经典译丛

WILEY

人工智能与智能系统

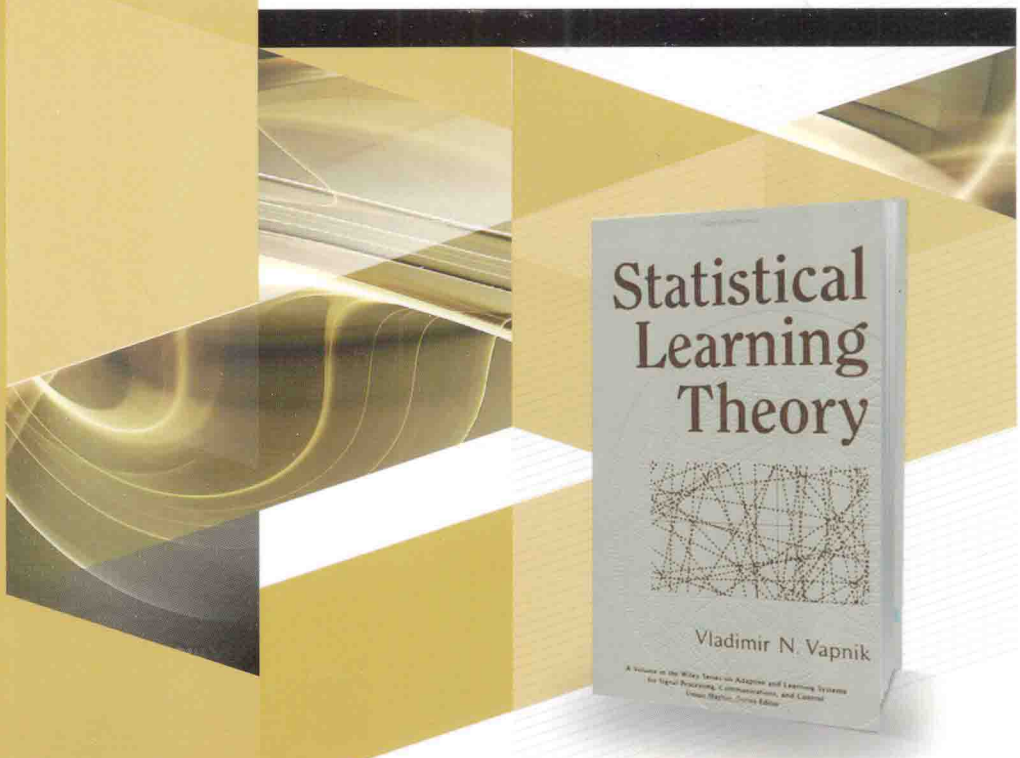
Statistical Learning Theory

统计学习理论

Statistical Learning Theory

【美】 Vladimir N. Vapnik 著

许建华 张学工 译



中国工信出版集团

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

经典译丛·人工智能与智能系统

统计学习理论

Statistical Learning Theory

[美] Vladimir N. Vapnik 著

许建华 张学工 译

電子工業出版社
Publishing House of Electronics Industry
北京·BEIJING

内 容 简 介

统计学习理论是研究利用经验数据进行机器学习的一种一般理论，属于计算机科学、模式识别和应用统计学相交叉与结合的范畴，其主要创立者是本书的作者 Vladimir N. Vapnik。统计学习理论的基本内容诞生于 20 世纪 60～70 年代，到 90 年代中期发展到比较成熟并受到世界机器学习界的广泛重视，其核心内容反映在 Vapnik 的两部重要著作中，本书即是其中一部，另一部是“The Nature of Statistical Learning Theory”（《统计学习理论的本质》）。由于较系统地考虑了有限样本的情况，统计学习理论与传统统计学理论相比有更好的实用性，在这一理论下发展出的支持向量机（SVM）方法以其有限样本下良好的推广能力而备受重视。

本书是对统计学习理论和支持向量机方法的全面、系统、详尽的阐述，是各领域中研究和应用机器学习理论与方法的科研工作者和研究生的重要参考资料。

版权贸易合同登记号　图字：01-2003-1537

图书在版编目（CIP）数据

统计学习理论/（美）瓦普尼克（Vapnik, V. N.）著；许建华，张学工译.
北京：电子工业出版社，2015.4
（经典译丛·人工智能与智能系统）
书名原文：Statistical Learning Theory
ISBN 978-7-121-25875-6

I. ①统… II. ①瓦… ②许… ③张… III. ①统计学-教材 IV. ①C8

中国版本图书馆 CIP 数据核字（2015）第 074542 号

策划编辑：马　岚
责任编辑：马　岚
印　　刷：三河市双峰印刷装订有限公司
装　　订：三河市双峰印刷装订有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编　100036
开　　本：787×1092　1/16　印张：36.5　字数：934 千字
版　　次：2015 年 4 月第 1 版
印　　次：2015 年 4 月第 1 次印刷
定　　价：99.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

再版译者序

从本书中译本2004年出版发行以来，得到了广大读者的喜爱和支持。十多年来，机器学习的理论和实践都得到了很大的发展，尤其是，最近几年有关大数据和深度学习等的研究，掀起了机器学习乃至整个人工智能领域的一轮新的高潮。

