Vladimir N. Vapnik 著

统计学习理论的本质

张学工　译

清 华 大 学 出 版 社
北 京

内 容 简 介

统计学习理论是针对小样本情况研究统计学习规律的理论，是传统统计学的重要发展和补充，为研究有限样本情况下机器学习的理论和方法提供了理论框架，其核心思想是通过控制学习机器的容量实现对推广能力的控制。在这一理论中发展出的支持向量机方法是一种新的通用学习机器，较以往方法表现出很多理论和实践上的优势。本书是该领域的权威著作，着重介绍了统计学习理论和支持向量机的关键思想、结论和方法，以及该领域的最新进展。本书的读者对象是在信息科学领域或数学领域从事有关机器学习和函数估计研究的学者和科技人员，也可作为模式识别、信息处理、人工智能、统计学等专业的研究生教材。

图书在版编目(CIP)数据

统计学习理论的本质/(美)瓦普尼克著；张学工译. —北京：清华大学出版社，2000
ISBN 7-302-03964-X

Ⅰ. 统… Ⅱ. ① 瓦… ② 张… Ⅲ. 统计学-研究 Ⅳ. C8

中国版本图书馆 CIP 数据核字(2000)第 35147 号

出 版 者：清华大学出版社　　地　　址：北京清华大学学研大厦
http://www.tup.com.cn　　邮　　编：100084
社 总 机：010-62770175　　客户服务：010-62776969
印 装 者：北京密云胶刷厂
发 行 者：新华书店总店北京发行所
开　　本：185×230　印张：15.5　字数：351 千字
版　　次：2000 年 9 月第 1 版　2004 年 6 月第 2 次印刷
书　　号：ISBN 7-302-03964-X/O·244
印　　数：4000～5000
定　　价：27.00 元

纪念我的母亲

——Vladimir N. Vapnik

译　　序

人的智慧中一个很重要的方面是从实例学习的能力,通过对已知事实的分析总结出规律,预测不能直接观测的事实。在这种学习中,重要的是要能够举一反三,即利用学习得到的规律,不但可以较好地解释已知的实例,而且能够对未来的现象或无法观测的现象做出正确的预测和判断。我们把这种能力叫做推广能力。

在人们对机器智能的研究中,希望能够用机器(计算机)来模拟这种学习能力,这就是我们所说的基于数据的机器学习问题,或者简单地称作机器学习问题。我们的目的是,设计某种(某些)方法,使之能够通过对已知数据的学习,找到数据内在的相互依赖关系,从而对未知数据进行预测或对其性质进行判断。同样,在这里,我们最关心的仍是推广能力问题。

统计学在解决机器学习问题中起着基础性的作用。但是,传统的统计学所研究的主要是渐近理论,即当样本趋向于无穷多时的统计性质。在现实的问题中,我们所面对的样本数目通常是有限的,有时还十分有限。虽然人们实际上一直知道这一点,但传统上仍以样本数目无穷多为假设来推导各种算法,希望这样得到的算法在样本较少时也能有较好的(至少是可接受的)表现。然而,相反的情况是很容易出现的。其中,近年来经常可以听到人们谈论的所谓神经网络过学习问题就是一个典型的代表:当样本数有限时,本来很不错的一个学习机器却可能表现出很差的推广能力。

人们对于解决此类问题的努力实际上一直在进行。但是,其中多数工作集中在对已有(基于传统统计学原则的)方法的改进和修正,或者利用启发式方法设计某些巧妙的算法。

在人类即将迈进一个新世纪的时候,人们开始逐渐频繁地接触到一个词,就是"统计学习理论"。这实际上是早在 20 世纪 70 年代就已经建立了其基本体系的一门理论,它系统地研究了机器学习的问题,尤其是有限样本情况下的统计学习问题。在 90 年代,这一理论框架下产生出了"支持向量机(SVM)"这一新的通用机器学习方法。或许是由于统计学习理论为人们系统研究有限样本情况下机器学习问题提供了有力的理论基础,或许更是因为在这一基础上的支持向量机方法所表现出的令人向往的优良特性,人们开始迅速重视起这一早在 20 年前就该重视的学术方向。

现在,越来越多的学者认为,关于统计学习理论和支持向量机的研究,将很快出现像在 80 年代后期人工神经网络研究那样的飞速发展阶段。然而,所不同的是,统计学习理论有完备的理论基础和严格的理论体系(相比之下神经网络有更多的启发式成分),而且其出发点是更符合实际情况的有限样本假设,因此,我们期望,统计学习理论的这个研究热潮将持续更长久,而且将在人们关于机器智能的研究中做出影响更深远的贡献。

本书的作者,Vladimir N. Vapnik 博士,就是统计学习理论的创立者之一,也是支持向量机方法的主要发明者。正如其书名所反映的那样,本书以最精炼的文字展示了统计学

习理论最核心的内容和思想,当然也包括在其中占重要位置的支持向量机的核心方法与思想,以及此领域最新的研究成果。

本书第一版于 1995 年出版。它为推动统计学习理论和支持向量机的研究起到了十分关键的作用,几乎成为此领域所有重要研究都必定引用的文献。从那时到现在,这一领域的发展是空前的。在 1998 年初我与 Vapnik 博士通信时,他告诉我,他将为在 1999 年出版的第二版增加一章约 40 页的新内容;而在 1999 年底我拿到最后定稿的第二版时却发现是增加了三章 120 多页的新内容,比第一版内容增加达 60% 之多。只从这一侧面,我们也可以看出这是一个正在快速发展的学科。

关于本书的主要内容和特点,我想,用 Springer 出版本书时在其封底上所写的简介作介绍是最合适的:

本书的目标是讨论关于学习和推广性的统计理论中的基本思想。它把学习问题看作是一个基于经验数据进行函数估计的一般问题。在本书中,作者省略了一些证明和技术细节,而把重点集中在对学习理论的主要结论的讨论,以及它们与统计学中的基本问题的联系。主要内容包括:

- 在基于经验数据最小化风险泛函的模型基础上对学习问题的表示。
- 对经验风险最小化原则的深入分析,包括其一致性的充分必要条件。
- 用经验风险最小化原则得到的风险的非渐近界。
- 在这些界的基础上,控制小样本学习机器的推广能力的原则。
- 支持向量机方法,它在用小样本估计函数时能够控制推广能力。

本书第二版包括了三章新内容,专门介绍了学习理论和 SVM 技术的新进展。内容包括:

- 用于实值函数估计的 SVM。
- 基于求解多维积分方程的直接学习方法。
- 经验风险最小化原则的一种扩展。

本书在写作风格上力求可读性与简明化,面向的读者包括统计学家、数学家、物理学家和计算机科学家。

Vladimir N. Vapnik 是 AT&T 实验室研究中心的技术领导,同时是伦敦大学教授。他是统计学习理论的创立人之一,SVM 的发明者。他已经用英文、俄文、德文和中文出版了 7 部专著。

初次接触到这本书,难免会让人感觉到有些数学的味道。这也是统计学习理论的一个基本特点,即希望能通过严格的数学推理找到机器学习问题的关键所在。但是,正如作者在第一版前言中所说的,书中略去了大量艰涩的数学推导,而只重点介绍其核心思想、结论与方法,学习本书并不需要专门的数学基础。从工程的角度看,本书有其不可多得的实用价值。其中关于学习过程一致性的原理,关于推广性的理论结果,关于控制学习机器容量(复杂度)的思想和方法,关于构造具有良好推广性的机器的原则与方法,关于支持向量机的思路、方法与应用,以及更深入的关于求解不适定积分方程的思想和关于局部化函数估计的思想,都将对我们更好地解决实际问题产生重要的指导作用。因此,作为一名从事模式识别方法与应用研究的青年学者,我认为,不论我们的目标是解决模式识别、函数估

计等领域的一系列实际问题,还是对机器学习的本质问题进行探讨,甚至是对理论或应用统计学本身进行研究,本书都是一个十分重要的思想来源。

在 1998 年我与作者探讨出版本书中译本的时候,国内在此领域尚很少有人研究(在 1988 年边肇祺等编写的《模式识别》教材中对统计学习理论当时已经取得的成果有较系统的介绍)。我们很高兴地看到,现在,国内同行已经开始注意到这一研究领域并积极开始自己的工作。出版本书中译本的目的,就是希望能为国内广大学者和研究生从事有关研究提供一本权威、系统的参考资料,为我国的机器学习理论、方法与应用研究做出应有的贡献。

在我本人学习统计学习理论的过程中,深深为 Vapnik 博士几十年来坚持在这个当时并未受到应有重视的方向上进行深入研究的精神所感动。我也希望能与本书的读者一起学习这种持之以恒的科学精神和严谨、系统的学术作风。

在本书的翻译中,我力求忠实、准确地反映原著的内容,同时也力求保留原著的风格。但由于本人水平有限,尤其是数学基础不甚扎实,因此难免会有错误和不准确之处,敬请广大同行读者不吝赐教。同时,对原著中的部分内容,译者根据自己的理解在必要时加入了一定的译注,其中多数内容并未与作者本人协商,错误和不当之处请读者一并指正。

致 谢

在这里,我首先要感谢 Vladimir N. Vapnik 博士同意我翻译出版这本重要著作,并感谢他在百忙之中先后给我寄来本书第一版和第二版的样书,以及他中间几次改稿的手稿。还要感谢他亲自帮助与 Springer 出版社联系。

感谢李衍达院士对我翻译本书和进行该领域研究给予的鼓励和支持。感谢边肇祺教授对出版本书中译本的大力推动和帮助。在翻译过程中,我在很多问题上先后向阎平凡教授、周杰副教授、季梁教授等求教,阎辉同学认真阅读了译文初稿并提出了一些宝贵的意见,在此一并表示感谢。在此我还要专门对邵旭辉博士表示感谢,因为最早正是他的帮助,才使我有机会接触到本领域的一些重要的著作,其中包括本书第一版的手稿。

本书翻译得到了国家自然科学基金的资助(项目编号 69885004)。

张学工

1999 年 12 月

北京 清华园

第二版前言

从本书第一版出版到现在已经过去 4 年了。这些年是统计推断中源于学习理论的新方法快速发展的一个阶段。

在这段时间里,出现了新的函数估计方法。这些新方法中,对高维未知函数并不总需要大量观测才能得到一个好的估计。这些新方法用容量因子来控制其推广性,所用的容量因子并不一定依赖于空间的维数。

在 VC 理论中知道这些因子已经很多年了。然而只是在最近,在支持向量机(SVM)出现之后,容量控制的实际重要性才变得明朗。在传统统计方法中,为了控制方法的性能,人们设法降低特征空间的维数;与此形成对照的是,SVM 方法却大大地增加维数,而依靠所谓的大间隔因子来控制方法的性能。

在本书第一版中,我们介绍了包括 SVM 方法在内的一般的学习理论。当时,SVM 学习方法是崭新的,其中某些方法是首次被介绍。现在,不论是在学习的理论还是其应用中,SVM 间隔控制方法已经成为当前最重要的研究方向之一。

本书的第二版增加了关于 SVM 方法的三章新内容,其中包括把 SVM 方法推广到用于估计实值函数,基于(用 SVM 方法)求解多维积分方程的直接方法,以及经验风险最小化原则的扩展及其在 SVM 中的应用。

从本书第一版的出版到现在的这段时光也改变了我们在对归纳问题之本质的理解中的一般思想。在 SVM 很多成功的实验之后,学者们变得更决心要在 Occam 剃刀原则[①]基础上对传统的推广性思想进行批评。

这种理智的决心本身也是一个非常重要的科学成就。注意,这些新的推理方法应该早在 20 世纪 70 年代[②]就诞生了:那时我们已经知道了统计学习理论和 SVM 算法的所有必要元素。但是,人们做出这一理智的决断却用了 25 年的时间。

现在,从纯理论问题出发对推广性的分析已经变成了一个非常实际的研究课题,而这一事实又为本书第一版所描述的计算机学习问题研究的一般情景增添了新的重要的细节。

Red Bank, New Jersey　　Vladimir N. Vapnik

1999 年 8 月

---

① Occam's razor,亦称 Ockham's razor,是科学和哲学中的一个原则,表明实体不应该被没有必要地增加。这一原则一般被解释为:在两个或多个相互竞争的理论中,最简单的一个是最可取的;对未知现象的解释应该首先尝试用已经知道的东西来进行。这一原则也称作吝啬定律(law of parsimony),是由英国经院哲学家 William of Ockham 提出的。译者认为,作者在这里说基于此原则对传统的推广性思想进行批判研究,指的是要通过控制学习机器的复杂性(容量)实现对推广性的控制。——译者

② 本书提及的年代多数均为 20 世纪中的年代,如 70 年代即指 1970 年—1979 年。为了叙述简练,我们后面一般不再一一注明是 20 世纪。——译者

第一版前言

在 1960 年至 1980 年间,统计学领域出现了一场革命:起源于 20 世纪二三十年代的 Fisher 理论体系被一种新的体系取代了。这个新的体系对一个基本的问题给出了新的回答,这个基本问题就是:

对于一种未知的依赖关系,为了以观测为基础对它进行估计,人们必须对这种依赖关系先验地知道些什么?

在 Fisher 的体系中,对这个问题的回答是有很强的限制性的:我们几乎必须知道一切。具体说就是,对于这个待求的依赖关系,除了有限个参数的取值外,我们必须知道它的所有其他信息。在这个体系中,依赖关系的估计问题被认为就是对这些未知参数取值的估计。

新的体系克服了这种苛刻的限制。它指出,要从观测数据对依赖关系进行估计,只要知道未知依赖关系所属的函数集的某些一般的性质就足够了。

这种新理论所研究的主题包括:确定在什么一般条件下可能对未知依赖关系进行估计,阐述能够指导人们找到对未知依赖关系的最佳估计的原则(归纳原则),并且最终发展出实现这些原则的有效算法。

引导这一革命的是 60 年代的四项发现:

(1) Tikhonov, Ivanov 和 Phillips 发现的关于解决不适定问题的正则化原则;

(2) Parzen, Rosenblatt 和 Chentsov 发现的非参数统计学;

(3) Vapnik 和 Chervonenkis 发现的在泛函数空间的大数定律以及它与学习过程的关系;

(4) Kolmogorov, Solomonoff 和 Chaitin 发现的算法复杂性及其与归纳推理的关系。

这四项发现也成为人们对学习过程的研究取得新进展的一个重要基础。

学习的问题是一个非常一般性的问题,在统计科学中研究的几乎所有问题都可以在学习理论中找到其对应。而且,一些十分重要的一般性结论也是首先在学习理论的范畴内被发现,然后再用统计学术语重新进行表达的。

尤其需要指出的是,学习理论[①]第一次强调了所谓小样本统计学的问题。研究表明,对很多函数的估计问题,通过考虑样本的数量,我们可以得到比基于传统统计技术的方法更好的解。

在这一新的理论框架中的小样本统计学,无论是在统计学习理论中,还是在理论和应用统计学中,都形成了一个前沿的研究方向。新理论框架下提出的统计推断规则,不但应该满足已有的渐近条件,而且应该保证在有限的可用信息情况下取得最好的结果。这一理论在各种统计问题中产生了一系列新的推理方法。

---

① 这里的“学习理论”并不是泛指一般的关于学习的理论,而是专指作为本书主题的统计学习理论,后面提到“学习理论”时多数也是如此;一些地方单独说“理论”也是指统计学习理论。—— 译者

为研究这些方法(它们可能经常与直觉相矛盾),建立起了一套完整的理论,其中主要包括以下四方面内容:

(1) 关于统计推断一致性的充分必要条件的一系列概念;

(2) 在这些概念基础上的反映学习机器推广能力的界;

(3) 在这些界的基础上的针对小样本数的归纳推理原则;

(4) 实现这种新的推理的方法。

当人们研究统计学习理论时,往往会遇到两个主要的困难:一个是技术性的,另一个是概念性的。前者指对有关证明的理解中存在的困难,后者则是要理解问题的本质和其思想体系。

为了克服这个技术性困难,读者须有一定的耐心,坚持弄懂有关形式化推导过程的各个细节。

而为了理解问题的本质、精髓和思想体系,就必须把这一理论作为一个整体来看待,而不能仅仅看成是其中不同部分的堆积。掌握问题的本质是非常重要的,它可以引导我们沿着正确的方向去寻找结果,防止误入歧途。

本书的目的在于阐述统计学习理论的本质。我将在书中说明抽象的理论研究是如何发展出新的算法的。为了使读者更容易理解,我把本书写得比较简短。

在本书中,我试图在概念上不做简化的情况下尽可能把内容写得简单一些,注重理论的本质和思想,而不包括理论的细节和定理的详细证明(这些内容中的一部分可以在我 1982 年由 Springer 出版的《Estimation of Dependencies Based on Empirical Data》一书中找到,更全面的内容则收录于我 1998 年由 J. Wiley 出版的《Statistical Learning Theory》一书中)。但是,为了在阐述有关思想时不过于简化,书中需要介绍一些重要的新概念(新的数学结构)。

本书内容包括:引言,一至五章,对每一章内容都有一个非正式推导和评述,最后是结论。

引言部分介绍了学习问题的研究历史,这对于阅读后面章节是一个必要的铺垫。

第一章专门讨论学习问题的表示,引出了根据经验数据的最小化风险函数的一般模型。

第二章对于理解统计学习理论的新思想可能是最重要的一章,同时可能也是最难读的一章。这一章阐述了关于学习过程的概念原理,其中包括一些重要概念,在这些概念上建立了学习过程一致性的充分必要条件。

第三章讨论了关于学习过程收敛速度的界的非渐近理论,这些关于界的理论是建立在从学习过程的概念模型得到的一些概念的基础之上的。

第四章重点研究小样本数理论。在这里,我们引出针对小样本数情况的归纳原则,它可以控制推广能力。

第五章描述了一种建立在小样本数理论基础之上的新的通用学习机器——支持向量机,它是与传统的神经网络一起讨论的。

在每章之后有一个评述,重点在于讨论在传统的数学统计学中的有关研究与统计学习理论中的研究之间的关系。

在结论部分中讨论了学习理论中一些尚需要进一步研究的问题。

本书是面向较广的读者范围而写的,包括不同领域的学生、工程师和科学家(统计学家、数学家、物理学家、计算机科学家等)。理解本书不需要数学的特殊分支的知识,但是,本书并不是一本简易读物,因为书中尽管没有去刻画众多的(数学的)树木,却的确展现了一片(概念的)森林。

在本书写作时我还抱有另外一个目标,就是希望通过本书强调抽象理论研究在实际中的巨大作用。之所以要强调这一点,是因为在过去几年里,在各种计算机科学的学术会议上,我不断听到这样的说法:

复杂的理论是没有用的,有用的是简单的算法。

本书的目的之一就是要说明,至少在统计推断问题中,这种说法是不正确的。我希望能够说明,在这一科学领域中,那句古老的原则仍然适用,就是:

没有什么比一个好的理论更实用了。

本书并不是对标准的理论的一个综述,而是试图宣传一种观点,这种观点不仅仅是关于学习和推广性问题的,而是对整个理论和应用统计学都适用。

我衷心希望读者能对这本书有兴趣并且发现它有用。

致 谢

由于 AT&T 贝尔实验室自适应系统研究部主任 Larry Jackel 的支持才使本书成为可能。

本书是在我的同事们的合作下完成的,他们包括 Jim Alvich, Jan Ben, Yoshua Bengio, Bernhard Boser, Léon Bottou, Jane Bromley, Chris Burges, Corinna Cortes, Eric Cosatto, Joanne DeMarco, John Denker, Harris Drucker, Hans Peter Graf, Isabelle Guyon, Patrick Haffner, Donnie Henderson, Larry Jackel, Yann LeCun, Robert Lyons, Nada Matic, Urs Mueller, Craig Nohl, Edwin Pednault, Eduard Sackinger, Bernhard Scholkopf, Patrice Simard, Sara Solla, Sandi von Pier 和 Chris Watkins 等。

Chris Burges, Edwin Pednault 和 Bernhard Scholkopf 等阅读了几个版本的手稿并且改进和简化了一些说明。

手稿完成后我把它给了 Andrew Barron, Yoshua Bengio, Robert Berwick, John Denker, Federico Girosi, Ilia Izmailov, Larry Jackel, Yakov Kogan, Esther Levin, Vincent Mirelly, Tomaso Poggio, Edward Reitman, Alexander Shustorovich 和 Chris Watkins 等以征求意见,他们的意见也改进了书中的阐述。

在此我向帮助我完成本书的所有人表示深深的感谢。

Red Bank, New Jersey　　Vladimir N. Vapnik

1995 年 3 月

目　　录

0 引 论

学习问题研究的四个阶段

学习问题的研究历史可以分为四个阶段,它们分别以下面四个重要事件为标志:

(1) 第一个学习机器的创立;

(2) 学习理论的基础的创立;

(3) 神经网络的创立;

(4) 神经网络的替代方法的创立。

在不同历史阶段有不同的研究主题和重点,所有这些研究共同勾画出了人们对学习问题进行探索的一幅复杂的(和充满矛盾的)图画。

0.1 Rosenblatt 的感知器(60 年代)

在 35 年多以前 F. Rosenblatt 提出了第一个学习机器的模型,称作感知器,这标志着人们对学习过程进行数学研究的真正开始[①]。从概念上讲,感知器的思想并不是新的,它已经在神经生理学领域中被讨论了多年。但是,Rosenblatt 做了一件不寻常的事,就是把这个模型表现为一个计算机程序,并且通过简单的实验说明这个模型能够被推广。感知器模型被用来解决模式识别问题,在最简单的情况下就是用给定的例子来构造一个把两类数据分开的规则。

0.1.1 感知器模型

为了构造这样一个分类规则,感知器利用了最简单的神经元模型的自适应特性(Rosenblatt, 1962)。每一个神经元都是一个 McCulloch-Pitts 模型,它有 n 个输入 $x=(x^1, \cdots, x^n)\in X \quad R^n$ 和一个输出 $y\in\{-1, 1\}$(图 0.1)。输出与输入之间通过下面的函数

① 注意,在 30 年代 Fisher 提出的判决分析实际上并没有考虑归纳推断的问题(即用例子估计判决规则的问题),这一问题是在 Rosenblatt 的工作之后才被考虑的。在 30 年代,判别分析被看作是利用已知的两类向量的概率分布函数来设计将两类向量分开的决策规则的问题。

依赖关系相连:

$$y = \mathrm{sgn}\{(w \cdot x) - b\},$$

其中,$(u \cdot v)$ 表示两个向量的内积,b 是一个域值,$\mathrm{sgn}(\cdot)$ 是符号函数,即如果 $u > 0$ 则 $\mathrm{sgn}(u) = 1$,如果 $u \leq 0$ 则 $\mathrm{sgn}(u) = -1$。

图 0.1　从几何上看,神经元定义了输入空间中取值为 -1 和 1 的两个区域,它们被超平面 $(w \cdot x) - b = 0$ 分开

从几何上说,神经元把空间 X 分为两个区域:一个是输出 y 取值为 1 的区域,另一个是输出 y 取值为 -1 的区域。这两个区域被超平面

$$(w \cdot x) - b = 0$$

分开。向量 w 和标量 b 决定了分类超平面的位置。在学习过程中,感知器为神经元选择适当的系数。

Rosenblatt 研究了一个由多个神经元组成的模型:他考虑了神经元的多层结构,其中前一层神经元的输出是下一层神经元的输入(一个神经元的输出可以是多个神经元的输入),最后一层只有一个神经元,因此这个(基本的)感知器有 n 个输入和一个输出。

从几何上说,感知器用分段线性的面把空间 X 分为两部分(图 0.2)。通过为所有神经

图 0.2　感知器是由多个神经元构成的;从几何上看,感知器定义了输入空间中取值为 -1 和 1 的两个区域,它们被分段线性的面分开

元选择适当的系数,感知器定义了 X 空间中的两个区域,它们被分段线性面(它们并不一定相连)分隔开。在这个模型中,学习就是用给定的训练数据寻找所有神经元的合适的系数。

在 60 年代,人们并不清楚如何同时为感知器的所有神经元选择参数(这个答案在 25 年之后才知道),因此,Rosenblatt 提出了这样的方案:除了最后一个神经元之外,固定其他所有神经元的系数,在学习过程中寻找最后一个神经元的系数。从几何上说,他提出的方法就是把输入空间 X 变换到一个新空间 Z(通过为除了最后一个神经元之外的所有其他神经元选择适当的系数),用训练数据在空间 Z 中构造分类超平面。

沿袭传统生理学中的关于带奖励和惩罚刺激的学习概念,Rosenblatt 提出了一种迭代地寻找系数的简单算法。

令输入空间中给定的训练数据为

$$(x_1, y_1), \cdots, (x_l, y_l),$$

并令

$$(z_1, y_1), \cdots, (z_l, y_l),$$

为对应的在空间 Z 中的训练数据(向量 z_i 是 x_i 的变换)。在每一步 k,把训练数据中的一个元素作用到感知器,用 w(k) 表示这时最后一个神经元的系数向量。算法有下面几部分组成:

(1) 如果训练数据中的下一个样本 z_{k+1}, y_{k+1} 被正确分类,即[①]

$$y_{k+1}(w(k) \cdot z_{k+1}) > 0,$$

则超平面的系数向量不变,即

$$w(k+1) = w(k).$$

(2) 如果下一个样本被错分了,即[②]

$$y_{k+1}(w(k) \cdot z_{k+1}) < 0,$$

则系数向量根据下面的规则进行修正:

$$w(k+1) = w(k) + y_{k+1} z_{k+1}.$$

(3) 初始的系数向量 w 是零:

$$w(1) = 0.$$

利用这一学习规则,感知器表现出了在简单例子上的推广能力。

0.1.2 对学习过程分析的开始

1962 年,Novikoff 证明了关于感知器的第一个定理,这一定理实际上是学习理论的开始。定理指出,如果

(1) 训练向量 z 的模以某个常数 R 为界($\|z\| \leq R$);

(2) 训练数据能够以间隔 ρ 被分开:

---

① 原文中错印为 $y_{k+1}(w(k) \cdots z_{k+1}) > 0$ ——译者

② 原文中错印为 $y_{k+1}(w_i(k) \cdot z_{k+1}) < 0$ ——译者

$$\sup_{w} \min_{i} y_i(z_i \cdot w) > \rho$$

(3) 对感知器进行足够多次训练过程，则在最多

$$N \leq \frac{R^2}{\rho^2}$$

次修正以后就可以构造出将这些训练数据分开的超平面。

这一定理在创建学习理论中起了十分重要的作用。它在一定意义上将导致机器具有推广能力的原因和最小化训练集上的错误数的原则联系了起来。我们在最后一章[①]将要看到，表达式$[R^2/\rho^2]$反映了对很广泛的一类学习机器都很重要的一个概念，利用它可以控制推广能力。

0.1.3 对学习过程的应用分析与理论分析

Novikoff 证明了感知器能够将训练数据分开。用同样的方法可以证明，如果数据是可分的，那么在有限次修正以后，感知器能够将任意无限长的数据序列分开(即在最后一次修正以后，剩下的无穷多数据都将能被正确地分开)。而且，如果感知器采用下面的训练停止规则，即若在第 k 次修正($k = 1, 2, \cdots$)以后，接下来训练数据中的 m_k 个样本都没有使决策规则改变(即它们都被正确分类)，则停止感知器的学习过程。其中

$$m_k = \frac{1 + 2\ln k - \ln \eta}{-\ln(1 - \varepsilon)},$$

那么，有如下的结论成立：

(1) 感知器会在前 l 步学习中停止学习过程，其中，

$$l \leq \frac{1 + 4\ln \frac{R}{\rho} - \ln \eta}{-\ln(1 - \varepsilon)} \frac{R^2}{\rho^2},$$

(2) 在学习停止时感知器已经建立了一个决策规则，它在测试集上的错误率小于 ε 的概率是 $1 - \eta$ 或者说它以概率 $1 - \eta$ 具有在测试集上小于 ε 的错误率(Aizerman, Braverman and Rozonoer, 1964)。

因为有这些结论，很多学者认为使学习机器具有推广性(即具有小的测试错误率)的唯一因素就是使它在训练集上的误差最小。因此，对学习过程的研究分化为两个分支，分别叫做对学习过程的应用分析和对学习过程的理论分析。

学习过程的应用分析学派的基本思想可以归纳为：

要得到好的推广性，只要选择使训练错误数最小的神经元系数就足够了。最小化训练错误数的原则是一个不言而喻的归纳原则，从实用角度看是不需要证明的。应用分析的主要目标就是寻找同时构造所有神经元的系数的方法，使所形成的分类面能够在训练数据上达到错误数最小。

学习过程的理论分析学派的思想是不同的：

最小化训练错误数的原则并不是不言而喻的，而是需要证明的。或许还存在另外的归

① 应该是本书的第五章，也就是第一版的最后一章。——译者

纳原理,能够达到更好的推广性能。学习过程理论分析的主要目标就是寻找能够达到最好的推广性能的归纳原则,并构造算法来实现这一原则。

本书将说明最小化训练错误数的原则的确不是不证自明的,存在另一个更智能化的归纳原则,它能够提供更好的推广性能。

0.2 学习理论基础的创立(60—70 年代)

关于感知器的实验广为知晓后,人们很快提出了一些其他类型的学习机器(如 B. Widrow构造的 Madaline 自适应学习机、K. Steinbuch 提出的学习矩阵等,实际上他们已经开始了构造特殊的学习机器硬件)。然而,与感知器不同的是,这些机器从一开始就被作为解决现实中实际问题的工具来研究,而没有被看作是学习现象的一般模型。

为了解决现实中的实际问题,人们还开发了很多计算机程序,包括创建各种类型的逻辑函数的程序(如最初为专家系统目的设计的决策树)、隐含马尔可夫模型(用于语音识别问题)等。这些方法也没有涉及到对一般学习现象的研究。

在感知器之后,关于构造一般性学习机器的研究的下一步是在 1986 年完成的,这就是用所谓后向传播[①]技术同时寻找多个神经元的权值。这一方法实际上开创了学习机器研究历史的一个新时代。我们将在下一节对它进行讨论,而本节将集中介绍学习理论基础的发展历史。

对于应用分析学派来说,在从构造感知器(1960 年)到实现后向传播(1986 年)之间的这段时间里,没有发生什么特别的事情。与此形成对比的是,这段时间里统计学习理论的发展却是成果累累。

0.2.1 经验风险最小化原则的理论

早在 1968 年,统计学习理论的基本思想就已经得到了很大的发展。对于指示函数集(即模式识别问题),已经提出了 VC 熵和 VC 维的概念,它们是这一新理论中的核心概念。利用这些概念,发现了泛函空间的大数定律(频率一致收敛于其概率的充分必要条件),研究了它与学习过程的联系,并且得到了关于收敛速率的非渐近界的主要结论(Vapnik and Chervonenkis, 1968);在 1971 年发表了这些工作的完全的证明(Vapnik and Chervonenkis, 1971)。所得到的这些界使得一个全新的归纳原则(即 1974 年提出的结构风险最小化归纳原则)成为可能,从而完成了模式识别学习理论。这一模式识别理论的新体系在一部专著中进行了总结[②]。

在 1976 年到 1981 年间,最初针对指示函数集得到的这些结论推广到了实函数集,主

---

① back-propagation,亦译反向传播,简称 BP。——译者

② Vapnik V and Chervonenkis A. Theory of Pattern Recognition(俄文版). Nauka, Moscow, 1974;德文翻译版: Wapnik W N, Tscherwonenkis A Ja. Theorie der Zeichenerkennung. Akademia-Verlag, Berlin, 1979

要内容有:大数定律(均值一致收敛于其期望的充分必要条件)、完全有界的函数集和无界函数集一致收敛速率的界,以及结构风险最小化原则。在 1979 年的一部专著[①] 中总结了这些成果,阐述了依赖关系估计的一般问题的新理论。

最后,在 1989 年发现了经验风险最小化归纳原则和最大似然方法一致性[②]的充分必要条件,完成了对经验风险最小化归纳推理的分析(Vapnik and Chervonenkis, 1989)。

在历时 30 年对学习过程的分析基础上,90 年代开始了对于能够控制推广性能的新的学习机器的合成。

这些成果是由对学习过程的深入研究引出的,它们是本书的主要内容。

0.2.2 解决不适定问题的理论

在 60 年代和 70 年代,在数学的各个不同分支中,产生了一些开创性的新理论,它们对于创造新的思想起了非常重要的作用。下面我们列出其中的一些理论,并将在后面各章的评述中对它们进行讨论。

让我们从旨在解决所谓不适定问题的正则化理论开始。

早在 20 世纪初,Hadamard 观察到在一些(很一般的)情况下,求解(线性)算子方程

$$Af = F, \quad f \in \mathfrak{F}$$

的问题(寻找满足这一等式的函数 $f \in \mathfrak{F}$)是不适定的:即使方程存在唯一解,如果方程右边有一个微小变动(如用 $\|F - F_\delta\| < \delta$ 任意小的 F_δ 取代 F),也会导致解有很大的变化(即可能导致 $\|f_\delta - f\|$ 很大)。

在这种情况下,如果方程右边的 F 是不准确的(比如等于 F_δ,而 F_δ 与 F 相差某个 δ 水平的噪声),那么使泛函

$$R(f) = \|Af - F_\delta\|^2$$

最小化的函数 f_δ 并不能保证在 δ 趋于 0 时是方程真实解的一个好的近似。

Hadamard 当时认为这种不适定问题是一个纯数学的现象,现实中的问题都是适定的。然而,在 20 世纪后半叶中,人们发现现实中的一些很重要的问题是不适定的。尤其是,当人们试图反演因果关系,即从已知的结果出发去寻找其未知原因时,就会出现不适定问题。即使这种因果关系形成了一对一的映射,它的反演问题仍然可能是不适定的。

对于我们的讨论来说,重要的是,根据数据估计密度函数这个统计学中的主要问题是不适定的。

60 年代中期,人们发现,如果不是最小化泛函 $R(f)$,而是最小化另一个称作正则化泛函的函数

$$R^*(f) = \|Af - F_\delta\|^2 + \gamma(\delta)\Omega(f),$$

其中 $\Omega(f)$ 是某个泛函数(属于一种特殊类型的泛函),$\gamma(\delta)$ 是某个适当的常数(它依赖于

---

① Vapnik V N. Estimation of Dependencies Based on Empirical Data(俄文版). Nauka, Moscow, 1979; 英文翻译版: Vapnik Vladimir. Estimation of Dependencies Based on Empirical Data. Springer, New York, 1982

② 指以概率收敛到最好的可能结果。一致性的确切定义将在 2.1 节给出。

噪声的水平),那么我们就可以得到一个解的序列,它在 δ趋于 0 时收敛于我们希望的解(Tikhonov, 1963; Ivanov, 1962 及 Phillips, 1962)。

正则化理论是表明智能推理方法存在的第一个信号,它表明了最小化泛函 R(f)这一看上去好像不言而喻的方法是行不通的,而并不是显而易见的最小化泛函 $R^*(f)$的方法却是可行的。

由解决不适定问题的理论而产生的思想的影响是十分深远的。这种正则化思想和正则化技术在很多学科中得到了广泛的传播,其中也包括统计学。

0.2.3 密度估计的非参数方法

作为一个特例,从一个范围较宽的密度的集合中估计密度函数的问题就是一个不适定问题。传统理论研究的主题是从一个范围较窄的密度集合中估计密度(比如从一个由有限个参数决定的密度集合即所谓密度的参数集合中进行估计),使用的是一种"不言而喻"类型的推理(最大似然方法)。如果把从中进行估计的密度集合加以扩展,就无法再使用这种推理方式。为了从范围宽的集合(非参数集合)中估计密度,必须采用某种新的推理方式,其中利用了正则化技术。在 60 年代,人们提出了几种此类(非参数)算法(Rosenblatt, 1956; Parzen, 1962 及 Chentsov, 1963),70 年代中期则发现了创建此类算法的一般途径,它是建立在解决不适定问题的标准做法基础上的(Vapnik and Stefanyuk, 1978)。

密度估计的非参数方法带来了新的统计学算法,它们可以克服传统体系的缺陷。现在人们可以从一个较宽范围的函数集中估计函数了。

然而我们必须注意到,这些方法的出发点是利用大量的样本来估计一个函数。

0.2.4 算法复杂度的思想

最后,在 60 年代,提出了统计学和信息论中最伟大的思想之一,就是算法复杂度的思想(Solomonoff, 1960; Kolmogorov, 1965 及 Chaitin, 1966)。激发这一思想的是两个看上去不同的基本问题:

(1) 归纳推理的本质是什么? (Solomonoff 研究的问题)

(2) 随机性的本质是什么? (Kolmogorov 及 Chaitin 研究的问题)

对这两个问题,Solomonoff、Kolmogorov 和 Chaitin 提出的答案开创了研究推理问题的信息论方法。

随机性概念的思想可以粗略地描述如下:对于一个长度为 l 的很长的数据串,如果不存在任何复杂度远小于 l 的算法能够产生出这个数据串,则它就构成了一个随机串。算法的复杂度是用实现这个算法的最小程序的长度来衡量的。已经证明,算法复杂度的概念是普遍的(它是确定的,除了一个反映计算机类型的加性常数外)。而且也已经证明,如果对串的描述不能被计算机压缩,则这个串具有一个随机序列的一切性质。

这一点说明了这样一个思想,如果我们可以在很大程度上压缩对一个给定串的描述,那么所使用的算法就描述了数据的内在性质。

在 70 年代, 在这些思想的基础上, Rissanen(1978) 提出了对于学习问题的最小描述长度(MDL) 归纳推理。

我们将在第四章中考虑这一原则。

所有这些新思想仍是处在发展中的。然而, 对于如何对付在有限数量的经验数据基础上进行依赖关系估计的问题, 这些思想的确改变了人们的主要认识。

0.3 神经网络(80 年代)

0.3.1 神经网络的思想

在 1986 年, 几个作者独立地提出了同时构造感知器所有神经元的向量系数的方法, 就是称作后向传播的方法(LeCun, 1986 及 Rumelhart, Hinton and Williams, 1986)。这一方法的思想是很简单的, 它不是用 McCulloch-Pitts 的神经元模型, 而是采用了一种稍加修改的模型, 在其中把原来的不连续函数 sgn{(w · x) - b} 替换为连续的所谓 sigmoid 函数[①]:

$$y = S\{(w \cdot x) - b\}$$

(这里的 S(u) 是一个满足性质

$$S(-\infty) = -1, \quad S(+\infty) = 1$$

的单调函数, 比如 S(u) = tanh(u), 见图 0.3), 这样, 新神经元合成的就是一个连续函数, 它对于任意固定的 x, 都存在对所有神经元的所有系数的梯度。在 1986 年找到了计算这个梯度的方法[②]。利用计算出的梯度, 人们可以应用任何基于梯度的方法来构造对预期函数的逼近。当然, 基于梯度的方法只能保证找到局部极小点。但是尽管如此, 看上去好像人们已经找到了学习过程应用分析的主要思想, 剩下的问题就只是它的实现了。

