BIG DATA SERIES 大数据丛书

AN ELEMENTARY INTRODUCTION TO STATISTICAL LEARNING THEORY

统计学习理论基础

【美】桑吉夫·库尔卡尼 SANJEEV KULKARNI
吉尔伯特·哈曼 GILBERT HARMAN 著

肖忠祥 闫效莺 段沛沛 程国建 译

WILEY

机械工业出版社 CHINA MACHINE PRESS

大数据丛书

An Elementary Introduction to Statistical Learning Theory

统计学习理论基础

【美】 桑吉夫·库尔卡尼（Sanjeev Kulkarni） ◎著
吉尔伯特·哈曼（Gilbert Harman）
肖忠祥　闫效莺　段沛沛　程国建 ◎译

机械工业出版社

本书原作者是美国普林斯顿大学电气工程系和哲学系的两位教授，本书是在普林斯顿大学“电气工程及原理”课程中关于“学习理论和认知论”的入门性课程基础上形成的。

全书共包含18章，从概率密度、贝叶斯决策理论引入样本学习的基本概念，进而介绍了最近邻域学习、核学习及神经网络学习，在此基础上探讨了PCA学习、VC维概念、函数估计问题等，最后重点介绍了非常实用的支持向量机（SVM）及Boosting算法。各章均包含小结、附录、问题及参考文献，非常适合高等院校计算机类及电气类、自动化类研究生及本科高年级学生作为参考书。

图书在版编目(CIP)数据

统计学习理论基础/（美）桑吉夫·库尔卡尼（Sanjeev Kulkarni），（美）吉尔伯特·哈曼（Gilbert Harman）著；肖忠祥等译. —北京：机械工业出版社，2016.12

书名原文：An Elementary Introduction to Statistical Learning Theory

ISBN 978-7-111-55522-3

Ⅰ. ①统… Ⅱ. ①桑…②吉…③肖… Ⅲ. ①统计学 Ⅳ. ①C8

中国版本图书馆CIP数据核字(2016)第287366号

机械工业出版社（北京市百万庄大街22号 邮政编码100037）

策划编辑：王 康 责任编辑：王 康 刘丽敏

封面设计：路恩中 责任校对：佟瑞鑫

责任印制：常天培

北京圣夫亚美印刷有限公司印刷

2017年3月第1版第1次印刷

169mm×239mm · 11.25印张 · 211千字

标准书号：ISBN 978-7-111-55522-3

定价：43.00元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

电话服务

服务咨询热线：010-88379833

读者购书热线：010-88379649

网络服务

机工官网：www.cmpbook.com

机工官博：weibo.com/cmp1952

教育服务网：www.cmpedu.com

金书网：www.golden-book.com

封面无防伪标均为盗版

译者序

在工程应用及商务分析中，非常重要的任务之一是能够从小样本数据中进行快速机器学习。统计学习理论（SLT）给出了从少量数据样本中抽取模式及其关系的理论基础，这个学习机理的核心是平衡所有解决方案之间的性能与复杂度从而找出最优的解决方案。

支持向量机（SVM）提供的学习能力来自于对统计学习理论的深度数学分析，其学习过程是基于有限的观测值来估计系统的未知关系及其结构的过程。统计学习理论给出了设计这样一个经验主义机器学习的数学条件，这为在精确地表达已有数据和处理未知数据之间保持最佳的平衡提供了解决方案。SVM 的主要优点表现在：①从小样本数据记录中学习；②模型复杂度可控（SVM 可以通过调整一些参数对模型的复杂性进行直接的控制）；③奇异点检测与数据压缩。SVM 的主要缺点表现在：①黑盒模型（SVM 模型的可解释性要比理解神经网络具有更大的挑战性）；②市场推广困难（解释 SVM 及其统计学习理论基础需要深厚的数学背景知识及模式识别经验，甚至对一个经验丰富的研究人员来说都是一个挑战）。

