和概率模型的编排次序有一定的逻辑，但这几章的大部分内容都是相互独立的，第10～12章关于特征、模型组合及机器学习实验的相关内容也是如此。

我还要说明的是，绪论和第1章属于导论，且都是自成体系的：尽管绪论中的确包含了一些技术细节，但对于大学预科以上水平的读者，理解起来不会有太大难度；第1章则对本书所覆盖的大部分内容给出了提纲挈领式的介绍。这两部分内容可从本书网站免费下载：www.cs.bris.ac.uk/~flach/mlbook；今后，我还将陆续添加一些其他材料，如讲义幻灯片等。考虑到本书所涉及内容的广泛性，出现一些小的错误在所难免，因此如果你希望了解已有的勘误列表，或提交新的勘误，欢迎你访问上述网站。

致谢

独自撰写一部著作难免孤独，但幸运的是，我得到了许多同事和朋友的热情帮助和鼓励。布里斯托尔的Tim Kovacs、鲁汶的Luc De Raedt以及波士顿的Carla Brodley组织了专门的阅读小组，给予了我极有价值的反馈。我还收到了来自Hendrik Blockeel、Nathalie Japkowicz、Nicolas Lachiche、Martijn van Otterlo、Fabrizio Riguzzi以及Mohak Shah的富有帮助的评论。还有许多人也以各种形式向我提供了帮助，在此我一并向他们表示感谢。

José Hernández-Orallo做了许多本职以外的工作，她不仅仔细阅读了我的手稿，而且还提出了诸多批评和出色的建议，我已一一采纳。José，我一定会找个机会请你吃饭。

感谢我在布里斯托大学的同事和合作者Tarek Abudawood、Rafal Bogacz、Tilo Burghartdt、Nello Cristianini、Tijl De Bie、Bruno Golénia、Simon Price、Oliver Ray以及Sebastian Spiegler，感谢他们与我一道工作，并开展了诸多富有启发的讨论。感谢我的国际合作者Johannes Fürnkranz、Cèsar Ferri、Thomas Gärtner、José Hernández-Orallo、Nicolas Lachiche、John Lloyd、Edson Matsubara以及Ronaldo Prati，本书中的许多内容都来自或受到我们合作研究的启发。有时，本书需要推进，多亏Kerry、Paul、David、Renée和Trijntje仗义出手，我方得以逃到某个僻静之所从容写作。

剑桥出版社的David Tranah对本书加工工作的启动给予了诸多帮助，封面上隐喻“理解数据”的点彩画正是他的建议（如果你正在琢磨这是谁的画像，那我得解释一下：据David自己说，这就是个普通的剪影画，并非特指某个人）。感谢Mairi Sutherland非常细致的编辑工作。

谨以此书献给先父，若他得知本书已经完成，定会开一瓶香槟来庆祝。他归纳问题的视角虽说有些病态，但也是发人深省的：每天用来喂鸡的那只手最终会将鸡的脖子拧断（这里我要向素食读者表示歉意）。感谢父母为帮助我找到自己的人生道路所提供的一切。

最后，千言万语也不足以表达我对妻子Lisa的感激之情。在我们新婚燕尔之际，我便开始筹划本书的撰写。我们当时谁都没有预料到这本书居然会花费将近四年的时间。后见之明实在太奇妙，因为事后再看，怀疑下面几件事无法同时进行当然是合理的，但事前还是会坚信，我在完成一本书的同时，还能去组织一场国际会议，并且监督房子的大规模翻新。不过，这也见证了Lisa对我的支持、鼓励和默默忍受。所幸这三件事都已圆满成功。Dank je wel, meisje！[①]

Peter Flach
于布里斯托大学

① 荷兰语，大意为“谢谢你，姑娘！”。——译者注

目 录

试读结束：需要全本请在线购买： www.ertongbook.com

绪 论

机器学习概述

你也许尚未意识到，自己很可能已经是一名机器学习技术的普通用户了。例如，目前绝大多数电子邮件客户端都采用了识别和滤除垃圾邮件(spam[①] e-mail)的算法。早期的垃圾邮件过滤器(spam filter)采用的是基于人工规则(如正则表达式)的模式匹配方法。但人们很快发现，这样的系统不仅难以维护，而且也缺乏灵活性。这很容易理解，因为绝对意义上的“垃圾”邮件是不存在的。有些邮件对某些人来说是“垃圾”，但对其他人来说则可能极有价值。如今，机器学习技术在邮件分类领域已得到了广泛应用，其所带来的自适应性和灵活性远非早期方法可以企及。