图 0.3 用光滑函数 S(u) 近似不连续函数 sgn(u) = ±1

0.3.2 理论分析目标的简化

后向传播技术的发现可以看作是感知器的第二次诞生。然而, 与它的第一次诞生相比, 这次新生所处的形势已经完全不同。从 60 年代以来出现了功能强大的计算机, 一些新的学科介入到学习问题的研究中来。这些使得研究的规模和形式有了很大的改观。

---

① sigmoid 函数是一种取值在 - 1 到 + 1(或者 0 到 1) 之间的平滑单调上升函数(如图 0.3 所示), 是神经网络中最常用的一种函数, 本书第 5.2 节中给出了一种 sigmoid 函数的定义。可译作 S 型函数, 但目前多数中文文献中都直接用 sigmoid 函数的说法, 因此我们也采用这种说法。——译者

② 后向传播方法实际上是在 1963 年在解决某些控制问题时被发现的(Brison, Denham and Dreyfuss, 1963), 在感知器中是重新被发现。

具有多个可调节神经元的感知器,与只有单个可调节神经元、但具有同样多的自由参数的感知器相比,是否具有更好的推广特性,人们对这一点实际上并不能断定,但是尽管如此,由于实验规模的原因,科学界还是对这种新方法表现出了更大的热情。

Rosenblatt 的第一个实验是针对数字识别问题的。为了说明感知器的推广能力,Rosenblatt 用了由几百个向量组成的训练数据,包含几十维坐标。在 80 年代乃至 90 年代,数字识别学习的问题一直是一个重要的研究内容。今天,为了得到好的决策规则,人们采用数万(甚至数百万)的观测向量,坐标达几百维。这要求计算过程要有特殊的组织方法。因此,在 80 年代,人工智能学者在计算学习领域中扮演了主要的角色。在这些人工智能研究者中间,一些较极端的学者有很大的影响(正是他们强调了"复杂的理论是没有用的,有用的是简单的算法")。

这些人工智能学者对于处理学习问题有丰富的经验,善于对一些理论上非常复杂的问题构造"简单的算法"。60 年代末,人们认为在几年之内就可以完成计算机自然语言翻译器(但时至今日我们尚离解决这个极其复杂的问题很远);在此之后的另一个计划是构造通用问题求解器;再之后又是建立大系统自动控制机的计划,等等。所有这些科研计划都没有取得成功。接下来要研究的问题就是如何建立一种计算学习技术。

这些学者首先做的是改变了所用的术语。特别地,感知器被改称为神经网络。然后这些研究被称作是与生理学家共同进行的,对学习问题的研究减少了一般性,增加了主观色彩。在 60 和 70 年代学习问题研究的主要目标是寻找从小数量样本出发进行归纳推理的最好途径,而到了 80 年代,目标变成了构造利用大脑来推广的模型[①]。

1984 年,提出了可能近似正确(probably approximately correct, PAC)模型[②],这是把理论介绍到人工智能领域中的一次尝试。这个模型是用统计学中常用的一致性概念的一种特例来定义的,其中结合了对计算复杂度的一些要求[③]。

事实上,PAC 模型的几乎所有结果都来自统计学习理论,它们构成了统计学习理论的四部分内容之一(即关于界的理论)的一些特例。尽管如此,这一模型仍对学习问题的研究起了无可置疑的作用,它使得人工智能界认识到了统计分析的重要性。但是,这还不足以导致新的学习技术的出现。

从感知器的第二次诞生到现在已经几乎过去 10 年了。从概念上看,感知器的这第二次诞生的重要性小于第一次。虽然在一些特殊领域中应用神经网络取得了很重要的成果,但是所得到的理论成果并没有对一般的学习理论带来多大贡献。而且,在神经网络的实验中也没有发现新的有意义的学习现象。实验中观察到的所谓过适应问题,实际上是在解决不适定问题的理论中称之为"错误结构"的现象。从解决不适定问题的理论中,人们得到了防止过学习的工具——在算法中采用正则化技术。

---

① 当然研究人是如何学习的是很有意义的,但是这并不一定是建立人工学习机器的最佳途径,正如人们对鸟类如何飞行的研究实际上对建造飞机并没有多少帮助一样。

② Valiant L G. 1984. A theory of learnability. Commun. ACM 27(11), 1134～1142

③ L Valiant 指出,"如果从定义中把对计算的要求去掉,那么正如 Vapnik 特别讨论的那样,我们得到的就是统计学意义上的非参数推理的概念。"(见 Valiant L. 1991. A view of computational learning theory. in the book: "Computation and Cognition", Society of Industrial and Applied Mathematics, Philadelphia, p. 36)

因此,差不多 10 年的神经网络研究没有从本质上推进对学习过程本质的认识[①]。

0.4 回到起点(90 年代)

在过去的几年里,与神经网络有关的一些事情发生了改变。

现在,更多的注意力放在了对神经网络的替代方法的研究上,比如,人们用很大的精力进行了对径向基函数模型的研究(参见文献(Powell, 1992)中的综述)。就像在 60 年代一样,神经网络又被重新叫做多层感知器。统计学习理论中的较高深的部分现在开始吸引更多的学者。尤其是在过去的几年里,结构风险最小化原则和最小描述长度原则成了人们分析研究的一个热点。与渐近的理论形成对照,关于小样本数理论的讨论广泛开展起来。

看起来好像所有事情都开始追本溯源。

而且,统计学习理论现在扮演的是一个更积极的角色:在完成了学习过程的一般分析后,已经开始了关于(对任意数目的观测能得到最高的推广能力)最优算法合成的研究。

然而,这些研究尚不属于历史,它们是当今研究活动的主题[②]。

---

① 本节指出了神经网络研究中的重要缺陷和不足。但是,译者认为,虽然缺乏理论上的重要进展,但神经网络的研究至少在唤起众多领域对机器学习的兴趣、鼓励人们尝试用学习机器来解决一些实际问题等方面仍起了十分重要的作用,而且在很多实际应用中神经网络都取得了很好的效果。从这个意义上看近十几年来的神经网络研究仍是富有成果的,在人们对学习问题研究的历程中所作出的贡献也是应该肯定的。另外,本书讨论的神经网络只是多层感知器模型,对于其他神经网络模型(比如自组织映射、反馈网络等)都没有涉及。——译者

② 这一评述是在 1995 年给出的。然而,在本书第一版出版之后,在计算机学习的新方法研究方面又发生了一些重要的变化。

在过去的 5 年里,在统计学习理论的激发下,在学习方法论上出现了一些新思想。与受学习过程的生物模拟所启发的构造学习算法的旧思想不同,新的思想是受最小化错误率的理论界限的努力激励的,这些界限是作为学习过程的形式化分析的结果得到的。这些思想(它们常常意味着与旧体系相矛盾的方法)产生了这样的新算法,它们不但有良好的数学性质(比如解的唯一性、处理大量样本的简单方法,以及不依赖于输入空间的维数),而且表现出出色的性能:它们超过了用旧方法得到的最先进的解。

现在,在学习问题上已经出现了一种新的方法论情形,即实用的方法是由对统计学界限深入的理论分析所得到的结果,而不是发明新的聪明的启发式方法的结果。

这一事实在很多方面改变了学习问题的特点。

第一章

学习问题的表示

在本书中,我们把学习问题看作是利用有限数量的观测来寻找待求的依赖关系的问题。

1.1 函数估计模型

我们用下面三个部分来描述从样本学习的一般模型(图 1.1):

(1) 产生器(G),产生随机向量 $x\in R^n$,它们是从固定但未知的概率分布函数 $F(x)$ 中独立抽取的。

(2) 训练器(S),对每个输入向量 x 返回一个输出值 y,产生输出的根据是同样固定但未知的条件分布函数[①] $F(y|x)$。

(3) 学习机器(LM),它能够实现一定的函数集 $f(x,\alpha)$, $\alpha\in\Lambda$,其中 Λ 是参数集合[②]。

图 1.1 根据样本学习的一个模型。在学习过程中,学习机器 LM 观察数据对 (x,y)(训练集)。在训练之后,学习机器必须对任意输入 x 给出输出 y。学习的目标是能够给出输出 y,使之接近训练器的响应 y

学习的问题就是从给定的函数集 $f(x,\alpha)$, $\alpha\in\Lambda$ 中选择出能够最好地逼近训练器响应的函数。这种选择是基于训练集的,训练集由根据联合分布 $F(x,y)=F(x)F(y|x)$ 抽取出的 l 个独立同分布(i.i.d.)观测

$$(x_1,y_1),\cdots,(x_l,y_l) \tag{1-1}$$

组成。

---

① 这是一般的情况,其中包括训练器采用某一函数 $y=f(x)$ 的情况。

② 注意,参数 $\alpha\in\Lambda$ 并不一定必须是向量,它们可以是任意的抽象参数,因此我们实际上考虑的是任意的函数集。

1.2 风险最小化问题

为了选择所能得到的对训练器响应最好的逼近,就要度量在给定输入 x 下训练器响应 y 与学习机器给出的响应 $f(x,\alpha)$ 之间的损失或差异 $L(y,f(x,\alpha))$。考虑损失的数学期望值

$$R(\alpha)=\int L(y,f(x,\alpha))\,dF(x,y), \tag{1-2}$$

它就是风险泛函[①]。学习的目标就是,在联合概率分布函数 $F(x,y)$ 未知、所有可用的信息都包含在训练集(1-1)式中的情况下,寻找函数 $f(x,\alpha)$,使它(在函数类 $f(x,\alpha)$,$\alpha\in\Lambda$ 上)最小化风险泛函 $R(\alpha)$。

1.3 三种主要的学习问题

学习问题的这种形式化表述的面是很广的。它包括了很多特殊的问题。我们考虑其中主要的问题:模式识别、回归函数估计和概率密度估计。

1.3.1 模式识别

令训练器的输出 y 只取两种值 $y=\{0,1\}$,并令 $f(x,\alpha)$,$\alpha\in\Lambda$ 为指示函数集合(指示函数即只有 0 或 1 两种取值的函数)。考虑下面的损失函数:

$$L(y,f(x,\alpha))=\begin{cases}0 & \text{若 } y=f(x,\alpha)\\ 1 & \text{若 } y\neq f(x,\alpha)\end{cases}. \tag{1-3}$$

对于这个损失函数,(1-2)式的泛函确定了训练器和指示函数 $f(x,\alpha)$ 所给出的答案不同的概率。我们把指示函数给出的答案与训练器输出不同的情况叫做分类错误。

这样,学习问题就成了在概率测度 $F(x,y)$ 未知,但数据(1-1)式已知的情况下,寻找使分类错误的概率最小的函数。

1.3.2 回归估计

令训练器的输出 y 为实数值,并令 $f(x,\alpha)$,$\alpha\in\Lambda$ 为实函数集合,其中包含着回归函数

---

① 在某些书籍和文献中,这里的风险泛函也常常被叫做"风险函数",即把它看作参数 α 的函数。在本书的体系中,强调不同的 α 代表不同的函数,α 可以是任何广义的参数,因此把 $R(\alpha)$ 看作是由 α 代表的学习函数的函数(即泛函)更为恰当。翻译中沿用了原著的叫法,即把 $R(\alpha)$ 叫做风险泛函(risk functional),而 $L(y,f(x,\alpha))$ 叫做损失函数(loss-function)。书中还有其他类似的地方我们不再一一说明。—— 译者

$$f(x,\alpha)=\int y\,dF(y|x).$$

我们知道,回归函数就是在损失函数[①]

$$L(y,f(x,\alpha))=(y-f(x,\alpha))^2 \tag{1-4}$$

下使泛函(1-2)式最小化的函数。

这样,回归估计的问题就是,在概率测度 $F(x,y)$ 未知但数据(1-1)式已知的情况下,对采用(1-4)式损失函数的风险泛函(1-2)式最小化。

1.3.3 密度估计(Fisher-Wald 表示)

最后,考虑从密度函数集 $p(x,\alpha),\ \alpha\in\Lambda$ 中估计密度函数的问题。对这个问题,考虑下面的损失函数:

$$L(p(x,\alpha))=-\log p(x,\alpha). \tag{1-5}$$

我们知道,待求的密度函数在损失函数(1-5)式下使风险泛函(1-2)式最小化。因此,从数据估计密度函数的问题就是,在相应的概率测度 $F(x)$ 未知、但给出了独立同分布数据

$$x_1,\ \cdots,\ x_n$$

的情况下,使风险泛函最小化。

1.4 学习问题的一般表示

学习问题可以一般地表示如下:设有定义在空间 Z 上的概率测度 $F(z)$。考虑函数的集合 $Q(z,\alpha),\ \alpha\in\Lambda$。学习的目标是最小化风险泛函

$$R(\alpha)=\int Q(z,\alpha)\,dF(z),\quad \alpha\in\Lambda, \tag{1-6}$$

其中概率测度 $F(z)$ 未知,但给定了一定的独立同分布样本[②]

$$z_1,\ \cdots,\ z_l. \tag{1-7}$$

这种一般问题就是在经验数据(1-7)式基础上最小化风险泛函(1-6)式,其中 z 代表了数据对 (x,y),$Q(z,\alpha)$ 就是特定的损失函数(比如(1-3)式、(1-4)式或(1-5)式之一),前面讨论的学习问题都是这一一般问题的特例。在本书后面的内容中,我们将讨论在学习问题的这种一般表示下得到的结论。要将这些结论应用到具体问题中,只需将相应的损失函数替换到有关公式中。

---

① 如果函数集 $f(x,\alpha),\ \alpha\in\Lambda$ 中没有包含回归函数 $f(x)$,则采用损失函数(1-4)式下使泛函(1-2)式最小的函数 $f(x,\alpha_0)$ 在如下的 $L_2(F)$ 度量

$$\rho(f(x),f(x,\alpha_0))=\int (f(x)-f(x,\alpha_0))^2 dF(x)$$

下距离回归函数最近。

② 原著中样本(sample)一词多数时候用来指给定数据的集合(即样本集),有时也用来指集合中的元素。根据上下文一般可以确定其具体含义,因此在翻译时遵循了原著中的说法。——译者

1.5 经验风险最小化归纳原则

为了在未知的分布函数 $F(z)$ 下最小化(1-6)式的风险泛函,可以采用下面的归纳原则:

(1) 把风险泛函 $R(\alpha)$ 替换为所谓的经验风险泛函

$$R_{emp}(\alpha) = \frac{1}{l}\sum_{i=1}^{l} Q(z_i, \alpha), \tag{1-8}$$

它是用训练集(1-7)式得到的。

(2) 用使经验风险(1-8)式最小的函数 $Q(z, \alpha_l)$ 逼近使风险(1-6)式最小的函数 $Q(z, \alpha_0)$。

这一原则称作经验风险最小化(empirical risk minimization)归纳原则,简称 ERM 原则。

对于一个归纳原则,如果对任何给定的观测数据,学习机器都依照这一原则来选择逼近,则我们说这一归纳原则定义了一个学习过程。在学习理论中,ERM 原则扮演了一个具有决定性的角色。

ERM 原则是非常一般性的。解决一些特殊的学习问题的很多传统方法,比如在回归估计问题中的最小二乘方法、概率密度估计中的最大似然方法(ML 法)等,都是 ERM 原则的具体实现,其中采用了前面我们讨论的损失函数。

实际上,在(1-8)式中代入(1-4)式定义的损失函数,则我们要进行最小化的泛函变成

$$R_{emp}(\alpha) = \frac{1}{l}\sum_{i=1}^{l} (y_i - f(x_i, \alpha))^2,$$

这就得到了最小二乘方法。而如果把(1-5)式定义的损失函数替换到(1-8)式中,则经验风险泛函变成

$$R_{emp}(\alpha) = -\frac{1}{l}\sum_{i=1}^{l} \ln p(x_i, \alpha).$$

最小化这一泛函就等价于最大似然方法(最大似然方法中上式的右边采用的是正号)。

1.6 学习理论的四个部分

学习理论需要研究的是下面四个问题:

(1) 一个基于 ERM 原则的学习过程具有一致性的条件(充分必要条件)是什么?

(2) 这个学习过程收敛的速度有多快?

(3) 如何控制这个学习过程的收敛速度(推广能力)?

(4) 怎样构造能够控制推广能力的算法?

对这些问题的回答构成了学习理论的四个部分:

(1) 学习过程一致性的理论。

(2) 学习过程收敛速度的非渐近理论。

(3) 控制学习过程的推广能力的理论。

(4) 构造学习算法的理论。

在后面的各章中将分别讨论这四部分。

非正式推导和评述——1

第一章中给出的学习问题的表示反映了两个主要的要求:

(1) 从一个宽的函数集合中估计待求的函数;

(2) 在有限数量的例子的基础上估计待求的函数。

在(创建于 20 年代和 30 年代的)传统理论体系中发展起来的方法没有考虑到这些要求。因此,在 60 年代,人们在两个方向上进行了很大的努力,一是把传统的结果推广到范围更宽的函数集合,二是针对小样本数目改进已有技术。下面我们将对其中的一些研究进行讨论。

1.7 解决学习问题的传统模式

在传统理论体系的框架中,函数估计的所有模型都是基于最大似然方法的。它成了传统体系下的一个归纳引擎。

1.7.1 密度估计问题(最大似然方法)

设 $p(x,\alpha),\ \alpha\in\Lambda$ 是一个密度函数的集合,其中,与本章给出的问题表示不同,这里的集合 Λ 必须是包含在 R^n 中的(即 α 是一个 n 维向量)。设未知的密度 $p(x,\alpha)$ 属于这个函数集合中。研究的问题是用(按未知密度分布的)独立同分布数据

$$x_1,\cdots,x_l$$

来估计这个密度函数。

在 20 年代,Fisher(1952)研究出了估计密度函数的未知参数的最大似然方法,提出了用使泛函

$$L(\alpha)=\sum_{i=1}^{l}\ln p(x_i,\alpha)$$

最大的参数取值来逼近未知的参数。

在一定的条件下,这种最大似然方法是一致的。在下一章中,我们将利用关于泛函空间中的大数定律的结论来描述最大似然法一致性的充分必要条件。在下面几小节中我们将说明如何利用最大似然法来估计待求函数。

1.7.2 模式识别(判别分析)问题

利用最大似然技术,Fisher 研究了模式识别问题(当时他称之为判别分析问题)。他提出的模型是:

存在两类数据,它们的分布服从两个不同的统计规律 $p_1(x, \overset{*}{\alpha})$ 和 $p_2(x, \overset{*}{\beta})$(即密度,属于参数化密度函数类)。设第一类数据出现的概率是 q_1,第二类的概率是 $1-q_1$。问题是寻找一个决策规则,使错误的概率最小。

如果知道这两个统计规律和概率 q_1 的值,可以立即构造出这样一个规则:若向量 x 属于第一类的概率不小于它属于第二类的概率,决策规则就认为这个向量属于第一类。这个决策规则可以取得最小的错误率。所谓 x 属于第一类的概率不小于它属于第二类的概率,就是下面的不等式成立:

$$q_1 p_1(x, \overset{*}{\alpha}) \geq (1-q_1) p_2(x, \overset{*}{\beta}).$$

这一决策规则可以表示成下面的等价形式:

$$f(x) = \operatorname{sgn}\left[\ln p_1(x, \overset{*}{\alpha}) - \ln p_2(x, \overset{*}{\beta}) + \ln \frac{q_1}{1-q_1}\right], \tag{1-9}$$

称作判别函数(判别规则),它把第一类的样本赋值为 1,而把第二类样本赋值为 -1。为了得到这一判别函数,必须估计两个概率密度[①]:$p_1(x, \overset{*}{\alpha})$ 和 $p_2(x, \overset{*}{\beta})$。在传统的体系中,人们用最大似然法来估计这两个密度中的参数 $\overset{*}{\alpha}$ 和 $\overset{*}{\beta}$。

1.7.3 回归估计模型

在传统体系中,回归估计是建立在另外一个模型基础上的。这个模型就是所谓的度量含有加性噪声的函数的模型。

设某个未知函数有下面的参数化形式:

$$f_0(x) = f(x, \alpha),$$

其中 $\alpha \in \Lambda$ 是未知的参数向量,另设在任意点 x_i 都可以度量这个函数带有噪声的取值:

$$y_i = f(x_i, \alpha) + \xi,$$

其中噪声 ξ 独立于 x_i 且其分布服从一个已知的密度函数 $p(\xi)$。问题是,利用通过度量受到加性噪声影响了的函数 $f(x, \alpha)$ 所得的数据,从集合 $f(x, \alpha)$, $\alpha \in \Lambda$ 中估计函数 $f(x, \alpha)$。

在这个模型中,利用观测数据对

$$(x_1, y_1), \cdots, (x_l, y_l),$$

---

① 原文中错印为 $p_1(x, \alpha)$ 和 $p_2(x, \beta)$。—— 译者

人们可以用最大似然法估计未知函数 $f(x, \alpha)$ 的参数 α，方法是最大化下面的泛函：

$$L(\alpha) = \sum_{i=1}^{l} \ln p(y_i - f(x_i, \alpha)).$$

（记住 $p(\xi)$ 是一个已知的函数且 $\xi = y - f(x, \alpha)$）。如果采用具有零均值和某个固定方差的正态分布作为噪声模型，

$$p(\xi) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\xi^2}{2\sigma^2}\right),$$

就得到最小二乘法：

$$L^*(\alpha) = -\frac{1}{2\sigma^2}\sum_{i=1}^{l}(y_i - f(x_i, \alpha))^2 - l\ln(\sqrt{2\pi}\sigma),$$

对参数 α 最大化 $L^*(\alpha)$ 等价于最小化所谓的最小二乘泛函

$$M(\alpha) = \sum_{i=1}^{l}(y_i - f(x_i, \alpha))^2.$$

选择其他形式的噪声模型 $p(\xi)$ 可以得到参数估计的其他方法[①]。

1.7.4 最大似然法的局限

从上面的讨论可见，在传统的体系中，对本章提到的所有依赖性估计问题的求解都是基于最大似然方法的。然而，哪怕是在最简单的情形下，这种方法也可能失败。下面我们将说明对于一个由几个正态密度混合而成的密度函数，不可能用最大似然法估计它的参数。为了说明这一点，只需分析下面的例子中给出的最简单情况就足够了。

例　对于一个由两个正态密度通过最简单的混合而形成的密度

$$p(x, \alpha, \sigma) = \frac{1}{2\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\alpha)^2}{2\sigma^2}\right) + \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right),$$

其中只有一个密度的参数 (α, σ) 是未知的，用最大似然方法不可能估计出这个密度函数。

事实上，对于任何数据 $x_1, \cdots, x_l$ 及任一给定的常数 A，总存在一个小的 $\sigma = \sigma_0$，使得对 $\alpha = x_1$ 似然函数值超过 A，即

$$\begin{aligned} L(\alpha = x_1, \sigma_0) &= \sum_{i=1}^{l} \ln p(x_i; \alpha = x_1, \sigma_0) \\ &> \ln \frac{1}{2\sigma_0\sqrt{2\pi}} + \sum_{i=2}^{l} \ln \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{x_i^2}{2}\right) \\ &= -\ln\sigma_0 - \sum_{i=2}^{l}\frac{x_i^2}{2} - l\ln 2\sqrt{2\pi} > A. \end{aligned}$$

从这个不等式可以得到结论：在这个例子中似然函数值的最大值不存在，因此最大似然方法无法给出估计参数 α 和 σ 的解。

---

① 在 1964 年，P. Huber 扩展了传统的回归估计的模型，提出了所谓鲁棒性回归估计模型。根据这个模型，人们不需要准确的噪声模型 $p(\xi)$，取而代之的是给出一个包含这个函数的密度函数集（满足一些很一般性的条件）。问题是，对给定的参数化函数集合和给定的密度函数集合，构造一个具有最小最大特性的估计器（最小最大特性即在密度函数集中最坏的密度下的最好逼近）。这个问题的求解实际上具有如下的形式：选择一个合适的密度函数，然后用最大似然法估计参数（Huber, 1964）。

由此看出,最大似然方法只能适用于非常有限的密度函数集。

1.8 密度估计的非参数方法

在 60 年代初,几位学者提出了密度估计的一些新方法,称作非参数方法。这些方法的目标在于从一个范围较宽的函数集中估计密度,而不限于参数化的函数集(M. Rosenblatt, 1956[①]; Parzen, 1962 及 Chentsov, 1963)。

1.8.1 Parzen 窗

在这些方法中,Parzen 窗法可能是最著名的。根据这种方法,我们需要首先确定一种核函数。为了简单起见我们考虑一种简单的核函数:

$$K(x, x_i; \gamma) = \frac{1}{\gamma^n} K\left(\frac{x - x_i}{\gamma}\right), \quad x \in R^n,$$

其中 $K(u)$ 是某一对称单峰密度函数。

用这个核函数就可以确定估计

$$p(x) = \frac{1}{l} \sum_{i=1}^{l} K(x, x_i; \gamma).$$

在 70 年代,对 Parzen 类型的非参数密度估计,建立起了一套完善的渐近理论(Devroye and Gyorfi, 1985),它包括以下两个重要的结论:

(1) 对于从一个非常宽的密度类中估计密度函数,Parzen 估计是一致的(在各种度量下一致)。

(2) 对于“平滑”的密度函数,Parzen 估计器的渐近收敛速度是最优的。

对其他类型的非参数估计也得到了同样的结果。

因此,对于两种传统的模型(判别分析和回归估计),如果观测数目足够多,用非参数方法替代参数方法可以得到对待求依赖关系的好的逼近。

然而,对非参数方法的实验研究却没有显示出它比以往方法有很大的优势,这说明当应用于有限数目的观测时,非参数方法出色的渐近特性不再成立。

1.8.2 密度估计的问题是不适定的

非参数统计是作为解决密度估计和函数回归估计问题的一个处方发展起来的。为了使理论更全面,有必要去寻找一个构造和分析各种非参数算法的一般性原则。在 1978 年我们发现了这一原则(Vapnik and Stefanyuk, 1978)。

根据定义,密度 $p(x)$(如果存在的话)是下面的积分方程的解:

$$\int_{-\infty}^{x} p(t)dt = F(x), \tag{1-10}$$

① 原著中引用的文献与文后的参考文献不完全相符,现按文后的参考文献为准,后面有类似情况,不再一一标注。——译者

其中 $F(x)$ 是概率分布函数。(回顾在概率论的理论中,首先确定的是概率分布函数,然后只有当分布函数绝对连续时,才可以定义密度函数。)

密度估计问题的一般形式化表示可以描述为:在给定的函数集 $\{p(t)\}$ 中,寻找作为积分方程(1-10)的解的函数,但方程中的概率分布函数 $F(x)$ 是未知的,已知的是一系列给定的独立同分布数据 $x_1, \cdots, x_l, \cdots$,它们是按照未知的分布函数得到的。

利用这些数据,可以构造出一个统计学中非常重要的函数,称作经验分布函数(图 1.2):

$$F_l(x) = \frac{1}{l}\sum_{i=1}^{l}\theta(x - x_i),$$

其中 $\theta(u)$ 是阶跃函数,当 $u \geq 0$ 时取值为 1,而其他情况下取值为 0。

图 1.2　从数据 $x_1, \cdots, x_l$ 构造出的经验分布函数 $F_l(x)$,它逼近概率分布函数 $F(x)$

经验分布函数 $F_l(x)$ 到待求函数 $F(x)$ 的一致收敛性

$$\sup_x \left| F(x) - F_l(x) \right| \xrightarrow{P}_{l \to \infty} 0$$

是理论统计学中最基本的事实之一。我们还将在第二章的评述和第三章的评述中再次讨论这一事实。

这样,(从密度定义出发的)密度估计问题的一般性表示如下:

在概率分布函数未知,但给出了遵循这一函数的独立同分布数据 $x_1, \cdots, x_l, \cdots$ 的情况下,求解积分方程(1-10)。

利用已知数据可以构造经验分布函数 $F_l(x)$。因此,我们不知道方程(1-10)右边的准确结果,但知道它的一种逼近,且随着观测数目的增加这个逼近一致地收敛于未知的函数,我们必须在这种情况下求解积分方程(1-10)。

需要注意的是,在一类范围很宽的函数 $\{p(t)\}$ 中求解这个积分方程的问题是不适定的。由此我们得到两个结论:

(1) 一般来说,估计一个密度是一个很难的(不适定的)计算问题。

(2) 为了较好地解决这个问题,必须采用正则化技术(而不是“显而易见的”方法)。

已经证明,已经提出的所有非参数算法都可以通过用标准的正则化技术(使用不同类型的正则化因子),并用经验分布函数代替未知分布函数来得到(Vapnik, 1979, 1988)。

1.9 用有限数量信息解决问题的基本原则

现在我们把用有限数量信息解决问题的基本原则表达为:

在解决一个给定的问题时,要设法避免把解决一个更为一般的问题作为其中间步骤。

这个原则是显然的,但尽管如此,遵循这一原则并非易事。对于我们讨论的依赖关系估计问题来说,这一原则意味着,当解决模式识别或回归估计问题时,我们必须设法去"直接"寻找待求的函数(其含义将在下一节定义),而不是首先估计密度,然后用估计的密度来构造待求的函数。

注意到,密度估计是统计学中的一个全能的问题(知道了密度就可以解决各种问题)。估计密度一般说来是一个不适定问题,因此需要大量观测才能较好地解决。与此相对比,我们实际上需要解决的问题(决策规则估计或回归估计)是很特殊的,通常只需要有某一合理数量的观测就可以解决。

为了说明这一观点,让我们来考虑下面的情形。假设我们要构造一个把两个向量集合分开的决策规则,两个集合分别遵循两个正态分布: $N(\mu_1, \Sigma_1)$ 和 $N(\mu_2, \Sigma_2)$。为了构造(1-9)式的判别函数,必须从数据中估计两个 n 维均值向量 μ_1 和 μ_2 及两个 $n\times n$ 矩阵 Σ_1 和 Σ_2。最后的结果是得到一个二次的分类多项式:

$$f(x) = \operatorname{sgn}\left[\frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2) - \frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - C\right],$$

$$C = \ln\frac{|\Sigma_1|}{|\Sigma_2|} - \ln\frac{q_1}{1-q_1},$$

其中含有 $n(n+3)/2$ 个系数[①]。为了从未知密度函数的参数构造一个好的判别规则,我们需要很精确地估计协方差矩阵参数,因为判别函数中使用了协方差矩阵的逆(一般来说,估计密度是一个不适定问题;在这里的参数化问题里它可能给出病态协方差矩阵)。为了较好地估计高维的协方差矩阵,我们需要非常大量的观测,所需的观测数目是不可预测的(取决于实际协方差矩阵的特性)。因此,在高维空间中,(从两个不同的正态密度构造的)一般的正态判别函数很少在实际中成功。在实际中,人们使用线性判别函数,这是当两个协方差矩阵相同 $\Sigma_1 = \Sigma_2 = \Sigma$ 时的情况:

$$f(x) = \operatorname{sgn}\left[(\mu_1-\mu_2)^T\Sigma^{-1}x + \frac{1}{2}(\mu_2^T\Sigma^{-1}\mu_2) - \frac{1}{2}(\mu_1^T\Sigma^{-1}\mu_1) + \ln\frac{q_1}{1-q_1}\right]$$

(在这种情况下我们只需要估计判别函数的 n 个参数)。

值得注意的是,Fisher(1952)建议即使在两个协方差矩阵不同的情况下也采用线性判别函数,并提出了一种构造这样的函数的启发式方法[②]。

---

① 式中的 q_1 是第一类的先验概率,后同。——译者

② 在60年代,(在二次函数是最优解的情况下)构造最好的线性判别函数的问题得到了解决(Anderson and Bahadur, 1966)。在解决现实问题时,人们往往采用线性判别函数,即使知道最优解属于二次判别函数。

在第五章中,我们将通过在高维(256 维)空间中构造分类多项式(最多到 7 阶)来解决一个特殊的模式识别问题。它就是通过避免解决不必要的一般性问题而得到的。

1.10 基于经验数据的风险最小化模型

下面我们将说明,本章给出的学习问题的表示不仅使我们能够考虑在任意给定的函数集中进行估计的问题,而且能够实现利用小样本解决问题的基本原则,即避免去解决不必要的一般性问题。

1.10.1 模式识别

对模式识别问题来说,泛函(1-2)式计算了容许函数集中任意函数的错误率。模式识别问题就是利用样本从容许函数的集合中寻找使错误率最小的函数。这正是我们要得到的。

1.10.2 回归估计

在回归估计中,我们在损失函数(1-4)式下最小化泛函(1-2)式。这一泛函可以等价地重写为

$$R(\alpha)=\int (y-f(x,\alpha))^2 dF(x,y)$$

$$=\int (f(x,\alpha)-f_0(x))^2 dF(x)+\int (y-f_0(x))^2 dF(x,y), \qquad (1\text{-}11)$$

其中 $f_0(x)$ 是回归函数。注意到公式(1-11)中的第二项与所选择的函数无关,因此最小化这个泛函等价于最小化下面的泛函:

$$R^*(\alpha)=\int (f(x,\alpha)-f_0(x))^2 dF(x).$$

这个新的泛函等于容许函数集中的一个函数到回归函数之间的平方 $L_2(F)$ 距离。因此,我们考虑下面的问题:利用样本在容许函数的集合中寻找与回归函数最近的函数(在 $L_2(F)$ 度量下)。

如果我们接受了 $L_2(F)$ 度量,那么对回归估计问题的这种形式化描述(即最小化 $R(\alpha)$)就是直接的。(它不需要求解一个更一般的问题,比如寻找 $F(x,y)$。)

1.10.3 密度估计

最后考虑泛函

$$R(\alpha)=-\int \ln p(t,\alpha)dF(t)=-\int p_0(t)\ln p(t,\alpha)dt.$$

让我们在这个泛函后面加上一项常数(一个与估计函数无关的泛函)

$$c = \int \ln p_0(t)\, dF(t),$$

其中 $p_0(t)$ 和 $F(t)$ 分别是待求的密度和它的概率分布函数。我们得到

$$R^*(\alpha) = -\int \ln p(t, \alpha)\, dF(t) + \int \ln p_0(t)\, dF(t)$$

$$= -\int \ln \frac{p(t, \alpha)}{p_0(x)} p_0(t)\, dt.$$

公式右边的表达式是所谓的 Kullback-Leibler 距离，在统计学中用来度量对一个密度的逼近与真实密度之间的距离。因此我们考虑下面的问题：用给定的样本在容许的密度函数集合中寻找离待求密度的 Kullback-Leibler 距离最近的函数。如果接受了 Kullback-Leibler 距离，那么这种形式化表示是顺理成章的。

所有这些问题表示的简要形式就是基于经验数据最小化风险泛函的一般模型。

1.11 随机逼近推理

为了基于经验数据最小化风险泛函，我们在第一章中考虑了经验风险最小化归纳原则。这里我们讨论另一种一般的归纳原则，就是所谓的随机逼近方法，它是在 50 年代由 Robbins 和 Monroe(1951) 提出的。

根据这一原则，为了用独立同分布数据

$$z_1, \cdots, z_l$$

对参数 α 最小化泛函

$$R(\alpha) = \int Q(z, \alpha)\, dF(z),$$

我们采用下面的迭代过程：

$$\alpha(k+1) = \alpha(k) - \gamma_k \operatorname{grad}_\alpha Q(z_k, \alpha(k)), \quad k = 1, 2, \cdots, l, \tag{1-12}$$

其迭代步数等于观测样本的数目。已经证明，在梯度 $\operatorname{grad}_\alpha Q(z, \alpha)$ 和 γ_k 取值的一些很一般性的条件下，这种方法是一致的。

受到 Novikoff 定理的启发，Ya. Z. Tsypkin 和 M. A. Aizerman 于 1963 年在莫斯科控制科学研究所的学术讨论会上开始了关于学习过程一致性的讨论。他们研究了能保证学习过程一致性的两种一般性归纳原则：

(1) 随机逼近的原则；

(2) 经验风险最小化原则。

这两种归纳原则都被应用到用经验数据使风险泛函(1-6)式最小化的一般问题上。研究的结果是，在 1971 年创立了两种不同的一般性学习理论：

(1) 对随机逼近归纳推理的一般性渐近学习理论[①] (Aizerman, Braverman and Rozonoer, 1965 及 Tsypkin, 1971, 1973)。

---

① 在 1967 年 S. Amari 也提出了这一理论。

(2) 对 ERM 归纳推理的一般性非渐近模式识别理论(Vapnik and Chervonenkis, 1968, 1971, 1974)。(到 1979 年,这一理论被推广到任意基于经验数据的风险最小化问题(Vapnik, 1979)。)

然而,随机逼近原则是过于浪费了:它每一步只用到训练数据中的一个元素(见(1-12)式)。为了使它更经济,我们可以多次使用训练数据(分多个时间阶段)。这种情况下就马上出现下面的问题:

什么时候必须停止训练过程?

有两种可能的回答:

(1) 当对训练数据中的所有元素梯度值都非常小,以至于学习过程无法继续时,停止训练过程。

(2) 当学习过程没有饱和,但达到了某种停止准则时,停止学习过程。

容易看到,在第一种情况下,随机逼近方法只不过是最小化经验风险的一种特殊做法。而第二种情况则形成了最小化风险泛函的一种正则化方法[①]。因此,在这种"不浪费的方式"下,随机逼近方法既可以解释为 ERM 方法的归纳特性,也可以解释为正则化方法的归纳特性。

要完成我们关于传统归纳推理的讨论,需要考虑贝叶斯推理。为了应用这种推理,我们必须拥有额外的先验信息,作为对包括待求函数在内的参数化函数集的补充。也就是说,我们必须知道一个分布函数,它描述了容许函数的集合中各个函数是待求函数的概率。因此,贝叶斯推理是基于使用很强的先验信息的(它要求待求的函数属于学习机器的函数集合)。在这个意义上,它不是一种一般性的推理方法。我们将稍后在第四章的评述中讨论这种推理。

总之,除了 ERM 归纳原则外,我们还可以采用其他的归纳原则。但是,(与其他原则相比)ERM 原则看起来更有鲁棒性(它更好地利用经验数据,不依赖先验信息,并且存在很清晰的实现方法)。

因此,在分析学习过程时,核心问题变成了对 ERM 原则的探索。

① 早在 50 年代,在解决不适定问题的正则化理论提出之前,人们已经发现了解决不适定问题的迭代过程的停止准则具有正则化特性。

第二章

学习过程的一致性

本章是统计学习理论的基本内容之一,目的在于描述基于经验风险最小化归纳原则的学习过程的概念模型。这部分理论要回答的核心问题是,对一个使经验风险最小的学习过程,它在什么时候能够取得小的实际风险(即能够推广),而什么情况下不能。换句话说,学习理论的这部分内容的目的是研究经验风险最小化学习过程一致性的充分必要条件。

人们也许会提出这样的问题:

既然我们的目标是建立从有限数量的观测进行学习的算法,为什么还要研究渐近理论(一致性是一个渐近概念)?