近年来，基于深度神经网络的机器学习理论研究风靡人工智能学术领域，在工业界的应用也崭露头角，引人瞩目的事件是 2016 年谷歌公司的 AlphaGo 战胜世界围棋冠军李世石。一个错误的观点是认为深度学习神经网络将要取代传统的浅层人工神经网络以及基于统计学习理论的 SVM。不可取代的原因在于：其一，科学理论的发展有其自身的规律性，成为学术热点并不代表着会成为核心应用技术；其二，深度学习对复杂异构大数据模式的机器学习较为有效，而对规范的小数据模式 SVM 则是利器；其三，统计学习理论与深度学习的融合或许会对大数据时代的人工智能发展起到一定的推动作用，已发表的基于 SVM 的深度学习及深度 SVM 就是例证。

本书以通俗易懂的方式提供了统计学习理论与机器学习及模式识别的基本概念及常用算法，这包括概率密度函数、贝叶斯决策、最近邻规则、人工神经网络、VC 维、函数估计问题、支持向量机、集成学习等，适合作为高校研究生及本科生的人工智能、机器学习等课程的教学用书及自学参考书。

本书的出版得到西安石油大学优秀学术著作出版基金的支持，在此表示感谢！

程国建

前 言

本书为新兴领域的统计学习理论提供了一个宽泛和易于理解的入门性介绍，这一领域的发展源于对模式识别和机器学习、非参数统计、计算机科学、语言学中的语言学习和认知心理学、哲学问题中的归纳法以及哲学和科学方法论等学科与技术的研究。

本书是“学习理论与认知论”课程的非常好的入门教材，目前已在普林斯顿大学电气工程专业的教学中使用。“学习理论与认知论”课程并没有特定的基础要求，向所有对其感兴趣的学生开放，包括新生、主修科学的高年级学生，以及来自工程、人文、社会科学的学生。虽然许多材料技术性较强，但是我们发现大部分学生可以体会和领悟本书的要点。

模式识别的工程研究关注的是基于一个有用的方法研发出的自动化系统来区分不同的输入模式。为邮局开发的系统用于如何扫描手写地址并将邮件排序，制造商关注如何设计一个计算机系统把普通的谈话内容进行文字转录，还有诸如计算机能否用来分析医学图像，进而做出诊断等此类问题。

机器学习提供了对一些模式识别问题进行求解的有效方法。它可能是采用受过训练的系统来识别手写邮政编码，或能使自动化系统与用户进行交互使其学会实现对语音的识别；也许是使用机器学习算法来开发一套医学图像分析系统。

机器学习和模式识别也关注学习系统所包含的一般原则。一种系统化的方法技术非常有用，因为我们并不是从无到有开发算法并在每个新的应用程序中特设某一种方式。评估一个学习系统的性能所采用的技术也是非常重要的。对学习算法的实践环节而言，知道什么是可实现的，什么是可用的评价基准，并提出新的技术也同等重要。

这些问题也出现在认知论与哲学问题中。我们能学到什么？以及我们如何进行学习？我们能够从其他思想和外部世界学到什么？通过归纳法我们又能学到什么？

哲学问题的归纳法关注的是如何在归纳推理的基础上学到一些新东西。而给定的事实是归纳推理前提的真实性无法保证其结论的真实性。这个问题没有

唯一解，这并不是因为无解，而是因为有太多解，这取决于采用什么学习方法。在本书中，我们解释了如何根据归纳形成各种不同的解决方案。

因此，我们希望本书能为广大读者在统计学习理论中提供一个简便的入门性介绍。对于那些对学习理论或实际算法的深入研究感兴趣的读者，我们希望本书提供给他们一个有益的出发点。而对于那些对一般的认知论和哲学感兴趣的读者，我们希望本书有助于他们从其他领域中领悟一些重要的想法。对其他读者而言，我们也希望本书有助于他们对统计学习理论有更深层次的理解，因为它揭示了学习的本质及其限制，这也是人工智能的核心进展。