机器学习是从经验数据中发现规律即依赖关系的过程。Vapnik等发展起来的以学习过程一致性理论、VC维理论和结构风险最小化原则为核心的统计学习理论，奠定了对各种经验学习过程的理论性质研究的基础。统计学习理论从对一个学习过程随着样本数增加的收敛情况的理论分析出发，逐步得出了关于有限样本下机器学习推广能力的一系列理论结论，并在此基础上发展出了结构风险最小化的原则和支持向量机等方法，使得对机器学习的研究脱离了经验性、启发式的发展模式，有了一个严谨的理论框架。在大数据时代，很多问题中的训练样本数目越来越大，但对统计学习过程理论性质的研究仍然是机器学习方法研究的基础。另一方面，还有很多大数据问题，其数据之大主要来源于数据维数庞大，样本数目与数据维数和所研究问题的复杂度相比远达不到充分大。对于这种超高维小样本的大数据，对它们的机器学习研究，关于推广性的理论就变得更加重要。在这种背景下，机器学习领域的研究者和学生对《统计学习理论》这本经典理论著作有了更大的需求。适应这种需求，电子工业出版社决定继续出版这本经典之作，相信这对广大读者在新形势下开展深入的机器学习理论、方法和应用研究将是很大的帮助和促进。

清华大学　张学工

2015年4月

译 者 序

1996 年，我有机会了解到统计学习理论当时的重要进展以及当时刚刚出现的支持向量机（Support Vector Machine，SVM），并得到了 Vapnik 教授于 1995 年出版的重要著作“The Nature of Statistical Learning Theory”。认识到统计学习理论和支持向量机将是机器学习领域的重要方向，而国内多数研究者无法看到该领域的权威著作，于是 1997 年底前后我起意翻译该书。当时 Vapnik 正在撰写该书的第二版，我和他讨论之后决定直接翻译第二版，争取使其第二版的中、英文版同时面世。这本书就是 2000 年初出版的《统计学习理论的本质》（英文版于 1999 年底出版）。

正如书名所反映的那样，《统计学习理论的本质》一书以极其精练的篇幅、系统而扼要地阐述了统计学习理论的核心思想、结论和方法，但却没有包含对它们的证明或展开论述。《统计学习理论的本质》在我国关于统计学习理论和支持向量机的研究方面起到了预期的推动作用，但不少读者来信表示希望能学习更深入的内容，这也是我们在清华大学开设统计学习理论方面的研究生课程的一个体会。Vapnik 于 1998 年出版的“Statistical Learning Theory”正是一本满足这种要求的权威著作，但可惜国内很多读者也无法看到这本书。由于此书巨大的篇幅和本人精力和水平所限，一直未敢设想翻译它，毕竟，翻译不是我的本业。

2003 年初，电子工业出版社的编辑找到我，说他们决定要翻译出版这本书，并已办理了有关版权事宜，想请我来翻译或组织翻译。盛情之下难以推托，也为出版社对严肃学术著作的热心所感动，便答应下来组织其翻译工作。此时正值许建华博士与我合作进行了几年的统计学习理论方面的研究，精力上相对比我更有保证一些，便请他执笔翻译，于是才有了现在的中译本《统计学习理论》。其中，引言、第 1 章到第 3 章由我们二人共同翻译，中文版序言由我翻译，其余各章全部由许建华博士翻译。翻译初稿请正在清华学习我的“统计学习理论”课程的几位研究生（凡时财、黄河、徐云鹏等）进行了校对。翻译时参照的是原书英文版第 4 次重印版本。

由于英文版原书出版已经 5 年多了，其间虽然统计学习理论的核心内容并无太大改变，但毕竟这几年是统计学习理论从较少人注意到受到广泛重视的几年，想必作者对原书内容会有很多新的看法。于是我恳请 Vapnik 为本书中文版专门撰写一个序言，他欣然同意，在此序言中从哲学的高度对学习和推广的问题进行了阐述，提出了有关研究的新趋势和一些更深层次的问题，还探讨了东、西方文化的差异以及这种差异在经验推理中的反映，这使得中译本比原著增添了新的学术价值，也为我们开展更具创新性的研究指出了一个方向。我要代表广大中文读者向 Vapnik 表示特别感谢！

读者可能会提出关于《统计学习理论的本质》与《统计学习理论》两部著作的比较的问题。我想，简略介绍一下这两本书的写作过程或许会给读者提供一些信息。统计学习理论的基础思想和框架其实在 20 世纪 60 和 70 年代已经奠定，到 90 年代发展到比较完善，且产生出了支持向量机这一新的通用机器学习方法。此时，作为这一理论的主要完成者，Vapnik 教授开始动笔写一本全面反映统计学习理论的著作，即 1998 年出版的本书。在写作过程中，他感觉到有必要把统计学习理论的核心内容更精练地、同时也是更及时地介绍出来，因此他在该书写作期

间先写出了《统计学习理论的本质》一书，于 1995 年出版。由于统计学习理论和支持向量机在 20 世纪 90 年代中期的快速发展，1998 年，在《统计学习理论》完成之后，他感到有必要对《统计学习理论的本质》一书增添很多新内容，于是在 1999 年底出版了该书的第二版（中文版即对应这个版本），比第一版增添了 60%以上的新内容，其中部分内容在《统计学习理论》一书中也没有涉及到。我个人认为，《统计学习理论》与《统计学习理论的本质》基本上可以说是互补的，前者内容全面而深入，但也正因为如此，利用该书来把握统计学习理论的主线和核心内容或许需要较长的时间；而后者则提炼了前者的精华，适宜在较短时间内领略统计学习理论的脉络和本质，但由于篇幅所限，难以通过该书对统计学习理论有更深入的学习和研究。如果从作为教材的角度考虑，我想不妨用后者作为教材而用前者作为主要参考书，或者按照作者在序言中所说，把本书作为两门课的教材。