SpamAssassin是一款非常流行的开源垃圾邮件过滤器，其工作机制是依据一组内置规则或“测试”(SpamAssassin的专用术语)为每封到来的邮件进行评分。如果某封邮件的得分不低于5分，则SpamAssassin会自动为其添加一个“junk”(即垃圾邮件)标签，并在该邮件标头(email header)中添加一段简要说明。下面展示了我收到的某封邮件的简要说明。

```
-0.1 RCVD_IN_MXRATE_WL          RBL: MXRate recommends allowing
                                [123.45.6.789 listed in sub.mxrate.net]
 0.6 HTML_IMAGE_RATIO_02        BODY: HTML has a low ratio of text to image area
 1.2 TVD_FW_GRAPHIC_NAME_MID    BODY: TVD_FW_GRAPHIC_NAME_MID
 0.0 HTML_MESSAGE               BODY: HTML included in message
 0.6 HTML_FONx_FACE_BAD         BODY: HTML font face is not a word
 1.4 SARE_GIF_ATTACH            FULL: Email has a inline gif
 0.1 BOUNCE_MESSAGE             MTA bounce message
 0.1 ANY_BOUNCE_MESSAGE         Message is some kind of bounce message
 1.4 AWL                        AWL: From:address is in the auto white-list
```

在该图中，从左到右依次为某项测试的评分、测试标识符以及包含对邮件相关部分引用的简短描述。可以看到，不同测试给出的评分可能为负(表明该邮件为普通邮件而非垃圾邮件)也可

① spam这个词原本是香料火腿(spice ham)的缩写，该肉制品之所以“声名狼藉”，完全要归咎于电视剧《蒙提·派森的飞行马戏团》(*Monty Python's Flying Circus*)在1970年某一集中的讽刺。

能为正。由于该邮件的综合得分为5.3分，表明该邮件有可能是一封垃圾邮件。这封邮件实际上传达的是来自中介服务器的一条通知，即另外一封邮件(其得分高达14.6分)被认定为垃圾邮件而被拒收。由于该条“退信”消息中包含了原始邮件，故继承了后者的某些特征，如低文本-图像比(text-to-image ratio)，所以它的得分超过了给定的阈值——5分。

下面再举一例。这次所涉及的是一封我本人期盼已久的重要邮件。但不幸的是，我后来在“垃圾箱”中找到了它。

```
2.5  URI_NOVOWEL            URI: URI hostname has long non-vowel sequence
3.1  FROM_DOMAIN_NOVOWEL    From: domain has series of non-vowel letters
```

这封邮件事关我与研究小组的一名同事提交到欧洲机器学习会议(ECML)和欧洲数据库中的知识发现原理与实践会议(PKDD)的一篇文章(自2001年起，这两个会议一直是联合举办)。虽然这两个会议在2008年所使用的域名ecmlpkdd2008.org对每位机器学习研究者来说几乎是耳熟能详，但却因连续出现了11个“非元音字符”而引起了SpamAssassin的怀疑！这个例子表明对于不同的用户，某项SpamAssassin测试的重要性可能不同。机器学习是一种可为用户量身打造程序的绝佳方法。

至此，你一定十分希望了解SpamAssassin确定每项测试分值或“权值”的依据到底是什么。我要告诉你的是，这正是机器学习的用武之地。假设现有一个规模较大的电子邮件“训练集”(training set)，该集合中的所有邮件都带有人工标注的类别标签，即“垃圾邮件”或“普通邮件”，且每封邮件的全部测试结果均已给出。我们的目标是为每项测试找到一个恰当的权值，以使所有垃圾邮件的得分均超过5分，而所有普通邮件的得分均低于5分。在本书后面的章节中，你将接触到大量适于解决该问题的机器学习技术。下面我们先通过一个简单的示例来阐述这些方法的主要思想。

例1(线性分类) 假设我们只考虑两项测试。给定的邮件训练集中共有4封邮件，其中含3封普通邮件和1封垃圾邮件，如表1所示。可以看出，对于垃圾邮件(对应表中第1行)，两项测试都是成功的；对其中一封普通邮件(表中第2行)，两项测试均失败；而对其余两封普通邮件(对应表中第3、4行)，两项测试均只有一项成功。显然，将这两项测试的权值全部设为4，便可以对上述4封邮件正确分类。依据“背景知识1”所引入的数学符号，我们可将上述分类器表示为$4x_1 + 4x_2 > 5$或$(4,4)\cdot(x_1,x_2) > 5$。实际上，采用任意介于2.5和5之间的权值都能够确保仅当两项测试均成功时，邮件的最终得分才高于阈值5。我们甚至可以考虑为两项测试分配不同的权值(只要满足每个权值均小于5，且总和超过5即可)，尽管也许很难从训练数据看出这种做法的合理性。