感谢普林斯顿大学本科教育创新课程发展 250 周年纪念基金的资助。Rajeev Kulkarni 对全书提供了非常有用的意见。Joel Predd 和 Maya Gupta 提供了许多宝贵的意见。此外，感谢 Joshua Harris 对本书的仔细审读。同时也感谢几年来，我的助教和学生们一起对该课程内容的讨论。谢谢！

目 录

第1章 引言：分类、学习、特征及应用

1.1 范围

本书主要关注模式分类——根据目标的几个观测值或测量值，将其划入其中一个范畴。最简单的情况是将一个对象划分为两类之一，但更常见的是类别数目不确定的情况。与之密切相关的第二个任务是对与目标属性实际数目的估计。如在分类中，存在对一些目标的观察值或测量值，我们的估计正是基于这些观察的结果进行的。

本书讨论的大部分问题来源于第一个任务，即分类。但偶尔也会讨论第二个任务——估计。这两种情况中，我们感兴趣的都是当已知观测值或测量值时，目标的分类规则或估计值，更具体地说，是用于分类和估计的学习规则的建立方法。

下面进一步讨论一些具体的例子。如考虑从视觉数据中识别手写字符、脸、其他物体或者语音等。尽管人类很擅长这种分类问题，但是却很难设计出能接近人类识别性能和鲁棒性的自动算法来完成这些任务。

经过半个多世纪的努力，在诸如电气工程、数学、计算机科学、统计、哲学和认知科学等领域中，人们仍在探寻最优的机器学习算法。也就是说，人们在学习理论的研究及应用等方面已取得了巨大的进步，在这一领域中有很多深入且实用的结论，它们与前述学科相关。这些理论的应用很广泛，但是，对这些理论的多数讨论相当先进，需要一定的技术背景以及专业知识支撑。

写作本书的目的是为读者提供一个对该领域易于理解的入门级介绍，既可以方便那些希望对这个问题做深层次研究的读者，也可以为那些渴望对基本概念做全面理解的读者提供一个基础。本书主要专注两类的模式分类问题。该问题来源于实际应用，同时也可用于解释该领域的许多关键问题，同时我们还去除了一些不必要的复杂内容。虽然还有一些学习相关的重要内容在本书中没有包含，但是我们的目标是致力于提供一个具有更大深度、更加广泛性的参考资料和模型。我们希望这本书将作为一个有价值的学习切入点。

1.2 为什么需要机器学习？

模式识别算法在很多实际问题中都很有用，它是人工智能的一个重要方面。

但你可能会问为什么我们需要设计一个用于分类的机器学习方法来学习好的规则，而不是仅仅为某一个给定的应用设计一个好的规则，并实现它。

主要原因是在许多应用中，我们可以找到一个好的规则的唯一方法是使用数据来学习。例如，要精确地描述一幅图像中的人脸是如何构成的是非常困难的，因此，直接设计分类规则来判断一个给定的图像中是否包含人脸是很困难的。但是，给出一个好的学习算法，我们就可以为算法提供具有人脸的图像和许多没有人脸的例子，然后让算法总结出能识别是否有人脸的规则。拥有一个学习算法的其他优点包括：对假设和建模误差具有鲁棒性，减少编程量，并能适用于环境的不断变化。

一般来说，对于一个分类问题，我们希望根据对象的某些测量值决定该对象的分类。通过使用多个具有类别标签的对象来学习规则，这时会出现以下问题：

1. 我们所说的“对象”和对象的“测量”是什么？
2. 在分类问题中，哪些是对象所属的类？
3. 在估计问题中，我们试图估计的值是什么？
4. 如何衡量一个分类或估计规则的质量，我们所期望的最好的规则是什么？
5. 哪些信息可用于学习？
6. 我们该如何去学习一个好的分类和估计规则？

本章我们将回答前 3 个问题。其余问题需要一些概率的背景知识，这些概率知识将在第 2 章和第 3 章提到。有了这些背景知识，我们将在第 4 章和第 5 章讨论第 4 个问题。第 6 章讨论第 5 个问题。本书的其余部分专门介绍最后一个问题的多方面应用和方法。