我们在统计学习理论方面的研究工作一直得到了国家自然科学基金的支持，本书的翻译工作也是如此，项目编号包括 69885004 和 60275007 等。

在中译本中，所有的人名都采用了原文，以便于读者根据人名（与时间）从参考文献中找到原出处。需要指出的是，本书内容涵盖广、深度大，由于我们的水平有限，可能会存在一些不确切甚至错误的译文，希望广大读者及时指正，以便我们在重印时更正，谢谢！

张学工

2003 年 12 月 26 日

于北京清华园

中文版序言[①]

从本书英文版出版到现在已经5年了。从技术的(数学的)角度来看,本书在内容上并没有太多值得增加的东西。然而,在这5年里,在统计学习理论方法的应用方面有了很大的进展,这些应用包括求解现实中的高维问题,理解该理论框架下发展出的方法所发挥的作用,以及创建新的关于推广性的哲学。

这些发展让人们从经验推理问题的更宽的视角来重新审视推广性问题。

推广性的两种模型:辨识和预测模型

推广性的问题已经有两千年历史了,它是学习的核心问题,也是关于自然科学的哲学的核心问题。现代对这一问题的研究始于20世纪前期,其间有两件重要的事情。K. Popper(1968)用不可证伪性的概念提出了他的关于归纳问题的理论,A. N. Kolmogorov(1933b)提出了概率论和统计学的公理体系。在这一公理体系中,有两种不同的推理的数学模型:演绎模型(概率的理论)和归纳模型(统计的理论)。从那时起,人们可以把统计学看成归纳推理的一个数学模型,其主要问题是:给定观测(数据)寻求推广(感兴趣的函数)。人们说明了,推广性的核心问题与用数据估计概率测度的某些性质的问题是相联系的。后来,在20世纪70年代早期,人们发现,对于很宽泛的一类归纳推理原则来说,有且仅有两种因素是影响推广性的,它们是:

1. 经验损失,它说明了被选中的函数多么忠实地刻画了观测。
2. 容量因素,它描述了从中选择解函数的函数集的多样性。

容量概念的引入是创立学习理论的主要工具。我们引入了3种主要的容量概念(VC熵、生长函数和VC维),它们描述了推广性的充分必要条件(不同的容量概念对应于学习问题的不同表示)。函数集VC维的概念可以看成是Popper的不可证伪性概念的数学体现。

在概率理论的公理化之后,紧接着,Glivenko,Cantelli和Kolmogorov证明了统计学主要的定理,这些定理指出,经验分布函数序列(随着用于估计的数据量的增加)收敛于真实分布函数(Glivenko-Cantelli定理),并且,这种收敛速率很快,具有指数型的速度(Kolmogorov界)。这些发现表明,归纳推理的一般问题是可以有一个有效的解的。

模型辨识

然而,对统计推理一般理论进行分析的进程由于R. Fisher的巨大影响而中断了将近30年。

---

① 此为原著作者 Vladimir N. Vapnik 专门为本书中译本出版而撰写的序言,反映了作者对统计学习理论的全面概括和最新见解,其中还特别谈到了推理与东、西方文化的问题。由于这是在原书出版时隔5年以后写的序言,可能个别用词上与正文并不严格一致,但根据上下文不难理解其含义。读者可能需要先阅读正文中的有关内容,然后才能更好地理解这里所谈到的某些内容。——译者注

Fisher的目标是建立用于分析简单实验的结果的工具("简单"是指可以用低维向量来描述)。为了达到这一目标,Fisher重新考虑了统计推理的一般问题。他把推理的问题简化为估计一个产生随机信号且属于一个已知函数族的密度函数的问题。他考虑的是下面的传统模型,即估计受(加性)噪声污染的信号:

- 信号是确定性和随机的两个成分之和。确定性部分是由某个函数的值定义的,这个函数除有限几个参数外是已知的,噪声部分是由一个已知密度函数定义的。Fisher把估计确定性部分的参数的问题看成是统计分析的目标。
- 为了估计这些参数,Fisher引入了最大似然方法。

因此,根据Fisher的观点,统计学的主要目标就是从一个给定的(简单)模型族中估计观测到的事件的模型。

从Fisher时代起,人们进行了很多努力来把Fisher的体系推广到能够包含"不太简单"的模型族,以及推广到不必给出噪声的准确模型。在这两个方向上都取得了一些重要的成果:

1. 人们不再只能考虑单个固定的噪声分布规律,而是可以考虑一系列候选的噪声分布模型(它们属于一个很宽的未知分布规律族)。这是鲁棒统计学的思想。
2. 人们也可以采用很宽的确定性函数集合,它们并不一定用参数模型来描述。这是非参数统计学的思想。

然而,由Fisher提出的核心思想仍然保持下来,这一核心思想就是估计产生信号的模型。因此,Fisher的统计学体系可以被称为"模型辨识"。模型估计的思想反映了传统的科学目标,这一目标在科学哲学中被描述为:发现存在的自然法则。

不适定问题

应该注意到,在20世纪30年代,当Fisher提出他的体系时,不适定问题的概念还没有发展起来,这是现代应用分析理论中最重要的概念之一。这一概念主要是在20世纪60年代发展起来的。不适定问题理论中的主要发现是这样一个事实,即存在很广泛的一类问题,它们的形式化的解存在,但却不能用有限的资源(计算资源或信息资源)得到。