表1 供SpamAssassin使用的一个规模较小的训练集。x_1和x_2两列分别对应两项测试的结果，第4列表明哪些邮件为垃圾邮件，而最右列则表明如果将5作为判别函数$4x_1 + 4x_2$的阈值，可轻易地将垃圾邮件分离出来

电子邮件	x_1	x_2	垃圾邮件	$4x_1 + 4x_2$
1	1	1	1	8
2	0	0	0	0
3	1	0	0	4
4	0	1	0	4

背景知识1　用数学语言描述SpamAssassin的工作原理

在本模块中，我们将借助数学语言来描述SpamAssassin的原理。在后面所有类似的模块中，我会对一些重要、有用的概念和符号做出提示。如果你对所涉及的知识点不够熟悉，则需要花费一些时间来回顾(既可借助书本，也可参考一些像维基百科或MathWorld这样的在线资源)，以便充分理解本书的其余内容。

借助数学语言，可以给出刻画SpamAssassin分类器的多种有用形式。若将对给定邮件的第i项测试结果记为x_i(如果测试成功，则$x_i=1$，否则$x_i=0$)，并记第i项测试的权值为w_i，则邮件的总分可表示为$\sum_{i=1}^{n} w_i x_i$。这里用到这样一个事实：仅当$x_i=1$(即对该封邮件的测试成功)时，w_i方计入总分。若用t表示阈值(即如果某封邮件的总分超过t，则将其判为垃圾邮件。本例中$t=5$)，则“决策规则”可表示为$\sum_{i=1}^{n} w_i x_i > t$。

可以看到，上述不等式左边的部分对变量x_i是线性的。这意味着如果x_i的增量为δ，则总和的增量为一个不依赖于x_i的数——$w_i\delta$。若x_i在该不等式中以平方项或其他指数不为1的形式出现，则该结论不成立。

借助线性代数的相关知识，我们可对该决策规则的符号做进一步简化。设$\boldsymbol{w}=(w_1,\cdots,w_n)$，$\boldsymbol{x}=(x_1,\cdots,x_n)$，则上述不等式可简化为内积形式：$\boldsymbol{w}\cdot\boldsymbol{x}>t$。将该不等式修改为等式，即可得到用于判决某封邮件是否为垃圾邮件的“决策边界”(或“决策面”)方程：$\boldsymbol{w}\cdot\boldsymbol{x}=t$。不难看出，由于该方程等号左端为线性项，故在由变量x_i所张成的空间中，决策边界为一平面。向量$\boldsymbol{w}$与该平面垂直，并指向垃圾邮件样本的方向，图1给出了数据点和决策面在2D空间中的可视化结果。

有时可通过引入额外的常量来进一步简化决策边界、决策规则的符号。记$\boldsymbol{x}^\circ=(1,x_1,\cdots,x_n)$，$\boldsymbol{w}^\circ=(-t,w_1,\cdots,w_n)$，即为$\boldsymbol{x}$引入额外分量1，将对应的权值设为$-t$。这样决策规则即为$\boldsymbol{w}^\circ\cdot\boldsymbol{x}^\circ>0$，上述决策面方程就可整理为更紧凑的形式：$\boldsymbol{w}^\circ\cdot\boldsymbol{x}^\circ=0$。借助齐次坐标，在新的坐标系中，决策面成为一个经过原点的平面，代价只是增加一个额外的维度(但请注意，这并不会对数据产生任何实际影响，因为所有的数据点和“真实”的决策边界均位于平面$x_0=1$中)。