对这个问题的回答是:

为了建立任何理论,我们都必须使用一些基本的概念,依据这些概念来发展理论。在统计学习理论中,使用描述一致性的充分必要条件的概念是十分重要的,这保证了所建立的理论具有一般性,并且不能从概念上进一步改进。

本章中最重要的内容是函数集的 VC 熵的概念,学习过程一致性的充分必要条件是建立在这一概念之上的。

利用这一概念,我们将在下一章得到关于学习过程收敛速度的定量特性,而后面将利用这些特性创建学习算法。

2.1 传统的一致性定义和非平凡一致性概念

设 $Q(z,\alpha)$ 是对给定的独立同分布观测 $z_1,\cdots,z_l$ 使经验风险泛函

$$R_{emp}=\frac{1}{l}\sum_{i=1}^{l}Q(z_i,\alpha)$$

最小化的函数。

定义　如果下面两个序列依概率收敛于同一个极限,即

$$R(\alpha_l)\xrightarrow[l\to\infty]{P}\inf_{\alpha\in\Lambda}R(\alpha), \tag{2-1}$$

$$R_{emp}(\alpha_l) \xrightarrow[l\to\infty]{P} \inf_{\alpha\in\Lambda} R(\alpha), \tag{2-2}$$

则我们说 ERM 原则(或方法)对函数集 $Q(z,\alpha)$, $\alpha\in\Lambda$ 和概率分布函数 $F(z)$ 是一致的(参见图 2.1 的示意)。

图 2.1　如果期望风险 $R(\alpha)$ 和经验风险 $R_{emp}(\alpha)$ 都收敛到最小可能的风险值 $\inf_{\alpha\in\Lambda} R(\alpha)$, 则学习过程是一致的

换句话说,一个 ERM 方法,如果它提供一个函数序列 $Q(z,\alpha_l)$, $l=1,2,\cdots$, 对这个序列来说期望风险和经验风险都收敛到最小可能的风险值,则这个 ERM 方法是一致的。(2-1)式保证了所达到的风险收敛于最好的可能值,而(2-2)式保证了可以在经验风险的取值基础上估计最小可能的风险。

本章的目的就是描述 ERM 方法一致性的条件。我们将从函数集和概率测度的一般特性出发来得到这些条件。

然而不幸的是,对于上面这种一致性的传统定义,得到这样的条件是不可能的,因为这种定义中包括了平凡一致性的情况。

什么是一致性的平凡情况呢?

假设我们已经建立了某个函数集 $Q(z,\alpha)$, $\alpha\in\Lambda$, 对这个函数集 ERM 方法是不一致的。考虑另一个扩展的函数集,它包括了这个函数集和一个额外的函数 $\varphi(z)$。假设这个额外的函数满足不等式

$$\inf_{\alpha\in\Lambda} Q(z,\alpha) > \varphi(z), \quad \forall z,$$

显然对这个扩展的函数集(包括函数 $\varphi(z)$ 在内)来说,ERM 方法就是一致的了(图 2.2)。实际上,对任何分布函数和对任意数量的观测,经验风险的最小值都将在函数 $\varphi(z)$ 上取得,而它也给出了期望风险的最小值。

这个例子说明存在平凡一致性的情况,在这种情况下一致性仅取决于函数集中是否包含一个最小化函数。

因此,任何采用传统定义的一致性理论必须能够确定其中是否可能有平凡一致性的情况。这就是说这种理论必须考虑到给定函数集中特定的函数。

为了建立一种 ERM 方法一致性的理论,使它不是依赖于函数集中个别元素的特性,而是依赖于函数集的整体特性(容量),我们需要修正一致性的定义,从中刨除平凡一致性的情况。

图 2.2　平凡一致性的情况。ERM 方法对函数集 $Q(z,\alpha),\ \alpha\in\Lambda$ 来说不一致，而对函数集 $\{\ (z)\}\cup Q(z,\alpha),\ \alpha\in\Lambda$ 则一致

定义　对函数集 $Q(z,\alpha),\ \alpha\in\Lambda$，定义其子集 $\Lambda(c)$ 如下：

$$\Lambda(c)=\int Q(z,\alpha)\,dF(z)>c,\quad \alpha\in\Lambda .$$

如果对函数集的任意非空子集 $\Lambda(c),\ c\in(-\infty,\infty)$ 都有

$$\inf_{\alpha\in\Lambda(c)} R_{emp}(\alpha) \xrightarrow[l\to\infty]{P} \inf_{\alpha\in\Lambda(c)} R(\alpha) \tag{2-3}$$

成立，则我们说 ERM 方法对函数集 $Q(z,\alpha),\ \alpha\in\Lambda$ 和概率分布函数 $F(z)$ 是非平凡一致的。

换句话说，一个 ERM 方法，如果把函数集中取得风险最小值的函数去掉后仍然能够满足(2-3)式的收敛关系，则这个 ERM 方法是非平凡一致的。

注意，在前面给出的传统的一致性定义中，我们用了(2-1)式和(2-2)式两个条件，而在这里的非平凡一致性定义中则只使用了(2-3)式一个条件。可以证明，在非平凡一致性的情况下条件(2-1)式将自动得到满足。

在本章中，我们将要研究的是非平凡一致性。为了简单起见，我们在后面说一致性就是指非平凡一致性。

2.2　学习理论的关键定理

下面的定理被称作学习理论的关键定理(Vapnik and Chervonenkis，1989)：

定理 2.1　设函数集 $Q(z,\alpha),\ \alpha\in\Lambda$ 满足条件

$$A\leqslant\int Q(z,\alpha)\,dF(z)\leq B\quad (A\leq R(\alpha)\leq B),$$

那么，ERM 原则一致性的充分必要条件是：经验风险 $R_{emp}(\alpha)$ 在函数集 $Q(z,\alpha),\ \alpha\in\Lambda$ 上在如下意义下一致收敛于实际风险 $R(\alpha)$：

$$\lim_{l\to\infty} P\left\{\sup_{\alpha\in\Lambda}(R(\alpha)-R_{emp}(\alpha))>\varepsilon\right\}=0,\quad \forall\varepsilon>0. \tag{2-4}$$

我们把这种一致收敛称作一致单边收敛[①]。

换句话说,根据学习理论的关键定理,ERM 原则的一致性等价于(2-4)式的一致单边收敛成立。

从概念的角度看,这个定理是十分重要的,因为它指出了 ERM 原则一致性的条件是必要地(和充分地)取决于函数集中"最坏"的函数(在(2-4)的意义上)的。也就是说,根据这一定理,对 ERM 原则的任何分析都必须是"最坏情况分析"[②]。

关于最大似然法

我们在第一章已经看到,ERM 原则中包括了最大似然法。但是,对最大似然法,我们定义另外一种非平凡一致性的概念。

定义　如果对给定密度集 $p(x,\alpha)$, $\alpha\in\Lambda$[③]中的任何密度 $p(x,\alpha_0)$,下面的依概率收敛关系

$$\inf_{\alpha\in\Lambda}\frac{1}{l}\sum_{i=1}^{l}(-\log p(x_i,\alpha)) \xrightarrow[l\to\infty]{P} \inf_{\alpha\in\Lambda}\int(-\log p(x,\alpha))p(x,\alpha_0)\,dx$$

成立,则我们说最大似然方法是非平凡一致的。其中,$x_1,\cdots,x_l$ 是根据密度 $p_0(x)$ 得到的独立同分布样本。

换句话说,如果最大似然方法对估计容许的密度集中的任意密度都一致,则我们定义它是非平凡一致的。

对于最大似然方法,下面的关键定理成立(Vapnik and Chervonenkis, 1989):

定理 2.2　最大似然方法在密度集

$$0<a\le p(x,\alpha)\le A<\infty,\quad \alpha\in\Lambda$$

上非平凡一致的充分必要条件是,损失函数集

$$Q(x,\alpha)=-\ln p(x,\alpha),\quad \alpha\in\Lambda,$$

一致单边收敛对某一(任一)概率密度 $p(x,\alpha)$, $\alpha\in\Lambda$ 成立。

2.3　一致双边收敛的充分必要条件

学习理论的关键定理把 ERM 方法一致性的问题转化为了一致收敛的问题((2-4)式)。为了研究一致收敛的充分必要条件,我们考虑两个随机过程,它们被称作经验过程。

---

① 与下式定义的所谓一致双边收敛相对应:

$$\lim_{l\to\infty}P\{\sup_{\alpha\in\Lambda}|R(\alpha)-R_{emp}(\alpha)|>\varepsilon\}=0,\quad\forall\varepsilon>0.$$

② 下面的事实确认了这一定理的重要性。在 80 年代末和 90 年代初,相对于这里的统计学习理论是一种"最坏情况分析"的理论的观点,人们曾经尝试了几种不同的方法。在这些方法中,人们希望发展一种"真实情况分析"的学习理论。但是,根据关键定理,对 ERM 原则这种理论是不可能的。

③ 原著中误写作 $p(x,\alpha)\in\Lambda$。——译者

考虑下面的随机变量序列

$$\xi = \sup_{\alpha\in\Lambda}\left| \int Q(z,\alpha)\,dF(z) - \frac{1}{l}\sum_{i=1}^{l} Q(z_i,\alpha)\right|, \quad l = 1,2,\cdots. \tag{2-5}$$

这一随机变量序列既依赖于概率测度 $F(z)$，也依赖于函数集 $Q(z,\alpha)$，$\alpha\in\Lambda$，我们称之为一个双边经验过程。要研究的问题是在什么条件下这个经验过程依概率收敛于零。过程(2-5)式依概率收敛是指等式

$$\lim_{l\to\infty} P\left\{\sup_{\alpha\in\Lambda}\left| \int Q(z,\alpha)\,dF(z) - \frac{1}{l}\sum_{i=1}^{l} Q(z_i,\alpha)\right| > \varepsilon\right\} = 0, \quad \forall\,\varepsilon > 0 \tag{2-6}$$

成立。

与经验过程 ξ 一起，我们也考虑单边经验过程，它是这样一个随机变量序列：

$$\xi' = \sup_{\alpha\in\Lambda}\left(\int Q(z,\alpha)\,dF(z) - \frac{1}{l}\sum_{i=1}^{l} Q(z_i,\alpha)\right)_+, \quad l = 1,2,\cdots, \tag{2-7}$$

其中我们记$(\ \ )_+$为

$$(u)_+ = \begin{cases} u & \text{如果 } u > 0 \\ 0 & \text{其他} \end{cases}.$$

我们要研究的问题是，在什么条件下随机变量序列 ξ' 依概率收敛于零。过程(2-7)式依概率收敛是指下面的等式成立：

$$\lim_{l\to\infty} P\left\{\sup_{\alpha\in\Lambda}\left(\int Q(z,\alpha)\,dF(z) - \frac{1}{l}\sum_{i=1}^{l} Q(z_i,\alpha)\right) > \varepsilon\right\} = 0, \quad \forall\,\varepsilon > 0. \tag{2-8}$$

根据关键定理，(2-8)式的一致单边收敛是 ERM 方法一致的充分必要条件。

我们将看到，一致双边收敛的条件在创建一致单边收敛的条件中起了非常重要的作用。

2.3.1 关于大数定律及其推广

注意到，如果函数集 $Q(z,\alpha)$，$\alpha\in\Lambda$ 中只包含一个元素，那么(2-5)式中定义的随机变量序列 ξ 总是依概率收敛到零。这一事实形成了统计学中的基本定律——大数定律：

随着(观测数量)l 的增加，随机变量序列 ξ 收敛于零。

对于函数集中包含有限个元素的情况，可以容易地把大数定律推广为：

如果函数集 $Q(z,\alpha)$，$\alpha\in\Lambda$ 包含有限数目 N 个元素，那么随机变量序列 ξ 依概率收敛于零。

这种情况可以解释为在 N 维向量空间中的大数定律(函数集中的每个函数对应于一维坐标；向量空间中的大数定律说明所有维坐标同时依概率收敛)。

当函数集 $Q(z,\alpha)$，$\alpha\in\Lambda$ 中有无限多个元素时问题就出现了。与有限元素情况下不同，对于包含无穷多元素的集合，随机变量序列 ξ 并不一定收敛于零。这时的问题是：

在函数集 $Q(z,\alpha)$，$\alpha\in\Lambda$ 和概率测度 $F(z)$ 的什么特性下，随机变量序列 ξ 依概率收敛于零。

这种情况下我们就需要泛函空间的大数定律(在函数 $Q(z,\alpha)$，$\alpha\in\Lambda$ 的空间)，即在一个给定的函数集上，存在均值到其数学期望的一致(双边)收敛。

因此,是否存在泛函空间的大数定律(均值到其期望[①]的一致双边收敛)的问题可以看作是传统的大数定律的推广。

注意到,在传统的统计学中,并没有考虑是否存在一致单边收敛的问题;这个问题之所以变得重要,是由于关键定理中所指出的分析 ERM 归纳原则一致性的方法。

一致单边收敛和一致双边收敛的充分必要条件都是在一个新概念的基础上得到的,这个概念叫做在 l 个样本上函数集 $Q(z, \alpha)$, $\alpha \in \Lambda$ 的熵。

为了简单起见我们将分两步介绍这个概念:首先对指示函数集的这一概念进行介绍(指示函数指只取 0 和 1 两个值的函数),然后再介绍有界实函数集情况。

2.3.2 指示函数集的熵

设 $Q(z, \alpha)$, $\alpha \in \Lambda$ 是一个指示函数集,考虑样本

$$z_1, \cdots, z_l.$$

定义一个量 $N^{\Lambda}(z_1, \cdots, z_l)$,它代表用指示函数集中的函数能够把给定的样本分成多少种不同的分类,我们用这个量来表征函数集 $Q(z, \alpha)$, $\alpha \in \Lambda$ 在给定的数据集上的多样性。

更形式化一点说,考虑 α 在 Λ 中取不同的值得到的 l 维二值向量的集合

$$q(\alpha) = (Q(z_1, \alpha), \cdots, Q(z_l, \alpha)), \quad \alpha \in \Lambda,$$

从几何上说,样本 $z_1, \cdots, z_l$ 的所有可能的分类情况可以构成一个 l 维超立方体,$N^{\Lambda}(z_1, \cdots, z_l)$ 就是用函数集 $Q(z, \alpha)$, $\alpha \in \Lambda$ 可以得到的这个立方体上不同的顶点数目(图 2.3)。

我们把值

$$H^{\Lambda}(z_1, \cdots, z_l) = \ln N^{\Lambda}(z_1, \cdots, z_l)$$

叫做随机熵,它描述了函数集在给定数据上的多样性。$H^{\Lambda}(z_1, \cdots, z_l)$ 是一个随机数,因为它是建立在独立同分布的数据之上的。考虑随机熵在联合分布函数 $F(z_1, \cdots, z_l)$ 上的期望:

$$H^{\Lambda}(l) = E\ln N^{\Lambda}(z_1, \cdots, z_l),$$

我们把这个量称作指示函数集 $Q(z, \alpha)$, $\alpha \in \Lambda$ 在数量为 l 的样本上的熵,它依赖于函数集 $Q(z, \alpha)$, $\alpha \in \Lambda$、概率测度以及观测数目 l,反映了给定指示函数集在数目为 l 的样本上期望的多样性。

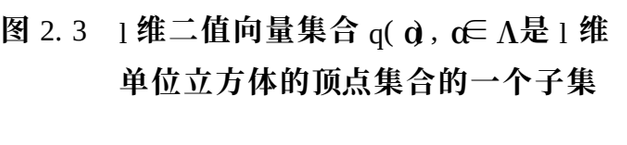

图 2.3 l 维二值向量集合 $q(\alpha)$, $\alpha \in \Lambda$ 是 l 维单位立方体的顶点集合的一个子集

2.3.3 实函数集的熵

下面我们把 l 个样本上指示函数集的熵的定义进行推广。

定义 设 $A \leq Q(z, \alpha) \leq B$, $\alpha \in \Lambda$ 是一个有界损失函数的集合,用这个函数集和训练

① 原著中这里误写为均值到其概率的一致双边收敛。—— 译者

集 $z_1, \cdots, z_l$, 可以构造下面的 l 维向量集合:

$$q(\alpha) = (Q(z_1, \alpha), \cdots, Q(z_l, \alpha)), \quad \alpha \in \Lambda. \tag{2-9}$$

这个向量集合处在 l 维立方体之中(图 2.4), 并且在 C 度量(或在 L_p 度量)下有一个有限的最小 ε网格[①]。令 $N = N^{\Lambda}(\varepsilon, z_1, \cdots, z_l)$ 是向量集 $q(\alpha)$, $\alpha \in \Lambda$的最小 ε网格的元素数目。

注意, $N^{\Lambda}(\varepsilon, z_1, \cdots, z_l)$ 是一个随机变量, 因为它是用随机向量 $z_1, \cdots, z_l$ 构造的。这个随机值的对数

$$H^{\Lambda}(\varepsilon, z_1, \cdots, z_l) = \ln N^{\Lambda}(\varepsilon, z_1, \cdots, z_l)$$

称作函数集 $A \leq Q(z, \alpha) \leq B$, $\alpha \in \Lambda$ 在样本 $z_1, \cdots, z_l$ 上的随机 VC 熵。随机 VC 熵的期望

$$H^{\Lambda}(\varepsilon, l) = EH^{\Lambda}(\varepsilon, z_1, \cdots, z_l)$$

称作函数集 $A \leq Q(z, \alpha) \leq B$, $\alpha \in \Lambda$ 在数量为 l 的样本上的 VC 熵[②], 这里的期望是对乘积测度 $F(z_1, \cdots, z_l)$ 进行的。

图 2.4 l 维向量集合 $q(\alpha)$, $\alpha \in \Lambda$处在一个 l 维立方体之中

注意, 上面给出的实函数的熵的定义是对指示函数集的熵定义的推广。实际上, 对于指示函数集, $\varepsilon < 1$ 的最小 ε网格不依赖于 ε 且是单位立方体的顶点的一个子集。因此, 对 $\varepsilon < 1$, 有

$$N^{\Lambda}(\varepsilon, z_1, \cdots, z_l) = N^{\Lambda}(z_1, \cdots, z_l),$$
$$H^{\Lambda}(\varepsilon, z_1, \cdots, z_l) = H^{\Lambda}(z_1, \cdots, z_l),$$
$$H^{\Lambda}(\varepsilon, l) = H^{\Lambda}(l).$$

下面我们将对有界实函数集给出有关理论。所得到的一般结论当然也适用于指示函数集。

2.3.4 一致双边收敛的条件

在关于函数集 $Q(z, \alpha)$, $\alpha \in \Lambda$ 可测性的一定的(技术性)条件下, 有下面的定理成立。

定理 2.3 一致双边收敛(式(2-6))的充分必要条件是等式

$$\lim_{l \to \infty} \frac{H^{\Lambda}(\varepsilon, l)}{l} = 0, \quad \forall \varepsilon > 0 \tag{2-10}$$

成立。

---

① 对向量集合 $q(\alpha)$, $\alpha \in \Lambda$, 如果

(i) 存在 $N = N^{\Lambda}(\varepsilon, z_1, \cdots, z_l)$ 个向量 $q(\alpha_1), \cdots, q(\alpha_N)$, 使得对任意向量 $q(\alpha^*)$, $\alpha^* \in \Lambda$, 我们可以在这 N 个向量中找到一个 $q(\alpha_i)$, 它 ε靠近向量 $q(\alpha^*)$(在某个给定的度量下)。在 C 度量下这就意味着

$$\rho_C(q(\alpha^*), q(\alpha_i)) = \max_{1 \leq i \leq l} |Q(z_i, \alpha^*) - Q(z_i, \alpha_i)| \leq \varepsilon$$

(ii) N 是具有这一特性的向量的最小数目, 则向量集合 $q(\alpha)$, $\alpha \in \Lambda$有一个最小 ε网格。

② VC 熵与传统度量的 ε熵

$$H(\varepsilon) = \ln N^{\Lambda}(\varepsilon)$$

在下列方面不同: $N^{\Lambda}(\varepsilon)$是函数集 $Q(z, \alpha)$, $\alpha \in \Lambda$的最小 ε网格的基数, 而 VC 熵是函数集在 l 个样本上的多样性的期望。

也就是说,随着观测数目的增加,VC 熵与观测数目的比值应该趋近于零。

推论 在指示函数集 $Q(z, \alpha)$, $\alpha \in \Lambda$ 可测性的一定条件下,一致双边收敛的充分必要条件是

$$\lim_{l \to \infty} \frac{H^{\Lambda}(l)}{l} = 0,$$

这是(2-10)式的一个特例。

一致双边收敛的这个条件是在 1968 年得到的(Vapnik and Chervonenkis, 1968, 1971)。它对有界实函数集的推广(定理 2.3)是在 1981 年发现的(Vapnik and Chervonenkis, 1981)。

2.4 一致单边收敛的充分必要条件

一致双边收敛可以描述为

$$\lim_{l \to \infty} P\{\sup_{\alpha}(R(\alpha) - R_{emp}(\alpha)) > \varepsilon \text{ 或 } \sup_{\alpha}(R_{emp}(\alpha) - R(\alpha)) > \varepsilon\} = 0 \qquad (2\text{-}11)$$

条件(2-11)式中包括了一致单边收敛,因此形成了 ERM 方法一致性的一个充分条件。然而注意到,在解决学习问题时,我们面对一种非对称情况:我们要求最小化经验风险时的一致性,却不关心最大化经验风险时的一致性。因此(2-11)等式左边 ERM 方法一致性的第二项条件可以不满足。

下面的定理将给出一个条件,在这个条件下一致性在最小化经验风险时成立,而在最大化经验风险时并不一定成立(Vapnik and Chervonenkis, 1989)。

我们考虑有界实函数集 $Q(z, \alpha)$, $\alpha \in \Lambda$ 和一个新的函数集 $Q^*(z, \alpha^*)$, $\alpha^* \in \Lambda^*$,这个新函数集满足一定的可测性条件和下面的条件:对 $Q(z, \alpha)$, $\alpha \in \Lambda$ 中的任意函数,在 $Q^*(z, \alpha^*)$, $\alpha^* \in \Lambda^*$ 中存在一个函数,使得

$$\begin{aligned} &Q(z, \alpha) - Q^*(z, \alpha^*) \geq 0, \quad \forall z, \\ &\int (Q(z, \alpha) - Q^*(z, \alpha^*))\, dF(z) \leq \delta \end{aligned} \qquad (2\text{-}12)$$

(见图 2.5)。

图 2.5 对任意函数 $Q(z, \alpha)$, $\alpha \in \Lambda$,考虑一个函数 $Q^*(z, \alpha^*)$, $\alpha^* \in \Lambda^*$,它不超出 $Q(z, \alpha)$ 而且与之非常接近

定理 2.4　对完全有界函数集 $Q(z,\alpha)$, $\alpha\in\Lambda$, 经验均值一致单边收敛于其期望((2-8)式) 的充分必要条件是: 对任意的正 δ η和 ε 存在一个满足(2-12) 式的函数集 $Q^*(z,\alpha^*)$, $\alpha^*\in\Lambda^*$, 使得函数集$Q^*(z,\alpha^*)$, $\alpha^*\in\Lambda^*$ 在l 个样本上的 ε熵[①]满足下面的不等式:

$$\lim_{l\to\infty}\frac{H^{\Lambda^*}(\varepsilon,l)}{l}<\eta \tag{2-13}$$

换句话说, 有界函数集 $Q(z,\alpha)$, $\alpha\in\Lambda$一致单边收敛的充分必要条件就是, 存在与 $Q(z,\alpha)$, $\alpha\in\Lambda$非常接近(在(2-12) 式意义上)的另一个函数集 $Q^*(z,\alpha^*)$, $\alpha^*\in\Lambda^*$, 对这个新函数集, 条件(2-13) 式成立。这里注意, 条件(2-13) 式比定理 2.3 中的条件(2-10) 式更弱。

根据关键定理, 这也是 ERM 方法一致性的充分必要条件。

2.5　不可证伪性理论

从形式上看, 定理 2.1, 2.3 和 2.4 给出了基于 ERM 归纳原则的学习的概念模型。但是, 为了证明定理 2.4, 为了更深入地理解 ERM 原则, 我们需要回答下面的问题:

如果不满足定理 2.4 的条件会怎么样?

为什么在这种情况下 ERM 方法不具一致性?

下面我们将说明, 如果存在一个 ε, 使得

$$\lim_{l\to\infty}\frac{H^{\Lambda}(\varepsilon,l)}{l}\neq 0,$$

那么采用函数 $Q(z,\alpha)$, $\alpha\in\Lambda$的学习机器就会遇到在科学哲学中称为不可证伪性理论的情况。

在给出这一理论的正式内容之前, 我们先来回顾一下关于不可证伪性的思想。

Kant 的区分问题和 Popper 的不可证伪理论

从古代哲学的年代开始, 人们就接受了两种推理模型, 即

(1) 演绎, 指从一般到特殊的推理;

(2) 归纳, 指从特殊到一般的推理。

对于演绎方法, 一种理想的模型是, 定义一个公理和推理规则系统, 利用这一系统得到一系列推理(结论)。演绎方法应保证我们能够从真实的前提得出真实的结论。

推理的归纳方法是由特殊的论断形成一般性判断。然而, 从真实的特殊论断得到的一般性判断却不总是真实的。不过, 人们假定存在这样的归纳推理, 它们得到的一般判断能够被证实。

最初由 I. Kant 提出的区分问题是归纳理论中的一个中心问题, 这个问题是:

---

① 定理中提到函数集在 l 个样本上的 ε熵就是前面定义的 VC 熵, 这里作者可能是使用了以前的叫法。——译者

一个归纳过程得到证实的情况与归纳步骤没有得到证实的情况之间的区别是什么?

这个区分问题通常是在自然科学的哲学领域中讨论的。自然科学的所有理论都是对观察到的真实事实的推广,因此理论是用归纳推理建立起来的。在自然科学的历史上,既有反映事实的真理论(比如化学),也有不反映事实的假理论(比如炼金术)。有时需要很多年的实验才能证明一种理论是错误的。

那么就有下面的问题:

是否存在一种形式化方法来区分真理论和假理论?

让我们假定气象学是一种真理论而占星学是一种假理论。它们之间形式上的区别是什么呢?

(1) 是它们模型的复杂程度不同吗?

(2) 是它们模型的预测能力不同吗?

(3) 是它们对数学的运用不同吗?

(4) 是它们推理的形式化层次不同吗?

这些因素中没有哪一项能够明确说明两种理论谁更有优势。

(1) 占星学模型的复杂程度不亚于气象学模型。

(2) 两种理论都在它们的一些预测中失败。

(3) 占星学通过求解微分方程来恢复行星的位置,它们并不比气象学中的基本方程简单。

(4) 最后,在两种理论中,推理的形式化层次相同,都包括对现实的形式化描述和非形式化解释两部分。

在 30 年代,K. Popper(波普)(1968)提出了他著名的区分真理论和假理论的准则。根据 Popper 的思想,一个理论可以被证实的一个必要条件是它存在被证伪的可能性。所谓一个理论被证伪,Popper 指的是存在这样一些特殊的论断,它们属于这一理论的范畴但却不能被这一理论解释。如果一个理论可能被证伪,则它满足了作为一个科学理论的必要条件。

让我们回到前面的例子中来。气象学和占星学都做天气预报,比如考虑下面的论断:

曾经有一次,在美国新泽西州的七月出现了热带暴风雨然后又下了雪。

假定根据气象学理论这种情况不可能出现,那么这个论断就证伪了这一理论,因为如果这种情况真正发生了(注意没有人能以概率 1 保证这是不可能的[①]),这一理论将无法对它进行解释。在这个例子中气象学满足了被看作一个科学理论的必要条件。

假定这个论断能够被占星学理论解释。(在繁星密布的天空中有很多元素,它们可以被用来建立一种解释。)这种情况下,这个论断就没有证伪这一理论。如果没有任何例子能够证伪占星学理论,那么根据 Popper 的观点,占星学就应该被视作一种非科学的理论。

在下一节里我们将描述不可证伪性的定理。我们将看到,如果对某一函数集一致收敛

① 回顾 Laplace(拉普拉斯)关于太阳到今天为止每天都升起的条件下太阳明天会升起的概率的计算。根据我们使用和相信的模型,太阳将肯定升起,但是,以概率 1,我们只能断定,在过去有历史记载的数千年中一直到现在,太阳每天都升起。

条件不成立,就会出现不可证伪的情况。

2.6 关于不可证伪性的定理

下面我们说明如果一致双边收敛不成立,则最小化经验风险的方法是不可证伪的。

2.6.1 完全(Popper)不可证伪的情况

为了清楚地解释为什么会出现这种情形,我们从最简单的情况开始。回顾前面的内容,根据 VC 熵的定义,对于一个指示函数集,下面的关系成立:

$$H^{\Lambda}(l) = E\ln N^{\Lambda}(z_1, \cdots, z_l) \text{ 且 } N^{\Lambda}(z_1, \cdots, z_l) \leq 2^l.$$

现在假设对指示函数集 $Q(z, \alpha)$, $\alpha \in \Lambda$ 的 VC 熵,下面的等式成立:

$$\lim_{l\to\infty} \frac{H^{\Lambda}(l)}{l} = \ln 2.$$

可以证明,当观测数目 l 增大时,熵与观测数目的比值 $H^{\Lambda}(l)/l$ 单调减小[①]。因此,如果熵与观测数目之比的极限趋近于 $\ln 2$,那么对任意有限数目 l,等式

$$\frac{H^{\Lambda}(l)}{l} = \ln 2$$

成立。

这意味着对几乎所有的样本集 $z_1, \cdots, z_l$(即除了零测度集合外的所有样本集),等式

$$N^{\Lambda}(z_1, \cdots, z_l) = 2^l$$

成立。

换句话说,对这个学习机器的函数集,几乎所有(任意数量 l 的)样本 $z_1, \cdots, z_l$ 都可以被其中的函数按全部可能的方式分开。这表明这个机器的最小经验风险等于零。我们说这个学习机器是不可证伪的,因为它能够对几乎所有数据给出一个一般的解释(函数)(图2.6)。

图 2.6 一个采用函数集 $Q(z, \alpha)$, $\alpha \in \Lambda$ 的学习机器,如果对样本产生器给出的几乎所有样本 $z_1, \cdots, z_l$ 以及对这些 z 的所有可能的标号 $\delta, \cdots, \delta$,机器中总包含一个函数 $Q(z, \alpha^*)$ 满足 $\delta = Q(z_i, \alpha^*)$, $i = 1, \cdots, l$,那么这个学习机器是不可证伪的

① 这一论断类似于说相对信息的值(相对于观测数目)不能随着观测数目的增加而增加。

注意,这时不论期望风险的取值如何,经验风险的最小值总等于零。

2.6.2 关于部分不可证伪的定理

在指示函数集的熵与观测数目的比值趋近于一个非零极限的情况下,下面的定理指出,存在原空间 $z \in Z$ 的一定的子空间,在其中这个学习机器是不可证伪的(Vapnik and Chervonenkis, 1989)。

定理 2.5 对指示函数集 $Q(z, \alpha)$, $\alpha \in \Lambda$,设成立收敛关系

$$\lim_{l \to \infty} \frac{H^{\Lambda}(l)}{l} = c > 0,$$

那么存在空间 Z 的一个概率测度是

$$P(Z^*) = a(c) \neq 0$$

的子空间 Z^*,使得对几乎所有训练集

$$z_1, \cdots, z_l$$

与集合 Z^* 的交集

$$z_1^*, \cdots, z_k^* = (z_1, \cdots, z_l) \cap Z^*,$$

以及任何给定的二值数序列

$$\delta, \cdots, \delta, \quad \delta \in \{0, 1\},$$

都存在一个函数 $Q(z, \alpha^*)$ 满足下面的等式:

$$\delta = Q(z_i^*, \alpha^*), \quad i = 1, 2, \cdots, k.$$

因此,如果一致双边收敛的条件不成立,那么存在某个输入空间的子空间,在其中学习机器是不可证伪的(图 2.7)。

图 2.7 一个采用函数集 $Q(z, \alpha)$, $\alpha \in \Lambda$ 的学习机器,如果存在一个非零测度的区域 $Z^* \in Z$,使得对样本产生器给出的几乎所有样本 $z_1, \cdots, z_l$ 及对这些 z 的所有可能的标号 $\delta, \cdots, \delta$,机器中总包含一个函数 $Q(z, \alpha^*)$ 对区域 Z^* 中的所有 z_i 满足 $\delta = Q(z_i, \alpha^*)$, $i = 1, \cdots, l$,那么这个学习机器是部分不可证伪的。(原著中误写为 $\delta = Q(z_i, \alpha)$ ——译者)

2.6.3 关于潜在不可证伪的定理

现在我们来考虑一致有界实函数集合

$$|Q(z,\alpha)|\leq C,\quad \alpha\in\Lambda$$

对这类函数集,存在一种更复杂的不可证伪性模型。因此我们给出下面的不可证伪性的定义。

定义 对一个有容许实函数集 $Q(z,\alpha)$, $\alpha\in\Lambda$ 的学习机器,输入产生器服从分布 $F(x)$,如果存在两个函数[①]

$$\psi(z)\geq\psi(z)$$

使得

(1) 存在一个正的常数 c,使得

$$\int(\psi(z)-\psi(z))dF(z)=c>0$$

成立(这个等式说明两个函数 $\psi(z)$ 和 $\psi(z)$ 是有实质性差别的);

(2) 对几乎任何样本

$$z_1,\cdots,z_l$$

和任何二值数序列

$$\delta(1),\cdots,\delta(l),\quad \delta(i)\in\{0,1\}$$

以及任何 $\varepsilon>0$,总可以在函数集 $Q(z,\alpha)$, $\alpha\in\Lambda$ 中找到一个函数 $Q(z,\alpha^*)$,使得不等式

$$|\psi_{\delta(i)}(z_i)-Q(z_i,\alpha^*)|<\varepsilon$$

成立,那么我们说这个学习机器对服从分布 $F(x)$ 的输入产生器是潜在不可证伪的。

在这种不可证伪性的定义中,我们用两个有实质性差别的函数 $\psi(z)$ 和 $\psi(z)$ 来产生函数对给定向量 z_i 的值 y_i。为了使这些值更有任意性,可以按任意的规则 $\delta(i)$ 在两个函数之间进行切换。考查函数集 $Q(z,\alpha)$, $\alpha\in\Lambda$,和根据分布函数 $F(z)$ 产生的输入向量,如果对基于随机向量 z_i 和切换规则 $\delta(i)$ 得到的 $(\psi_{\delta(i)}(z_i),z_i)$ 对的几乎任何序列,都能在函数集中找到一个函数,它能以很高的精度描述这些取值对,那么我们就说这个函数集形成了一个对根据分布函数 $F(z)$ 产生的输入向量潜在不可证伪的机器(图 2.8)。

注意到,这种不可证伪性的定义推广了 Popper 的概念:

(1) 在第 2.6.1 小节考虑的最简单的例子里,对指示函数集 $Q(z,\alpha)$, $\alpha\in\Lambda$,我们实际上在 $\psi(z)=1$ 和 $\psi(z)=0$ 的情况下使用了这一概念;

(2) 在定理 2.5 中,我们可以把这两个函数定义成

$$\psi(z)=\begin{cases}1 & \text{若 } z\in Z^*\\ Q(z) & \text{若 } z\mid Z^*\end{cases},\qquad \psi(z)=\begin{cases}0 & \text{若 } z\in Z^*\\ Q(z) & \text{若 } z\mid Z^*\end{cases},$$

其中 $Q(z)$ 是某一指示函数。

基于这种潜在不可证伪性的定义,我们给出下面的一般性定理,它对任意一致有界函数的集合(包括指示函数集)成立(Vapnik and Chervonenkis, 1989)。

定理 2.6 设对一致有界实函数集合 $Q(z,\alpha)$, $\alpha\in\Lambda$,存在一个 ε,使收敛

$$\lim_{l\to\infty}\frac{H^{\Lambda}(\varepsilon,l)}{l}=c^*>0$$

成立,那么采用这个函数集的学习机器是潜在不可证伪的。

这样,如果定理 2.4 的条件不成立(当然此时定理 2.3 的条件也不成立),那么学习机

---

① 这两个函数不一定属于集合 $Q(z,\alpha)$, $\alpha\in\Lambda$

图 2.8　一个函数集为 $Q(z, \alpha)$, $\alpha \in \Lambda$ 的学习机器, 如果对任何 $\varepsilon > 0$, 存在两个有实质性差别的函数 $\psi(z)$ 和 $\psi(z)$, 使得对样本产生器给出的几乎所有样本 $z_1, \cdots, z_l$, 和用这两个函数及规则 $u_i = \psi(z_i)$ (z_i) 构造的任何取值 $u_1, \cdots, u_l$ (其中 $\delta(z) \in \{0, 1\}$ 是任一二值函数), 总有一个函数 $Q(z, \alpha^*)$ 满足不等式 $|\psi(z_i) - Q(z_i, \alpha^*)| \leq \varepsilon$, $i = 1, \cdots, l$, 那么这个学习机器是潜在不可证伪的

器就是不可证伪的。这就是为什么 ERM 原则可能不一致的主要原因。

在继续讨论统计学习理论的其他内容之前, 我想来看一下 Popper 的思想是多么神奇。早在 20 世纪 30 年代 Popper 提出了确定(在很宽的哲学意义上的)推广能力的一般概念, 而到了 90 年代它成了在分析 ERM 归纳原则的一致性时最重要的概念之一。

2.7　学习理论的三个里程碑

下面我们再来考虑指示函数集 $Q(z, \alpha)$, $\alpha \in \Lambda$ (即考虑模式识别问题)。正如上面已经指出的, 在指示函数集 $Q(z, \alpha)$, $\alpha \in \Lambda$ 情况下, 如果 $\varepsilon < 1$, 向量 $q(\alpha)$, $\alpha \in \Lambda$ 的最小 ε 网格(参见 2.3.3 小节)不依赖于 ε, 在最小 ε 网格中的元素数

$$N^{\Lambda}(z_1, \cdots, z_l) = N^{\Lambda}(\varepsilon; z_1, \cdots, z_l)$$

等于用函数集 $Q(z, \alpha)$, $\alpha \in \Lambda$ 对数据 $z_1, \cdots, z_l$ 不同的划分数。

对这个函数集, VC 熵也不依赖于 ε

$$H^{\Lambda}(l) = E\ln N^{\Lambda}(z_1, \cdots, z_l),$$

其中的数学期望是对于 $(z_1, \cdots, z_l)$ 进行的。

考虑在 $N^{\Lambda}(z_1, \cdots, z_l)$ 值的基础上构造的两个新的概念:

(1) 退火的 VC 熵

$$H^{\Lambda}_{\text{ann}}(l) = \ln EN^{\Lambda}(z_1, \cdots, z_l);$$

(2) 生长函数

$$G^{\Lambda}(l) = \ln \sup_{z_1, \cdots, z_l} N^{\Lambda}(z_1, \cdots, z_l).$$

这些概念的定义方法使得对任何 l, 都有不等式

$$H^{\Lambda}(l) \leq H^{\Lambda}_{ann}(l) \leq G^{\Lambda}(l)$$

成立。

在这些函数的基础上建立了学习理论的主要里程碑。

在 2.3.4 小节我们介绍了等式

$$\lim_{l\to\infty} \frac{H^{\Lambda}(l)}{l} = 0,$$

它描述了 ERM 原则一致性的一个充分条件(充分必要条件由与本式略有不同的(2-13)式给出)。这一等式是学习理论中的第一个里程碑: 我们要求所有最小化经验风险的机器都满足它。

但是, 这一等式对所得到的风险 $R(\alpha)$ 收敛到最小值 $R(\alpha)$ 的速度没有给出任何说法。有可能构造出这样一些例子, 即 ERM 原则是一致的, 但风险收敛的渐近速度却非常慢。

这就提出了下面的问题:

在什么条件下收敛的渐近速度是快的?