1.3 一些应用

在讨论详细细节之前，先准备一些具体的例子可能会对后期的理解有帮助。关于学习、分类和估计的应用实例有很多，下面列举几个例子。

1.3.1 图像识别

在许多应用中，分类的对象是数字图像。在这种情况下，“测量”可以描述图像中各像素的输出，如在黑白图像中，每个像素的亮度可作为一个测量。如果图像有 $N \times N$ 个像素，则像素总数为 N^2。在彩色图像中，每一个像素有 3 个测量值，其对应于 3 种颜色分量的强度，即 RGB 值。因此，对于一个 $N \times N$ 的彩色图像，有 $3N^2$ 个测量值。

根据不同应用，许多分类任务以这些测量值为基础。如人脸检测或识别就

是一个常见且有用的应用。在这种情况下，“类别”就是有人脸和没有人脸，或者有可能是对数据库中每个人脸都有单独的定义。

另一个实例是字符识别。在这种情况下，笔迹可以被分割成只包含一个单一字符的小图像，这时类别可能包括字母表的26个字母（如果区分大小写字母，则是52个字母）、10个数字和一些特殊字符（句号、问号、逗号、冒号等）。

另一类应用，其图像可能来自工业界，分类的任务就是判断当前图像是否有缺陷。

1.3.2 语音识别

在语音识别中，我们对识别说话者的语义感兴趣。在这些应用中，测量可能是一组表示语音信号的数字。首先，将信号分割成独立的字或音素。在每段中，语音信号可使用各种方式来表示。例如，信号可以使用不同时间频带的能量和强度来表示。虽然对信号描述细节的讨论不在本书的讨论范围，但是可以将信号最终表示为一组实际值。

最简单的情况，类别可以简单地表示为“yes”与“no”，稍微复杂些的任务可能是判断发音是10个数字中的哪一个，或者是在一个可接受的较大范围内判断发音归属哪个单词。

1.3.3 医学诊断

在医疗诊断中，我们感兴趣的是疾病判别，对每种疾病都有一个单独的分类。

在这些应用中，测量通常是某些医疗检查的结果（例如，血压、温度和各种血液测试）或医疗诊断（如医学图像），各种症状的存在/不存在及强度、以及一些患者的基本信息（年龄、性别、体重等）。

在测量结果的基础上，我们要判断有哪些疾病（如果有的话）。

1.3.4 统计套利

金融学中，统计套利通常是指一些具有代表性的、涉及一系列证券的短期自动贸易策略。在该策略中，人们试图基于更多证券间的历史相关性、近期价格变化以及一些常用的经济/金融变量等为一套证券设计一种交易算法。这些可以被认为是“测量”，预测可被投射为一个分类或估计问题。在分类中，类别可以是“买入”“卖出”或“什么都不做”。在估计问题中，人们可能会预测未来的某个时间范围内每种证券的预期回报是多少。之后，根据对预期收益的估计，做出交易决策（买入、卖出等）。

1.4 测量、特征和特征向量

正如在1.1节和1.3节对目标分类的讨论，为了进行判定我们使用了目标的观测值。例如，当人们要对目标进行分类时，可能会采用观察、挑选、感知、倾听等方法，或者会使用一些仪器测量目标的其他属性，如尺寸、重量、温度等。

类似地，当设计一种机器对目标进行自动分类（或学习分类）时，我们假设机器可以对访问对象的各种属性进行测量。这些测量值是传感器从对象目标获取的一些感兴趣的物理变量或特性。

考虑描述的简单性，本书中我们将每个测量（或特征）建模为一个实数。虽然在一些应用中，某些特征可能不能用数表示，但是这个假设在绝大多数常见的应用中是可行的。

我们假设对象的所有相关的，以及可获取的特性可以用有限的测量/特征值表示。则这些特征值可以放在一起，形成一个特征向量。假设有 d 个特征，特征值分别为 x_1，x_2，…，x_d，则特征向量可以表示为 $\bar{x}=(x_1,x_2,\cdots,x_d)$。这个特征向量可以被看作是 d 维空间 R^d 中的一个点或一个向量，我们将 R^d 称为特征空间。特征向量的各个分量表示相应特征的特征值，即特征空间中某一维的一个值。