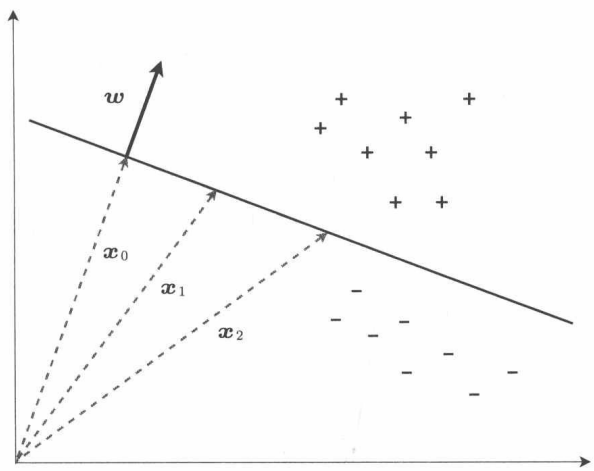


图1　2D空间中的线性分类示例。图中将正例(“+”)和反例(“−”)分离的直线(决策边界)方程为$\boldsymbol{w}\cdot\boldsymbol{x}_i=t$，该式中$\boldsymbol{w}$为一个与决策边界垂直的向量，其方向指向正例，$t$为决策阈值，而$\boldsymbol{x}_i$对应于决策边界上的某一个点。特别地，$\boldsymbol{x}_0$的方向与$\boldsymbol{w}$一致，故有$\boldsymbol{w}\cdot\boldsymbol{x}_0=\|\boldsymbol{w}\|\cdot\|\boldsymbol{x}_0\|=t$($\|\boldsymbol{x}\|$表示向量$\boldsymbol{x}$的长度)，因此决策边界可表示为$\boldsymbol{w}\cdot(\boldsymbol{x}-\boldsymbol{x}_0)=0$。该形式在一些场合中使用时尤为方便，特别是这种符号能够更清晰地表达如下重要事实——决定决策边界位置的是$\boldsymbol{w}$的方向，而非其长度

你可能会感到疑惑，上面这些内容究竟与学习有什么关系？毕竟，这只是一个数学问题。你这样想似乎有些道理，但我们完全可以换一种说法——“SpamAssassin通过正例和反例来学习如何识别垃圾邮件”，这听起来似乎也有道理。况且，如果能够预先获得更多的训练数据，SpamAssassin的表现会更为优异。依据经验提升性能几乎是各种形式的机器学习方法的核心。下面我们给出机器学习的一般定义：*机器学习是对依据经验提升自身性能或丰富自身知识的各种算法和系统的系统性研究*。在上述SpamAssassin的例子中，“经验”对应一组正确标注的训练数据，而“性能”对应识别垃圾邮件的能力。图2给出了机器学习如何介入垃圾邮件分类这项任务的原理性示意。在不同的机器学习任务中，“经验”往往具有不同的形式，如对错误的纠正、实现某个目标后的奖励等。此外还需注意，与人类的学习类似，在某些任务中，机器学习的目的可能不是针对特定任务取得性能提升，而是在整体上使知识得到提升。

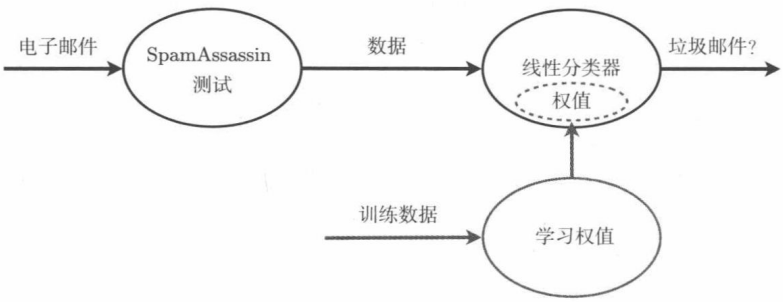


图2 从该图顶部可以看到SpamAssassin完成垃圾邮件分类任务的流程：通过SpamAssassin的内置测试，将每封邮件内容转化为一个数据点，然后运用线性分类器进行决策。在该图的底部可见机器学习的主要工作

我们已经看到，一个机器学习问题往往存在多种解决方案，即便对于像例1这样的简单问题也不例外。这就引发了一个很重要的问题：我们如何从众多候选方法中做出抉择？考虑这个问题的一种方式是意识到，我们其实对算法在训练数据上的性能并不关心，毕竟我们已经知道其中哪些邮件为垃圾邮件。我们真正关心的其实是即将到来的邮件能否被正确分类。这听起来有点像“先有鸡还是先有蛋”的问题——要想知道一封邮件是否被正确分类，必须事先获知其类别；如果我们已经知道该邮件的类别，则无须再进行分类决策。但是，我们必须牢记在训练数据上取得优异性能只是手段，而非目的。实际上，如果一味追求系统在训练数据上的性能，很容易造成一种貌似喜人、实则存在巨大隐患的现象——过拟合(overfitting)。

例2(过拟合) 假设你正在备战“机器学习101”课程的考试。为了帮助复习，Flach教授将历年真题及答案共享给了大家。做题时，你一开始还尝试自己求解答案，然后与参考答案进行比对。但不幸的是，做着做着你就走火入魔了，将全部时间都花在了机械地记忆参考答案上。如果即将到来的考试完全采用过去考过的题目，你一定会取得非常优异的成绩。然而，如果考试中出现了考察相同知识点、但具有不同表述形式的新题目，那么跟老老实实复习相比，你势必会因准备不足而得到低得多的分数。这种情形下，我们就可以说，你对历年考试产生了“过拟合”：机械记忆所获取的知识无法推广到新考题上。