我们说收敛的渐近速度快, 是指对任何 $l > l_0$, 都有下面的指数界成立:

$$P\{R(\alpha) - R(\alpha) > \varepsilon\} < e^{-c\varepsilon^2 l},$$

其中 $c > 0$ 是某个常数。

于是我们得出, 等式

$$\lim_{l\to\infty} \frac{H^{\Lambda}_{ann}(l)}{l} = 0$$

是收敛速度快的一个充分条件[①]。这一等式是学习理论的第二个里程碑: 它保证了收敛有快的渐近速度。

到这里, 我们已经考虑了两个问题: 一个是 ERM 方法一致性的充分必要条件, 另一个是 ERM 方法收敛速度快的充分条件。这两个等式都是对一个给定的概率测度 $F(z)$ 有效的(VC 熵 $H^{\Lambda}(l)$ 和 VC 退火熵 $H^{\Lambda}_{ann}(l)$ 都是用这个测度构造的)。然而, 我们的目标是建立一个学习机器, 使它能够解决很多不同的问题(对于很多不同的概率测度)。

于是就有问题:

在什么条件下, 不依赖于概率测度, ERM 原则是一致的且同时有快的收敛速度?

下面的等式给出了对任何概率测度 ERM 具有一致性的充分必要条件:

$$\lim_{l\to\infty} \frac{G^{\Lambda}(l)}{l} = 0.$$

而且, 如果这个条件成立, 则收敛的速度是快的。

这个等式就是学习理论中的第三个里程碑。它描述了在什么充分必要条件下, 一个履行 ERM 原则的学习机器有一个快的收敛的渐近速度, 而不管所用的概率测度如何(即不管所要解决的问题如何)。

这些里程碑构成了后面建立学习机器收敛速度的界的基础, 包括与分布无关的界和严格依赖于分布的界。这些我们将在第三章中进行介绍。

---

① 收敛速度快的这一条件的必要性仍是一个未解决的问题。

非正式推导和评述——2

在引论和第一章中,我们讨论了经验风险最小化方法和密度估计的方法;但是我们将不用它们来构造学习算法。在第四章中将介绍另一种归纳原则,我们将在第五章中用它来构造学习算法。另一方面,在 1.11 节中我们还介绍了随机逼近归纳原则,我们并没有把它看得很重要,尽管事实上一些学习过程(比如神经网络中)是基于这一原则的。

于是人们很自然要问:

为什么 ERM 原则和密度估计的方法这样重要?

为什么我们用了这么多时间来讨论 ERM 原则一致性的充分必要条件?

在本章的评述中我们将尝试说明,在某种意义上,函数估计问题的两种方法,一种是基于密度估计方法的,另一种是基于 ERM 方法的,它们反映了统计推断中的两个最一般性的思想。

为了说明这一点,我们把统计学的一般问题形式化为用数据估计未知概率测度的问题。我们将区分两种方式的概率测度估计,即所谓强估计和弱估计。我们将说明,提供强估计的方法是基于密度估计方法的,而提供弱估计的方法则是基于 ERM 方法的。

概率测度的弱估计构成了统计学的基础中最重要的问题之一,即所谓广义 Glivenko-Cantelli 问题。在第二章中讨论的结果提供了对这一问题的一个全面的解决方法。

2.8 概率论和统计学的基本问题

在 30 年代,Kolmogorov(1933)介绍了概率论的一种公理化理论,从此概率论成为了一种纯数学的(即演绎的)学科:在这一理论中的任何分析,都可以基于从给定公理出发的形式化推理来实现。这使得对概率论和统计学的深入分析得以发展。

概率论的公理

根据 Kolmogorov 的公理化概率理论, 对每一个随机实验, 存在一个基本事件 z 的集合 Z, 其中定义了实验的所有可能结果(基本事件)。在基本事件集合 Z 上, 定义一个子集 A Z 的系统{A}, 叫做事件。作为一个事件来看, 集合 Z 决定了一种对应于必然事件(总会发生的事件)的情况。同时认为集合 A 中还包含了空集 , 它是永远不会发生的事件。

对{A}中的元素定义了并集、补集和交集的运算。在集合 Z 上定义了事件{A}的一个 σ代数 F [①]。如果

(1) $Z \in F$;

(2) 若 $A \in F$, 则 $A \in F$;

(3) 若 $A_i \in F$, 则 $\bigcup_{i=1}^{\infty} A_i \in F$,

那么, Z 的子集的集合 F 被称作事件 $A \in F$ 的一个 σ代数。

例 让我们来考虑下面的随机实验: 某人掷两个骰子, 一个红色, 一个黑色, 观察实验的结果。这个实验的基本事件空间 Z 可以用一对整数来描述: 第一个数代表红骰子的点数, 第二个数代表黑骰子的点数。这个实验的事件可以是这个基本事件集合的任意子集, 比如, 可以是所有两个点数相加等于 10 的基本事件的子集 A_{10}, 也可以是红骰子点数大于黑骰子点数的所有基本事件的子集 $A_{r>b}$, 等等(图 2.9)。

图 2.9 双骰子实验的基本事件空间, 其中标出了事件 A_{10} 和 $A_{r>b}$

由集合 Z 和事件 $A \in F$ 的 σ代数 F 组成的(Z, F)对是随机实验的定性方面的一种理想化表达。

实验的定量方面是由定义在集合 F 的元素 A 上的概率测度 P(A)决定的。如果

(1) $P(A) \geq 0$;

(2) $P(Z) = 1$;

(3) $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$, 若 $A_i, A_j \in F$ 且 $A_i \cap A_j = $, " i, j, 那么在元素 $A \in F$ 上定义的函数 P(A)被叫做 F 上的一个可数可加性概率测度, 简称为概率测度。

如果确定了由三元组(Z, F, P)定义的概率空间, 则我们说一个实验的概率模型就确定了。

---

① 关于 σ代数, 读者可以参考概率论的任何高级教材(比如, 可以参考 Schiryaev A N. Probability. Springer, New York, p. 577)。这个概念使得人们可以用测度理论中发展的形式化工具来构建概率论的基础。

例　在上面的实验中,让我们考虑均匀的骰子,即其中所有基本事件都同样可能(概率都是 1/36),那么所有事件的概率就都得到了定义(比如事件 A_{10} 的概率是 3/36,事件 $A_{r>b}$ 的概率是 15/36)。

在概率论和统计学理论中,独立试验的概念[①] 起着非常关键的作用。

考虑包含 l 次不同试验的实验,概率空间为 $(Z, \mathcal{F}, P)$,设

$$z_1, \cdots, z_l \tag{2-14}$$

是这些试验的结果。对有 l 次试验的实验,可以考虑模型 $(Z^l, \mathcal{F}^l, P^l)$,其中 Z^l 是所有可能的结果(2-14)式的空间,$\mathcal{F}^l$ 是 Z^l 上的 σ 代数,它包括集合 $A_{k_1} \times \cdots \times A_{k_l}$,$P^l$ 是定义在 σ 代数 $\mathcal{F}^l$ 的元素上的概率测度。

如果对任意 $A_{k_1}, \cdots, A_{k_l} \in \mathcal{F}$,等式

$$P^l\{z_1 \in A_{k_1}; \cdots; z_l \in A_{k_l}\} = \prod_{i=1}^{l} P\{z_i \in A_{k_i}\}$$

都成立,那么我们说(2-14)式是一个 l 次独立试验的序列。

设(2-14)式是模型为 $(Z, \mathcal{F}, P)$ 的 l 次独立试验的结果。考虑为一个固定的事件 $A \in \mathcal{F}$ 定义的随机变量 $v(z_1, \cdots, z_l; A)$:

$$v_l(A) = v(z_1, \cdots, z_l; A) = \frac{n_A}{l},$$

其中 n_A 是集合 $z_1, \cdots, z_l$ 中属于事件 A 的元素的数目。随机变量 $v_l(A)$ 叫做在一个 l 次独立随机试验的序列中事件 A 出现的频率。

利用这些概念,我们可以形式化地给出概率论和统计学理论的基本问题。

1. 概率理论的基本问题

给定模型 $(Z, \mathcal{F}, P)$ 和一个事件 A^*,估计事件 A^* 在一系列 l 次独立随机试验中发生的频率的分布(或它的某些特性)。形式化地说,这意味着寻找分布函数

$$F(\xi A^*, l) = P\{v_l(A^*) < \xi\} \tag{2-15}$$

(或依赖于这个函数的某些泛函)。

例　在我们上面双骰子的例子中,相应的问题就可能是:掷 l 次骰子,事件 A_{10}(点数和为 10)的频率小于 ξ 的概率是多少?

在统计学理论中我们面对的是反问题。

2. 统计学理论的基本问题

给定一个随机实验 $(Z, \mathcal{F})$ 的定性模型,并给定独立同分布数据

$$z_1, \cdots, z_l, \cdots,$$

它们是根据一个未知的概率测度 P 出现的,估计在所有子集 $A \in \mathcal{F}$ 上定义的这个概率测度 P(或者依赖于这个函数的某些泛函)。

---

① 实际上是独立试验的概念使得概率理论不同于测度理论。如果没有独立试验的概念,则概率论的公理定义了测度理论中的一个模型。

例　设我们的两个骰子是不均匀的,且它们之间是互相联系的(比如用一条线连在一起),问题是,给定 l 次(l 对)试验的结果,估计对所有事件(子集) $A \in F$ 的概率测度。

在本书中我们考虑这样的基本元素集合 $Z \subset R^n$,其中 σ代数 F 定义为包含 Z 上所有的 Borel 集合[①]。

2.9　估计概率测度的两种方式

我们可以定义估计概率测度的两种方式:一种强方式和一种弱方式。

定义:

(1) 对估计器

$$E_l(A) = E_l(z_1, \cdots, z_l; A), \quad A \in F,$$

如果

$$\sup_{A \in F} |P(A) - E_l(A)|_{l \to \infty} \xrightarrow{P} 0 \tag{2-16}$$

则我们说估计器 $E_l(A)$ 以强方式估计概率测度 P。

(2) 如果对某个子集 $F^* \subset F$,

$$\sup_{A \in F^*} |P(A) - E_l(A)|_{l \to \infty} \xrightarrow{P} 0, \tag{2-17}$$

则我们说估计器 $E_l(A)$ 以取决于子集 $F^* \subset F$ 的弱方式估计概率测度 P,其中(集合 F 的)子集 F^* 不一定形成一个 σ代数。

对我们的推导来说,重要的是,如果我们可以以两种方式之一来估计概率测度(对弱方式是相对于下面讨论的一种特殊的 F^* 集合),那么就可以在一个给定的函数集中最小化风险泛函。

实际上,考虑有界风险函数 $0 \leq Q(z, \alpha) \leq B$ 的情况。让我们利用 Lebesgue 积分的定义把风险泛函重写为一种等价的形式(图 2.10):

$$R(\alpha) = \int Q(z, \alpha) dP(z) = \lim_{m \to \infty} \sum_{i=1}^{m} \frac{B}{m} P\left\{ Q(z, \alpha) > \frac{iB}{m} \right\}. \tag{2-18}$$

如果估计器 $E_l(A)$ 以强方式很好地逼近 $P(A)$,即对任何事件 A(包括事件 $A^*_{\alpha i} = \{Q(z, \alpha) > iB/m\}$)都一致地很好地逼近概率,那么在从数据估计的概率测度 $E_l(A)$ 基础上构造的泛函

$$R^*(\alpha) = \lim_{m \to \infty} \sum_{i=1}^{m} \frac{B}{m} E_l\left\{ Q(z, \alpha) > \frac{iB}{m} \right\} \tag{2-19}$$

将(对任意 α)一致地很好地逼近风险泛函 $R(\alpha)$。因此它可以用于选择最小化风险的函数。在第一章和第二章中讨论的经验风险泛函 $R_l(\alpha)$ 就对应于这种情况,其中(2-19)式中的估计器 $E_l(A)$ 从给定的数据估算事件 A 的频率。

然而注意,为了在给定的函数集 $Q(z, \alpha)$, $\alpha \in \Lambda$ 上用(2-19)式逼近(2-18)式,我们不必

① 我们考虑包含所有开平行六面体的最小 σ代数。

图 2.10　(2-18)式定义的 Lebesgue 积分是一个乘积之和的极限,其中乘子 $P\{Q(z,\theta)>\frac{iB}{m}\}$ 是集合 $\{z:Q(z,\theta)>\frac{iB}{m}\}$ 的(概率)测度,而乘子 $\frac{B}{m}$ 是切片的高度

要在 σ代数的所有事件 A 上一致逼近 P,而是只需要在事件

$$A^{*}_{\alpha i}=\{Q(z,\theta)>\frac{iB}{m}\}$$

上一致逼近(在估算风险(2-18)式时只有这些事件参与)。因此,为了寻找使得风险泛函最小的函数,相对于事件集合

$$\{Q(z,\theta)>\frac{iB}{m}\},\quad \alpha\in\Lambda$$

对概率测度的弱方式逼近就足够了。

因此,为了寻找在未知概率测度 $P\{A\}$ 下使风险((2-18)式)最小化的函数,我们可以最小化泛函(2-19)式,其中用一个逼近 $E_l(A)$ 来代替 $P\{A\}$,它以强方式和弱方式收敛于 $P\{A\}$(对弱方式是相对于事件 $A^{*}_{\alpha i},\ \alpha\in\Lambda,\ i=1,\cdots,m$ 的)。

2.10　概率测度的强方式估计与密度估计问题

遗憾的是,没有一种估计器能够以强方式对一个任意的概率测度进行估计。如果对一个概率测度存在密度函数(Radon-Nikodym 微分),那么我们可以估计这个概率测度。假定密度函数 $p(z)$ 存在,并设 $p_l(z)$ 是对这个密度函数的一个逼近。考虑估计器

$$E_l(A)=\int_A p_l(z)dz.$$

根据 Scheffe 定理,对这个估计器,下面的界成立:

$$\sup_A|P_l(A)-E_l(A)|\le\int|p(z)-p_l(z)|dz,$$

即概率测度的逼近与实际测度之间的强方式距离以密度的逼近与实际密度之间的 L_1 距离为上界。

因此,为了以强方式估计概率测度,只要对密度函数进行估计就足够了。在 1.8 节中,我们强调了从数据估计一个密度函数将产生不适定问题,因此一般来说,我们不能保证从一个固定数目的观测能够得到好的逼近。

幸运的是,正如上面我们看到的,为了估计使风险泛函最小的函数,我们并不一定需要估计密度。只要在弱方式下对概率测度进行近似就足够了,其中的事件集合 F^* 依赖于容许的函数集 $Q(z, \alpha)$, $\alpha \in \Lambda$: 它必须包含事件

$$Q(z, \alpha) > \frac{iB}{m}, \quad \alpha \in \Lambda, \ i = 1, \cdots, m.$$

所考虑的容许事件的集合"越小",则弱估计时必须考虑的事件集合 F^* 也"越小",因此(后面我们将会看到)在一个较小的函数集上最小化风险需要较少的观测。在第三章中我们将讨论一致收敛速度的界,它们依赖于容许事件集合的容量。

2.11 Glivenko-Cantelli 定理及其推广

在 30 年代,Glivenko 和 Cantelli 证明了一个定理,可以把这个定理看作是统计学基础中最重要的结论。他们证明了一个随机变量 ξ 的概率分布函数

$$F(z) = P\{\xi < z\}$$

可以用经验分布函数任意好地逼近,这个经验分布函数是:

$$F_l(z) = \frac{1}{l} \sum_{i=1}^{l} \theta(z - z_i),$$

其中,$z_1, \cdots, z_l$ 是依据未知密度得到的独立同分布数据[①](见图 1.2)。更确切地说,Glivenko-Cantelli 定理说明了,对任何 $\varepsilon > 0$,等式

$$\lim_{l \to \infty} P\{\sup_z |F(z) - F_l(z)| > \varepsilon\} = 0$$

(以概率收敛[②])成立。

让我们用另外一种方式来表述一下 Glivenko-Cantelli 定理。考虑事件集合

$$A_z = \{z: z < z\}, \quad z \in (-\infty, \infty) \tag{2-20}$$

(在指向 $-\infty$ 直线上的射线的集合)。对这个事件集合中的任意事件 A_z,我们可以估算它的概率

$$P(A_z) = \int_{-\infty}^{z} dF(z) = F(z). \tag{2-21}$$

利用数目为 l 的独立同分布样本,我们也可以估计事件 A_z 在独立试验中出现的频率

$$v(A_z) = \frac{n_{A_z}}{l} = F_l(z). \tag{2-22}$$

在这些表达下,Glivenko-Cantelli 定理断定了(2-22)式的估计相对于事件集合(2-20)式以

---

① 稍晚些时候得到了对 $n > 1$ 个变量的推广。

② 实际上,有一种更强方式的收敛成立,即所谓"几乎必然"收敛(或译殆必收敛)。

弱方式收敛于概率测度(2-21)式(因为只考虑了所有事件的一个子集,所以是弱方式收敛)。

为了对各种指示函数集(即对模式识别问题)证明ERM归纳原则是正确的,在本章中我们建立了一个在任意事件集合上频率一致收敛于概率的一般理论。这一理论完成了由针对特殊事件集合的Glivenko-Cantelli理论开始的对概率测度的弱方式逼近的分析。

在1981年,这些结论推广到了在函数集上均值到数学期望的一致收敛,这实际上开始了关于一般类型的经验过程的研究。

2.12 归纳的数学理论

尽管人们在理论统计学的基础中取得了一系列重要的成果,但是,学习理论的主要概念问题在二十多年的时间里(从1968年到1989年)仍然一直没有解决。这个问题就是:

均值到期望的一致收敛是否构成ERM归纳原则一致性的一个充分必要条件,还是这个条件只是充分的?如果是后者,是否可能存在其他的更少限制性的充分条件?

对这个问题的回答并不是显然的。事实上,一致收敛构成了函数集的一个全局的特性,而我们或许会认为ERM原则的一致性应该是由函数集中一个接近待求函数的子集的局部特性决定的。

利用非平凡一致性的概念,我们在1989年证明了一致性是容许函数集的一个全局特性,它取决于单边一致收敛(Vapnik and Chervonenkis, 1989)。我们找到了单边收敛的充分必要条件。

这些条件的证明基于一套新的思想——在对归纳推理的哲学讨论中出现的关于不可证伪性的思想。然而,在这些讨论中,归纳没有被作为统计推断的一部分来考虑,而是被看作在比统计学模型更一般的框架中的一个推理工具。

第三章

学习过程收敛速度的界

在本章中我们讨论一致收敛速度的界。这里我们只考虑上界(下界虽然也存在(Vapnik and Chervonenkis, 1974),但是对于控制学习过程来说它们不像上界那么重要)。

利用第二章介绍的容量的两个不同概念(退火熵函数和生长函数),我们将讨论收敛速度的两种类型的界:

(1) 依赖于分布的界(基于退火熵函数);

(2) 与分布无关的界(基于生长函数)。

然而,这些界都不是构造性的,因为它们并没有给出计算退火熵函数和生长函数的明确的方法。

因此,我们将介绍关于函数集容量的一个新的特性(函数集的 VC 维),它是一个标量值,对学习机器可用的任何函数集都可以估算这一特性。

在 VC 维概念的基础上,我们将得到

(3) 构造性的与分布无关的界。

用等价的方式重写这些界,就可以找到学习机器所取得的风险的界(即可以估计学习机器的推广能力)。在第四章中我们将用这些界来控制学习机器的推广能力。

3.1 基本不等式

我们将从 $Q(z,\alpha),\alpha\in\Lambda$ 是一个指示函数集的情况开始介绍关于界的理论结果,然后再将它们推广到实函数集。

设 $Q(z,\alpha),\alpha\in\Lambda$ 是一个指示函数集,$H^{\Lambda}(l)$ 是对应的 VC 熵,$H^{\Lambda}_{ann}(l)$ 是其退火熵,$G^{\Lambda}(l)$ 是其生长函数(参阅 2.7 节)。

下面给出两个关于一致收敛速度的界,它们构成了界的理论中的基本不等式(Vapnik and Chervonenkis, 1968, 1971 及 Vapnik, 1979, 1996)。

定理 3.1　下面的不等式成立:

$$P\left\{\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)\,dF(z)-\frac{1}{l}\sum_{i=1}^{l}Q(z_i,\alpha)\right|>\varepsilon\right\}$$

$$\le 4\exp\left\{\frac{H^{\Lambda}_{ann}(2l)}{l}-\varepsilon^2 l\right\}. \tag{3-1}$$

定理 3.2　下面的不等式成立:

$$P\left\{\sup_{\alpha\in\Lambda}\frac{\int Q(z,\alpha)\,dF(z)-\frac{1}{l}\sum_{i=1}^{l}Q(z_i,\alpha)}{\sqrt{\int Q(z,\alpha)\,dF(z)}}>\varepsilon\right\}$$

$$\le 4\exp\left\{\frac{H^{\Lambda}_{ann}(2l)}{l}-\frac{\varepsilon^2}{4}l\right\}. \tag{3-2}$$

如果

$$\lim_{l\to\infty}\frac{H^{\Lambda}_{ann}(l)}{l}=0,$$

则这些界是非平凡的(即对任意 $\varepsilon>0$, 当观测数目 l 趋于无穷时不等式右边都趋于零)(2.7 节中我们把这一条件叫做学习理论的第二个里程碑)。

为了讨论这两个界之间的区别, 我们先来回顾前面已经指出的一点, 即对任意指示函数 $Q(z,\alpha)$, 风险泛函

$$R(\alpha)=\int Q(z,\alpha)\,dF(z)$$

描述了事件 $\{z:\ Q(z,\alpha)=1\}$ 的概率, 而经验风险泛函 $R_{emp}(\alpha)$ 则描述了这个事件的频率。

定理 3.1 针对频率与概率的偏差的模给出了对一致收敛速度的估计。显然, 最大的偏差更容易在拥有最大方差的事件上出现。对这种伯努利情况, 方差等于

$$\sigma=\sqrt{R(\alpha)(1-R(\alpha))},$$

因此方差的最大值是在有概率 $R(\alpha^*)\approx 1/2$ 的事件上取得的。换句话说, 最大偏差是与使风险较大的函数相关联的。

在第 3.3 节中, 利用这一收敛速度的界, 我们将得到风险的一个界, 其置信范围[①]是由一致收敛速度决定的, 也就是由使风险 $R(\alpha^*)\approx 1/2$ 的函数(函数集中"最坏"的函数)决定的。

为了得到一个更小的置信范围, 我们可以尝试用另一种一致收敛来构造风险的界, 这种一致收敛就是一致相对收敛

$$P\left\{\sup_{\alpha\in\Lambda}\frac{|R(\alpha)-R_{emp}(\alpha)|}{\sqrt{R(\alpha)(1-R(\alpha))}}\ge\varepsilon\right\}<\Phi(\varepsilon l),$$

其中, 用方差对真实风险与经验风险的偏差进行了归一化。一致相对收敛的上确界可以在任何函数 $Q(z,\alpha)$ 上得到, 包括有较小风险的函数。

然而, 从技术上说, 很难对这个界的右边部分进行较好的估计。我们可以对更简单的

---

① 原文是 confidence interval, 即置信区间, 但因为从本书后面的内容看这里的含义与统计学中置信区间的定义有所不同(原著中并没有明确给出这一概念的定义), 因此我们翻译为置信范围, 它指的是用某个值(比如经验风险)来作为对另一个值(比如实际风险)的估计或近似所可能带来的误差上限(以一定的概率)。——译者

情况得到一个好的界,即不是用方差进行归一化,而是考虑用函数 $\overline{R(\alpha)}$ 进行归一化。当 $R(\alpha)$ 比较小时这个函数与方差接近(而我们感兴趣的正是 $R(\alpha)$ 比较小的情况)。为了得到界中更好的系数,我们考虑在分子上直接采用差而不是差的模。这就是定理 3.2 所研究的相对一致收敛的情况。

在 3.4 节中我们将看到,用定理 3.2 得到的风险上界比在定理 3.1 基础上得到的风险上界要好很多。

定理 3.1 和 3.2 中得到的界是依赖于分布的:它们对于给定的观测分布函数 $F(z)$ 成立(在构造退火熵函数 $H^{\Lambda}_{ann}(l)$ 时用到了这个分布)。

要构造与分布无关的界,只要注意到下面的关系就足够了,即对任何分布函数 $F(z)$,生长函数不小于退火熵:

$$H^{\Lambda}_{ann}(l) \leq G^{\Lambda}(l).$$

因此,对任何分布函数 $F(z)$,都有下面的两个不等式成立:

$$P\left\{\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)\,dF(z) - \frac{1}{l}\sum_{i=1}^{l} Q(z_i,\alpha)\right| > \varepsilon\right\} \leq 4\exp\left\{\frac{G^{\Lambda}(2l)}{l} - \varepsilon^2 l\right\}. \tag{3-3}$$

$$P\left\{\sup_{\alpha\in\Lambda}\frac{\int Q(z,\alpha)\,dF(z) - \frac{1}{l}\sum_{i=1}^{l} Q(z_i,\alpha)}{\sqrt{\int Q(z,\alpha)\,dF(z)}} > \varepsilon\right\} \leq 4\exp\left\{\frac{G^{\Lambda}(2l)}{l} - \frac{\varepsilon^2}{4} l\right\}. \tag{3-4}$$

如果

$$\lim_{l\to\infty}\frac{G^{\Lambda}(l)}{l} = 0, \tag{3-5}$$

则上述不等式是非平凡的(2.7 节中我们把这一等式称作学习理论的第三个里程碑)。

特别需要注意到,条件(3-5)式是与分布无关的一致收敛(3-3)式的充分必要条件。特别地,如果(3-5)式的条件不满足,则存在 Z 上的概率测度 $F(z)$ 使得一致收敛

$$\lim_{l\to\infty} P\left\{\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)\,dF(z) - \frac{1}{l}\sum_{i=1}^{l} Q(z_i,\alpha)\right| > \varepsilon\right\} = 0$$

不成立。

3.2 对实函数集的推广

有几种方法可以把指示函数集的这些结果推广到实函数集上,下面我们讨论一种最简单也是最有效的方法(它给出更好的界并且对无界实函数也有效)(Vapnik 1979, 1998)。

现在,设 $Q(z,\alpha),\ \alpha\in\Lambda$ 是一个实函数集合,且

$$A=\inf_{\alpha,z}Q(z,\alpha)\le Q(z,\alpha)\le\sup_{\alpha,z}Q(z,\alpha)=B$$

(这里 A 可以是 $-\infty$, B 可以是 ∞)。用 $\square$ 来代表开区间 (A,B),我们构造实函数集 $Q(z,\alpha),\ \alpha\in\Lambda$ 的一个指示器集合(图 3.1):

$$I(z,\alpha,\beta)=\theta\{Q(z,\alpha)-\beta\},\quad \alpha\in\Lambda,\ \beta\in\square.$$

对任何给定的函数 $Q(z,\alpha^*)$ 和一个给定的 β^*,指示器 $I(z,\alpha^*,\beta^*)$ 用 1 来指示 $Q(z,\alpha^*)\ge\beta^*$ 的 $z\in Z$ 区域,用 0 来指示 $Q(z,\alpha^*)<\beta$ 的 $z\in Z$ 区域。

在 $Q(z,\alpha),\ \alpha\in\Lambda$ 是指示函数的情况下,指示器集合 $I(z,\alpha,\beta),\ \alpha\in\Lambda,\ \beta\in(0,1)$ 与这个 $Q(z,\alpha),\ \alpha\in\Lambda$ 集合重合。

对任意给定的实函数集 $Q(z,\alpha),\ \alpha\in\Lambda$,我们将通过考虑相应的指示器集合 $I(z,\alpha,\beta),\ \alpha\in\Lambda,\ \beta\in\square$ 来推广上一节的结果。

图 3.1　函数 $Q(z,\alpha)$ 的 β 水平指示器指示出了 z 为哪些值时,函数 $Q(z,\alpha)$ 超出了 β 为哪些值时则没有超出。函数 $Q(z,\alpha)$ 可以用其所有指示器的集合来描述

设 $H^{\Lambda}(l)$ 是指示器集合的 VC 熵,$H^{\Lambda}_{ann}(l)$ 是它的退火熵,$G^{\Lambda}(l)$ 是其生长函数。

利用这些概念,我们得到对实函数集的基本不等式,它们是不等式(3-1)和(3-2)的推广,这些推广是分以下三种情况进行的:

(1) 完全有界函数 $Q(z,\alpha),\ \alpha\in\Lambda$。

(2) 完全有界的非负函数 $Q(z,\alpha),\ \alpha\in\Lambda$。

(3) 非负(不一定有界)函数 $Q(z,\alpha),\ \alpha\in\Lambda$。

下面就对这三种情况下的界进行介绍。

(1) 设 $A\le Q(z,\alpha)\le B,\ \alpha\in\Lambda$ 是完全有界函数的集合,那么下面的不等式成立:

$$P\left\{\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)\,dF(z)-\frac{1}{l}\sum_{i=1}^{l}Q(z_i,\alpha)\right|>\varepsilon\right\}$$

$$\le 4\exp\left\{\frac{H^{\Lambda}_{ann}(2l)}{l}-\frac{\varepsilon^2}{(B-A)^2}l\right\}.\qquad(3\text{-}6)$$

(2) 设 $0\le Q(z,\alpha)\le B,\ \alpha\in\Lambda$ 是完全有界非负函数的集合,那么下面的不等式成立:

$$P\left\{\sup_{\alpha\in\Lambda}\frac{\int Q(z,\alpha)\,dF(z)-\frac{1}{l}\sum_{i=1}^{l}Q(z_i,\alpha)}{\sqrt{\int Q(z,\alpha)\,dF(z)}}>\varepsilon\right\}$$

$$\le 4\exp\left\{\frac{H^{\Lambda}_{ann}(2l)}{l}-\frac{\varepsilon^2}{4B}l\right\}. \tag{3-7}$$

这两个不等式是定理 3.1 和 3.2 中对指示函数集得到的不等式的直接推广，当 $Q(z,\alpha)\in\{0,1\}$ 时它们与不等式(3-1)和(3-2)相同。

(3) 设 $0\le Q(z,\alpha),\ \alpha\in\Lambda$ 是这样一个函数集合，它使得对某个 $p>2$，随机变量 $\xi_\alpha=Q(z,\alpha)$ 的 p 阶归一化矩[①]

$$m_p(\alpha)=\sqrt[p]{\int Q^p(z,\alpha)\,dF(z)}$$

存在，那么下面的不等式成立：

$$P\left\{\sup_{\alpha\in\Lambda}\frac{\int Q(z,\alpha)\,dF(z)-\frac{1}{l}\sum_{i=1}^{l}Q(z_i,\alpha)}{\sqrt[p]{\int Q^p(z,\alpha)\,dF(z)}}>a(p)\varepsilon\right\}$$

$$\le 4\exp\left\{\frac{H^{\Lambda,}_{ann}(2l)}{l}-\frac{\varepsilon^2}{4}l\right\}, \tag{3-8}$$

其中

$$a(p)=\sqrt[p]{\left(\frac{1}{2}\right)^{p-1}\frac{p-1}{p-2}} \tag{3-9}$$

如果

$$\lim_{l\to\infty}\frac{H^{\Lambda,}_{ann}(l)}{l}=0,$$

则(3-6)式、(3-7)式和(3-8)式的界是非平凡的。

3.3 主要的与分布无关的界

(3-6)式、(3-7)式和(3-8)式的界是依赖于分布的：不等式右边采用了退火熵 $H^{\Lambda}_{ann}(l)$，它是建立在分布函数 $F(z)$ 基础上的。要得到与分布无关的界，我们需在(3-6)式、(3-7)式和(3-8)式界的右边用生长函数 $G^{\Lambda}(l)$ 代替退火熵 $H^{\Lambda}_{ann}(l)$。因为对任何分布函数，生长函数 $G^{\Lambda}(l)$ 都不小于退火熵 $H^{\Lambda}_{ann}(l)$，所以新的界将不依赖于分布函数 $F(z)$。

因此，我们可以由各种类型的一致收敛的速度得到下面的与分布无关的界：

(1) 对完全有界函数集 $-\infty<A\le Q(z,\alpha)\le B<\infty$，有

① 这里我们考虑 $p>2$ 只是为了简化公式。对 $p>1$ 有类似的结果成立(Vapnik, 1979, 1998)。

$$P\left\{\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)\,dF(z)-\frac{1}{l}\sum_{i=1}^{l}Q(z_i,\alpha)\right|>\varepsilon\right\}$$

$$\le 4\exp\left\{\left(\frac{G^{\Lambda}(2l)}{l}-\frac{\varepsilon^2}{(B-A)^2}\right)l\right\}. \tag{3-10}$$

(2) 对非负的完全有界函数集 $0\le Q(z,\alpha)\le B<\infty$, 有

$$P\left\{\sup_{\alpha\in\Lambda}\frac{\int Q(z,\alpha)\,dF(z)-\frac{1}{l}\sum_{i=1}^{l}Q(z_i,\alpha)}{\sqrt{\int Q(z,\alpha)\,dF(z)}}>\varepsilon\right\}$$

$$\le 4\exp\left\{\left(\frac{G^{\Lambda}(2l)}{l}-\frac{\varepsilon^2}{4B}\right)l\right\}. \tag{3-11}$$

(3) 对存在某个 $p>2$ 阶归一化矩的非负实函数集 $0\le Q(z,\alpha)$, 有

$$P\left\{\sup_{\alpha\in\Lambda}\frac{\int Q(z,\alpha)\,dF(z)-\frac{1}{l}\sum_{i=1}^{l}Q(z_i,\alpha)}{\sqrt[p]{\int Q^p(z,\alpha)\,dF(z)}}>a(p)\varepsilon\right\}$$

$$\le 4\exp\left\{\left(\frac{G^{\Lambda}(2l)}{l}-\frac{\varepsilon^2}{4}\right)l\right\}. \tag{3-12}$$

如果

$$\lim_{l\to\infty}\frac{G^{\Lambda}(l)}{l}=0, \tag{3-13}$$

则这些界是非平凡的。

利用这些不等式, 我们可以建立不同学习机器的推广能力的界。

3.4 学习机器推广能力的界

为了描述采用 ERM 原则的学习机器的推广能力, 我们需要回答两个问题:

(A) 得到最小经验风险 $R_{emp}(\alpha)$ 的函数 $Q(z,\alpha)$ 所提供的真实风险 $R(\alpha)$ 是什么?

(B) 对于给定的函数集, 这个风险与最小可能的风险 $\inf_{\alpha}R(\alpha),\ \alpha\in\Lambda$ 有多么接近?