事件模型的辨识问题属于不适定问题。这一点也是在快速计算机诞生的20世纪60年代被发现的。人们发现,Fisher的应用统计学对于解决高维问题来说不是一个合适的工具,因为它存在"维数灾难"的问题。

模型辨识体系属于不适定问题,这一事实决定了经典(Fisher的)统计学理论的特点:经典统计学理论的多数结果或者本质上是渐进的,或者需要很强的先验假设才能对有限的样本数目得到结果。

预测统计学的VC体系

在20世纪60年代后期,为了克服模式识别问题中的"维数灾难",Vapnik和Chervonenkis提出了一种不同的方法(VC理论),这实际上开始了一个新的称为"预测统计学"的体系。预测统计学的目标是很好地预测事件,但不一定通过对事件模型的辨识来进行预测。当然,如果我

们可以辨识事件的模型，自然可以用这个模型来很好地预测事件。但是，估计事件的模型的问题是困难的（不适定的），而寻找一个用于预测的规则的问题则简单许多（问题更适定）。很可能有这样的情况，有很多不同的规则可以对事件的结果进行很好的预测，而它们与事件的模型不同。尽管这些规则不能辨识出模型，但它们可以是很有用的预测工具。因此，利用有限数量的信息，我们经常可以找到一个能够预测得不错的规则，即使它并不能很好地估计事件的模型[①]。这就是在估计一个好的预测模型的问题中有可能解决“维数灾难”问题的原因[②]。

构造预测模型的 VC 理论是 Glivenko-Cantelli-Kolmogorov 这一系列对归纳的分析的延续[③]。这一理论的核心是一些定义函数集容量的新概念：函数集的 VC 熵、生长函数和 VC 维，它们描述了由给定数目的点定义的函数集的多样性。

在 VC 理论体系中有两点是非常重要的：

1. 对于很多经验推理原则来说，容量不但决定了可学习性的**充分**条件，而且也决定了其**必要**条件。因此，要构造预测学习方法，人们不可避免地要利用 VC 有关的概念来研究。
2. 容量的概念不直接取决于维数，而且根据 VC 界的一系列结论，推广性取决于容量而不是维数。

支持向量机

在 20 世纪 90 年代中期，我和我的同事们发现了如何在高维空间中有效地控制容量，并提出了支持向量机（SVM）这种构造预测规则的通用方法。使用支持向量机，人们可以在很高维的空间里构造好的预测规则[④]（下面将给出在高于 100 000 维的空间里这样一个规则的例子[⑤]）。在很大程度上，机器学习的成功得益于支持向量方法的贡献。这个方法是：

- **通用的**（能够在很广的各种函数集中构造函数）
- **鲁棒的**（不需要微调[⑥]）
- **有效的**（在解决实际问题中总是属于最好的方法之一，在很多问题上取得最好的历史记录）
- **计算简单的**（方法的实现只需要利用简单的优化技术[⑦]）

---

① 事实上，不适定问题的实质可以描述如下，即在所有能够很好地预测的规则中，寻找能够很好地描述模型的规则。如果我们的目标是预测，就不需要解决这一（困难的）辨识问题。

② 注意，解决“维数灾难”问题的法宝并不是对经典体系的更精细的分析（现有的分析已经很完美了），而是对问题的重新设置，这种新的问题提法比经典的提法要求低，它是建立在一种不同的哲学思路上的。在这种新的思路里，我们放弃了辨识模型这一目标。

③ 事实上，它是从求解一般的 Glivenko-Cantelli 问题开始的。

④ 在高维空间中定义一个低 VC 维的函数集是可能的。因为 VC 维（而不是维数）控制了推广性，我们可能克服“维数灾难”。

⑤ 但是本序言中并没有给出这样的例子。实际上这里可能是指在 USPS 手写数字识别中的例子，见 12.2 节。——译者注

⑥ 原文是 does not require fine tuning，作者没有展开描述其含义，译者理解是指 SVM 不需要针对具体问题做很多算法的调整。译者认为，与其他一些方法相比，SVM 的确表现出更好的对问题的适应性，但针对一个实际问题，仍需具体问题具体分析才能取得更好的效果。——译者注

⑦ SVM 的求解只涉及一个不等式约束的二次优化问题，但此处说只需要简单的优化技术，译者认为只是理论上正确。实际上，即使这样的二次优化问题在某些情况下（比如样本量较大时）也并不容易解决。事实上，SVM 的高效实现算法仍然是很多人研究的一个课题。——译者注

● **理论上完善的**(基于 VC 推广性理论的框架)

大约从 2000 年开始,SVM 方法成为了非常受欢迎的学习方法。

转导类型的推理

统计学习理论考虑两种不同类型的推理:归纳推理和转导推理[①]。

转导推理的目的是估计某一未知预测函数在给定兴趣点上的值(而不是在该函数的全部定义域上的值)。关键是,通过求解要求较低的问题,我们可以得到更精确的解[②]。我们发展出了一套关于转导推理的一般理论,它说明,转导推理的推广性的界要好于归纳推理的相应的界。

我们还看到,在解决实际问题中,转导推理比归纳推理有很大的优势。

例 1 为药物发现而预测分子的活性(Weston et al.,2001)

在 CUP-2001 数据分析方法竞赛中,要求构造用 DuPont 制药公司提供的数据来预测分子活性的规则。数据属于一个 139 351 维的二值空间,包括有 1909 个向量的训练集和有 634 个向量的测试集。