推广性(generalization)[①]可能是机器学习中最基础的概念。如果SpamAssassin从训练数据中提炼的知识能够在你的邮件中得到运用，你一定十分满意；否则，你可能会寻找替代SpamAssassin的垃圾邮件过滤器。但过拟合并不是系统在新数据上表现得不尽人意的唯一可能原因。另一种可能的原因是SpamAssassin程序的作者设定权值时使用的训练数据不符合你电子邮件的实际情况。幸运的是，这类问题是存在解决方案的——你可选用符合你邮件特点的不同样本作为训练数据。如果可能的话，你甚至可以使用自己实际收到的垃圾邮件和普通邮件作为训练集。机器学习是一项能够改变软件行为以自动适应你个人情况的伟大技术，目前已有许多垃圾邮件过滤系统允许用户为其定制训练数据。

可见，当一个问题存在多种解决方案时，你必须谨慎选择以避免产生过拟合。本书后面的章节将向你介绍一些选择方法。现实中，我们有时也会遇到另一种情形，即无法找到一种在训练数据上表现良好的解决方案。这种情况下，我们该如何应对？例如，假设例1中的第2封邮件(两次测试均失败)是垃圾邮件，则任意一条直线(决策面)都无法将垃圾邮件和普通邮件分离(你可在由x_1和x_2组成的2D坐标系中画出这四封邮件所对应的点来验证这一说法)。在这种情况下，存在几种可能的解决方案。一种是无视这一点，认为第二封邮件可能是非典型的或存在标注错误(也称*噪声*)；另一种方案则是尝试描述能力更强的分类器。例如，我们可引入垃圾邮件的第二条决策规则：除已有的$4x_1+4x_2>5$外，我们再添加规则$4x_1+4x_2<1$。注意，采用这种方法会学习出不同的阈值，甚至不同的权值向量。一般仅在训练数据集容量较大、能够保证可靠地学到这些参数的情况下才会选择这种方法。

SpamAssassin风格的线性分类很适合入门讲解。但你有必要了解，它并不是机器学习中的唯一方法，否则本书的篇幅就要短得多了。如果学习的任务不仅包括测试的权值，而且包括测试本身，我们该如何应对？如何确定文本-图像比是不是合适的测试？最重要的是，我们如何首先设计出这样的测试？令人欣慰的是，这些都是机器学习最擅长解决的。

你可能已经注意到，截至目前，SpamAssassin的测试看似并未对邮件本身的内容给予足够的关注。诚然，像“Viagra”(万艾可/伟哥)、“free iPod”(免费iPod)或“affirm your account details”(确认你的账户细节)这样的词汇或短语都可作为很好的垃圾邮件指示信号，而其他某些词汇(如只有朋友才使用的特定昵称)则指向普通邮件。为此，许多垃圾邮件过滤器都运用了文本分类技术。通常，这些技术都会维护一个可作为垃圾邮件或普通邮件指示信号的词汇/短语表。例如，假设我们在4封垃圾邮件和1封普通邮件中发现了“Viagra”这个词。当新邮件到来时，如果它包含了“Viagra”这个词，我们会推断它有4:1的几率成为垃圾邮件，或者说该邮件为垃圾邮件的概率为0.8，为普通邮件的概率为0.2(可参考“背景知识2”来回顾概率论中的一些基本概念)。

背景知识2　概率论基础

概率涉及描述“事件”结果的随机变量。这里的“事件”通常是以“假设”的形式出现，因此需要用估计的方法来得出其概率。例如，考虑这样一个假设：“英国有42%的公民支持现任首相”。检验该假设的唯一方法是逐一询问每一位英国公民，但做如此全面的调查显然是不可行的。因此我们转而抽取(希望)具有代表性的公民样本进行调查，该假设的更准确的说法应为“在抽样调查中发现，有42%的英国公民支持现任首相”或“支持英国现任首相的公民比例约为42%”。请注意，这些假设都是基于比例或“相对频率”而形成的，对应于概率的相应假设则为“随机抽取一名英国公民，则该人支持现任首相的概率约为0.42”。显然，此例中的“事件”为“随机抽到的公民支持现任首相”。

① 也称为泛化能力。——译者注

试读结束：需要全本请在线购买：　www.ertongbook.com