对这两个问题的回答都可以利用上面描述的界得到。下面我们分别对实现完全有界函数集、完全有界非负函数集和任意非负函数集的学习机器给出它们的推广能力的界。这些界是上一节给出的界的另一种书写形式。

为了描述这些界, 我们引入符号

$$E=4\frac{G^{\Lambda}(2l)-\ln(\eta/4)}{l}, \tag{3-14}$$

并注意到当 $E<1$ 时这些界是非平凡的。

情况 1 完全有界函数集

设 $A\le Q(z,\alpha)\le B,\ \alpha\in\Lambda$ 是完全有界函数的集合, 那么

(A) 下面的不等式以至少 $1-\eta$ 的概率同时对 $Q(z,\alpha),\ \alpha\in\Lambda$ 的所有函数(包括使经

验风险最小的函数)成立:

$$R(\alpha) \leq R_{emp}(\alpha) + \frac{(B-A)}{2}\sqrt{E}, \tag{3-15}$$

$$R_{emp}(\alpha) - \frac{(B-A)}{2}\sqrt{E} \leq R(\alpha).$$

(这些界与(3-10)式的一致收敛速度的界是等价的。)

(B) 下面的不等式以至少 $1-2\eta$ 的概率对使经验风险最小的函数 $Q(z,\alpha)$ 成立:

$$R(\alpha) - \inf_{\alpha\in\Lambda} R(\alpha) \leq (B-A)\sqrt{\frac{-\ln\eta}{2l}} + \frac{(B-A)}{2}\sqrt{E}. \tag{3-16}$$

情况 2　完全有界非负函数集

设 $0\leq Q(z,\alpha)\leq B$, $\alpha\in\Lambda$ 是有界非负函数的集合, 那么

(A) 下面的不等式以至少 $1-\eta$ 的概率同时对 $Q(z,\alpha)\leq B$, $\alpha\in\Lambda$ 的所有函数(包括使经验风险最小的函数)成立:

$$R(\alpha) \leq R_{emp}(\alpha) + \frac{BE}{2}\left(1+\sqrt{1+\frac{4R_{emp}(\alpha)}{BE}}\right). \tag{3-17}$$

(这个界与一致收敛速度的界(3-11)式是等价的。)

(B) 下面的不等式以至少 $1-2\eta$ 的概率对使经验风险最小的函数 $Q(z,\alpha)$ 成立:

$$R(\alpha) - \inf_{\alpha\in\Lambda} R(\alpha) \leq B\sqrt{\frac{-\ln\eta}{2l}} + \frac{BE}{2}\left(1+\sqrt{1+\frac{4}{E}}\right). \tag{3-18}$$

情况 3　无界非负函数集

最后, 考虑无界非负函数集合 $0\leq Q(z,\alpha)$, $\alpha\in\Lambda$。

容易看到(通过举例), 如果不提供关于无界函数集和/或概率测度的额外信息, 我们不可能得到描述学习机器推广能力的不等式。下面我们作出如下的假设, 即有一个 (p,τ) 对, 使得不等式

$$\sup_{\alpha\in\Lambda}\frac{\left(\int Q^p(z,\alpha)\,dF(z)\right)^{1/p}}{\int Q(z,\alpha)\,dF(z)} \leq \tau < \infty \tag{3-19}$$

成立①, 其中 $p>1$。

学习理论关于无界函数集合的主要结论就是下面的论断, 为了简单起见我们这里只给出对 $p>2$ 情况的结论(在 $p>1$ 情况下的结论可以在文献(Vapnik, 1979, 1998)中找到):

(A) 下面的不等式以至少 $1-\eta$ 的概率同时对满足(3-19)式的所有函数成立:

$$R(\alpha) \leq \frac{R_{emp}(\alpha)}{(1-a(p)\tau\sqrt{E})_+}, \tag{3-20}$$

① 这个不等式描述了按 $F(z)$ 产生的随机变量 $\xi_\alpha = Q(z,\alpha)$ 分布函数的某种一般性质, 即其"分布的尾部"(随机变量 ξ_α 取大值的概率)。如果对 $p\geq 2$ 不等式(3-19)成立, 则称分布是"轻尾的"(ξ_α 的大值不经常发生), 在这种情况下收敛速度快是可能的。但是, 如果不等式(3-19)只对 $p<2$ 成立(ξ_α 的大值经常发生), 则收敛速度将会很慢(如果 p 充分接近 1 则收敛速度将任意慢)。

其中, $(u)_+ = \max(u, 0)$

$$a(p) = \sqrt[p]{\frac{1}{2}\left(\frac{p-1}{p-2}\right)^{p-1}}$$

(这个界是一致收敛速度的界(3-12)式和约束(3-19)式的推论。)

(B) 不等式

$$\frac{R(\alpha) - \inf_{\alpha\in\Lambda} R(\alpha)}{\inf_{\alpha\in\Lambda} R(\alpha)} \leq \frac{\tau a(p)\,\overline{E}}{(1 - \tau a(p)\,\overline{E})_+} + O\left(\frac{1}{l}\right) \tag{3-21}$$

以至少 $1-2\eta$ 的概率对使经验风险最小的函数 $Q(z, \alpha)$ 成立。

不等式(3-15)、(3-17)和(3-20)限定了集合 $Q(z, \alpha)$, $\alpha\in\Lambda$ 中所有函数的风险,其中也包括最小化经验风险的函数 $Q(z, \alpha)$。不等式(3-16)、(3-18)和(3-21)则评估了用 ERM 原则得到的风险与最小可能的风险之间的接近程度。

注意到,如果 $E<1$,那么从相对偏差一致收敛速度得到的界(3-17)式要比从一致收敛速度得到的界(3-15)式好得多:对较小的经验风险值,界(3-17)式的置信范围更小,量级是 E,而在(3-15)式中却是 $\sqrt{E}$ 量级的。

3.5 生长函数的结构

上面给出的学习机器推广能力的界主要是概念性的而不是构造性的,不能直接用来构造算法。为了使它们具有构造性,我们必须找到对给定的函数集 $Q(z, \alpha)$, $\alpha\in\Lambda$ 计算其退火熵 $H^{\Lambda}_{ann}(l)$ 和/或其生长函数 $G^{\Lambda}(l)$ 的途径。

我们将利用函数集 $Q(z, \alpha)$, $\alpha\in\Lambda$ 的 VC 维的概念找到构造性的界(VC 维是 Vapnik-Chervonenkis 维的缩写)。

VC 维的概念与生长函数之间重要的联系是在 1968 年被发现的(Vapnik and Chervonenkis, 1968, 1971)。

定理 3.3 任何生长函数,它或者满足等式

$$G^{\Lambda}(l) = l\ln 2,$$

或者受下面的不等式约束:

$$G^{\Lambda}(l) \leq h\left(\ln\frac{l}{h} + 1\right),$$

其中 h 是一个整数,使得当 $l=h$ 时,有

$$G^{\Lambda}(h) = h\ln 2,$$

$$G^{\Lambda}(h+1) < (h+1)\ln 2.$$

也就是说,生长函数要么是线性的,要么以一个对数函数为上界。(例如,生长函数不可能是 $G^{\Lambda}(l) = c\sqrt{l}$ 的形式,如图 3.2。)

定义 如果指示函数集 $Q(z, \alpha)$, $\alpha\in\Lambda$ 的生长函数是线性的,则我们说这个函数集的 VC 维是无穷大。

图 3.2　生长函数或者是线性的,或者以一个对数函数为界。
比如,它不可能像图中的点划线那样

如果指示函数集 $Q(z, \alpha), \alpha \in \Lambda$ 的生长函数以参数为 h 的对数函数为界,则我们说这个指示函数集的 VC 维是有限的且等于 h。

因为有下面的不等式成立:

$$\frac{H^{\Lambda}(l)}{l} \leq \frac{H_{ann}^{\Lambda}(l)}{l} \leq \frac{G^{\Lambda}(l)}{l} \leq \frac{h\left(\ln\frac{l}{h}+1\right)}{l}, \quad (l > h),$$

所以,学习机器所实现的指示函数集的 VC 维有限就是 ERM 方法一致性的一个充分条件,这一条件不依赖于概率测度。而且,一个有限的 VC 维意味着快的收敛速度。

VC 维有限也是 ERM 学习机器具有与分布无关的一致性的充分必要条件。下面的论断成立(Vapnik and Chervonenkis, 1974):

如果在某个事件集合(指示函数集)上,频率到概率的一致收敛对任何分布函数 $F(x)$ 成立,那么这个函数集的 VC 维是有限的。

3.6　函数集的 VC 维

下面我们首先给出指示函数集 VC 维的等价定义,然后再将这个定义推广到实函数集上。这些定义更强调了计算 VC 维的方法。

1. 指示函数集的 VC 维(Vapnik and Chervonenkis, 1968, 1971)

一个指示函数集 $Q(z, \alpha), \alpha \in \Lambda$ 的 VC 维,是能够被集合中的函数以所有可能的 2^h 种方式分成两类[①]的向量 $z_1, \cdots, z_h$ 的最大数目 h(也就是能够被这个函数集打散的向量的最大数目)。如果对任意的 n,总存在一个 n 个向量的集合可以被函数集 $Q(z, \alpha), \alpha \in \Lambda$ 打散,那么函数集的 VC 维就是无穷大。

① 任意指示函数都把一个给定的向量集合分成两个子集:使指示函数取值为 0 的向量的子集和使指示函数取值为 1 的向量子集。

2. 实函数集的 VC 维(Vapnik, 1979)

设 $A \leq Q(z, \alpha) \leq B$, $\alpha \in \Lambda$ 是一个以常数 A 和 B 为界的实函数集合(A 可以是 $-\infty$, B 可以是 ∞)。

我们与实函数集 $Q(z, \alpha)$, $\alpha \in \Lambda$ 一起考虑其指示器集合(图 3.1):

$$I(z, \alpha, \beta) = \theta\{Q(z, \alpha) - \beta\}, \quad \alpha \in \Lambda, \quad \beta \in (A, B), \tag{3-22}$$

其中, $\theta(z)$ 是阶跃函数

$$\theta(z) = \begin{cases} 0 & \text{若 } z < 0 \\ 1 & \text{若 } z \geq 0 \end{cases}.$$

实函数集 $Q(z, \alpha)$, $\alpha \in \Lambda$ 的 VC 维定义为相应的指示器集合(3-22)式的 VC 维, 其中的参数 $\alpha \in \Lambda$, $\beta \in (A, B)$。

例 1

(1) n 维坐标空间 $Z = \{z_1, \cdots, z_n\}$ 中的线性指示函数集合

$$Q(z, \alpha) = \theta \sum_{p=1}^{n} \alpha_p z_p + \alpha_0$$

的 VC 维是 $h = n + 1$, 因为用这个集合中的函数可以最多打散 $n + 1$ 个向量(图 3.3)。

图 3.3 平面中直线的 VC 维等于 3, 因为它们能打散 3 个向量而不能打散 4 个: 例如向量 z_2, z_4 不能被直线与向量 z_1, z_3 分开

(2) n 维坐标空间 $Z = \{z_1, \cdots . z_n\}$ 中的线性函数集合

$$Q(z, \alpha) = \sum_{p=1}^{n} \alpha_p z_p + \alpha_0, \quad \alpha_0, \cdots, \alpha \in (-\infty, \infty)$$

的 VC 维是 $h = n + 1$, 因为它对应的指示器函数的 VC 维等于 $n + 1$(注意: 用 $\alpha_0 - \beta$ 代替 α_0 并不改变指示函数集合)。

注意, 对线性函数来说, VC 维等于自由参数 $\alpha_0, \alpha, \cdots, \alpha$ 的个数, 但是这一规律对一般情况并不成立。

例 2

函数集合

$$f(z, \alpha) = \theta(\sin(\alpha z)), \quad \alpha \in R^1$$

的 VC 维是无穷大: 直线上的下列点:

$$z_1 = 10^{-1}, \cdots, z_l = 10^{-l}$$

可以被这个集合中的函数打散。

事实上, 要把这些数据分为由序列

$$\delta, \cdots, \delta, \quad \delta \in \{0, 1\}$$

确定的两类, 只要选择参数

$$\alpha = \pi \sum_{i=1}^{l} (1 - \delta) 10^i + 1$$

即可。这个例子反映了这样一个事实, 即只要选择适当的参数 α 我们可以对任意数目的适当选择的数据点用 sin(αx) 逼近以(-1, +1)为界的任何函数(图 3.4)。

图 3.4 用一个高频函数 sin(αx), 我们可以在任何 l 个适当选择的点上很好地逼近任何函数 $-1 \leq f(z) \leq 1$ 的取值

在第五章中我们将讨论一种 VC 维远小于其自由参数数目的函数集。

因此, 一般来说函数集的 VC 维与其自由参数的数目不同, 它可以大于自由参数个数(如例 2 中), 或者小于自由参数个数(我们将在第五章中用这种类型的函数集来构造一种新的学习机器)。

在下一节中我们将看到, 是函数集的 VC 维(而不是其自由参数个数)影响了学习机器的推广性能。这给我们克服所谓"维数灾难"创造了一个很好的机会: 用一个包含很多参数但却有较小 VC 维的函数集为基础实现较好的推广性。

3.7 构造性的与分布无关的界

在本节中我们给出的风险泛函的界将在第四章中用来构造控制学习机器推广能力的方法。

考虑有限 VC 维 h 的函数集。在这种情况下, 定理 3.3 指出, 界

$$G^{\Lambda}(l) \leq h \ln \frac{l}{h} + 1, \quad l > h \tag{3-23}$$

成立。因此,在 3.4 节[①]的所有不等式中,可以使用下面的构造性表达式:

$$E = 4\,\frac{h\left(\ln\frac{2l}{h}+1\right)-\ln\frac{\eta}{4}}{l}. \tag{3-24}$$

我们还将考虑损失函数集合 $Q(z,\alpha),\ \alpha\in\Lambda$ 中包含有限数目 N 个元素的情况。在这种情况下,我们可以使用表达式

$$E = 2\,\frac{\ln N-\ln\eta}{l}. \tag{3-25}$$

这样,就有下面的构造性界成立,其中在 VC 维有限[②]情况下使用(3-24)式的 E,而在集合中函数数目有限的情况下使用(3-25)式的 E。

情况 1　完全有界函数集

设 $A\le Q(z,\alpha)\le B,\ \alpha\in\Lambda$ 是完全有界函数的集合,那么

(A) 下面的不等式以至少 $1-\eta$ 的概率同时对 $Q(z,\alpha),\ \alpha\in\Lambda$ 的所有函数(包括使经验风险最小的函数)成立:

$$R(\alpha)\le R_{emp}(\alpha)+\frac{(B-A)}{2}\sqrt{E}, \tag{3-26}$$

$$R(\alpha)\ge R_{emp}(\alpha)-\frac{(B-A)}{2}\sqrt{E}.$$

(B) 下面的不等式以至少 $1-2\eta$ 的概率对使经验风险最小的函数 $Q(z,\alpha)$ 成立:

$$R(\alpha)-\inf_{\alpha\in\Lambda}R(\alpha)\le(B-A)\sqrt{\frac{-\ln\eta}{2l}}+\frac{(B-A)}{2}\sqrt{E}. \tag{3-27}$$

情况 2　完全有界非负函数集

设 $0\le Q(z,\alpha)\le B,\ \alpha\in\Lambda$ 是有界非负函数的集合,那么

(A) 下面的不等式以至少 $1-\eta$ 的概率同时对 $Q(z,\alpha)\le B,\ \alpha\in\Lambda$ 的所有函数(包括使经验风险最小的函数)成立:

$$R(\alpha)\le R_{emp}(\alpha)+\frac{BE}{2}\left(1+\sqrt{1+\frac{4R_{emp}(\alpha)}{BE}}\right). \tag{3-28}$$

(B) 下面的不等式以至少 $1-2\eta$ 的概率对使经验风险最小的函数 $Q(z,\alpha)$ 成立:

$$R(\alpha)-\inf_{\alpha\in\Lambda}R(\alpha)\le B\sqrt{\frac{-\ln\eta}{2l}}+\frac{BE}{2}\left(1+\sqrt{1+\frac{4}{E}}\right). \tag{3-29}$$

情况 3　无界非负函数集

最后考虑无界非负函数集合 $0\le Q(z,\alpha),\ \alpha\in\Lambda$。

(A) 下面的不等式以至少 $1-\eta$ 的概率同时对满足(3-19)式的所有函数成立:

$$R(\alpha)\le\frac{R_{emp}(\alpha)}{(1-a(p)\tau\sqrt{E})_+}, \tag{3-30}$$

其中,$(u)_+=\max(u,0)$

---

① 原著中误写作 3.3 节。——译者

② 且函数集中的函数数目无限。——译者

$$a(p)=\sqrt[p]{\frac{1}{2}\left(\frac{p-1}{p-2}\right)^{p-1}}.$$

(B) 对使经验风险最小的函数 $Q(z,\alpha)$,不等式

$$\frac{R(\alpha)-\inf_{\alpha\in\Lambda}R(\alpha)}{\inf_{x\in\Lambda}R(\alpha)}\leq\frac{\tau a(p)\overline{E}}{(1-\tau a(p)\overline{E})_+}+O\left(\frac{1}{l}\right)\tag{3-31}$$

以至少 $1-2\eta$ 的概率成立。

这些界是不能被显著地改进的[①]。

3.8 构造严格的(依赖于分布的)界的问题

要构造风险的严格的界,我们必须考虑关于概率测度的信息。设 Ω_0 是 Z^l 上所有概率测度的集合,Ω 是集合 Ω_0 的一个子集。如果知道一个包含 $F(z)$ 的子集 Ω,则说我们拥有关于未知概率测度 $F(z)$ 的先验知识。

考虑下面对生长函数的推广:

$$\Lambda(l)=\ln\sup_{F\in\Omega}E_F N^{\Lambda}(z_1,\cdots,z_l).$$

在 $\Omega=\Omega_0$ 的极端情况下,广义生长函数 $\Lambda(l)$ 就是生长函数 $G^{\Lambda}(l)$,因为在 $z_1,\cdots,z_l$ 上赋予概率 1 的测度是包含在 Ω 中的。在另一种极端情况下,Ω 只包含一个函数 $F(z)$,这时广义生长函数就是退火 VC 熵。

风险的严格界可以从广义生长函数的角度得到。它们与(3-15)、(3-17)和(3-21)式的与分布无关的界具有相同的泛函形式,只是 E 的表达式不同。E 的新的表达式是

$$E=4\frac{\Lambda(2l)-\ln\frac{\eta}{4}}{l}.$$

但是,这些界是非构造性的,因为尚没有找到计算广义生长函数的一般方法(与此相对照,对原来的生长函数,我们在函数集 VC 维的基础上得到了构造性的界)。

要找到严格的构造性的界,我们必须找到对不同概率测度集合 Ω 计算其广义生长函数的方法。这里主要的问题是找到一个与 Ω_0 不同的子集 Ω,使其广义生长函数可以在某种构造性概念的基础上进行估算(就像生长函数是用函数集的 VC 维来估算一样)。

① 存在一致收敛速度的下界,其数量级与上界的数量级接近(在上界中是 $\sqrt{(h/l)\ln(l/h)}$,而在下界中是 h/l;关于下界读者可以参阅文献(Vapnik and Chervonenkis, 1974)。)

非正式推导和评述——3

本章得到界的一种特殊情况在传统统计学中已经得到了研究，这就是 Kolmogorov-Smirnov 分布，它们在应用统计学和理论统计学中都有广泛的应用。

在学习理论中得到的界与传统的结果有两方面的不同：

(1) 它们更有一般性(对任何有限 VC 维的指示函数集都成立)。

(2) 它们对有限数目的观测成立(而传统的界是渐近的)。

3.9 Kolmogorov-Smirnov 分布

在人们发现了 Glivenko-Cantelli 定理之后，很快 Kolmogorov(1933) 就得到了对经验分布函数一致收敛到真实函数的速度的渐近准确估计。他证明了，如果一个标量随机变量的分布函数 $F(z)$ 是连续的且 l 充分大，那么对任意的 $\varepsilon > 0$，下面的等式成立：

$$P\left\{\sup_z \left|F_l(z) - F(z)\right| > \varepsilon\right\} = 2\sum_{k=1}^{\infty}(-1)^{k-1}\exp\{-2\varepsilon^2 k^2 l\}. \tag{3-32}$$

这一等式描述了一个基本的统计定律，根据这一定律，随机变量

$$\xi = \sup_z \left|F_l(z) - F(z)\right|$$

的分布不依赖于分布函数 $F(z)$，而具有(3-32)式的形式。

与此同时，Smirnov(1933) 发现了经验分布函数与真实函数的单边偏差的分布函数。他证明了，对连续的 $F(z)$ 和充分大的 l，下面的等式渐近地成立：

$$P\left\{\sup_z (F(z) - F_l(z)) > \varepsilon\right\} = \exp\{-2\varepsilon^2 l\},$$

$$P\left\{\sup_z (F_l(z) - F(z)) > \varepsilon\right\} = \exp\{-2\varepsilon^2 l\}.$$

随机变量

$$\xi = \sqrt{l}\left|F_l(x) - F(x)\right|$$

$$\xi = \sqrt{l}\,(F(x) - F_l(x))$$

被称作 Kolmogorov-Smirnov 统计量。

在将 Glivenko-Cantelli 定理推广到多维分布函数[①] 时，证明了对任意的 $\varepsilon > 0$，存在一个充分大的 l_0，使得对 $l > l_0$，不等式

$$P\left\{\sup_z \left|F_l(z) - F_l(z)\right| > \varepsilon\right\} < 2\exp\{-a\varepsilon^2 l\}$$

成立，其中 a 是任意的小于 2 的常数。

学习理论中得到的结论在两个方向上推广了 Kolmogorov 和 Smirnov 的结论：

(1) 所得到的界是对任意事件集合成立的(而不是像 Glivenko-Cantelli 情况下那样只是对射线的集合)。

(2) 所得到的界对任何 l 都成立(而不是只渐近地对充分大的 l 成立)。

3.10 在常数上的竞赛

注意到，在学习理论中得到的结论是不等式形式的，而 Kolmogorov 和 Smirnov 对一种特殊情况得到的结论是等式形式。对这种特殊情况，可以评价所得到的一般的界与精确值之间有多么接近。

设 $Q(z, \alpha)$，$\alpha \in \Lambda$ 是指示函数集，其 VC 维是 h。我们把(3-3)式的界用下面的形式重写：

$$P\left\{\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)dF(z) - \frac{1}{l}\sum_{i=1}^{l} Q(z_i,\alpha)\right| > \varepsilon\right\}$$

$$< 4\exp\left\{-a\varepsilon^2 - \frac{h\left(\ln\frac{2l}{h}+1\right)}{l}l\right\}, \tag{3-33}$$

其中系数 a 等于 1。在 Glivenko-Cantelli 情况下(此时 Kolmogorov-Smirnov 界成立)，我们实际上是在考虑指示函数集 $Q(z, \alpha) = \theta(z-\alpha)$。(对这些指示函数，

$$F(\alpha) = \int \theta(z-\alpha)dF(z),$$

$$F_l(\alpha) = \frac{1}{l}\sum_{i=1}^{l}\theta(z_i-\alpha),$$

其中 $z_1, \cdots, z_l$ 是独立同分布数据。)注意到，对这个指示函数集，VC 维等于 1：用(一维方向的)射线指示器只能打散一个点。因此，对充分大的 l，在(3-33)式右边的指数项里的第二项为任意小，界是由指数的第一项决定的。在这个一般表达式里的这一项几乎与 Kolmogorov-Smirnov 公式里的主项相同，除了一个常数：在这里的 $a = 1$，而在

---

① 对一个 n 维向量空间 Z，随机向量 $z = (z^1, \cdots, z^n)$ 的分布函数是这样确定的：

$$F(z) = P\{z^1 < z^1, \cdots, z^n < z^n\}$$

经验分布函数 $F_l(z)$ 估计事件 $A_z = \{z^1 < z^1, \cdots, z^n < z^n\}$ 发生的频率。

Kolmogorov-Smirnov 界里常数[①] a= 2。

在 1988 年 Devroye(1988)发现了一种在常数 a= 2 下得到一个非渐近界的方法。但是,在公式右边的指数项中,第二项是

$$\frac{h\ \ln\frac{l^2}{h}+\ 1}{l}$$

而不是

$$\frac{h\ \ln\frac{2l}{h}+\ 1}{l}. \tag{3-34}$$

对在实际中比较重要的情况,即

$$-\ \ln\eta<\ h(\ln h-\ 1)$$

的情况来说,本章所讨论的采用系数 a= 1 和(3-34)式的界更好。

3.11 经验过程的界

对实函数集得到的界是对指示函数集得到的界的推广,这些推广是在对实函数集构造的一个推广的 VC 维概念基础上得到的。

然而,对于对实函数集推广 VC 维的概念并推导相应的界,可以有几种方法。

一种推广方法是基于由 Dudley(1978)提出的 VC 子图概念之上的(在人工智能的文献中,这一概念被叫做伪维数)。利用 VC 子图的概念,Dudley 对有界实函数得到了在 ε熵度量上的一个界。在这个界的基础上,Pollard(1984)推导出了均值一致收敛于其期望的速度的界。Haussler 把这个界用到了学习机器上[②]。

注意,本章讨论的实函数集 VC 维概念,与 Dudley 的 VC 子图相比,对函数集容量的要求略强一些。但另一方面,利用 VC 维的概念,我们得到了更有吸引力的界:

(1) 它们有一个具有清楚物理含义的形式(它们依赖于比值 l/h)。

(2) 更重要地,利用这一概念,我们既可以对有界函数集合得到一致相对收敛的界,也可以对无界函数集合得到一致相对收敛的界。对无界损失函数集合的经验风险一致收敛(或一致相对收敛)到真实风险的速度,是对回归问题进行分析的基础。

一致相对收敛的界在传统统计学中没有与之类似的研究。它们是在学习理论中为了得到风险的严格的界而第一次被推导出的。

---

① 起初在 1968 年得到的结果中这个常数是 a= 1/8(Vapnik and Chervonenkis, 1968, 1971),然后在 1979 年被改进为 a= 1/4(Vapnik, 1979),在 1991 年,L. Bottou 给我看了一个 a= 1 的证明。J. M. Parrondo 和 C. Van den Broeck(1993)也得到了这个界。

② Haussler D. 1992. Decision theoretic generalization of the PAC model for neural net and other applications. Inform. Comp. 100(1):78～150

第四章

控制学习过程的推广能力

本章讨论控制学习机器推广能力的理论,这一理论的目的是致力于构造一种利用小样本训练实例来最小化风险泛函的归纳原则。

对数目为 l 的样本,如果比值 l/h(训练模式数目与学习机器函数的 VC 维的比值)较小,比如 l/h< 20,则我们就认为样本数是少的,即认为这种样本集是小样本。

为了构造针对小样本数的方法,我们需要利用前面得到的关于学习机器推广能力的界的结论,包括采用完全有界非负函数集的学习机器推广能力的界:

$$R(\alpha) \leq R_{emp}(\alpha) + \frac{B\varepsilon}{2}\left(1 + \sqrt{1 + \frac{4R_{emp}(\alpha)}{B\varepsilon}}\right), \tag{4-1}$$

和采用无界函数集的学习机器推广能力的界:

$$R(\alpha) \leq \frac{R_{emp}(\alpha)}{(1 - a(p)\tau\sqrt{\varepsilon})_+}, \tag{4-2}$$

$$a(p) = \sqrt[p]{\left(\frac{1}{2}\right)^{p-1}\frac{p-1}{p-2}},$$

其中,如果函数集 $Q(z,\alpha)$, $i = 1,\cdots,N$ 包含 N 个元素,则

$$\varepsilon = 2\frac{\ln N - \ln\eta}{l},$$

而如果函数集 $Q(z,\alpha)$, $\alpha\in\Lambda$ 包含无限多个元素且 VC 维 h 有限,则

$$\varepsilon = 4\frac{h\ln\frac{2l}{h} + 1 - \ln\frac{\eta}{4}}{l}.$$

这些界都是以至少 $1-\eta$ 的概率成立。

4.1 结构风险最小化归纳原则

ERM 原则是从处理大样本数问题出发的,这一原则的合理性可以通过考虑不等式

(4-1)和(4-2)来证明。

当 l/h 较大时，Φ 就较小，因此，不等式(4-1)右边的第二项((4-2)式分母中的第二项)就变得较小，于是实际风险就接近经验风险的取值。在这种情况下，较小的经验风险值就能够保证(期望)风险的值也较小。

然而如果 l/h 较小，那么一个小的 $R_{emp}(\alpha)$ 并不能保证小的实际风险值。在这种情况下，要最小化实际风险 $R(\alpha)$，我们必须对不等式(4-1)(或(4-2))右边的两项同时最小化。但是需要注意，不等式(4-1)右边的第一项取决于函数集中的一个特定的函数，而第二项则取决于整个函数集的 VC 维。因此要对风险的界(4-1)式(或(4-2)式)右边的两项同时最小化，我们必须使 VC 维成为一个可以控制的变量。

下面我们将要给出一个一般的原则，称作结构风险最小化(SRM)归纳原则[①]。这一原则旨在针对经验风险和置信范围这两项最小化风险泛函(Vapnik and Chervonenkis, 1974)。

设函数 $Q(z, \alpha)$，$\alpha\in\Lambda$ 的集合 S 具有一定的结构，这一结构是由一系列嵌套的函数子集 $S_k=\{Q(z, \alpha), \alpha\in\Lambda_k\}$ 组成的(图 4.1)，它们满足

$$S_1 \subset S_2 \subset \cdots \subset S_n \subset \cdots, \qquad (4\text{-}3)$$

其中，结构的元素满足下面两个性质：

(1) 每个函数集 S_k 的 VC 维 h_k 是有限的[②]，因此，

$$h_1 \leq h_2 \leq \cdots \leq h_n \leq \cdots.$$

(2) 结构的任何元素 S_k 或者包含一个完全有界函数的集合

$$0 \leq Q(z, \alpha) \leq B_k, \quad \alpha\in\Lambda_k,$$

或者包含对一定的 (p, τ) 对满足下列不等式的函数集合：

$$\sup_{\alpha\in\Lambda_k} \frac{\left(\int Q^p(z, \alpha)\,dF(z)\right)^{1/p}}{\int Q(z, \alpha)\,dF(z)} \leq \tau, \quad p > 2. \qquad (4\text{-}4)$$

图 4.1 函数集的结构是由嵌套的函数子集确定的

我们把这种结构叫做一个容许结构。

对一个给定的观测集 $z_1, \cdots, z_l$，SRM 原则在使保证风险最小的子集 S_k 中选择使经验风险最小的函数 $Q(z, \alpha_l^k)$。(保证风险[③]是根据情况由不等式(4-1)的右边或不等式(4-2)的右边确定的。)

SRM 原则定义了在对给定数据逼近的精度和逼近函数的复杂性之间的一种折衷。随着子集序号 n 的增加，经验风险的最小值减小，但决定置信范围的项(不等式(4-1)右边的

① 结构风险最小化原文为 structural risk minimization，也可译作有序风险最小化，比如在边肇祺等编著《模式识别》(清华大学出版社，1988)中就采用这样说法。——译者

② 然而集合 S 的 VC 维可以是无穷大。

③ 这里所谓的保证风险(guaranteed risk)实际就是指风险的上界。——译者

第二项或不等式(4-2)中的乘子)却增加(图 4.2)。SRM 原则通过选择子集[①] S_k 将这两者都考虑在内,子集 S_k 的选择是使得在这个子集中,最小化经验风险会得到实际风险的最好的界。

图 4.2　风险的界是经验风险与置信范围之和。随着结构元素序号的增加,经验风险将减小,而置信范围将增加。最小的风险上界是在结构的某个适当的元素上取得的

4.2　收敛速度的渐近分析

用 S^* 来表示函数集

$$S^* = \bigcup_{k=1}^{\infty} S_k.$$

假设函数集 S^* 在 S(回顾 $S=\{Q(z,\alpha),\alpha\in\Lambda\}$)中对于度量

$$\rho(Q(z,\alpha),Q(z,\alpha)) = \int |Q(z,\alpha) - Q(z,\alpha)|\,dF(z)$$

是处处稠密的[②]。

要对 SRM 原则进行渐近分析,我们考虑是这样一种规律,它对任意给定的 l,确定结构(4-3)式中的元素 S_n 的序号

---

① 原著将 S_k 误写为 S_n。——译者

② 处处稠密的含义是,如果对任意的 $\varepsilon>0$ 和任意的 $Q(z,\alpha^*)$,在函数集 $R(z,\beta)$, $\beta\in B$ 中总可以找到一个函数 $R(z,\beta^*)$,使得不等式

$$\rho(Q(z,\alpha^*),R(z,\beta^*)) \le \varepsilon$$

成立,则函数集 $R(z,\beta)$, $\beta\in B$ 在集合 $Q(z,\alpha)$, $\alpha\in\Lambda$ 中在度量 $\rho(Q,R)$ 下是处处稠密的。

$$n = n(l), \tag{4-5}$$

我们将在这样确定的元素 S_n 中最小化经验风险。在这方面，有下面的定理成立[①]。

定理 4.1 如果规律 $n = n(l)$ 使得

$$\lim_{l\to\infty} \frac{T^2_{n(l)} h_{n(l)} \ln l}{l} = 0, \tag{4-6}$$

其中的 T_n 是：

(1) 若所考虑的结构的子集 S_n 中是完全有界函数 $Q(z, \alpha) \leq B_n$，则

$$T_n = B_n;$$

(2) 若所考虑的是一个其元素满足(4-4)等式的结构，则 $T_n = \tau$。

那么，SRM 方法将得到这样一系列逼近 $Q(z, \alpha^{n(l)})$，对这些逼近，风险序列 $R(\alpha^{n(l)})$ 将收敛于最小风险

$$R(\alpha) = \inf_{\alpha\in\Lambda} \int Q(z, \alpha) dF(z),$$

且收敛的渐近速度[②] 是

$$V(l) = r_{n(l)} + T_{n(l)} \sqrt{\frac{h_{n(l)} \ln l}{l}}. \tag{4-7}$$

其中，$r_{n(l)}$ 是下列逼近的速度：

$$r_n = \inf_{\alpha\in\Lambda_n} \int Q(z, \alpha) dF(z) - \inf_{\alpha\in\Lambda} \int Q(z, \alpha) dF(z). \tag{4-8}$$

要得到最好的收敛速度，我们必须知道所选结构的逼近速度 r_n。对函数集的不同结构估计 r_n 的问题是传统函数逼近理论研究的对象，我们将在下一节中讨论这一问题。如果我们知道了逼近速度 r_n，就可以先验地找到规律 $n = n(l)$，它能够通过最小化等式(4-7)的右边部分提供最好的收敛渐近速度。

例 设 $Q(z, \alpha)$，$\alpha \in \Lambda$ 是这样一个函数集，其中的函数对 $p > 2$ 和 $\tau < \tau^* < \infty$ 满足不等式(4-4)。考虑一个 $n = h_n$ 的结构。设逼近的渐近速度是由下面的规律描述的：

$$r_n = \left(\frac{1}{n}\right)^c.$$

(这一规律描述了逼近理论中主要的经典结论，见下一节。)那么，如果

$$n(l) = \left(\frac{l}{\ln l}\right)^{\frac{1}{2c+1}},$$

其中[a]表示 a 的整数部分，这时收敛的渐近速度是

$$V(l) = \left(\frac{\ln l}{l}\right)^{\frac{c}{2c+1}}. \tag{4-9}$$

---

① 由于语言习惯的不同，在翻译本书时改变了原文中定理 4.1 的陈述顺序，这里的公式(4-6)、(4-7)在原文中分别是(4-7)和(4-6)。——译者

② 如果存在一个常数 C，使得

$$V^{-1}(l)\,\|\xi - \xi_l\|^p \xrightarrow[l\to\infty]{} C,$$

则我们说随机变量 ξ_l，$l = 1, 2, \cdots$ 以渐近速度 $V(l)$ 收敛于值 ξ。

4.3 学习理论中的函数逼近问题

定理 4.1 中描述了一种收敛速度的渐近理论,这一理论有吸引力的特性是,我们可以先验地(在学习过程开始之前)找到能够提供最好的(渐近)收敛速度的规律 $n = n(l)$,并且可以先验地估计收敛的渐近速度值[①]。这个速度既依赖于容许结构的构造(即依赖于对序列(h_n, T_n), $n = 1, 2, \cdots$),也依赖于逼近速率 r_n, $n = 1, 2, \cdots$。

在这一信息的基础上,我们可以通过最小化(4-7)式来估算收敛的速度。注意,(4-7)式中的第二项决定了学习过程的随机行为,而这一项是由风险的非渐近界确定的(见(4-1)和(4-2)式)。然而,该式中的第一项(它描述了学习过程的确定性成分)通常只有一个渐近的界。

考虑用一种结构对函数进行逼近的问题,这种结构中的元素 S_n 包含 n 次(代数或三角)多项式,或者在其他有 n 项的序列上的展开。传统的逼近理论研究了函数的平滑性质与这种逼近速度之间的联系。通常,未知函数的平滑性是用它所存在的导数的阶数 s 来表征的。典型的逼近渐近速度有如下的形式:

$$r_n = n^{-\frac{s}{N}}, \tag{4-10}$$

其中,N 是输入空间的维数(Lorentz, 1966)。注意,这意味着只有非常平滑的函数才能保证在高维空间中有高的渐近收敛速度[②]。

在学习理论中,我们希望找到在下面情况下的逼近速度:

(1) $Q(z, \alpha)$, $\alpha \in \Lambda$ 是一个高维函数的集合。

(2) 结构的元素 S_k 不一定是线性流形(它们可以是任意具有有限 VC 维的函数的集合)。

而且,我们关心的是逼近速度高的情况。

因此,在学习理论中,我们面对的是描述在什么情况下可能有高的逼近速度的问题。这就需要描述不同的"平滑"函数的集合以及为 r_n 提供界 $O\left(\frac{1}{n}\right)$(即收敛速度快)的这些集合的结构。

在 1989 年 Cybenko 证明了用 sigmoid 函数[③](神经元)的叠加可以逼近任何平滑函数(Cybenko, 1989)。

在 1992 年—1993 年,Jones(1992), Barron(1993)和 Breiman(1993)描述了不同函数集上的具有快的逼近速度的结构。

他们考虑了如下平滑函数的概念。设$\{f(x)\}$是一个函数集,设$\{\tilde{f}(\omega)\}$是他们的傅里叶变换的集合。用下面的量:

---

① 注意,收敛的高的渐近速度并不一定意味着在有限样本数目上的高的收敛速度。

② 设如果 $r_n \leq n^{-1/2}$,则认为收敛速度是高的。

③ 见引论 0.3.1 节中的译注(第 8 页)。——译者

$$\int \|\omega\|^{d} |\tilde{f}(\omega)| d\omega = C_d(f) < \infty, \quad d \geq 0 \tag{4-11}$$

来表征函数 $f(x)$ 的平滑性。按照这个概念，有下面的关于逼近速度 r_n 的定理。

定理 4.2 (Jones, Barron 和 Breiman) 设函数集中的函数 $f(x)$ 满足(4-11)式，那么如果下面的条件之一成立，则用结构的元素中最好的函数逼近待求函数的速度以 $O\left(\frac{1}{n}\right)$ 为界：

(1) 函数集 $\{f(x)\}$ 是由(4-11)式在 $d=0$ 下确定的，并且结构的元素 S_n 包含函数

$$f(x, \alpha, w, v) = \sum_{i=1}^{n} \alpha \sin[(x \cdot w_i) + v_i], \tag{4-12}$$

其中 α 和 v_i 是任意值，w_i 是任意向量(Jones, 1992)。

(2) 函数集 $\{f(x)\}$ 是由(4-11)式在 $d=1$ 下确定的，并且结构的元素 S_n 包含函数

$$f(x, \alpha, w, v) = \sum_{i=1}^{n} \alpha S[(x \cdot w_i) + v_i], \tag{4-13}$$