下表描述的是 Weston 等人的结果(Weston et al.,2001),他们没有参加那个竞赛,但后来把本书描述的算法(SVM)运用到这个数据集上。他们同时采用了归纳形式和转导形式。下表包括了竞赛优胜者的结果(共 119 组参赛者),SVM 归纳推理和 SVM 转导推理的结果。

竞赛优胜者的准确度	68.1%
归纳方式 SVM 的准确度	74.5%
转导方式 SVM 的准确度	82.3%

通过采用新的推理哲理(以转导代替归纳),所带来的性能提高比在构造归纳预测规则中采用新技术所带来的提高要大,这一点很值得注意。

例 2 文本分类(T. Jochim,1999)

在(T. Jochim,1999)描述的文本分类问题中,用转导推理代替归纳推理,把错误率从 30% 降低到了 15%。

发现转导推理和它相对于归纳推理的优势不仅仅是一个技术上的进步,而且是在推广性的哲理上的突破。

直到今天,传统的推理方法是**归纳 - 演绎**方法,人们首先用已有信息定义一个一般规则,然后用这个规则来推断所需要的答案。也就是说,首先从**特殊到一般**,然后从**一般到特殊**。

---

① 原文为 transductive inference。transduction 直译为"转导",在生物学中指信息从一个细胞传递给另一个细胞,或者指能量从一种形式转变为另一种形式。在统计学习理论中,transductive 推理是指从给定数据直接求得未知函数在某些感兴趣的点上的值,而不去求未知函数,读者可以参阅本书第 8 章有关内容及《统计学习理论的本质》一书第 9 章的讨论。在《统计学习理论的本质》中,我们把 transduction 译为转导推理,有时也称之为直推,在本书中我们沿用这种译法。——译者注

② 注意,转导推理相对于归纳推理的优势与预测推理相对于归纳推理的优势所基于的是同样的哲理,即,不要去求解比你所需要的要求更高的问题。

在转导模式中，我们进行直接的从**特殊到特殊**的推理，避免推理问题中的不适定部分（从特殊到一般的推理）。因此，在转导推理中，我们执行一种不直接依赖于推广性思想的经验推理。

在很多方面，转导推理与传统的科学哲学的主流思想有矛盾。在科学哲学中，发现一般的自然规律的问题只被看成是一个关于推理的科学问题，因为被发现的规律是允许“客观验证”的。在转导推理中，“客观验证”不是直截了当的。因此，在西方哲学的观点下，这种推理有其不好的方面[①]。

关于转导模式推理所存在的优势，以及两种归纳模式的存在（辨识模式和预测模式），相应的理论和实验证据形成了一种重要的推动力，即，要对传统科学哲学和关于学习和认知的基础重新进行审视。

简单世界与复杂世界

传统的科学哲学有一个很宏伟的目标：发现普遍的自然规律。这在一个简单世界中是可行的。简单世界是指可以只用几个变量描述的世界（比如物理学[②]）。

然而，这一目标在一个需要用很多变量来描述的复杂世界中不一定可行。在这样的世界中寻找规律，可能是一个不适定的问题。

因此，在一个复杂世界中，我们需要放弃寻找一般规律的目标，而考虑其他的目标。

复杂世界中推理的法则

在我 1995 年的“The Nature of Statistical Learning Theory”一书[③]中，对复杂世界中的推理提出了如下法则：

> 在解决一个感兴趣的问题时，不要把解决一个更一般性的问题作为一个中间步骤。要试图得到所需要的答案，而不是更一般的答案。很可能你拥有足够的信息来很好地解决一个感兴趣的特定问题，但却没有足够的信息来解决一个一般性的问题。

这些法则告诫人们要避免某些类型的推理，例如：

- 如果目标是估计函数，就不要去估计密度（即，要采用预测模型而不要采用辨识模型）。
- 如果目标是估计一些给定的兴趣点上的值，就不要估计函数（即，要采用转导推理而不要采用归纳推理）。
- 如果目标是选择几个最好的代表，就不要估计函数在点上的值（即，要采用选择而不要采用转导推理；选择问题将在下面讨论）。

---

① 由于转导推理是从特殊到特殊的推理，而不是对一般性的推理，因此难以直接进行客观验证，因此按照传统西方哲学的观点，这种推理是受到质疑的。——译者注

② 回顾 L. Landau 的话：“用 4 个变量，我可以解释几乎任意的物理现象”。

③ 《统计学习理论的本质》，第一版 1995 年出版，第二版 1999 年出版，2000 年由清华大学出版社出版第二版的中译本，张学工译。——译者注

这些法则构成了关于简单世界和复杂世界的科学哲学的方法论差别。因此，在对待一个复杂世界时，主要的问题就是确定关键的且有适定解的低要求问题，并寻找方法来解决这些问题。

传统自然哲学（简单世界哲学）已经发展了很多世纪，它深深地植根于物理学的成功中。复杂世界哲学则刚刚开始发展，它的灵感来自于对机器学习问题的理论分析和相应的计算机实验[①]。

关于高维学习问题的推广能力的最初分析，引发了针对复杂世界的对传统科学哲学的修正。尤其是：

- 业已证明，Occam 剃刀原则“实体不应超出需要地增加”不能作为关于归纳推理的一般原则。推广能力取决于容量因素，而不是实体数目（空间维数）。一些表现出最好的推广性的方法，比如 SVM，升压助推（Boosting），神经网络等，都通过增加实体数目来取得好的推广性。
- 转导推理被证明比归纳 - 演绎推理更准确。
- 有例子表明，为了良好的行动，并不一定需要预测良好。