在一个 $N \times N$ 维图像的目标识别中，对黑白图像，其特征数为 N^2 个，对于彩色图像，特征值有 $3N^2$ 个。

在语音识别中，特征数等于用于表示分类的语音分段实际值的数量。

1.5 概率的需要

在大多数应用中，通过特征向量的值，并不能唯一或明确地知道对象的类别。对于这种情况，有几个基本原因。首先，如果能捕捉用于目标分类的所有重要特征固然很好，但是这在通常的情况下并不可能。通常无法捕捉到测量特征的一些重要细节。这点在前面例子中很明显。

第二，根据应用和具体的测量值，特征值可能是噪声。也就是说，特征的观察值可能存在一些内在的不确定性和随机性，因此即使是相同的对象在不同的场合也可能出现不同的值。

基于这些原因，使用工具从概率的角度对问题进行精确建模，是一个不错的解决方案，在第2章和第3章中我们将回顾本书其余部分中会涉及的概率相关知识及工具。

1.6 监督学习

有了必备的概率知识，我们将在第4章对模式识别问题建模。理想（即非正常）情况下，基本概率结构是已知的，因此，由统计的结果可得到分类问题的正常结果，这部分将在第5章中讨论。

然而，实际应用中更典型的情况是底层的概率分布未知。关于这种情况，我们将诉诸于第6章讨论的借助样本标签来克服概率未知的知识不足问题。第6章中讨论的学习问题，与众多机器学习问题中涉及的如从例子学习、监督学习、统计模式分类、统计模式识别和统计学习类似。

术语“监督学习”来源于那些带有正确标签，即有“监督者”或“老师”的例子，这是与“非监督学习”相区别的。在“非监督学习”中，只有目标的例子，其所属的分类是未知的。同时还有很多其他的机器学习架构，如半监督学习、强化学习以及在统计、计算机科学和其他领域中的相关问题。本书重点关注监督学习。

1.7 小结

本章，我们描述了分类和估计中涉及的一些通用问题，讨论了一些具体和重要的应用实例。然后介绍了术语特征、特征向量和特征空间，并引入了概率和学习的相关概念。

这里我们同时提到了分类和估计。但是本书主要关注分类，一些讨论会扩展到估计。

在接下来的两章中，我们将回顾本书后面会用到的概率的重要公式。之后对分类（或模式识别）问题进行公式化描述，在讨论具体的学习方法和结果之前，先讨论从数据中学习的基本问题。

1.8 附录：归纳法

在每一章结尾的附录部分，主要讨论其他方面的问题，如可能是一个哲学本原问题。

本书我们主要关注归纳学习而不是演绎学习。演绎学习是由一般性原理、原则推演出有关个别性的知识，其思维过程是由前提推导结论，通过前提的正确性确保结论的正确。例如，通过矩形面积的计算方法以平行四边形面积与矩形面积之间的关系，你会发现平行四边形的面积等于底乘以高。同时你还可以通过三角形是矩阵的一半，推导出三角形的面积是底乘以高的一半。

归纳学习是由个别或特殊的知识概括出一般性的结论，其思维过程是由个别到一般，由证据来推导结论，它并不能保证结论的正确性。例如，你从邮件总是在星期六中午前交付的事实，可能会推断出邮件将会在下星期六中午前交付。这是一种归纳推理，因为数据不能保证结论的正确性。有时，即使推理的“前提”都是正确的，但是归纳推理的结论是错误的。

“归纳问题”的哲学本原是问一个人如何能由真实的前提完全相信归纳的结论。如果它的前提正确，也不可能由此证明归纳结论是真实的，因为典型的归纳推理不提供这样的保证。即使你由归纳得知你的邮件将会在下星期六中午之前交付，这与你的邮件没有在下星期六中午之前交付是兼容的。归纳推理不是演绎推理的特例。