其中 α 和 v_i 是任意值，w_i 是任意向量，$S(u)$ 是一个 sigmoid 函数(一个单调上升的函数，且有 $\lim_{u \to -\infty} S(u) = -1$, $\lim_{u \to \infty} S(u) = 1$)(Barron, 1993)。

(3) 函数集 $\{f(x)\}$ 是由(4-11)式在 $d=2$ 下确定的，并且结构的元素 S_n 包含函数

$$f(x, \alpha, w, v) = \sum_{i=1}^{n} \alpha (x \cdot w_i + v_i)_+, \quad (u)_+ = \max(0, u), \tag{4-14}$$

其中 α 和 v_i 是任意值，w_i 是任意向量(Breiman, 1993)。

尽管事实上，在这个定理中的平滑性概念与有界导数的阶数是不同的，但我们在这里仍可以看到与传统情况下相似的现象：要在某一高维空间中得到高的收敛速度，那么随着空间维数的增高，必须同时增加函数的平滑性。这一点从(4-11)式就可以直接得到。Girosi 和 Anzellotti(1993)发现，在 $d=1$ 和 $d=2$ 下，满足(4-11)式的函数集可以重写为

$$f(x) = \frac{1}{\|x\|^{n-1}} * \lambda(x), \quad f(x) = \frac{1}{\|x\|^{n-2}} * \lambda(x),$$

其中，$\lambda(x)$ 是任意傅里叶变换可积的函数，$*$ 表示卷积运算。在这种形式下更明显可以看出，由于 $1/\|x\|^{n-1}$ 项的迅速衰减，满足(4-11)式的函数随着维数的增高越来越受到限制。

同样的现象在 Mhaskar(1993)的结论中也很清楚。Mhaskar 证明了，用结构(4-13)式逼近有 s 阶连续导数的函数的收敛速度是 $O(n^{-\frac{s}{N}})$。

因此，如果待求函数不是非常平滑，我们无法保证函数收敛到未知函数的渐近速度是高的。

在第 4.5 节中我们将讨论一种新的学习模型，它基于对待求函数局部逼近的思想(而不是像上面考虑的那样全局逼近)。我们将考虑对待求函数在感兴趣的点的某个邻域内逼近，而邻域的半径可以随着观测数目的增加而减小。

局部逼近的速度可以高于全局逼近的速度，从而使学习机器有更好的推广能力。

4.4 神经网络的子集结构举例

SRM 的一般原则可以用很多不同的方法实现。这里我们讨论针对神经网络所实现的函数集的三种不同结构[①]的例子。

1. 由神经网络的构造所形成的一种结构

考虑一个全连接的前馈神经网络的集合,其中某个隐层中的节点数目是单调增加的。随着隐单元数目的增加,这些神经网络可以实现的函数集合就定义了一种结构(图 4.3)。

图 4.3 由隐单元数目确定的一种结构

2. 由学习过程给出的一种结构

考虑由一个固定构造的神经网络所能实现的函数集合 $S=\{f(x,w),w\in W\}$。参数 $\{w\}$ 是神经网络的权值。通过定义 $S_p=\{f(x,w),\|w\|\leq C_p\}$ 且 $C_1<C_2<\cdots<C_n$,就引入了一种结构。在损失函数集的一些很一般的条件下,在这个结构的元素 S_p 内,经验风险的最小化可以通过适当选择 Lagrange 乘子 $\gamma_1>\gamma_2>\cdots>\gamma_n$ 时最小化

$$E(w,\gamma_p)=\frac{1}{l}\sum_{i=1}^{l}L(y_i,f(x_i,w))+\gamma_p\|w\|^2$$

得到。众所周知的“权值衰减”过程就涉及了对这一泛函的最小化。

3. 由预处理给出的一种结构

考虑一个固定构造的神经网络,其输入经过了一个变换 $z=K(x,\beta)$,其中的参数 β 控制了由这种变换引入的退化程度(比如 β 可以是某个平滑核函数的宽度)。

由 $\beta\geq C_p$ 和 $C_1>C_2>\cdots>C_n$ 就定义了函数集 $S=\{f(K(x,\beta),w),w\in W\}$ 的一种结构。

要用这些结构执行 SRM 原则,我们必须知道(估计)结构的任意元素 S_k 的 VC 维,而

① 这里的结构(structure)指神经网络所实现的函数集的有序子集结构。为了区分,我们把神经网络本身的结构(即神经元节点的数目和神经元之间的连接方式)叫做神经网络的构造,原文使用的是 architecture。——译者

且必须能够对任意 S_k 找到使经验风险最小的函数。

4.5 局部函数估计的问题

我们考虑一个基于经验数据(在某一给定点 x_0 的邻域内)最小化局部风险的模型。用一个非负函数 $K(x, x_0; \beta)$ 来表达邻域的概念,这个函数依赖于点 x_0 和一个“局部性”参数 $\beta \in (0, \infty)$,并且满足下面两个条件:

$$0 \leq K(x, x_0; \beta) \leq 1,$$
$$K(x_0, x_0; \beta) = 1. \tag{4-15}$$

比如,“硬限”近邻函数(图 4.4(a))

$$K_1(x, x_0; \beta) = \begin{cases} 1 & \text{若 } \| x - x_0 \| < \frac{\beta}{2} \\ 0 & \text{其他} \end{cases} \tag{4-16}$$

和“软限”近邻函数(图 4.4(b))

$$K_2(x, x_0; \beta) = \exp - \frac{(x - x_0)^2}{\beta^2} \tag{4-17}$$

都满足这些条件。

图 4.4 近邻函数的例子

我们定义一个值:

$$K(x_0, \beta) = \int K(x, x_0; \beta) dF(x). \tag{4-18}$$

对函数集 $f(x, \alpha)$, $\alpha \in \Lambda$,考虑损失函数集合 $Q(z, \alpha) = L(y, f(x, \alpha))$, $\alpha \in \Lambda$。我们的目标是最小化如下的局部风险泛函:

$$R(\alpha, \beta, x_0) = \int L(y, f(x, \alpha)) \frac{K(x, x_0; \beta)}{K(x_0; \beta)} dF(x, y), \tag{4-19}$$

这种最小化是在概率测度 $F(x, y)$ 未知,但给定了独立同分布的样本

$$(x_1, y_1), \cdots, (x_l, y_l)$$

的情况下,在函数集 $f(x, \alpha)$, $\alpha \in \Lambda$ 和 x_0 点的不同近邻(由参数 β 所确定)上进行的。注意,在经验数据基础上局部风险最小化的问题是全局风险最小化问题的一个推广。(在后

一问题中,我们须对 $K(x, x_0; \beta) = 1$ 最小化泛函(4-19)式。)

对局部风险最小化问题,我们可以推广全局风险最小化问题中得到的界,即同时对所有有界函数 $A \leq L(y, f(x, \alpha)) \leq B$, $\alpha \in \Lambda$ 和所有函数 $0 \leq K(x, x_0; \beta) \leq 1$, $\beta \in (0, \infty)$, 下面的不等式以概率 $1 - \eta$ 成立:

$$R(\alpha, \beta, x_0) \leq \frac{\frac{1}{l}\sum_{i=1}^{l} L(y_i, f(x_i, \alpha)) K(x_i, x_0; \beta) + (B - A) E(l, h_\Sigma)}{\frac{1}{l}\sum_{i=1}^{l} K(x_i, x_0; \beta) - E(l, h_\beta)}_{+},$$

$$E(l, h) = \frac{h \ln \frac{2l}{h} + 1 - \ln \frac{\eta}{2}}{l},$$

其中, h_Σ 是函数集 $L(y, f(x, \alpha)) K(x, x_0; \beta)$, $\alpha \in \Lambda$, $\beta \in (0, \infty)$ 的 VC 维, h_β 是函数集 $K(x, x_0; \beta)$ 的 VC 维(Vapnik and Bottou, 1993)。

现在,利用 SRM 原则,我们可以对三个参数最小化上述不等式的右边部分,这三个参数是:经验风险值、VC 维 h_Σ 和近邻 β 的值(VC 维 h_β)。

在某些情况下,基于函数集给定的结构,利用给定数目的观测不可能较好地逼近待求函数。在这种情况下,局部风险最小化的方法有它的优势,这就是,它可能得到对待求函数在任意所关心的点上一个较好的局部逼近(图 4.5)。

图 4.5 用线性函数,我们可以在所关心的任意点的近邻内估计未知的平滑函数

4.6 最小描述长度与 SRM 原则

SRM 归纳原则是基于对经验过程收敛速度的统计分析之上的,与它有联系的,还有另外一种针对小样本数的归纳推理原则,就是所谓的最小描述长度(MDL)原则,它是基于对随机性概念的信息论分析之上的。在本节中,我们来考虑 MDL 原则,并针对模式识别问题指出 SRM 原则和 MDL 原则之间的联系。

在 1965 年,Kolmogorov 用算法复杂度的概念定义了一个随机串。

他定义,一个对象的算法复杂度是描述这个对象的最短的二进制计算机程序的长度,他并且证明,算法复杂度的值是不依赖于计算机的类型的,至多差一个加性常数。因此,这是对象的一个通用的特性。

Kolmogorov 的主要思想是:

对描述一个对象的串,如果对象的算法复杂度高,即描述它的串不能被显著地压缩,则认为这个串是随机的。

在算法复杂度的概念提出 10 年后,Rissanen(1978)提出利用 Kolmogorov 的概念作为学习机器归纳推理的主要工具,提出了所谓的 MDL 原则[①]。

4.6.1 MDL 原则

假设给定了一个训练数据对的集合

$$(\omega, x_1), \cdots, (\omega, x_l).$$

(数据对是根据某一未知概率测度独立地随机抽取的。)考虑两个串:一个是二值串

$$\omega, \cdots, \omega \tag{4-20}$$

另一个是向量串

$$x_1, \cdots, x_l \tag{4-21}$$

要研究的问题是:给定(4-21)式,串(4-20)式是一个随机的对象吗?

要回答这一问题,让我们用 Solomonoff-Kolmogorov 的思想分析串(4-20)式的算法复杂度。因为 $\omega, \cdots, \omega$ 都是二值的,因此串(4-20)式是由 l 个比特位描述的。

为了确定这个串的复杂度,我们尝试对它的表示进行压缩。既然训练对是随机独立抽取的,因此 ω 的值可能依赖于向量 x_i 但不依赖于向量 x_j, $i \neq j$(当然前提是这种依赖关系存在)。

考虑下面的模型:假设我们有给定的某个固定的码本 C_b,其中有 $N \geqslant 2^l$ 个不同的码表 T_i, $i = 1, \cdots, N$。每一个码表 T_i 描述了从 x 到 ω 的某个函数[②]。

我们来尝试在码本 C_b 中寻找一个码表 T,它以可能的最好方式来描述串(4-20)式,也就是说,这个码表对给定的串(4-21)式返回的二值串

$$\omega^*, \cdots, \omega^* \tag{4-22}$$

与串(4-20)式之间的 Hamming 距离最小(即用这个码表解码串(4-20)式的错误数最小)。

假如我们找到了一个完美的码表 T_0,使得所产生的串(4-22)式与串(4-20)式之间的 Hamming 距离为零,这个码表就解译了串(4-20)式。

因为码本 C_b 是固定的,因此要描述串(4-20)式,只要给出在这个码本中码表 T_0 的序号就足够了。描述 N 个码表中的任一个所需的最小比特数是$\lceil \log_2 N \rceil$,其中$\lceil A \rceil$表示不

① 在 Kolmogorov 提出他的随机性模型之前,Solomonoff 就考虑了用算法复杂度来作为一个一般的归纳原则。因此,描述复杂度的原则被叫做 Solomonoff-Kolmogorov 原则。但是,只是在 Rissanen 的工作之后,这一原则才被看作学习理论中推理的一个工具。

② 严格地说,要在码本中得到有限长度的码表,输入向量 x 必须是离散的。但是,正如我们将要看到的,量化的级数并不影响推广能力的界。因此我们可以考虑任意程度上的量化,甚至给出有无穷多项的码表。

小于 A 的最小的整数。因此,在这种情况下,描述串(4-20)式需要$\lceil \log_2 N \rceil$个(而不是 l 个)比特位。于是,利用一个包含有完美码表的码本,我们就可以把串(4-20)式的描述长度压缩一个比例

$$K(T_o) = \frac{\lceil \log_2 N \rceil}{l}. \tag{4-23}$$

我们把 $K(T_o)$ 叫做对串(4-20)式的压缩系数。

现在考虑一般的情况:码本 C_b 中不包含完美码表。设两个串(产生的串(4-22)式与待求的串(4-20)式)之间最小的 Hamming 距离为 $d \geq 0$。不失一般性我们可以假设 $d \leq l/2$。(否则的话,我们可以不用最小距离,而找最大 Hamming 距离的码表,在解码时将 1 变成 0、0 变成 1。这将导致在编码方案中多用一位。)这意味着要正确描述这个串,我们需要对从码本中选出的码表所给的结果进行 d 处修正。

对固定的 d,对长度为 l 的串有 C_l^d 种可能的不同修正。要确定其中的一种(即确定 C_l^d 个取值中任一个的序号),我们需要$\lceil \log_2 C_l^d \rceil$个比特。

因此,要描述串(4-20)式我们需要:$\lceil \log_2 N \rceil$比特来定义码表的序号和$\lceil \log_2 C_l^d \rceil$比特来描述修正。我们还需要$\lceil \log_2 d \rceil + \Delta_1$比特来确定修正的数目 d,其中 $\Delta_1 < 2\log_2 \log_2 d$,$d > 2$。把这些算在一起,我们需要$\lceil \log_2 N \rceil + \lceil \log_2 C_l^d \rceil + \lceil \log_2 d \rceil + \Delta_1$比特来描述串(4-20)式。这个数目应该与 l 相比较,l 是描述任意二值串(4-20)式所需的比特数。因此压缩系数是

$$K(T) = \frac{\lceil \log_2 N \rceil + \lceil \log_2 C_l^d \rceil + \lceil \log_2 d \rceil + \Delta_1}{l}. \tag{4-24}$$

如果压缩系数 K(T) 较小,那么根据 Solomonoff-Kolmogorov 的思想,这个串就不是随机的,而是在一定程度上依赖于输入向量 x。在这种情况下,解码码表 T 就在一定程度上逼近 x 和 ω之间的函数关系。

4.6.2 对于 MDL 原则的界

一个重要的问题是:

压缩系数 K(T) 是否确定了用码表 T 对向量 x 分类(解码)的测试错误概率?

回答是肯定的。

为了证明这一点,我们来比较一下在最简单的模型(有限函数集的学习机器)中 MDL 原则得到的结果和 ERM 原则得到的结果。

在这一章的开头,我们考虑了对模式识别问题学习机器推广能力的界(4-1)式。对学习机器有有限数目 N 个函数的特殊情况,我们得到,以至少 $1-\eta$的概率,不等式

$$R(T_i) \leq R_{emp}(T_i) + \frac{\ln N - \ln \eta}{l}\left(1 + \sqrt{1 + \frac{2R_{emp}(T_i)l}{\ln N - \ln \eta}}\right) \tag{4-25}$$

对给定函数集中的所有 N 个函数(对给定码本中的 N 个码表)同时成立。我们把不等式的右边进行适当的变换,这要用到压缩系数的概念和下面的关系:

$$R_{emp}(T_i) = \frac{d}{l}.$$

注意到,对 $d \leq l/2$ 和 $l > 6$,有下面的不等式成立:

$$\frac{d}{l} + \frac{\ln N - \ln \eta}{l}\left(1 + \sqrt{1 + \frac{2d}{\ln N - \ln \eta}}\right) < 2\,\frac{\lceil \ln N \rceil + \lceil \ln C_l^d \rceil + \lceil \log_2 d \rceil + \Delta_d}{l} - \frac{\ln \eta}{l} \tag{4-26}$$

(读者可以容易地验证这一不等式)。现在,让我们把不等式(4-26)的右边用压缩系数(4-24)式改写为

$$2\ln 2\,\frac{\lceil \log_2 N \rceil + \lceil \log_2 C_l^d \rceil}{l} + \frac{\lceil \log_2 d \rceil + \Delta_d}{l} - \frac{\ln \eta}{l} \leq 2\ K \ln 2 - \frac{\ln \eta}{l}.$$

因为不等式(4-25)以至少 $1-\eta$ 的概率成立,不等式(4-26)以概率 1 成立,因此不等式

$$R(T_i) < 2\ K(T_i)\ln 2 - \frac{\ln \eta}{l} \tag{4-27}$$

以至少 $1-\eta$ 的概率成立。

4.6.3 SRM 和 MDL 原则

现在,假设我们共有 M 个码本,它们有如下的结构:码本 1 包含较少的码表,码本 2 包含这些码表和多一些的码表,依此类推。

在这种情况下,我们可以用更复杂的编码方案来描述串(4-20)式:首先,描述码本的序号 m(这需要 $\lceil \log_2 m \rceil + \Delta_m$, $\Delta_m < 2\lceil \log_2 \log_2 m \rceil$ 个比特),然后用这个码本来描述串(如上面所述,这需要 $\lceil \log_2 N \rceil + \lceil \log_2 C_l^d \rceil + \lceil \log_2 d \rceil + \Delta_d$ 个比特)。

这种情况下所需要的总的描述长度是 $\lceil \log_2 N \rceil + \lceil \log_2 C_l^d \rceil + \lceil \log_2 d \rceil + \Delta_d + \lceil \log_2 m \rceil + \Delta_m$,压缩系数是

$$K(T) = \frac{\lceil \log_2 N \rceil + \lceil \log_2 C_l^d \rceil + \lceil \log_2 d \rceil + \Delta_d + \lceil \log_2 m \rceil + \Delta_m}{l}.$$

对这种情况,存在一个与(4-27)式类似的不等式。因此,对用来压缩串(4-20)式的描述码表,错误概率以不等式(4-27)为界。

这样,对 $d \leq l/2$ 和 $l > 6$,我们证明了下面的定理:

定理 4.3 如果在码本的一个给定结构上,我们以比例 $K(T)$ 用码表 T 压缩了串(4-20)式的描述,那么,以至少 $1-\eta$ 的概率,我们可以断定码表 T 发生错误的概率以下式为界:

$$R(T) < 2\ K(T)\ln 2 - \frac{\ln \eta}{l},\ l > 6. \tag{4-28}$$

注意压缩系数的概念有多么大的作用:要得到错误概率的界,我们实际上只需要关于这个系数的信息①,并不需要另一些细节,比如:

(1) 使用了多少个样本;

① 不等式右边的第二项 $-\ln \eta / l$ 实际上是十分简单的:对合理的 η 和 l,与第一项相比它可以忽略,但是它防止了人们选择过小的 η 和/或过小的 l。

(2) 码本的结构是怎样组织的;

(3) 采用了哪个码本;

(4) 在码本中有多少个码表;

(5) 用这个码表造成了多少训练错误。

尽管如此,界(4-28)式并不比在一致收敛理论的基础上得到风险的界(4-25)式差多少。后者有一个更复杂的结构,并且利用了函数集中函数(码表)数目的信息、训练集上错误数目的信息以及训练集中元素数目的信息。

同时注意到,界(4-28)式不能以大于乘子 2 的幅度被改进:容易证明,当码本中存在一个完美码表时,可以以乘子 1 使得等式成立。

这个定理证明了 MDL 原则:要使错误概率最小化,必须最小化压缩系数。

4.6.4 MDL 原则的一个弱点

然而,在 MDL 原则中存在一个弱点。

前面曾提到,MDL 原则使用包含有限个码表的码本。因此,要处理由一个连续值域的参数所确定的函数的集合,我们就必须做出有限个码表。

这可以用很多方法完成。问题在于:

对给定的函数集,什么才是一个"巧妙的"码本?

换言之,对一个给定的函数集,怎样才能构造一个包含小数目的码表、却有好的逼近能力的码本?

一个"巧妙的"量化有可能大大减少码本中码表的数目,这会影响到压缩系数。遗憾的是,寻找一个"巧妙的"量化是一个非常难的问题。这就是 MDL 原则中的弱点。

在下一章里,我们将考虑一种在十分高维的空间中正规化的线性函数集合(在我们的实验里,用了 $N \approx 10^{13}$ 维空间中的线性函数)。我们将说明,限制了其模的函数子集的 VC 维依赖于对模的约束值。它可以是较小的(在我们的实验中 $h \approx 10^2 \sim 10^3$)。如果这个集合中的一个函数将样本数为 l 的一个训练集没有错误地分开,那么就可以保证测试错误的概率与 $h \ln l / l$ 成正比。

对这个指示函数集用 MDL 方法的问题是:如何构造约有 l^h 个码表(而不是约有 l^N 个码表)的码本,使之能够很好地逼近这个线性函数集。

当构造合理的码本这一问题有一个明显的解决方法时,MDL 原则可以很好地发挥作用。但是即使在这种情况下,它也并不比 SRM 原则更好。这一点只要回顾一下 MDL 原则的界(它不能只用压缩系数的概念进一步改进)是通过粗化 SRM 原则的界得到的就清楚了。

非正式推导和评述——4

在计算数学和统计学的各个领域中,改进方法性能的很多努力都基本上引向了同一个思想,就是我们所称的结构风险最小化归纳原则。

这一思想首先出现在解决不适定问题的方法中:

(1) 拟解[①]的方法(Ivanov, 1962),

(2) 正则化方法(Tikhonov, 1963)。

之后,它出现在非参数密度估计方法中:

(1) Parzen 窗(Parzen, 1962),

(2) 投影方法(Chentsov, 1963),

(3) 条件最大似然方法(筛法(Grenander, 1981)),

(4) 最大惩罚似然方法(Tapia and Thompson, 1978),等等。

这一思想后来又在回归估计的方法中出现:

(1) 岭回归(Hoerl and Kennard, 1970),

(2) 模型选择(参见 Miller(1990)的综述)。

最后,它出现在对模式识别和回归估计算法的正则化技术中(Poggio and Girosi, 1990)。

当然,对利用容许函数集的一个结构来寻找解的思想,曾经有多次尝试以证明其正确性。但是,在传统方法的框架下,只有一些特殊的问题和渐近的情况得到了这种证明。

在依据经验数据最小化风险的模型中,SRM 原则提供了对容量(VC 维)的控制,可以适应有限数目观测的情况。

4.7 解决不适定问题的方法

1962 年,Ivanov 提出了一种为解决不适定问题,寻找线性算子方程

$$Af = F, \quad f \in M \tag{4-29}$$

① 原文为 quasi-solution,也可译为准解或近似解。——译者

的拟解的思想。(方程中的算子 A 把有度量 ρ_1 的度量空间 M ⊂ E_1 中的元素,映射到有度量 ρ_2 的度量空间 N ⊂ E_2 中的元素。)他提出,考虑一个嵌套的凸紧子集的集合

$$M_1 \subset M_2 \subset \cdots \subset M_k \cdots \tag{4-30}$$

$$\overline{\bigcup_{i=1}^{\infty} M_i} = M, \tag{4-31}$$

对其中每一个子集 M_i,找到一个函数 $f_i^* \in M_i$,使得距离

$$\rho = \rho_2(Af, F)$$

最小。Ivanov 证明了在一些一般的条件下,解序列

$$f_1^*, \cdots, f_k^*, \cdots$$

收敛于所期望的解。

拟解方法是与 Tikhonov 提出的正则化技术同时提出的;实际上,这两者是等价的。在正则化技术中,我们引入一个非负的(下)半连续泛函 $\Omega(f)$,它有以下的特性:

(1) 这个泛函的定义域与 M(方程(4-29)的解所属的域)重合。

(2) 使不等式

$$M_j = \{f: \Omega(f) \leq d_j\}, \quad d_j > 0$$

成立的区域形成了空间 E_1 度量下的一个紧统。

(3) (4-29)式的解属于某个 M_i^*:

$$\Omega(f) \leq d^* < \infty.$$

Tikhonov 建议寻找对不同的 γ,使泛函

$$\Phi_\gamma(f) = \rho_2^2(Af, F) + \gamma\Omega(f)$$

最小的函数 f_γ 的序列。他证明了当 γ 收敛于 0 时,f_γ 收敛于待求的解。

Tikhonov 还讨论了在算子方程的右边只是以某一精度 δ 给出的情况,即

$$\rho_2(F, F_\delta) \leq \delta,$$

他指出,即使在这种情况下也可以采用正则化技术。在这种情况下,如果

$$\lim_{\delta \to 0} \gamma(\delta) = 0,$$

$$\lim_{\delta \to 0} \frac{\delta^2}{\gamma(\delta)} = 0,$$

那么,通过最小化泛函

$$\Phi^*(f) = \rho_2^2(Af, F_\delta) + \gamma(\delta)\Omega(f), \tag{4-32}$$

可以得到一个解序列 f_δ,当 $\delta \to 0$ 时(在 E_1 的度量下)它收敛于所期望的解 f_0。

在这两种方法中,正式的收敛证明中都没有显式地包含“容量控制”。然而,关键点是在 Ivanov 方法中的任何子集 M_i 和在 Tikhonov 方法中的任何子集 $M = \{f: \Omega(f) \leq c\}$ 都是紧的。这就意味着它们有有界的容量(一个 ε 熵度量)。

因此,这两种方法都执行了 SRM 原则:首先在容许函数集合上定义一个结构,使得结构中的每个元素都有有限的容量,其容量随着元素序号递增。然后,在结构的任一元素上找到提供对方程右边最好逼近的函数。所得到的解序列收敛于待求的解。

4.8 随机不适定问题和密度估计问题

在 1978 年,我们把正则化理论推广到了随机不适定问题上(Vapnik and Stefanyuk, 1978)。我们考虑了在下述情况下求解算子方程(4-29)的问题:方程的右边是未知的,但给定了一个逼近序列 F_δ,它具有如下性质:

(1) 其中的每一个逼近 F_δ 是一个随机函数①。

(2) 当 δ 收敛于零时逼近序列依概率收敛于未知函数 F(在空间 E_2 的度量下)。换言之,随机函数序列 F_δ 具有下面的特性:

$$P\{\rho_2(F, F_\delta) > \varepsilon\} \underset{\delta \to 0}{\to} 0, \quad \forall \varepsilon > 0$$

利用 Tikhonov 的正则化技术,我们可以在随机函数 F_δ 的基础上得到一个对(4-29)式的解的逼近序列 f_δ。

我们证明了对任意 $\varepsilon > 0$,存在 $\gamma_0 = \gamma_0(\varepsilon)$,使得对任意的 $\gamma(\delta) \le \gamma_0$,使泛函(4-32)式最小的函数满足下面的不等式:

$$P\{\rho_1(f, f_\delta) > \varepsilon\} \le 2P\{\rho_2^2(F, F_\delta) > \gamma(\delta)\varepsilon\}. \tag{4-33}$$

也就是说,我们把方程右边的逼近与其准确值之间的随机偏差的分布(在 E_2 度量下)和用正则化方法所得的解与所期望的解之间的偏差的分布(在 E_1 度量下)联系了起来。

特别地,这一结论给了我们一个机会,使得我们可以去寻找一个构造各种密度估计方法的一般途径。

如 1.8 节中已经指出的,密度估计要求我们用独立同分布数据 $x_1, \cdots, x_l, \cdots$ 求解下面的积分方程:

$$\int_{-\infty}^{x} p(t)\,dt = F(x),$$

其中 $F(x)$ 是一个未知的概率分布函数。

我们构造一个经验分布函数

$$F_l(x) = \frac{1}{l}\sum_{i=1}^{l} \theta(x - x_i),$$

它是对 $F(x)$ 的一个随机逼近,因为它是用随机数据 $x_1, \cdots, x_l$ 构造的。

在 3.9 节中,我们发现,差 $\sup_x |F(x) - F_l(x)|$ 是由 Kolmogorov-Smirnov 界描述的。利用这个界,我们得到

$$P\left\{\sup_x (F(x) - F_l(x)) > \varepsilon\right\} < 2e^{-2\varepsilon^2 l}.$$

因此,如果我们最小化正则泛函②

---

① 随机函数就是用某个随机事件的实现定义的函数。关于随机函数的定义,读者可以参考任一概率论的高级教材,比如 Schiryaev A N. Probability. Springer, New York。

② 原文中(4-34)式误写为 $R(p) = \rho_{E_2}^2\left(\int_{-\infty}^{x} p(t)\,dt, F_l(x)\right) + \gamma\Omega(p)$。—— 译者

$$R(p) = \rho^2 \int_{-\infty}^{x} p(t)\,dt, F_l(x) + \gamma\Omega(p), \tag{4-34}$$

那么,根据不等式(4-33),我们得到的估计 $p_l(t)$ 与所期望的解的偏差可以用下面的关系描述:

$$P\{\rho_1(p, p_l) > \varepsilon\} \leq 2\exp\{-2\varepsilon\gamma\}.$$

因此,所得到的估计子一致性的条件是

$$\begin{aligned} &\gamma \underset{l\to\infty}{} 0, \\ &l\gamma \underset{l\to\infty}{} \infty. \end{aligned} \tag{4-35}$$

于是,在 (4-35) 式的条件下最小化(4-34)式类型的泛函就得到一致的估计子。利用不同的范数 E_2 和不同的泛函 $\Omega(p)$,我们可以得到各种类型的密度估计子(包括所有传统的估计子①)。就我们所关心的来说,重要的是所有的非参数密度估计子都遵循了 SRM 原则。通过选择泛函 $\Omega(p)$,我们就在容许解的集合上定义了一个结构(由常数 c 确定的嵌套函数集 $M_c = \{p: \Omega(p) \leq c\}$);利用法则 γ 可以确定结构中的适当的元素。

在第七章中,我们将利用这种方法建立估计密度、条件密度和条件概率的直接方法。

4.9 回归的多项式逼近问题

构造回归函数多项式逼近的问题在 70 年代十分流行,这一问题对于理解小样本数统计学里出现的问题起着重要的作用。

为简单起见,考虑用多项式估计一个一维回归函数的问题。令回归 $f(x)$ 是一个平滑函数。假设给定了对这一函数在有噪声情况下的有限数目的测量

$$y_i = f(x_i) + \xi, \quad i = 1, \cdots, l,$$

(在问题的不同表示中,利用了关于未知噪声不同类型的信息;在这里带噪声测量的模型中,我们假设噪声 ξ 的值不依赖于 x_i,而且测量的点 x_i 是根据一个未知的概率分布 $F(x)$ 随机选择的)。

要研究的问题是寻找与未知的回归函数 $f(x)$ 最接近(比如在 $L_2(F)$ 度量下)的多项式。与 1.7.3 小节介绍的传统的回归问题不同,这里要从中逼近回归的函数集是很宽的(任意阶多项式),而观测数目是固定的。

解决这个问题的过程在理解小样本数问题的本质方面给统计学家们上了一课。人们首先考虑了这个问题的一个简化版本:回归函数本身就是一个多项式(但其阶数未知),而且噪声服从零均值正态密度。对这个特殊的问题,采用了传统的渐近方法:在假设检验技术的基础上,估计回归多项式的阶数,然后再估计多项式的系数。然而实验证明,对小样本

---

① 这里顺便指出,如果用经验分布函数 $F_l(x)$ 逼近未知分布函数 $F(x)$,我们可以得到所有传统的估计子。但是经验分布函数并不是分布函数最好的逼近,因为,根据定义,分布函数应该是一个绝对连续函数,而经验分布函数是不连续的。利用绝对连续逼近(比如在一维情况下的多项式(译注:这里原文中误写作多边形)),我们可以得到更好的估计子,它们不但有好的渐近特性(这一点与传统的估计子相同),而且还拥有一些从观测数目有限的角度出发十分有用的特性(Vapnik, 1988)。

数,这种思想是错误的:即使知道回归多项式的实际阶数,我们也经常不得不选择一个小一些的阶数来逼近,这取决于可用的观测数目。

所以,人们提出了估计逼近多项式阶数的许多思想,其中包括文献(Akaike, 1970)和(Schwartz, 1978)中提出的思想(参见 Miller, 1990)。但是,这些思想只在渐近情况下得到了证实。

4.10 容量控制的问题

4.10.1 选择多项式的阶数

在回归问题中选择适当的多项式阶数 p 的问题可以在 SRM 原则的基础上进行研究,其中多项式集合采用最简单的结构:结构的第一个元素中包含一阶多项式:

$$f_1(x, \alpha) = \alpha x + \alpha, \quad \alpha = (\alpha, \alpha) \in R^2,$$

第二个元素包含二阶多项式:

$$f_2(x, \alpha) = \alpha x^2 + \alpha x + \alpha, \quad \alpha = (\alpha, \alpha, \alpha) \in R^3,$$

依此类推。

要选择多项式的最佳阶数,我们可以最小化下面的泛函(界(3-30)的右边部分):

$$R(\alpha, m) = \frac{\frac{1}{l}\sum_{i=1}^{l}(y_i - f_m(x_i, \alpha))^2}{1 - c\sqrt{E_l}}, \tag{4-36}$$

$$E_l = 4\frac{h_m \ln\frac{2l}{h_m} + 1 - \ln\frac{\eta}{4}}{l},$$

其中,h_m 是损失函数集合

$$Q(z, \alpha) = (y - f_m(x, \alpha))^2, \quad \alpha \in \Lambda$$

的 VC 维,c 是一个常数,它决定了"分布的尾部"(参见 3.4 节和 3.7 节)。

我们可以证明,对实函数集

$$Q(z, \alpha) = F(|g(z, \alpha)|), \quad \alpha \in \Lambda,$$

其中 F(u)是任意固定的单调函数,它的 VC 维不超过 eh^*,其中 $e < 9.34$,而 h^* 是函数集

$$I(z, \alpha, \beta) = \theta(g(x, \alpha) - \beta), \quad \alpha \in \Lambda, \ \beta \in R^1$$

的 VC 维。因此对我们的损失函数,其 VC 维有下面的界:

$$h_m \leq e(m + 1).$$

要知道最优的逼近多项式,我们需要选择多项式的阶数 m 和系数 α,使得泛函[①] (4-36)式最小。

---

① 我们在选择最优逼近多项式阶数的几个基准测试研究中使用了这个泛函(采用常数 $c = 1$ 及 $E_l = [m(\ln(l/m) + 1) - \ln\eta]/l$,其中 $\eta = l^{-1/2}$)。对小样本数,所得结果通常好于基于传统方法得到的结果。

4.10.2 选择最优的稀疏代数多项式

现在我们介绍代数多项式集合的另一种结构: 让结构的第一个元素包含多项式 $P_1(x, \alpha) = \alpha x^d$, $\alpha \in R^1$(阶数 d 任意), 其中只有一个非零项; 让第二个元素包含多项式 $P_2(x, \alpha) = \alpha x^{d_1} + \alpha x^{d_2}$, $\alpha \in R^2$, 其中有两个非零项; 依此类推。要研究的问题是选择最好的稀疏多项式 $P_m(x)$ 来逼近一个平滑回归函数。

为了做到这一点, 我们需要估计损失函数

$$Q(z, \alpha) = (y - P_m(x, \alpha))^2$$

的 VC 维, 其中 $P_m(x, \alpha)$, $\alpha \in \Lambda$ 是包含 m 项的任意阶多项式的集合。我们考虑单变量 x 的情况。

这个损失函数集的 VC 维 h 可以以 $2h^*$ 为界, 其中 h^* 是下面的指示器的 VC 维:

$$I(y, x) = \theta(y - P_m(x, \alpha) - \beta), \quad \alpha \in R^m, \quad \beta \in R^1.$$

Karpinski 和 Werther(1989)证明, 这一指示器集合的 VC 维 h^* 满足下面的界:

$$3m \leq h^* \leq 4m + 3.$$

因此我们的损失函数集的 VC 维小于 e(4m+3)。这一估计可以用来寻找使得泛函(4-36)式最小的稀疏代数多项式。

4.10.3 三角多项式集合上的结构

下面考虑三角多项式集合的结构。首先我们考虑一种由多项式的阶数确定的结构[①]。对 m 阶三角多项式, 我们的损失函数集的 VC 维小于 $h = 4m + 2$。因此, 要选择最优三角逼近, 我们可以最小化泛函(4-36)式。对这种结构, 三角多项式与代数多项式没有区别。

当我们构造稀疏三角多项式的结构时区别就出现了。稀疏代数多项式的结构中任意元素都有有限的 VC 维, 而稀疏三角多项式的结构中任意元素的 VC 维都是无穷大。

这个结论是从下面指示函数集的 VC 维是无穷大这一事实得到的:

$$f(x, \alpha) = \theta(\sin(\alpha x)), \quad \alpha \in R^1, \quad x \in (0, 1)$$

(参见 3.6 节例 2)。

4.10.4 特征选择的问题

选择稀疏多项式的问题在学习理论中起着十分重要的作用, 因为这个问题的推广就是一个利用经验数据进行特征选择(特征构造)的问题。

正如在例子中看到的, 上面的特征选择问题(稀疏多项式的项可以看作是特征)是比

① 阶数为 m 的三角多项式有下面的形式[②]:

$$f_m(x) = \sum_{k=1}^{m} (a_k \sin(kx) + b_k \cos(kx)) + a_0.$$

② 原著误写为 $f_p(x)$。—— 译者

较棘手的。为了避免在稀疏三角多项式中遇到的问题,我们需要先验地构造一个结构,其中元素的 VC 维是有界的,然后从这个结构的函数中选择决策规则。

构造学习算法的一个结构来选择(构造)特征并控制容量通常是一个很难的组合问题。

80 年代,在应用统计学中进行了一些努力,试图寻找能够控制容量的选择非线性函数的可靠方法。特别是,统计学家们开始研究在下面的函数集中的函数估计问题:

$$y = \sum_{j=1}^{m} \alpha_j K(x, w_j) + \alpha_0,$$

其中 $K(x, w)$ 是一个关于向量 x 和 w 的对称函数,$w_1, \cdots, w_m$ 是未知向量,$\alpha_1, \cdots, \alpha_m$ 是未知标量(Friedman and Stuetzle, 1981 及 Breiman, Friedman, Olshen and Stone, 1984)(这与 70 年代所研究的方法形成对比,那时研究的是估计对参数呈线性的函数(Miller, 1990))。在这一类函数中对函数 $K(x, w_j)$, $j = 1, \cdots, m$ 的选择可以解释为特征选择。