因此，随着推广性理论的发展，我们清楚地看到，在对自然的分析中有两种目标：

1. 要理解我们的世界的模型（简单世界的科学的目标）。
2. 要在这个世界中行动良好（复杂世界的科学的目标）。

根据一个人认为自然规律有多么复杂，他需要选择传统体系或者选择新体系[②]。

要发展这一新体系，除了数学分析之外，很重要的是考察 3 种不同类型的推理在哲学上的或文化上的依据，并且在研究方法和经验推理的理论时把这些依据考虑在内。

选择推理的问题

更严格地遵循上述法则，就得出了选择推理问题，它比转导推理的要求更低（因此可以更精确地解决）。下面，我们讨论选择推理问题的提出，并且给出一个需要选择推理的问题的例子[③]：

1. 转导选择问题 给定训练样本(x_i, y_i)，$x_i \in R^n$，$y_i \in \{-1, +1\}$，$i = 1, \cdots, \ell$，并给定一个工作集 x_j^*，$j = 1, \cdots, m$，在工作集中寻找 k 个以最大概率属于第一类的元素。例如：

---

① 在这一哲学中，计算机实验扮演着与遗传学中的果蝇实验同样的角色。

② T. Hestie, R. Tibshirani, and J. Friedman, *The Elements of Statistical learning: Data Mining, Inference, and Predication*, New York, Springer Verlag, 2001.
（本书中译本《统计学习基础——数据挖掘、推理与预测》已由电子工业出版社于 2004 年 1 月出版。——编者注）书中有个表考虑了很宽范围的数据挖掘算法及其特性。在这些算法的特性中，有下列两个特性：“推广的质量”和“可解释性”。从该表中可以立刻看到预测良好的算法没有良好的“可解释性”，反之，具有良好的可解释性的算法不能很好地预测。

③ V. Vapnik, *Estimation of Dependencies Based on Empirical Data*, Springer Verlag, 1982. 本书是我 1979 年的专著的英文译本，选择推理是在这本著作里提出的。

- **药物发现**。在这个问题中，已知一些有效药物的样本$(x_i, +1)$和无效药物的样本$(x_s, -1)$，在给定的候选样本$(x_j^*, \cdots, x_m^*)$中寻找某一确定数目的候选者，设该数目为 k，它们属于有效药物一类的概率是最大的。
- **国家安全**。在这个问题中，已知一些良民的例子(描述)$(x_i, +1)$和恐怖分子的例子$(x_s, +1)$，在给定的候选者$(x_j^*, \cdots, x_m^*)$中寻找某一确定数目的人，设该数目为 k，他们属于恐怖分子一类的概率是最大的。

注意，与一般的转导推理相对比，这种问题的提法不要求对所有候选样本的分类。如果这样分类，最大的困难是对“边界候选样本”的分类。在选择问题中，我们避免了这一任务中最困难的部分(即对边界处的样本的分类)。这里，我们再次得到了用预测策略替代模型辨识策略以及用转导推理替代预测推理所带来的好处，即：我们把一个不十分适定的问题替换为一个更适定的问题。

经验推理的问题与人类文化

在 20 世纪 60 年代快速计算机的出现，改变了两千年来关于推理的哲学的方法论，使之从纯粹的思索变为自然科学的方法。现在，任何关于推理的思想都可以用计算机实验进行验证，以评估其推广的质量。

最初的此类实验即表明，在主流哲学中有些东西是有偏差的，这就是对转导类型的推理的过低估计，事实上它可以比归纳更有效。对最简单的推理模型的数学分析支持了这一结论。

复杂世界的法则指出了以下几点：

- 辨识事件模型作为推理的一般路线是有局限性的。
- 认识到有多种特殊类型的推理(预测推理、转导推理和选择推理，可能还有其他多种推理)是很重要的。在这些推理中，由于采用简化的问题设置和放弃对一般性问题的求解，得到了更好的推理准确度。
- 新类型的推理没有直接的客观验证。

这些认识要求我们从一个新的角度来重新审视所继承的关于推理的哲学和文化遗产。在这个传统中，以前也存在着对推理的直接过程的讨论。在很多情况下，这一过程基于某种(可能是神学的)宇宙(它有充分的“延展结构”)在现实中的投影，以及寻求答案的某些步骤。

有很多文化大量使用直接推理的因素[①] 来得到结果，这种推理在经典意义上是非科学的，这样取得的结果并不比用科学推理所得到的结果更不精确[②]。这样就产生了一个问题：什么情况下(在经典意义上)“非科学”的推理方式更有效？

---

① 对可用信息的非传统形式的分析，与断言存在特殊的信息是非常不同的，区分二者非常重要。这里我们只讨论非传统形式的信息分析。

② 在 *The Feyman Lectures on Physics*(1975)中，Feynman 写到：“我们必须从一开始就搞清楚，如果一件事情不是科学的，它并不一定是坏的……因此，如果某件事被说成不是科学的，并不意味着说有什么地方不对了，这只是说它不是一种科学”。

近来在文献中开始了一场关于在西方世界和东方世界中不同的思维方式的讨论[①]。西方世界是基于归纳类型的推理,东方世界是基于转导类型的推理。这里是 Nisbett 在他的网页上给出的几段话:

"更多关于推理的近期工作对东亚人和西方人进行了比较。很早已经有人主张,西方人解析地推理,即,他们注重对象(无论是物理的还是社会的)及其特性,用特性来分类对象,并且用基于这种分类的规则来预测和解释对象的行为。形式逻辑在推理、类别构建和规则验证中发挥了作用。"

"与此形成对比,东亚人整体地推理,即,他们注重在其所处环境中的对象,很少关心类别或普适规则,基于在特定时刻施加于对象个体上的各种作用来解释其行为。没有太多地采用形式逻辑,而常常采用各种辩证推理规则,包括综合、超越和归一。"

"来自我们的实验室的最新证据支持了这些观点。西方人把注意力集中到对象上,经常看不到在刺激领域内的相互作用,往往用假定的秉性来解释对象的行为(而这样经常是错误的)。他们还在归纳推理中大量使用类别,轻易地学习类别,利用(有时是误用)形式逻辑规则来推理。东亚人把他们的注意力集中在对象所处的刺激领域上,他们对相互作用敏感,倾向于用领域中的条件或状况来解释对象的行为。相对来说,他们较少使用类别来归纳,他们发现类别学习相对比较困难,经常利用(有时是误用)各种辩证策略。"

经验推理的模型

当人们研究现实生活现象时,必须:

1. 引入现象的模型。
2. 用严格的(如果可能,应该是数学的)工具来分析这些模型。
3. 对分析的结果给出一种解释。

在上述这三项中,有两项不是数学研究的课题(而是模型和对分析结果的解释)。它们与哲学、现有经验和文化传统有关。

在本书中,我们采用一个简单模型并对它进行严格的数学分析。然而,对这种分析的结果的解释(比如转导推理的解释),使我们超出了标准的关于推广性的哲学,而发展到更一般的关于经验推理的哲学。新的关于推理的哲学(而不是数学)形成了经验推理研究发展的主要推动力。

在本书中,我们把经验风险最小化归纳原则(以及作为它的推广的结构风险最小化原则)看成是不证自明的。这些原则可能概括了关于推理的一般思想。然而,很有可能存在更好的归纳原则,它们可能考虑一些会带来更好结果的特殊细节(例如像邻域风险最小化原则中那样考虑输入空间的拓扑结构,参见《统计学习理论的本质》一书)。

本书中考虑的主要算法工具包含了两个主要思想:输入空间中函数集的线性参数化和通过控制大间隔来控制容量因素。很可能存在比大间隔更好的因素,它对线性参数化函数能够更好地控制推广能力。

---

① Nisbett, *The Geography of Thought, How Asians and Westerns Think Differently — and Why?*, Nichols Breadley Publishing, 2003.

但是，我们可以期望的真正的进展来自于对经验推理的哲学的改变。到现在为止，我们知道几种推理问题：归纳、转导、选择等，还有别的吗？我们用很少几个推理原则来研究这些问题，包括经验风险最小化原则以及它的推广（即结构风险最小化）。我们还研究了邻域风险最小化。还有别的吗？我们通过严格的分析定义了大间隔因素，它应该被最大化，以在（特征空间中的）线性参数化函数集上得到合适的推理。有什么能比线性化函数集中的大间隔更好吗？

我想，对经验推理的研究刚刚开始，在这一研究中的一个很大的挑战是能够通过严格的数学分析对经验推理的主要法则给出开放的（不含偏见的）实现。

致谢

我非常高兴地获悉，在我的“The Nature of Statistical Learning Theory”一书的中译本《统计学习理论的本质》出版后，“Statistical Learning Theory”一书也将在中国翻译出版。前者呈现了关于学习和推广性问题的主要思想，但没有给出证明；后者是一个内容上更深入的版本，其中包含了所有证明。我希望这两本书的中文版的出版不仅推动技术上的发展，而且能够推动关于经验推理分析的概念上的进步，进而产生出反映丰富的中国哲学文化传统的推理模型。

非常感谢张学工教授为翻译这两部著作所做的杰出工作。我希望这些译本能够促进对经验推理问题的研究，而这种研究实际上是在试图回答已经有两千年历史的年轻问题[①]：

人类智慧的基础是什么？

我在此谨祝中国的研究者们工作顺利！

Vladimir N. Vapnik

2003 年 9 月 29 日

① 由于 Vapnik 的原稿中的录入错误，导致本书第一次印刷中此处翻译有误，后经译者与 Vapnik 本人当面核实，此处作者的意思是这个问题虽然已经研究了两千多年，但对人类来说仍然是一个年轻的问题。——译者注

序　言

本书专门讨论统计学习理论，该理论研究从给定数据集中估计函数依赖关系的方法。这是一个非常普遍的问题，涵盖了经典统计学的若干重要论题，特别是判别分析、回归分析和密度估计问题。

在本书中，我们重点考察解决这类问题的新体系：过去 30 年里发展起来的被称为学习的体系。与针对大数据样本集发展起来的统计学和基于各种先验信息的统计学相比，我们的新理论是专门针对小数据样本集发展起来的，并且不依赖于对所解问题的先验知识，而是只考虑学习机器所实现的函数集的一种结构（函数的嵌套子集的集合），并且在结构上定义了一种子集容量的特定度量。