在过去，通常认为对过去的归纳总能推导出正确的结论，因此有理由肯定，对未来的归纳也将能推导出正确的结论。而循环推理对此假设不予认同，即我们假设要以正确的归纳性来证明归纳的正确性。

另一个方面，可以提供一个演绎的非循环论证吗？没有任何理由来采取演绎论证的形式更何况是循环推理呢？

本书将要阐明的是，在给定某些假设的情况下，统计学习理论为某些归纳方法提供了部分演绎的数学证明。

1.9 问题

1. 什么是特征空间？空间的维数代表什么？什么是向量？什么是特征向量？

2. 为了对对象进行分类，如果我们使用 F 个不同的特征值，其中每个特征可以取 G 个不同值，那么什么是特征空间的维数呢？

3. 对一个 12×12 的灰度图像（256 个灰度级），特征向量有多少维？存在多少个不同的特征向量？

4. 分类是估计的一个特例吗？典型的分类案例和典型估计案例的差别是什么？

5. 关于归纳的几个问题：

（a）归纳法的难点是什么？

（b）如何比较归纳与推导的可靠性？

（c）统计学习理论如何涉及归纳的可靠性？

1.10 参考文献

近半个世纪以来，统计模式识别作为一个独特的领域一直很活跃，虽然其基础知识是范围更为宽泛的概率和统计。统计模式识别（或统计学习）是广泛

的机器学习领域的一部分，其跨越了许多学科，如数学、概率、统计、电气工程、计算机科学、认知科学、计量经济学和哲学。许多会议、期刊和书籍专题都在讨论机器学习，其中很多材料讨论统计学习。

Mitchell（1997）是第一个对机器学习一般问题进行讨论的学者。Vickers（2010）是最近深入讨论归纳问题的学者。下面的其他参考资料是讨论统计模式识别及相关领域的经典参考文献。

[1] Bishop C. Pattern recognition and machine learning. New York: Springer; 2006.

[2] Bongard M. Pattern recognition. Washington (DC): Spartan Books; 1970.

[3] Devijver PR, Kittler J. Pattern recognition: a statistical approach. Englewood Cliffs (NJ): Prentice-Hall; 1982.

[4] Devroye L, Györfi L, Lugosi G. A probabilistic theoryof pattern recognition. New York: Springer Verlag; 1996.

[5] Duda RO, Hart PE. Pattern classification and scene analysis. New York: Wiley; 1973.

[6] Duda RO, Hart PE, Stork DG. Pattern classification. 2nd ed. New York: Wiley; 2001.

[7] Fukunaga K. Introduction to statistical pattern recognition. 2nd ed. San Diego (CA): Academic Press; 1990.

[8] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer; 2009.

[9] Ho YC, Agrawala A. On pattern classification algorithms: introduction and survey. Proc IEEE 1968; 56: 2101-2114.

[10] Kulkarni SR, Lugosi G, Venkatesh S. Learning pattern classification-A survey. IEEE Trans Inf Theory 1998; 44 (6): 2178-2206.

[11] Mitchell T. Machine learning. Boston (MA): McGraw-Hill; 1997.

[12] Nilsson NJ. Learning machines. New York: McGraw-Hill; 1965.

[13] Schalkoff RJ. Pattern recognition: statistical, structural, and neural approaches. New York: Wiley; 1992.

[14] Theodoridis S, Koutroumbas K. Pattern recognition. 4th ed. Amsterdam: Academic Press; 2008.

[15] Theodoridis S, Pikrakis A, Koutroumbas K, Cavouras D. Introduction to pattern recognition: a matlab approach. Amsterdam: Academic Press; 2010.

[16] Vapnik VN. The nature of statistical learning theory. New York: Springer; 1999.

[17] Vickers J. The Problem of Induction, in The Stanford Encylopedia of Philosophy; 2010, http: //plato. stanford. edu/entries/induction-problem/.

[18] Watanabe MS. Knowing and guessing. New York: Wiley; 1969.