在下一章我们将看到,对这一类型的函数集,可以有效地控制决定推广容量的两个因素——经验风险值和 VC 维。

4.11 容量控制的问题与贝叶斯推理

4.11.1 学习理论中的贝叶斯方法

在函数估计的传统体系中,贝叶斯方法占有一个重要的位置(Berger, 1985)。

根据贝叶斯公式,两个事件 A 和 B 通过下面的等式联系起来:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

人们用这个公式来修改函数估计的最大似然方法(我们在第一章的评述中讨论过)。

为了简单起见,考虑根据带有加性噪声的观测

$$y_i = f(x, \alpha) + \xi$$

进行回归估计的问题。为了用最大似然方法估计回归函数,我们必须知道一个参数化函数集 $f(x, \alpha)$, $\alpha \in \Lambda \subset R^n$,其中包含回归函数 $f(x, \alpha)$,而且我们必须知道噪声的模型 $P(\xi)$。

在贝叶斯方法中,我们还须拥有额外的信息:必须知道先验密度函数 $P(\alpha)$,它定义了参数化函数集 $f(x, \alpha)$, $\alpha \in \Lambda$ 中的任意函数是回归函数的概率。如果 $f(x, \alpha)$ 是回归函数,那么训练数据

$$[Y, X] = (y_1, x_1), \cdots, (y_l, x_l)$$

的概率等于

$$P([Y, X]|\alpha) = \prod_{i=1}^{l} P(y_i - f(x_i, \alpha)).$$

在得到了这些数据后,我们可以后验地估计以参数 α 确定了回归函数的概率:

$$P(\alpha|[Y, X]) = \frac{P([Y, X]|\alpha)P(\alpha)}{P([Y, X])}. \qquad (4\text{-}37)$$

我们可以用这个表达式来选择对回归函数的逼近。

我们考虑最简单的办法: 选择逼近函数 $f(x, \alpha^*)$, 使所得的这个条件概率最大[①]。寻找使得这个概率最大的 α^* 等价于最大化下面的泛函:

$$\Phi(\alpha) = \sum_{i=1}^{l} \ln P(y_i - f(x_i, \alpha)) + \ln P(\alpha). \tag{4-38}$$

$$P(\xi) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{\xi}{2\sigma^2}\right).$$

于是从(4-38)式[②]我们得到下面的泛函:

$$\Phi^*(\alpha) = \frac{1}{l}\sum_{i=1}^{l}(y_i - f(x_i, \alpha))^2 - \frac{2\sigma^2}{l}\ln P(\alpha), \tag{4-39}$$

为找到逼近函数, 我们必须使它关于 α 最小化。这个泛函的第一项就是经验风险的值, 而第二项可以解释为一个带有显式正则参数的正则化项。

因此, 贝叶斯方法把我们引到了与 SRM 或 MDL 推理中相同的方案。

然而, 本节评述的目的却是描述贝叶斯方法与 SRM 或 MDL 之间的一个不同点。

4.11.2 贝叶斯方法与容量控制方法的讨论

贝叶斯方法唯一的(但却是重要的)缺点是, 它局限于学习机器的函数集与机器所要解决的问题集重合的情况。严格来说, 它不能应用于容许问题集与学习机器的容许函数集不同的情况。例如, 如果回归函数不是多项式, 则贝叶斯方法不能被应用到用多项式逼近回归函数的问题中, 因为容许的多项式集合中的任何函数是回归函数的先验概率 $P(\alpha)$ 都等于零。因此, 对学习机器的任何函数后验概率(4-37)式都是零。要应用贝叶斯方法, 我们必须有很强的先验信息:

(1) 学习机器的给定函数集与要解决的问题集重合。

(2) 问题集上的先验分布是由给定的表达式 $P(\alpha)$ 描述的[③]。

与贝叶斯方法相对比, SRM 或 MDL 容量控制(复杂性控制)方法只使用很弱的(定性的)关于实际问题的先验信息: 它们使用容许函数集的一个结构(函数集按照关于其函数的有用性的某种思想进行排序); 这种先验信息不包括任何对实际问题的定量描述。因此, 利用这些方法, 我们可以逼近一个与学习机器的容许函数集不同的函数集。

因此, 归纳推理的贝叶斯方法是基于现实的强(定量)先验信息(以及训练数据)的, 而

---

① 另外一种估计子是在后验概率

$$_0(x\mathbb{C}[Y, X]) = \int f(x, \alpha)P(\alpha\mathbb{C}[Y, X])\,d\alpha$$

的基础上构造的, 具有如下值得注意的特性: 它使得与容许回归函数的均方偏差

$$R(\ ) = \int (f(x, \alpha) - (x\mathbb{C}[Y, X]))^2 P([Y, X]\mathbb{C}\alpha)P(\alpha)\,dx\,d([Y, X])\,d\alpha$$

最小。要以显式找到这个估计子, 我们必须解析地计算出这个积分(由于 α 的维数很高, 因此数值地计算这个积分是不可能的)。不幸的是, 这个积分的解析计算基本上还是一个没有解决的问题。

② 原文中误写作(4-37)式。——译者

③ 这一部分先验信息不如第一部分重要。人们可以证明, 随着观测数目的增加, 对 $P(\alpha)$ 的不精确描述所带来的影响将会减小。

SRM 或 MDL 方法中的归纳推理则是基于现实的弱(定性)先验信息(以及训练数据)的,但利用了容量(复杂性)控制。

一些贝叶斯主义的主张者在待解决的问题集与学习机器的容许函数集不符合的情况下仍利用这一方法,在与他们的讨论中,人们可以听到这样的说法:

贝叶斯方法在一般情况下也奏效。

的确,在一般情况下(当机器所实现的函数不一定与要逼近的函数重合时)贝叶斯方法有时也奏效,对这一事实有下面的解释。贝叶斯推理具有外在形式的容量控制。它包括两个阶段:一个非形式阶段,在这一阶段,我们对所面对的问题选择一个函数来描述(定量的)先验信息 $P(\theta)$;另一个是形式阶段,我们通过最小化泛函(4-38)式来寻找解。通过选择分布 $P(\theta)$就控制了容量。

因此,在一般情况下,贝叶斯体系实现的是一个解决所面对的问题的人-机过程,其中的容量控制是通过人为地选择正则化因素 $\ln P(\theta)$ 来完成的。

与贝叶斯推理相比,SRM 和 MDL 推理是解决问题的纯机器方法。对任意的 l,它们采用同样的容许函数集结构和同样的形式化机制进行容量控制。

第五章

模式识别的方法

要在学习算法中执行 SRM 归纳原则,我们必须在一个给定的函数集中使风险最小化,这要通过控制两个因素来完成:经验风险的值和置信范围的值。

发展这样的方法就是我们的理论在构造学习算法方面的目标。

在本章中,我们将描述对模式识别的学习算法并考虑它们对回归估计问题的推广[①]。

5.1 为什么学习机器能够推广?

学习机器的推广能力是基于两个因素的,在控制学习过程推广能力的理论中对这两个因素进行了介绍。根据这一理论,要保证学习过程高的推广能力,我们须在损失函数集 $S=\{Q(z,\alpha),\alpha\in\Lambda\}$ 上构造一个结构

$$S_1 \quad S_2 \quad \cdots \quad S_n$$

然后选择这一结构的一个适当的元素 S_k 和这个元素中的一个函数 $Q(z,\alpha_l^k)\in S_k$,使得相应的界最小,比如使界(4-1)式最小。界(4-1)式可以重写为下面的简单形式:

$$R(\alpha_l^k) \le R_{emp}(\alpha_l^k) + \Phi\left(\frac{l}{h_k}\right), \tag{5-1}$$

其中第一项是经验风险,第二项是置信范围。

有两种最小化不等式(5-1)右边的构造性方法。

在第一种方法中,在设计学习机器时确定一个 VC 维为某个 h^* 的容许函数集。对于给定数量 l 的训练数据,h^* 的值确定了机器的置信范围 $\Phi\left(\frac{l}{h^*}\right)$。因此选择一个适当的结构元素就是一个对特定数目的数据设计学习机器的问题。

在学习过程中这个机器最小化界(5-1)式中的第一项(在训练集上的错误数)。

---

① 在本书第一版中,本章在介绍了对模式识别的学习算法之后讨论了它们对回归估计问题的推广,但在第二版中,有关回顾估计的问题已经改在专门的第六章中讨论。——译者

如果对一个给定数目的训练数据,我们设计了一个过于复杂的机器,置信范围 $\Phi\frac{l}{h^*}$ 将会很大。这时即使我们可以把经验风险最小化为零,在测试集上的错误数目仍可能很大。这种现象叫做过学习或过适应。

为避免过学习(即为了得到小的置信范围),我们必须构造 VC 维小的学习机器。另一方面,如果函数集的 VC 维小,那么就难以逼近训练数据(难以得到不等式(5-1)第一项的小值)。要在得到小的逼近误差的同时保持小的置信范围,我们必须适当选择机器的构造,使之反映关于所面对问题的先验知识。

于是,要用这类机器解决面对的问题,我们首先必须找到学习机器的适当构造(这是在过学习和不良逼近之间折衷的结果),然后在这个机器中寻找使在训练数据上错误数最小的函数。这种最小化不等式(5-1)右边的方法可以描述如下:

保持置信范围固定(通过选择一个适当构造的机器)并最小化经验风险。

最小化不等式(5-1)右边的第二种方法可以描述为:

保持经验风险值固定(比如等于零)并最小化置信范围。

下面我们将讨论实现这两种方法的两类学习机器:

(1) 神经网络(实现第一种方法),

(2) 支持向量机(实现第二种方法)。

这两类学习机器都是 60 年代构造的采用线性指示函数集的学习机器的推广。

5.2 指示函数的 sigmoid 逼近

考虑在线性指示函数集合

$$f(x,w) = \mathrm{sgn}\{(w\cdot x)\},\ w\in R^n \tag{5-2}$$

上最小化经验风险的问题,其中$(w\cdot x)$表示向量 w 和 x 之间的内积。

令

$$(x_1,y_1),\cdots,(x_l,y_l)$$

是一个训练集,其中 x_j 是一个向量,$y_j\in\{1,-1\}$,$j=1,\cdots,l$。

目标是找到使得经验风险泛函

$$R_{emp}(w) = \frac{1}{l}\sum_{j=1}^{l}(y_j - f(x_j,w))^2 \tag{5-3}$$

最小的参数向量 w_0(权值)。

如果训练集可以被没有错误地分开(即经验风险可以为零),那么存在一个有限步骤的过程,比如 Rosenblatt 为感知器提出的过程(参见引论),使我们找到这样一个向量 w_0。

当训练集不能被没有错误地分开时就出现了问题。在这种情况下,以最少的错误数目划分训练集的问题是 NP 完全的,而且我们不能用常规的基于梯度的算法来找到泛函(5-3)式的一个局部极小点,因为对这一泛函来说,梯度要么等于零,要么不存在。

因此,人们提出了用所谓 sigmoid 函数(见图 0.3)

$$\bar{f}(x, w) = S\{(w \cdot x)\} \tag{5-4}$$

来逼近指示函数(5-2)式,其中 $S(u)$ 是一个平滑单调函数,满足

$$S(-\infty) = -1, \quad S(+\infty) = 1,$$

比如,

$$S(u) = \tanh u = \frac{\exp(u) - \exp(-u)}{\exp(u) + \exp(-u)}.$$

对 sigmoid 函数集,经验风险泛函

$$R_{emp}(w) = \frac{1}{l}\sum_{j=1}^{l}(y_j - S\{(w \cdot x_j)\})^2$$

对 w 是平滑的,它有梯度

$$\mathrm{grad}_w R_{emp}(w) = -\frac{2}{l}\sum_{j=1}^{l}[y_j - S\{(w \cdot x_j)\}]S'\{(w \cdot x_j)\}x_j^T,$$

因此可以用标准的基于梯度的方法来进行最小化,比如用梯度下降方法:

$$w_{new} = w_{old} - \gamma(\cdot)\mathrm{grad}R_{emp}(w_{old}),$$

其中 $\gamma(\cdot) = \gamma(n) \geq 0$ 是一个依赖于迭代次数 n 的值,w_{old} 是本次迭代之前的权值,w_{new} 是更新后的权值。梯度下降方法收敛于局部极小点的充分条件是梯度值是有界的,且系数 $\gamma(n)$ 满足以下条件:

$$\sum_{n=1}^{\infty}\gamma(n) = \infty, \quad \sum_{n=1}^{\infty}\gamma^2(n) < \infty.$$

于是,解决问题的思想就是在参数估计阶段用 sigmoid 函数逼近,而在识别阶段则对最后一个神经元用阈值函数(采用前一阶段所得到的参数)。

5.3 神经网络

在这一节里我们考虑传统的神经网络[①],它执行的是第一种策略:保持置信范围固定而最小化经验风险。

在神经网络中,上述思想被用来估计多层感知器(神经网络)的所有神经元的权值。人们在网络中不是采用线性指示函数(单个的神经元),而是考虑 sigmoid 函数集。

对神经网络的 sigmoid 逼近计算经验风险梯度的方法叫做后向传播方法,它是在 1986 年提出的(Rumelhart, Hilton and Williams, 1986 及 LeCun, 1986)。利用这个梯度,可以在标准的基于梯度算法的基础上迭代地修正神经网的系数(权值)。

5.3.1 后向传播方法

为了描述后向传播方法,我们采用下面的表示法(图 5.1):

---

① 正如我们在第 10 页脚注①中指出的,这里讨论的神经网络实际上指的仅仅是采用 BP 学习算法的前馈型网络(多层感知器),没有涉及其他神经网络模型。——译者

(1) 神经网包含 m+ 1 个层: 第一层 $x(0)$ 描述输入向量 $x=(x^1,\cdots,x^n)$。我们记输入向量为

$$x_i(0)=(x_i^1(0),\cdots,x_i^n(0)),\quad i=1,\cdots l,$$

记输入向量 $x_i(0)$ 在第 k 层上的映射为

$$x_i(k)=(x_i^1(k),\cdots,x_i^{n_k}(k)),\quad i=1,\cdots l,$$

其中, n_k 表示向量 $x_i(k)$, $i=1,\cdots,l$ 的维数(n_k, $k=1,\cdots,m-1$ 可以是任意数目, 但 $n_m=1$)。

(2) 第 k- 1 层与第 k 层通过($n_k\times n_{k-1}$) 矩阵 $w(k)$ 相连接:

$$x_i(k)=S\{w(k)x_i(k-1)\},\quad k=1,\cdots,m,\quad i=1,\cdots,l, \tag{5-5}$$

其中, $S\{w(k)x_i(k-1)\}$ 定义了向量

$$u_i(k)=w(k)x_i(k-1)=(u_i^1(k),\cdots,u_i^{n_k}(k))$$

的 sigmoid 函数, 它由向量的每一维分量的 sigmoid 函数组成:

$$S(u_i(k))=(S(u_i^1(k)),\cdots,S(u_i^{n_k}(k))).$$

我们的目标是在条件(5-5)式下最小化泛函

$$I(w(1),\cdots,w(m))=\sum_{i=1}^{l}(y_i-x_i(m))^2. \tag{5-6}$$

图 5.1　一个神经网络是几层 sigmoid 单元的一种组合,
一层的输出构成了下一层的输入

这一优化问题是通过采用标准的等式约束下的拉格朗日乘子技术来解决的。我们将最小化拉格朗日函数

$$L(W,X,B)=\frac{1}{l}\sum_{i=1}^{l}(y_i-x_i(m))^2-\sum_{i=1}^{l}\sum_{k=1}^{m}(b_i(k)\{x_i(k)-S\{w(k)x_i(k-1)\}\}),$$

其中, $b_i(k)\geq 0$ 是对应约束条件(5-5)式的拉格朗日乘子, 约束(5-5)式描述了向量 $x_i(k-1)$ 与向量 $x_i(k)$ 之间的联系。

我们知道,

$$\Delta L(W,X,B)=0$$

是在约束(5-5)式下性能函数(5-6)式取得局部极小点的必要条件(函数对所有参数 $b_i(k)$, $x_i(k)$, $w(k)$, $i=1,\cdots,l$, $k=1,\cdots,m$ 的梯度为零)。

这个条件可以分成三个子条件:

(i) $\dfrac{L(W,X,B)}{b_i(k)}=0,\ \forall\ i,k,$

(ii) $\dfrac{L(W,X,B)}{x_i(k)}=0,\ \forall\ i,k,$

(iii) $\dfrac{L(W,X,B)}{w(k)}=0,\ \forall\ w(k).$

这些方程的解确定了一个平稳点 (W_0,X_0,B_0), 其中包括了所期望的权值 $W_0=(w^0(1),\cdots,w^0(m))$。下面我们用显式形式重写这 3 个条件:

1. 第一个子条件

第一个子条件给出了下面的一组方程:

$$x_i(k)=S\{w(k)x_i(k-1)\},\quad i=1,\cdots,l,\quad k=1,\cdots,m,$$

它们的初始条件是

$$x_i(0)=x_i,$$

这组方程称作前向动力。

2. 第二个子条件

我们对两种情况考虑第二个子条件: 对 $k=m$(最后一层)的情况和 $k\neq m$(隐层)的情况。

对最后一层, 我们得到

$$b_i(m)=2(y_i-x_i(m)),\quad i=1,\cdots,l.$$

对一般情况(隐层), 我们得到

$$b_i(k)=w^{\mathrm{T}}(k+1)\Delta S\{w(k+1)x_i(k)\}b_i(k+1),$$
$$i=1,\cdots,l,\quad k=1,\cdots,m-1,$$

其中, $\Delta S\{w(k+1)x_i(k)\}$ 是一个 $n_{k+1}\times n_{k+1}$ 对角阵, 其对角元素为 $S'(u_r)$, u_r 是(n_{k+1} 维的)向量 $w(k+1)x_i(k)$ 的第 r 个元素。这些方程描述了后向动力。

3. 第三个子条件

遗憾的是, 第三个子条件并没有给出计算权值矩阵 $w(k)$, $k=1,\cdots,m$ 的直接方法。因此, 为估计这些权值, 人们采用最陡梯度下降方法:

$$w(k) \leftarrow w(k) - \gamma(\eta) \frac{L(W, X, B)}{w(k)}, \quad k = 1, \cdots, m.$$

以显式表示,这一方程是

$$w(k) \leftarrow w(k) - \gamma(\eta) \sum_{i=1}^{l} b_i(k) \Delta S\{w(k) x_i(k-1)\} w(k) x_i^T(k-1),$$

$$k = 1, 2, \cdots, m.$$

此公式描述了权值更新的规则。

5.3.2 后向传播算法

后向传播算法包括三个部分:

(1) 前向传递:

$$x_i(k) = S\{w(k) x_i(k-1)\}, \quad i = 1, \cdots, l, \quad k = 1, \cdots, m,$$

边界条件是:

$$x_i(0) = x_i, \quad i = 1, \cdots, l.$$

(2) 后向传递:

$$b_i(k) = w^T(k+1) \Delta S\{w(k+1) x_i(k)\} b_i(k+1),$$

$$i = 1, \cdots, l, \quad k = 1, \cdots, m-1,$$

边界条件是:

$$b_i(m) = 2(y_i - x_i(m)), \quad i = 1, \cdots, l.$$

(3) 权值更新: 对权值矩阵 $w(k)$, $k = 1, 2, \cdots, m$:

$$w(k) \leftarrow w(k) - \gamma(\eta) \sum_{i=1}^{l} b_i(k) \Delta S\{w(k) x_i(k-1)\} w(k) x_i^T(k-1).$$

应用后向传播技术,我们可以求得经验风险泛函的一个局部极小点。

5.3.3 用于回归估计问题的神经网络

要使神经网络适合解决回归估计问题,只要在最后一层中使用线性函数来代替 sigmoid 函数就足够了。这意味着对上面描述的算法只需做如下的修改:

$$x_i(m) = w(m) x_i(m-1),$$

$$\Delta S\{w(m) x_i(m-1)\} = 1, \quad i = 1, \cdots, l.$$

5.3.4 关于后向传播方法的讨论

这种神经网络方法的主要问题是:

(1) 经验风险泛函可能有很多个局部极小点,标准的优化过程只保证收敛到其中的一个。所得解的质量取决于很多因素,尤其是与权值矩阵 $w(k)$, $k = 1, 2, \cdots, m$ 的初始值有关。

要得到一个“小的”局部极小点,需对初始参数做适当选择,而这种选择是基于启发式的。

(2) 基于梯度算法的收敛是比较慢的。有一些加快收敛速度的启发式方法。

(3) Sigmoid 函数有一个尺度因子,它将影响逼近的质量。这个尺度因子的选择是在逼近质量和收敛速度之间的一种折衷。有一些关于选择尺度因子的经验性建议。

因此,神经网络并不是有良好控制的学习机器。尽管如此,在很多实际应用中,神经网络显示出了很好的结果。

5.4 最优分类超平面

下面我们讨论一种新的通用学习机器,它执行的是第二种策略,即保持经验风险值固定而最小化置信范围。

像在神经网络中一样,我们从考虑线性决策规则(分类超平面)开始。但是,与前面的考虑不同,我们采用一种特殊类型的超平面,即所谓最优分类超平面(Vapnik and Chervonenkis, 1974 及 Vapnik, 1979)。首先我们考虑训练数据是线性可分情况下的最优分类超平面,然后,在 5.5.1 小节中我们把最优分类超平面的思想推广到不可分数据的情况。利用构造最优超平面的技术,我们将描述一种新型的通用学习机器——支持向量机。最后,我们构造用于解决回归估计问题的支持向量机。

5.4.1 最优超平面

假定训练数据

$$(x_1, y_1), \cdots, (x_l, y_l), \quad x \in R^n, \quad y \in \{+1, -1\}$$

可以被一个超平面

$$(w \cdot x) - b = 0 \tag{5-7}$$

分开。如果这个向量集合被超平面没有错误地分开,并且离超平面最近的向量与超平面之间的距离是最大的,则我们说这个向量集合被这个最优超平面(或最大间隔超平面)分开(图 5.2)。

为了描述分类超平面,我们使用下面的形式:

$$(w \cdot x_i) - b \geq 1, \quad \text{若 } y_i = 1.$$

$$(w \cdot x_i) - b \leq -1, \quad \text{若 } y_i = -1.$$

在后面,我们采用这些不等式的一种紧凑的形式:

$$y_i[(w \cdot x_i) - b] \geq 1, \quad i = 1, \cdots, l. \tag{5-8}$$

容易验证,最优超平面就是满足条件(5-8)式并且使得

$$\Phi(w) = \|w\|^2 \tag{5-9}$$

最小化的超平面。(最小化是关于向量 w 和标量 b 进行的。)

图 5.2 最优分类超平面是以最大间隔将数据分开的超平面

5.4.2 Δ-间隔分类超平面

一个超平面[①]

$$(w^* \cdot x) - b = 0, \quad \| w^* \| = 1$$

如果它以如下的形式将向量 x 分类:

$$y = \begin{cases} 1 & \text{若}(w^* \cdot x) - b \geq \Delta \\ -1 & \text{若}(w^* \cdot x) - b \leq -\Delta' \end{cases}$$

则我们称之为 Δ-间隔分类超平面。

容易验证,用正规形式(5-8)式定义的最优超平面是 Δ-间隔分类超平面,其 $\Delta = 1/\| w \|$。有下面的定理成立:

定理 5.1 设向量 $x \in X$ 属于一个半径为 R 的球中,那么 Δ-间隔分类超平面集合的 VC 维 h 以下面的不等式为界:

$$h \leq \min\left[\frac{R^2}{\Delta^2}, n\right] + 1.$$

在 3.5 节中我们指出了分类超平面集合的 VC 维等于 n+1,其中 n 是空间的维数。但是 Δ-间隔分类超平面的 VC 维可以更小[②]。

推论 以概率 1-η我们可以断定,测试样本不能被 Δ-间隔超平面正确分类的概率有下面的界:

$$P_{error} \leq \frac{m}{l} + \frac{E}{2}\left(1 + \sqrt{1 + \frac{4m}{lE}}\right),$$

其中,

$$E = 4\,\frac{h\left(\ln\frac{2l}{h} + 1\right) - \ln\frac{\eta}{4}}{l},$$

m 是没有被 Δ-间隔超平面正确分类的训练样本数目,h 是定理 5.1 中给出的 VC 维的界。

在这一定理的基础上,我们可以构造 SRM 方法,其中为了得到好的推广性,我们选择适当的 Δ 值。

5.5 构造最优超平面

要构造最优超平面,我们必须用系数的模最小的超平面把属于两个不同类 $y \in \{-1, 1\}$ 的样本集

$$(y_1, x_1), \cdots, (y_l, x_l)$$

中的向量 x_i 分开。

---

① 原著中,向量的模有时用 ‖·‖ 表示,有时用 ©¦ ©表示,翻译时统一改用 ‖·‖ 表示,不再一一注明。——译者

② 在 5.7 节中我们将讨论在 10^{13} 维空间中的一个分类超平面,它具有相对较小的估计 VC 维($\approx 10^3$)。

要找到这个超平面,我们需要求解下面的二次规划问题:最小化泛函

$$\Phi(w) = \frac{1}{2}(w \cdot w), \tag{5-10}$$

约束条件为不等式类型:

$$y_i[(x_i \cdot w) - b] \geq 1, \quad i = 1, 2, \cdots, l. \tag{5-11}$$

这个优化问题的解是由下面的拉格朗日泛函(拉格朗日函数)的鞍点给出的:

$$L(w, b, \alpha) = \frac{1}{2}(w \cdot w) - \sum_{i=1}^{l} \alpha\{[(x_i \cdot w) - b] y_i - 1\}, \tag{5-12}$$

其中,α 为拉格朗日乘子。我们需要把对拉格朗日函数关于 w, b 求其最小值和关于 $\alpha \geq 0$ 求其最大值。

在鞍点上,解 w_0, b_0 和 α^0 必须满足以下条件:

$$\frac{L(w_0, b_0, \alpha^0)}{b} = 0,$$

$$\frac{L(w_0, b_0, \alpha^0)}{w} = 0.$$

以显式重写这些方程,我们得到最优超平面的下列特性:

(1) 对最优超平面,系数 α^0 必须满足约束

$$\sum_{i=1}^{l} \alpha^0 y_i = 0, \quad \alpha^0 \geq 0, \quad i = 1, \cdots, l \tag{5-13}$$

(第一方程)

(2) 最优超平面(向量 w_0)是训练集中的向量的线性组合:

$$w_0 = \sum_{i=1}^{l} y_i \alpha^0 x_i, \quad \alpha^0 \geq 0, \quad i = 1, \cdots, l \tag{5-14}$$

(第二方程)

(3) 进一步,只有所谓的支持向量可以在 w_0 的展开中具有非零的系数 α^0。支持向量就是使得不等式(5-11)中的等式成立的向量。因此我们得到

$$w_0 = \sum_{\text{支持向量}} y_i \alpha^0 x_i, \quad \alpha^0 \geq 0. \tag{5-15}$$

这一点是从传统的 Kühn-Tucker 条件得到的。根据 Kühn-Tucker 条件可知,最优超平面的充分必要条件是分类超平面满足条件:

$$\alpha^0\{[(x_i \cdot w_0) - b_0] y_i - 1\} = 0, \quad i = 1, \cdots, l. \tag{5-16}$$

把 w_0 的表达式代入拉格朗日函数中,并考虑到 Kühn-Tucker 条件,我们得到下面的泛函:

$$W(\alpha) = \sum_{i=1}^{l} \alpha - \frac{1}{2} \sum_{i,j=1}^{l} \alpha\alpha y_i y_j (x_i \cdot x_j). \tag{5-17}$$

问题变为在非负象限

$$\alpha \geq 0, \quad i = 1, \cdots, l \tag{5-18}$$

中最大化这一泛函,并服从约束条件

$$\sum_{i=1}^{l} \alpha y_i = 0. \tag{5-19}$$

根据(5-15)式,拉格朗日乘子和支持向量决定了最优超平面,因此要构造最优超平面,我们需要解决的是一个简单的二次规划问题:在约束条件(5-18)和(5-19)式下最大化(5-17)式的二次型①。

设 $\alpha = (\alpha_1^0, \cdots, \alpha_l^0)$ 为这个二次优化问题的解,那么与最优超平面对应的向量 w_0 的模等于:

$$\| w_0 \|^2 = 2W(\alpha) = \sum_{\text{支持向量}} \alpha_i^0 \alpha_j^0 (x_i \cdot x_j) y_i y_j .$$

基于最优超平面的分类规则就是下面的指示函数:

$$f(x) = \operatorname{sgn} \sum_{\text{支持向量}} y_i \alpha_i^0 (x_i \cdot x) - b_0 , \tag{5-20}$$

其中 x_i 是支持向量,α_i^0 是对应的拉格朗日系数,b_0 是常数(阈值),

$$b_0 = \frac{1}{2}[(w_0 \cdot x^*(1)) + (w_0 \cdot x^*(-1))],$$

其中,我们用 $x^*(1)$ 表示属于第一类的某个(任意一个)支持向量,用 $x^*(-1)$ 表示属于第二类的一个支持向量(Vapnik and Chervonenkis, 1974 及 Vapnik, 1979)。

不可分情况下的推广

为了在数据为线性不可分的情况下构造最优型的超平面,我们引入非负变量 $\xi \geq 0$ 和函数

$$F_\sigma(\xi) = \sum_{i=1}^{l} \xi_i^\sigma,$$

其中参数 $\sigma > 0$。

我们在约束条件

$$y_i((w \cdot x_i) - b) \geq 1 - \xi_i, \quad i = 1, 2, \cdots, l \tag{5-21}$$

和另一个条件

$$(w \cdot w) \leq \Delta^{-2} \tag{5-22}$$

下最小化泛函 $F_\sigma(\xi)$。

对足够小的 $\sigma > 0$,这个优化问题的解定义了这样一个超平面,它的参数属于(5-22)式定义的子集(即属于由常数 $c_n = \Delta^{-2}$ 决定的结构

$$S_n = \{(w \cdot x) - b : (w \cdot w) \leq \Delta^{-2}\}$$

的元素),在这种情况下这个超平面使得训练错误数最小。

为了计算上的原因,我们考虑 $\sigma = 1$ 的情况,这是计算上比较简单的最小的 $\sigma > 0$。我们把这个超平面称作 Δ-间隔分类超平面。

---

① 这个二次规划问题是简单的,因为它的约束条件比较简单。关于这个问题的求解,我们可以用文献(More and Toraldo, 1991)中的特殊方法,它比较快而且可以适应支持向量数较多(约 10^4 个支持向量)时的情况。注意,在训练数据中,支持向量只占所有训练向量的一小部分(在我们的实验中为 3%～5%)。

1. 构造 Δ-间隔分类超平面

可以证明(用上面介绍的技术), Δ-间隔超平面是由向量

$$w = \frac{1}{C^*}\sum_{i=1}^{l} \alpha_i y_i x_i$$

决定的, 其中参数 α_i, $i = 1, \cdots, l$ 和 C^* 是下面的凸优化问题的解:

最大化泛函

$$W(\alpha, C^*) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2C^*}\sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \frac{C^*}{2\Delta^2},$$

约束条件为

$$\sum_{i=1}^{l} y_i \alpha_i = 0, \quad C^* \geq 0,$$

$$0 \leq \alpha_i \leq 1, \quad i = 1, \cdots, l.$$

2. 构造软间隔分类超平面

为了简化计算, 我们可以引入以下的(略加修改的)软间隔最优超平面的概念(Cortes and Vapnik, 1995)。软间隔超平面(也称为广义最优超平面)是由在约束条件(5-21)式下使泛函

$$\Phi(w, \xi) = \frac{1}{2}(w \cdot w) + C\sum_{i=1}^{l} \xi_i$$

最小化的向量 w 决定的(这里的 C 是一个给定的值)。

求解这个二次优化问题的技术与在可分情况下采用的技术几乎相同: 要找到广义最优超平面的系数

$$w = \sum_{i=1}^{l} \alpha_i y_i x_i,$$

我们必须找到一系列参数 α_i, $i = 1, \cdots, l$, 使得下列与可分情况下相同的二次型最大:

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j),$$

只是约束条件略有不同:

$$0 \leq \alpha_i \leq C, \quad i = 1, \cdots, l,$$

$$\sum_{i=1}^{l} y_i \alpha_i = 0.$$

就像在可分情况下一样, 这里也只有部分系数 α_i, $i = 1, \cdots, l$ 不为零, 它们确定了支持向量。

注意, 如果在泛函 $\Phi(w, \xi)$ 中的系数 C 等于使泛函 $F_1(\xi)$ 最小化的最优参数值 C^*:

$$C = C^*,$$

则对上述两个优化问题(分别由泛函 $F_1(\xi)$ 和泛函 $\Phi(w, \xi)$ 定义)的解是重合的。

5.6 支持向量机

支持向量(SV)机[①]实现的是如下的思想:它通过某种事先选择的非线性映射将输入向量 x 映射到一个高维特征空间 Z,在这个空间中构造最优分类超平面(图 5.3)。

例 要构造一个与二阶多项式对应的决策面,我们可以构造一个特征空间 Z,它有如下的 $N=\frac{n(n+3)}{2}$ 个坐标:

$z^1=x^1,\cdots,z^n=x^n$, n 个坐标,

$z^{n+1}=(x^1)^2,\cdots,z^{2n}=(x^n)^2$, n 个坐标,

$z^{2n+1}=x^1x^2,\cdots,z^N=x^nx^{n-1}$, $\frac{n(n-1)}{2}$ 个坐标,

其中,$x=(x^1,\cdots,x^n)$。在这个空间中构造的分类超平面就是在输入空间中的二阶多项式。要在 n 维空间中构造阶数 d 的 n 的多项式,我们需要多于约 $(n/d)^d$ 个特征。

图 5.3 支持向量机把输入空间映射到一个高维特征空间,然后在这个特征空间中构造最优超平面

在上面的方法中有两个问题:一个是概念上的,另一个是技术上的。

(1) 怎样找到一个推广性好的分类超平面?(概念上的问题。)

特征空间的维数将会很高,将训练数据分开的一个超平面不一定能够很好地推广[②]。

(2) 怎样在计算上处理如此高维的空间?(技术上的问题。)

要在一个 200 维空间中构造一个 4 或 5 阶的多项式,需要构造一个上十亿维的特征空间。如何克服这种“维数灾难”?