为了控制这种新体系框架下学习机器的推广性，必须考虑两个因素，一是所选函数对给定数据的逼近程度，二是从中选取逼近函数的函数子集的容量大小。

本书介绍这一类推理方法（学习过程）的综合性研究成果，具体包括：

- 通用的定性理论，包含学习过程一致性的充分必要条件
- 通用的定量理论，包含学习过程收敛速率（推广速率）的界
- 小数据集函数估计的原则，它们的基础就是所提出的新理论
- 函数估计方法及其在实际问题中的应用，这些方法的基础是上述的原则

本书由三部分组成：学习和推广性理论、函数的支持向量估计和学习理论的统计学基础。

在第一部分中，分析了与推广性相关的因素，并说明如何控制这些因素，以获得好的推广性。

第一部分包括 8 章内容。第 1 章介绍了学习问题的两种不同的研究方法。第一种方法将学习表达成最小化期望风险泛函问题，前提条件是未知定义风险的概率测度但给出了独立同分布（i.i.d）观测样本。为了得到这一方法框架下问题的解，人们必须给出一定的归纳原则，即，定义这样一个构造性的泛函（代替期望风险泛函），最小化这一泛函即可找到一个确保期望损失较小的函数。第二种方法将学习表达成所求函数的辨识问题，即利用观测数据，找到一个与所求函数相近的函数。一般而言，这一方法必须求解所谓的不适定问题。

第 2 章讨论学习理论的主要问题与统计学基础问题之间的联系，即从数据估计概率测度的问题。这一章介绍了两种估计概率测度的方法。一种方法基于概率测度估计的弱方式收敛，另一种方法基于强方式收敛。这两种估计未知测度的方法对应着第 1 章中学习问题的两种研究方法。

第 3 章专门讨论学习过程的定性模型，即基于经验风险最小化归纳原则的学习过程一致性理论。这一章将证明，对于基于这一归纳原则的学习过程一致性问题，某些经验过程的收敛性是充分必要的（即存在着一致大数定律）。第 3 章中主要讨论了这些条件。相应的定理将在本书的第三部分给予证明。

第4章和第5章估计经验过程收敛速率的界。利用这些界，对于最小化经验风险泛函的函数，我们得出了风险的界。在第4章中，我们得到了指示函数集(对应于模式识别问题)的界。在第5章中，我们将这些界推广到实函数集(对应于回归估计问题)。界与两个因素有关：经验风险值和最小化经验风险的函数所在的函数集的容量。

第6章介绍一种新的归纳原则，即所谓的结构风险最小化原则，它最小化第4章和第5章中介绍的与经验风险值和容量这两个因素有关的界。这一原则能够使我们找到一个函数，它用有限数量的观测样本达到期望风险所保证的最小点。

第7章专门讨论求解随机不适定问题，包括密度估计、条件密度估计和条件概率估计问题。为了解这些估计问题，我们采用正则化方法(其思路与结构风险最小化原则基本相同)。利用正则化方法，可以得到求解问题的经典方法和新方法。

在第8章中，我们考虑一种新的学习问题表达方式，介绍估计给定点上函数值的问题。对于有限的经验数据，估计给定点上函数值的直接方法的推广性要比估计函数的方法的推广性好。因此，我们考虑直接估计给定点上函数值的方法，它们不是建立在估计函数依赖关系的基础上的。

本书的第二部分介绍了一类方法，即，在从有限数据集估计多维函数时具有较好推广性的方法。

这一部分包括5章。第9章介绍经典的算法：感知器、神经网络和径向基函数。

第10章到第13章专门讨论求解依赖性估计问题的新方法，即所谓的支持向量方法。第10章讨论估计指示函数(对应于模式识别问题)的支持向量机。第11章讨论估计实函数的支持向量机。

第12章和第13章讨论用支持向量机解决实际问题。第12章讨论模式识别问题。第13章讨论各种实函数估计问题，如函数逼近、回归估计和解反问题。

本书的第三部分研究一致大数定律，它使学习机器具有推广性。

这一部分包括3章，每一章研究不同的经验过程：在给定事件集合上频率一致收敛于概率(见第14章)、在给定函数集上均值一致收敛于期望(见第15章)、在给定函数集上均值一致单边收敛于期望(见第16章)。这些过程的收敛性构成了学习过程理论和理论统计学的基础。

本书在最后部分给出了参考文献、发展历史和一般性的评述，反映了作者对统计学习理论和相关学科发展过程的观点。

本书的前两部分按照作为统计学、数学、工程学、物理学、计算机科学研究生“学习理论”课程的教材的层次撰写，也可以供工程师了解学习理论或用做解决实际问题的新方法。第三部分以更高的层次撰写，可以作为数学和统计学博士生“经验过程”专业课的教材。

由于AT&T贝尔实验室自适应系统研究部主任Larry Jackel和AT&T研究实验室图像处理研究部主任Yann LeCun的支持，才使本书的出版成为可能。与同事们的合作促进了本书的完成和出版，他们是Youshua Bengio, Bernhard Boser, Leon Bottou, Chris Burges, Eric Cosatto, John Denker, Harris Drucker, Alexander Gammerman, Hans Peter Craft, Isabella Guyon, Patrick Haffner, Martin Hasler, Larry Jackel, Yann LeCun, Esther Levin, Robert Lyons, Nada Matic, Craig Nohl, Edwin Pednault, Edward Sackinger, Bernard Schölkopf, Alex Smola, Patrice Simard, Sara Solla, Vladimir Vovk以及Chris Watkins等。

我还与Leo Breiman, Jerry Friedman, Federico Girosi, Tomaso Poggio, Yakov Kogan以及Alexander