5.6.1 高维空间中的推广

上述的概念性问题可以通过构造 Δ-间隔分类超平面和软间隔分类超平面来解决。

根据定理 5.1,Δ 值大的 Δ-间隔分类超平面集合的 VC 维是小的,因此根据定理 5.1 的推论,所构造的超平面将有较高的推广能力。

对最大间隔超平面,有下面的定理:

定理 5.2 如果包含 l 个样本的训练集被最大间隔超平面分开,那么测试错误概率的数学期望(对训练集)是以下面 3 个值中最小者的期望为界的,这 3 个值是:比值 m/l,m 为支持向量的个数;比值 $[R^2\|w\|^2]/l$,其中 R 是将数据包含其中的超球的半径,$\|w\|^{-2}$ 是间隔值;以及比值 n/l,n 是输入空间的维数,即

① 本书中把支持向量机(support vector machine)简写为 SV 机,而现在文献中更常用的是把支持向量机简写为 SVM。——译者

② 回顾在传统的判别分析中,Fisher 对用少量的数据构造二次判别函数的担心(第 1.9 节)。

$$EP_{error} \leq E \min \left\{ \frac{m}{l}, \frac{[R^2 \| w \|^2]}{l}, \frac{n}{l} \right\}, \tag{5-23}$$

其中 P_{error} 是测试错误概率。

式(5-23)中:

(1) 因为数据压缩的期望是大的①。

(2) 因为分类间隔的期望是大的。

(3) 因为输入空间是小的。

传统方法忽略了推广性的前两个原因而只依赖第三个原因。在支持向量机中,我们忽略维数因素而依赖前两个因素。

5.6.2 内积的回旋②

然而,即使最优超平面有好的推广性并且理论上可以被找到,仍然存在如何处理高维特征空间的技术问题。

在 1992 年我们发现(Boser, Guyon and Vapnik, 1992),为了在特征空间 Z 中构造最优分类超平面,并不需要以显式形式来考虑特征空间,而只需要能够计算支持向量与特征空间中向量的内积((5-17)式和(5-20)式)。

考虑在 Hilbert 空间中内积的一个一般表达③

$$(z_i \cdot z) = K(x, x_i),$$

其中 z 是输入空间中的向量 x 在特征空间中的像。

根据 Hilbert-Schmidt 理论,$K(x, x_i)$ 可以是满足下面一般条件的任意对称函数(Courant and Hilbert, 1953)。

定理 5.3(Mercer) 要保证 L_2 下的对称函数 $K(u, v)$ 能以正的系数 $a_k > 0$ 展开成

$$K(u, v) = \sum_{k=1}^{\infty} a_k \psi(u) \psi(v) \tag{5-24}$$

(即 $K(u, v)$ 描述了在某个特征空间中的一个内积),充分必要条件是,对使得

$$\int g^2(u)du < \infty$$

的所有 $g \neq 0$,条件

---

① 我们可以将这个定理的结果与对下述压缩方案的分析结果进行比较。要构造最优分类超平面,我们只需在训练数据中指出支持向量和它们的分类。这样需要的描述长度是:约$\lceil \log_2 m \rceil$比特用于表示支持向量的数目 m,$\lceil \log_2 C_l^m \rceil$比特用于指定支持向量,还有$\lceil \log_2 C_m^{m_1} \rceil$比特用来在支持向量中确定 m_1 个(译注:原文中误写为 m_i)第一类的代表。因此,对 $m \ll l$ 和 $m_1 \approx m/2$,压缩系数是

$$K \approx \frac{m \log_2 \frac{l}{m} + 1}{l}.$$

根据定理 4.3,对一般压缩方案来说错误概率是与 K 成比例的。从定理 5.2 知 $EP_{error} \leq Em/l$。

因此,即使(5-23)式中的随机值 m 总是其中最小的,对 SV 机得到的界也比对一般压缩方案得到的界好得多。

② 原文为 convolution of the inner product,指把变换空间中的内积转化为原空间中的某个函数进行计算,从而避免直接在变换空间运算。作者把这个在原空间中用来计算变换空间中内积的函数称作内积的 convolution,我们译作内积的回旋。在更多的文献和本书后面的讨论中,这种函数常被称作核函数。——译者

③ 这一思想由 Aizerman、Braverman 和 Rozonoer(1964, 1965)用在对位势函数方法的收敛特性的分析中。这与提出最优超平面方法在同一时间(1965 年)(Vapnik and Chervonenkis, 1965)。然而,将这两种思想结合起来,从而引出 SV 机,却是 1992 年的事情。

$$\int K(u,v)g(u)g(v)dudv > 0$$

成立。

5.6.3 构造 SV 机

内积的回旋使得我们可以构造在输入空间中非线性的决策函数

$$f(x) = \operatorname{sgn}\left(\sum_{\text{支持向量}} y_i \alpha_i K(x_i, x) - b\right), \tag{5-25}$$

它们等价于在高维特征空间 $\psi_1(x), \cdots, \psi_N(x)$ 中的线性决策函数($K(u,v)$ 是这个特征空间中内积的一种回旋)。

在可分情况下(不可分情况类似),要求得系数 α_i,只要寻找泛函

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{5-26}$$

的最大值,约束条件为

$$\sum_{i=1}^{l} \alpha_i y_i = 0, \quad \alpha_i \geq 0, \quad i = 1, 2, \cdots, l. \tag{5-27}$$

除了内积的形式外,这个泛函与寻找最优超平面的泛函相同:在(5-17)式中是内积 $(x_i \cdot x_j)$,而这里我们使用内积的回旋 $K(x_i, x_j)$。

构造(5-25)式类型的决策函数的学习机器叫做支持向量机(SV 机)。(我们用这个名字来强调在支持向量上展开解的思想。在 SV 机中,构造的复杂程度取决于支持向量的数目,而不是特征空间的维数。)图 5.4 是 SV 机的图解。

图 5.4 两层的 SV 机是在高维特征空间 Z 中
最优超平面的一个紧凑的实现[①]

① 原著中将 $K(x_i, x)$ 误写为 $F(x_i, x)$。——译者

5.6.4 SV 机的例子

采用不同的函数作为内积的回旋 $K(x, x_i)$, 我们可以构造实现输入空间中不同类型的非线性决策面的学习机器。下面我们考虑 3 种类型的学习机器:

(1) 多项式学习机器;

(2) 径向基函数机器;

(3) 两层神经网络。

为了简单起见, 这里我们考虑训练向量被没有错误地分开的情况。

注意到, 支持向量机执行了 SRM 原则。这可以用下面的讨论说明。设

$$\Phi(x) = (\psi(x), \cdots, \psi_N(x))$$

是一个特征空间, $w = (w_1, \cdots, w_N)$ 是确定了这个空间中的一个超平面的权值向量。考虑超平面集合的一种结构, 其元素 S_k 包含的函数满足

$$R^2 \| w \|^2 \leq k$$

其中 R 是包含向量 $\Phi(x)$ 的最小超球的半径, $\| w \|$ 是权值的模(我们采用特征空间中对于向量 $z = \Phi(x_i)$ 的正规超平面, x_i 是训练数据中的元素)。

根据定理 5.1(现在应用于特征空间), k 给出了对函数集 S_k 的 VC 维的一种估计。

SV 机把训练数据没有错误地分开:

$$y_i[(\Phi(x_i) \cdot w) - b] \geq 1, \quad y_i = \{+1, -1\}, \quad i = 1, 2, \cdots, l$$

并且有最小模 $\| w \|$。

换句话说就是, SV 机用估计 VC 维最小的元素 S_k 中的函数把训练数据分开。

回顾在特征空间中, 有等式

$$\| w_0 \|^2 = \sum_{i=1}^{l} \alpha_i^0 \alpha_j^0 K(x_i, x_j) y_i y_j = \sum_{i=1}^{l} \alpha_i^0 \tag{5-28}$$

成立。为了控制机器的推广能力(使测试错误概率最小), 我们必须构造使得泛函

$$\Phi(R, w_0, l) = \frac{R^2 \| w_0 \|^2}{l} \tag{5-29}$$

最小的分类超平面。把数据没有错误地分开的超平面以概率 $1-\eta$ 有如下的测试错误的界:

$$E = 4 \frac{h \ln \frac{2l}{h} + 1 - \ln \frac{\eta}{4}}{l},$$

其中 h 是超平面集合的 VC 维。我们用 $h_{est} = R^2 \| w_0 \|^2$ 来近似最大间隔超平面的 VC 维 h。要估计这一泛函, 只要估计 $\| w_0 \|^2$(比如用(5-28)式)并通过寻找

$$R^2 = R^2(K) = \min_a \max_{x_i} [K(x_i, x_j) + K(a, a) - 2K(x_i, a)] \tag{5-30}$$

估计 R^2 即可[①]。

---

① 在(5-30)式中, a 为包含所有样本的最小超球的中心, 这在本书第一版第一次印刷时曾在 5.4.2 小节中明确说明过, 后来再版时修改了那部分内容而可能忘记了在此处对 a 进行说明。——译者

1. 多项式学习机器

要构造 d 阶多项式决策规则,我们可以用下面的函数作为内积的回旋:

$$K(x, x_i) = [(x \cdot x_i) + 1]^d. \tag{5-31}$$

这个对称函数满足定理 5.3 的条件,因此它描述了特征空间中内积的一种回旋,其中包含所有直到 d 阶的 $x_i \cdot x_j \cdot x_k$ 乘积。用上面讨论的技术,我们可以构造如下形式的决策函数:

$$f(x, \alpha) = \operatorname{sgn}\left\{\sum_{\text{支持向量}} y_i \alpha_i [(x_i \cdot x) + 1]^d - b\right\},$$

它是 n 维输入空间中的 d 阶多项式的一种因子分解。

尽管特征空间的维数很高(n 维输入空间中的 d 阶多项式有 $O(n^d)$ 个自由参数),解决现实问题的多项式子集的估计 VC 维却可以是低的。

正如上面讨论的,要估计结构中从中选择决策函数的元素的 VC 维,我们必须估计包含训练数据的最小超球半径 R 以及在特征空间中权值的模(定理 5.1)。

注意,半径 R = R(d) 和特征空间中权值的模都依赖于多项式的阶数。

为了得到在某个兴趣点 x_0 邻域内的一个局部多项式逼近,我们来考虑硬限邻域函数(4-16)式。根据局部算法的理论,我们在点 x_0 周围选择一个半径为 R_β 的球,训练集中有 l_β 个元素落到这个球内,然后只用这些数据,构造一个决策函数,使得在选定邻域内的错误概率最小。这个问题的解是一个使得泛函

$$\Phi(R_\beta, w_0, l_\beta) = \frac{R_\beta^2 \| w_0 \|^2}{l_\beta} \tag{5-32}$$

最小的半径 R_β(参数 $\| w_0 \|$ 也依赖于所选的半径)。这一泛函描述在所选的半径 R_β、模 $\| w_0 \|$ 的最小值和落入半径 R_β 内的训练向量数目 l_β 之间的一种折衷。

2. 径向基函数机器

传统的径向基函数(RBF)机器采用下面的决策规则集合:

$$f(x) = \operatorname{sgn}\left\{\sum_{i=1}^{N} a_i K_\gamma(\| x - x_i \|) - b\right\}, \tag{5-33}$$

其中 $K_\gamma(\| x - x_i \|)$ 依赖于两个向量之间的距离 $\| x - x_i \|$。关于 RBF 机器的理论,读者可以参考文献(Micchelli, 1986)和(Powell, 1992)。

对任意固定 γ,函数 $K_\gamma(\| x - x_i \|)$ 是一个非负的单调函数;当 z 趋向于无穷大时它趋向于零[①]。这种类型的函数中最常用的是

$$K_\gamma(\| x - x_i \|) = \exp\{-\gamma \| x - x_i \|^2\}. \tag{5-34}$$

要构造(5-33)式的决策规则,我们需要估计:

(1) 参数 γ 的值;

(2) 中心 x_i 的数目 N;

① 这里应当是指 $K_\gamma(z)$ 是一个单调函数,当 z 趋向于无穷大时它趋向于零,即 $z = \| x - x_i \|$。——译者

(3) 描述各中心的向量 x_i;

(4) 参数 a_i 的值。

在传统的 RBF 方法中,前三步(决定参数 γ、N 和中心向量 x_i, $i = 1, \cdots, N$)是基于启发式知识的,只有第四步(在找到上述参数之后)是通过最小化经验风险泛函确定的。

可以把径向基函数选作 SV 机中内积的回旋函数。在这种情况下,SV 机将构造(5-33)式集合中的一个函数。可以证明(Aizerman, Braverman and Rozonoer, 1964, 1965),径向基函数(5-34)式满足定理 5.3 的条件。

与传统的 RBF 方法对比,在 SV 技术中,所有四种类型的参数都是通过由控制泛函(5-29)式中的参数 R, w_0 来最小化测试错误概率的界确定的。通过最小化泛函(5-29)式,可以确定

(1) N,支持向量的数目;

(2) x_i,支持向量(的原像);

(3) $a_i = \alpha y_i$,展开式的系数;

(4) γ,核函数的宽度参数[①]。

3. 两层神经网络

最后,我们可以通过选择核函数

$$K(x, x_i) = S[v(x \cdot x_i) + c]$$

定义两层的神经网络,其中 $S(u)$ 是 sigmoid 函数。与多项式机器和径向基函数机器中的核不同,它们总是满足 Mercer 条件的,而 sigmoid 核 $\tanh(vu + c)$, $\forall v \leq 1$ 只是对参数 v, c 的某些值满足 Mercer 条件。对参数的这些取值,我们可以构造实现下述规则的 SV 机:

$$f(x, \alpha) = \operatorname{sgn}\left\{\sum_{i=1}^{N} \alpha_i S(v(x \cdot x_i) + c) + b\right\}.$$

利用上面介绍的技术,可以自动地找到下列内容:

(1) 两层机器的构造,即确定隐层单元的数目 N(支持向量的数目);

(2) 第一层(隐层)神经元(支持向量)的权值向量 $w_i = x_i$;

(3) 第二层的权值向量(α值)。

5.7 SV 机的实验

下面,我们将给出两种类型的构造模式识别问题中决策规则的实验[②]:

(1) 平面上人工数据的实验,其数据是可视的;

(2) 实际数据的实验。

---

① 实际上,译者认为,支持向量机方法目前并没有给出易实现的选择内积回旋函数(核函数)中某些参数的一般方法,比如径向基函数的宽度、多项式核函数的阶数等,这是一个值得进一步研究的问题。——译者

② 这些实验是在 AT&T 贝尔实验室的自适应系统研究部进行的。

5.7.1 平面上的实验

为了示范 SV 技术,我们首先给出一个人工的例子(图 5.5)。

图 5.5 SV 机对人工数据分差的例子

图中两类向量分别用黑点和白圈代表,决策边界是用一个 d = 2 的多项式型内积构造的。在图中,样本不能被没有错误地分开;我们用叉号标出了错分的样本,用套在样本点上的圆圈标出样本中的支持向量。

注意,在两个例子中,支持向量的数目相对训练数据来说都是很少的,而且对二阶的多项式来说训练错误的数目是最少的。

5.7.2 手写数字识别

自从 Rosenblatt 的第一次实验后,人们对学习识别手写数字的兴趣一直很浓厚。下面我们将介绍利用不同的 SV 机学习识别手写数字的实验结果。我们还将与其他分类器得到的结果进行比较。在这些实验中,采用的是美国邮政服务(U.S. Postal Service)数据库(LeCun et al, 1990)。它包括7 300个训练模式和2 000个测试模式[①],它们是从现实生活中的邮政编码采集的。数据的分辨率是 16× 16 像素,隐层输入空间的维数是 256。图 5.6 给出了数据库中的一些例子。

表 5.1 给出了在美国邮政服务数据库上解决数字识别问题时的人工表现和各种学习机器的表现。

我们采用三种类型的 SV 机来构造决策规则[②]:

(1) 多项式机器,采用回旋函数

$$K(x, x_i) = \frac{(x \cdot x_i)}{256}^{d}, \quad d = 1, \cdots, 7.$$

① 这里训练模式和测试模式指的就是训练样本和测试样本。在模式识别中,一些人把模式一词用来指样本的类别,也有人用来指个体的数据样本,只要稍加注意即可区分(参见边肇祺等编著. 模式识别. 北京:清华大学出版社, 1988)。——译者

② 这些结果是由 C. Burges, C. Cortes 和 B. Scholkopf 取得的。

图 5.6　美国邮政服务数据库中模式的例子(带有标号)

表 5.1　各种分类器解决这一问题的性能表现[①]

分 类 器	粗错误率
人工表现	2.5%
决策树 C4.5	16.2%
最好的两层神经网络	5.9%
五层神经网络(LeNet 1)	5.1%

(2) 径向基函数机器,采用回旋函数[②]:

① 人工识别的表现是由 J. Bromley 和 E. Sackinger 报道的; C4.5 的结果是 C. Cortes 得到的; 两层神经网络的结果是 B. Scholkopf 得到的; 专门为此用途的五层神经网络结构(LeNet 1)的结果是由 Y. LeCun 等人取得的。

② 原文中将 $\| x - x_i \|^2$ 误写为 $(x - x_i)^2$。——译者

$$K(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{256\sigma^2}\right).$$

(3) 两层神经网络机器，采用回旋函数：

$$K(x, x_i) = \tanh\left(\frac{b\|x \cdot x_i\|}{256} - c\right).$$

所有机器都构造 10 个分类器，每一个把 10 个数字中的一类与其他类分开。选择分类器输出值最大分类为 10 类。

实验结果在表 5.2 中给出。对不同类型的 SV 机，表 5.2 给出了机器的最佳参数(第二列)、支持向量的平均数目(对每个分类器的平均)以及机器的性能表现。其中的支持向量数目指的是对每个分类器的平均。

表 5.2　用不同的 SV 机对美国邮政服务数据库进行数字识别的实验结果

SV 分类器类型	分类器参数	支持向量数	粗错误率
多项式	d= 3	274	4.0%
RBF 分类器	σ^2= 0.3	291	4.1%
神经网络	b= 2, c= 1	254	4.2%

注意，对这个问题，所有 3 种类型的 SV 机表现出了大致相同的性能。与在美国邮政服务数据库上构造整个决策规则来解决数字识别问题的任何其他类型的学习机器[①] 相比，这个性能更好。

在这些实验中，我们观察到一个重要的特性：不同类型的 SV 机利用的支持向量集合大致相同。3 种不同的分类器中共同的支持向量的比例超过 80%。

表 5.3 给出了 3 种类型机器的 10 个分类器的支持向量总数，表中 Poly 代表多项式机器，RBF 代表径向基函数机器，NN 代表神经网络机器。表中还给出了 3 种机器中共同的支持向量的数目。

表 5.3　各种 SV 机的支持向量总数和共同的支持向量所占的比例

	Poly	RBF	NN	共同
支持向量总数	1677	1727	1611	1377
共同的支持向量的比例	82%	80%	85%	100%

表 5.4 给出了各种 SV 机之间共同的支持向量的具体比例，其中的符号含义与表 5.3 中相同，表中数值的含义是：对竖直方向上列出的分类器，它们的支持向量中也是水平方向上所列分类器的支持向量的部分所占的比例。

实验中观察到的这一事实，如果对较宽范围的实际问题都成立的话，将是一个十分重要的特性。

① 注意，利用 4.5 节中讨论的局部逼近方法(它不是构造整个决策规则，而是在任意兴趣点上逼近决策规则)，我们可以得到更好的结果：3.3% 的错误率(L. Bottou and V. Vapnik, 1992)。

对这个数据库，最好的结果是错误率 2.7%，是由 P. Simard, Y. LeCun, J. Denker(1993)得到的，并没有采用任何学习方法。他们提出了一种特殊的 7200 个模板的弹性匹配方法，其中采用了一种聪明的距离概念(称作 tangent 距离)，它考虑了对字符的小的变换、旋转、变形等的不变性。

表 5.4　每两种 SV 机中共同的(总)支持向量的百分比　　%

	Ploy	RBF	NN
Poly	100	84	94
RBF	87	100	88
NN	91	82	100

5.7.3　一些重要的细节

在这一小节里,我们给出在用多项式 SV 机解决数字识别问题中的一些重要的细节。

训练数据是线性不可分的。对线性决策规则,训练集中总的错分数目为 340 个(约 5% 的错误率)。对二阶多项式分类器,训练集中的总错分数减少为 4 个。图 5.7 中给出了这 4 个错分的样本(并标出了期望的类别标号)。从三阶多项式开始,训练数据就是可分的了。

图 5.7　二阶多项式的训练错误样本及其标号

表 5.5 描述了用不同阶数的决策多项式进行实验的结果(10 个决策多项式,每一次实验中每个分类器为一个多项式)。表中给出的支持向量数目是对 10 个分类器的均值。

注意到,随着多项式阶数的增高支持向量的数目慢慢地增加。七阶多项式比三阶多项式只多 50% 的支持向量[①]。

然而对应七阶多项式分类器的特征空间的维数却比三阶多项式分类器对应的特征空间维数高 10^{10} 倍。注意,分类器的性能并没有随着空间维数的增高而有大的变化,说明不存在过学习问题。

表 5.5　不同阶次多项式的实验结果

多项式阶数	特征空间维数	支持向量	粗错误率
1	256	282	8.9%
2	≈33000	227	4.7%
3	≈1.1×10^{6}	274	4.0%
4	≈1.1×10^{9}	321	4.2%
5	≈1.1×10^{12}	374	4.3%
6	≈1.1×10^{14}	377	4.5%
7	≈1.1×10^{16}	422	4.5%

为对一个特定的分类器选择最佳的多项式阶数,我们估计了所构造的所有多项式(从 2 阶到 7 阶)的 VC 维(用[R^2A^2]估计)。用这种办法我们对 10 个两类问题找到了 10 个最

① 线性分类器相对较多的支持向量是由于不可分性造成的:数目 282 中既包含了支持向量,也包含了错分的数据。

佳分类器(不同阶次的多项式)。图 5.8 展示了这些估计,图中对所有 10 个两类决策规则,画出了估计 VC 维与多项式阶数的关系。那么问题是:

是否估计 VC 维最小的多项式会给出最好的分类器?

为回答这个问题,我们构造了表 5.6,其中列出了每一阶次的多项式分类器的性能。

图 5.8　对各个两类数字识别问题(区分某指定数字与其他),(对在相应的特征空间中定义在正规超平面集合上的)结构中最佳元素的 VC 维估计,与多项式阶数之间的关系

表 5.6 中,每一行对应一个两类分类器,它把一个数字与其他数字分开(这个数字在表的第一列中标明)。

表 5.6　选择最佳多项式阶数

数字	选中的分类器			测试错误数目						
	deg	dim	h_{est}	1	2	3	4	5	6	7
0	3	$\sim 10^6$	530	36	14	11	11	11	12	17
1	7	$\sim 10^{16}$	101	17	15	14	11	10	10	10
2	3	$\sim 10^6$	842	53	32	28	26	28	27	32
3	3	$\sim 10^6$	1 157	57	25	22	22	22	22	23
4	4	$\sim 10^9$	962	50	32	32	30	30	29	33
5	3	$\sim 10^6$	1 090	37	20	22	24	24	26	28
6	4	$\sim 10^9$	626	23	12	12	15	17	17	19
7	5	$\sim 10^{12}$	530	25	15	12	10	11	13	14
8	4	$\sim 10^9$	1 445	71	33	28	24	28	32	34
9	5	$\sim 10^{12}$	1 226	51	18	15	11	11	12	15

表中其他行的含义依次是:

· deg 即用所讨论的方法选出的多项式阶数(从 2 阶到 7 阶中选出);

· dim 即对应特征空间的维数,也是在这个空间中的线性分类器最大可能的 VC 维;

· h_{est} 即所选多项式的 VC 维估计(它远小于自由参数个数);

· 测试错误数目即用构造的各个阶数的多项式得到的测试错误数目;框起来的数字表示选中的多项式的错误数。

这样,表 5.5 说明了对 SV 多项式机器,不存在随着多项式维数增高的过学习问题,而表 5.6 则说明,即使在最好和最差解之间差别小(对从 2 阶到 7 阶的多项式)的情况下,我们的理论也给出了一个逼近最佳解的方法(寻找最佳多项式阶数的方法)。

同时还注意到,表 5.6 也说明了数字识别的问题本质上是非线性的。最好的多项式分类器与线性分类器之间错误数目的差别最大可以到四倍(如对数字 9)。

5.8 关于 SV 机的讨论

任何学习机器的品质都是由 3 个主要因素表征的,这 3 个因素是:

(1) 这个学习机器的通用性如何?

即它能够逼近的函数集有多么丰富?

(2) 这个机器的推广性如何?

即这个机器(通过实现某个给定的函数集和函数集的某个给定的结构)所到达的错误率上界与最小可能值有多么接近?

(3) 这个机器的学习过程的收敛有多快?

即利用给定数目的观测,需要多少次运算才能找到决策规则?

下面我们就逐一对这 3 个因素进行讨论。

(1) SV 机实现的是下面的函数集:

$$f(x, \alpha, w) = \operatorname{sgn}\left\{\sum_{i=1}^{N} \alpha_i K(x, w_i) - b\right\}, \tag{5-35}$$

其中,N 是任意整数($N < l$),α_i, $i = 1, \cdots, N$ 是任意标量,w_i, $i = 1, \cdots, N$ 是任意向量。核 $K(x, w)$ 可以是满足定理 5.3 条件的任意对称函数。

正如前面所看到的,对这些函数集,最好的有保障的损失是在权值向量 $w_1, \cdots, w_N$ 等于训练数据中的某些向量 x(支持向量)时取得的。

利用函数集

$$\bar{f}(x, \alpha, w) = \sum_{\text{支持向量}} y_i \alpha_i K(x, w_i) - b,$$

在其中采用多项式、径向基函数或者神经网络类型的回旋函数,我们可以以任意的精度逼近一个连续函数。

注意,对 SV 机来说,我们不需要通过事先选择数目 N 来建立机器的构造(而在传统的神经网络或者传统的径向基函数机器中这一点是需要的)。

而且,在 SV 机中,只要改变函数 $K(x, w)$,我们就可以改变学习机器的类型(即逼近函数的类型)。

(2) 对特征空间中的函数集上给定的结构,SV 机使得错误率的上界最小化。对最好

的解,(5-35)式中的向量 w_i 必须与训练数据中的某些向量重合(即支持向量)[①]。SV 机从(5-35)式的集合中寻找这样的函数,它们把训练数据分开,并且属于有最小 VC 维界的子集。(在更一般的情况下,它们最小化(5-1)式风险的界。)

(3) 最后,要找到所求的函数,SV 机必须在非负象限中最大化一个非负的二次型。这个问题是一种特殊的二次规划问题的一个特例,即在约束条件

$$a_i \leq x^i \leq b_i, \quad i = 1, \cdots, n$$

下最大化非负二次型 $Q(x)$,其中 x^i, $i = 1, \cdots, n$ 是向量 x 的各个坐标,a_i, b_i 是给定的界限。对这个特别的二次规划问题存在快速算法[②]。

5.9 SVM 与 Logistic 回归

5.9.1 Logistic 回归

经常地,我们不但需要构造一个决策规则,而且需要找到这样一个函数,它对任意给定的向量 x 确定它属于第一类的概率 $P\{y = 1 | x\}$。与构造一个性能良好的决策规则相比,这个问题更具一般性。知道了条件概率函数,我们可以构造贝叶斯(最优)决策规则:

$$r(x) = \operatorname{sgn} \ln \frac{P\{y = 1 | x\}}{1 - P\{y = 1 | x\}} .$$

下面,我们考虑以下的(参数化)条件概率估计问题[③]。假设下面两个概率的比的对数是给定的参数集 $f(x, w)$, $w \in W$ 中的一个函数 $f(x, w_0)$:

$$\ln \frac{P\{y = 1 | x\}}{1 - P\{y = 1 | x\}} = f(x, w_0) .$$

从这个方程可以得到,条件概率函数 $P\{y = 1 | x\}$ 有如下的形式:

$$P\{y = 1 | x\} = \frac{e^{f(x, w_0)}}{1 + e^{f(x, w_0)}} . \tag{5-36}$$

函数式(5-36)称作 logistic 回归。

我们的目标是,给定数据

$$(y_1, x_1), \cdots, (y_l, x_l),$$

估计 logistic 回归的参数 w_0[④]。首先,我们来说明,泛函

$$R_x(w) = E_y \ln(1 + e^{-yf(x, w)}) \tag{5-37}$$

的最小点确定了待求的参数(E_y 是在一个固定的 x 值下对 y 求期望)。

---

① 这一论断是 5.4 节描述求解二次优化问题的 Kühn-Tucker 条件必要性的一个直接推论。Kühn-Tucker 条件是这个问题的解的充分必要条件。

② 尽管理论上如此,但 SVM 的实现算法在很多情况下仍存在大量问题需要研究,比如当样本数目(尤其是支持向量数目)比较多时算法往往会需要很大的内存,而且寻优速度慢。如何建立更实用的 SVM 算法正是当前急需研究的问题之一。——译者

③ 我们将在第七章中讨论这一问题更一般的非参数表示。

④ 注意,(5-36)式是第 5.2 节中考虑的 sigmoid 函数的一种形式。因此,一个采用(5-36)式的 sigmoid 函数的一层神经网络经常被认为是 logistic 回归的一种估计。

我们知道,最小点的必要条件是

$$\frac{R_x(w)}{w} = \frac{}{w} E_y \ln(1 + e^{-yf(x,w)}) \Big|_{w_0} = 0.$$

对 w 求导,并利用表达式(5-36),我们得到①

$$\frac{R(w)}{w} = \frac{}{w} E_y \ln(1 + e^{-yf(x,w)})$$

$$= \frac{-f_w'(x,w)e^{-f(x,w)}}{1 + e^{-f(x,w)}} P\{y = 1 \mid x\} + \frac{f_w'(x,w)e^{f(x,w)}}{1 + e^{f(x,w)}} P\{y = -1 \mid x\}$$

$$= \frac{-f_w'(x,w)e^{-f(x,w)}}{1 + e^{-f(x,w)}} \cdot \frac{e^{f(x,w_0)}}{1 + e^{f(x,w_0)}} + \frac{f_w'(x,w)e^{f(x,w)}}{1 + e^{f(x,w)}} \cdot \frac{1}{1 + e^{f(x,w_0)}}$$

这一表达式在 $w = w_0$ 时等于 0。这就证明了,泛函(5-37)式的最小点确定了 logistic 回归的参数。

下面,我们假定所求的 logistic 回归是一个线性函数

$$f(x, w) = (x \cdot w_0) + b,$$

我们将通过用观测

$$(y_1, x_1), \cdots, (y_l, x_l)$$

最小化泛函

$$R(w) = E_{y,x} \ln(1 + e^{-y[(x \cdot w) + b]}) \tag{5-38}$$

来估计这个 logistic 回归函数的参数 w_0 和 b。

为了最小化泛函(5-38)式,我们用结构风险最小化方法,采用如下定义的结构:

$$(w \cdot w) \leq r.$$

我们以下面的形式来考虑这个最小化问题:最小化泛函②

$$R_{emp}(w, b) = \frac{1}{2}(w \cdot w) + C \sum_{i=1}^{l} \ln(1 + e^{-y_i[(w \cdot x_i) + b]}) \tag{5-39}$$

可以证明,(5-39)式的最小值定义了对 logistic 回归的如下逼近:

$$P\{y = 1 \mid x\} = \frac{\exp\left[\sum_{i=1}^{l} y_i \alpha_i^0 (x_i \cdot x) + b_0\right]}{1 + \exp\left[\sum_{i=1}^{l} y_i \alpha_i^0 (x_i \cdot x) + b_0\right]}, \tag{5-40}$$

其中,系数 α_i^0 和 b_0 是下列方程的解:

$$\alpha_i = \frac{\exp\left[-y_i \sum_{j=1}^{l} C\alpha_j y_j (x_j \cdot x_i) + b\right]}{1 + \exp\left[-y_i \sum_{j=1}^{l} C\alpha_j y_j (x_j \cdot x_i) + b\right]},$$

① 原著中公式有误。——译者

② 原著中误写为$(w, w), (w, x_i)$。——译者

$$\sum_{i=1}^{l} y_i \frac{\exp\left(-y_i\left(\sum_{j=1}^{l} C\alpha_j y_j (x_j \cdot x_i) + b\right)\right)}{1+\exp\left(-y_i\left(\sum_{j=1}^{l} C\alpha_j y_j (x_j \cdot x_i) + b\right)\right)} = 0.$$

事实上，点(w_0, b_0)使得泛函(5-39)式最小的一个必要条件是

$$\left.\frac{R(w,b)}{w}\right|_{w_0,b_0} = w - C\sum_{i=1}^{l} y_i x_i \left.\frac{\exp\{-y_i[(w\cdot x_i)+b]\}}{1+\exp\{-y_i[(w\cdot x_i)+b]\}}\right|_{w_0,b_0} = 0,$$

$$\left.\frac{R(w,b)}{b}\right|_{w_0,b_0} = -C\sum_{i=1}^{l} y_i \left.\frac{\exp\{-y_i[(w\cdot x_i)+b]\}}{1+\exp\{-y_i[(w\cdot x_i)+b]\}}\right|_{w_0,b_0} = 0. \tag{5-41}$$

记

$$\frac{\exp\{-y_i[(w_0\cdot x_i)+b_0]\}}{1+\exp\{-y_i[(w_0\cdot x_i)+b_0]\}} = \alpha_i^0, \tag{5-42}$$

我们可以把表达式(5-41)重写如下：

$$w_0 = C\sum_{i=1}^{l} y_i \alpha_i^0 x_i,$$

$$\sum_{i=1}^{l} y_i \alpha_i^0 = 0. \tag{5-43}$$

把表达式(5-43)代回到(5-37)式中，我们就得到了逼近(5-40)式。

注意，从(5-42)式和(5-43)式，我们有

$$0 < \alpha_i^0 < 1.$$

也就是说，这个解不是稀疏的。

为了找到 logistic 回归，我们可以把泛函(5-39)式(用表达式(5-43))重写为等价形式，即

$$R_{emp}(\alpha, b) = \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i\alpha_j y_i y_j (x_i \cdot x_j) + C\sum_{i=1}^{l} \ln\left(1+\exp\left\{-y_i\sum_{j=1}^{l} y_i y_j \alpha_j (x_i \cdot x_j) + b\right\}\right).$$

因为这个泛函相对参数 α和 b是凸的，我们可以用梯度下降方法找到其最小点。

5.9.2 SVM 的风险函数

让我们引入下面的记号

$$z = (w \cdot x) + b.$$

利用这一记号，我们可以把对 logistic 回归的风险泛函重写如下：

$$Q(z) = \ln(1+e^{-yz}).$$

考虑损失函数

$$Q^*(z) = c_1(1-z)_+, \tag{5-44}$$

其中，c_1是某个常数(在构造 SVM 时我们使用了 $c_1 = 1$)，$(a)_+ = \max(0, a)$(有一个结点的线性样条函数，关于样条逼近的更多内容请参见第 6.3 节)。

图 5.9　logistic 损失函数(虚线)和用一个结点的
线性样条对它的逼近(实线)

图 5.9 显示了在 $c_1 = 0.8$ 下的这一损失函数(图中实线)和 logistic 损失(图中的虚线)。

容易看到, SVM 使下面的泛函最小化:

$$R_{emp}(w, b) = \frac{1}{2}(w \cdot w) + C\sum_{i=1}^{l}(1 - y_i[(w \cdot x_i) + b])_+ . \tag{5-45}$$

事实上, 我们用 ξ 来代表表达式

$$\xi = (1 - y_i[(w \cdot x_i) + b])_+ ,$$

它等价于不等式

$$y_i[(w \cdot x_i) + b] \geq 1 - \xi. \tag{5-46}$$

现在, 可以把我们的优化问题(5-45)式重写如下: 在约束条件(5-46)式和约束条件

$$\xi \geq 0$$

下, 最小化泛函

$$R(w, b) = \frac{1}{2}(w \cdot w) + C\sum_{i=1}^{l}\xi. \tag{5-47}$$

这个问题与我们在第 5.1 节中为了对不可分情况构造最优分类超平面而提出的问题相同。

5.9.3　Logistic 回归的 SVM$_n$ 逼近

我们可以用 $n > 1$ 个结点的线性样条函数构造对 logistic 损失函数更好的 SVM 逼近。

假设我们对 logistic 损失给出了下面的样条逼近:

$$F(z) = \sum_{k=1}^{n} c_k(a_k - z)_+ ,$$

其中,

$$z = y[(w \cdot x) + b],$$

a_k, $k = 1, \cdots, n$ 是样条的结点, 且 $c_k \geq 0$, $k = 1, \cdots, n$ 是样条的系数。(因为 logistic 损失函数是凸单调函数, 我们可以用一个系数 c_k 非负的线性样条以任意的精确度来逼近它。)

图 5.10 显示了对 logistic 损失(图中为虚线)的逼近, 其中(a)是用两个结点的样条函数的逼近, (b)是用三个结点的样条函数的逼近(图中实线)。

让我们最小化泛函

图 5.10　logistic 损失函数(虚线)及其逼近(实线)

$$R(w,b) = \frac{1}{2}(w\cdot w) + C\sum_{i=1}^{l}\sum_{k=1}^{n} c_k(a_k - z_i)_+ ,$$

它是我们对泛函(5-38)式的逼近。

记　$(a_k - z_i)_+ = (a_k - y_i[(w\cdot x_i)+b])_+ = \xi_i^k$,

$$\xi_i^k \geq 0,\ k = 1,\cdots,n,\ i = 1,\cdots,l,$$

利用这种记号,我们可以把问题重写如下:

最小化泛函

$$R(w,b) = \frac{1}{2}(w\cdot w) + C\sum_{i=1}^{l}\sum_{k=1}^{n}\xi_i^k$$

约束条件是

$$y_i[(w\cdot x_i) + b] \geq a_k - \xi_i^k,\quad i = 1,\cdots,l,\quad k = 1,\cdots,n,$$

和

$$\xi_i^k \geq 0,\ i = 1,\cdots,l,\ k = 1,\cdots,n.$$

像以前一样,要在对偶空间中求解这个二次优化问题,我们构造拉格朗日函数

$$L = \frac{1}{2}(w\cdot w) + C\sum_{i=1}^{l}\sum_{k=1}^{n}\xi_i^k$$

$$-\sum_{i=1}^{l}\sum_{k=1}^{n}\beta_i^k(y_i[(w\cdot x_i) + b] - a_k + \xi_i^k) - \sum_{i=1}^{l}\sum_{k=1}^{n}\xi_i^k r_i^k$$

对 w, b 和 ξ 求最小,我们得到

$$w = \sum_{i=1}^{l}\sum_{k=1}^{n}\beta_i^k y_i x_i, \tag{5-48}$$

$$\sum_{i=1}^{l}\sum_{k=1}^{n}\beta_i^k y_i = 0, \tag{5-49}$$

$$0 \leq \beta_i^k \leq Cc_k,\quad k = 1,\cdots,n. \tag{5-50}$$

把 w 的表达式代回到拉格朗日函数中,并考虑到(5-49)式,我们得到泛函

$$W(\beta) = \sum_{i=1}^{l}\sum_{k=1}^{n}\beta_i^k a_k - \frac{1}{2}\sum_{i,j=1}^{l}\sum_{k=1}^{n}\beta_i^k\sum_{k=1}^{n}\beta_j^k y_i y_j (x_i\cdot x_j), \tag{5-51}$$

其中,$a_1,\cdots,a_n$ 是我们对 logistic 损失函数的样条逼近中的结点。

参数 $\beta_i^1,\cdots,\beta_i^n$, $i = 1,\cdots,l$ 定义了最优向量 w 的展开式(5-48),要得到这些参数,我们须在约束条件(5-49)式和(5-50)式下最大化泛函(5-51)式。

我们也可以从下面的 Kuhn-Tucker 条件得到参数 b:

$$\beta_i^k\{y_i[(w\cdot x_i)+b]-a_k+\xi_i^k\}=0,\quad i=1,\cdots,l,\quad k=1,\cdots,n.$$

利用这些参数,可以构造线性函数

$$l(x)=\sum_{j=1}^{l}y_j\sum_{k=1}^{n}\beta_i^k(x_j\cdot x)+b, \tag{5-52}$$

它定义了对 logistic 回归(5-36)式的逼近

$$P\{y=1\mid x\}=\frac{\exp\left\{\sum_{j=1}^{l}y_j\sum_{k=1}^{n}\beta_j^k(x_j\cdot x)+b\right\}}{1+\exp\left\{\sum_{j=1}^{l}y_j\sum_{k=1}^{n}\beta_j^k(x_j\cdot x)+b\right\}}. \tag{5-53}$$

像以前一样,要定义 logistic 回归的指数中的向量 w,我们需要计算两个向量 x 之间的内积。因此,利用满足 Mercer 条件的核 $K(x,x_i)$,我们可以构造一个如下形式的对 logistic 回归的逼近:

$$P\{y=1\mid x\}=\frac{\exp\left\{\sum_{j=1}^{l}y_j\sum_{k=1}^{n}\beta_j^k K(x_j,x)+b\right\}}{1+\exp\left\{\sum_{j=1}^{l}y_j\sum_{k=1}^{n}\beta_j^k K(x_j,x)+b\right\}},$$

其中的系数 β_j^k 是下面的二次优化问题的解:最大化泛函

$$W(\beta)=\sum_{i=1}^{l}\sum_{k=1}^{n}\beta_i^k a_k-\frac{1}{2}\sum_{i,j=1}^{l}\sum_{k=1}^{n}\beta_i^k\sum_{k=1}^{n}\beta_j^k y_i y_j K(x_i,x_j), \tag{5-54}$$

约束条件是

$$\sum_{i=1}^{l}\sum_{k=1}^{n}\beta_i^k y_i=0,$$

$$0\le\beta_i^k\le C c_k,\quad k=1,\cdots,n.$$

注意,在 logistic 损失的逼近中使用的结点数目越多,则在构造相应的超平面中使用的支持向量数目也就越多。随着逼近精度(结点数目)的提高,SVM_n 失去了稀疏性。

然而,随着 SVM_n 中 n 的增加,我们并不能保证用给定的样本数目能得到更好的性能。与估计一个好的决策规则的问题相比,较好地估计 logistic 回归的问题更为一般性,因此,要较好地解决它需要更多的数据。

我们的实验没有显示出 SVM_n 比 SVM_1 对 logistic 回归更有优势。

5.10 SVM 的组合

在 1996 年,Y. Freund 和 R. Schapire 提出了在一个线性决策规则中组合几种弱规则[①] (特征)的 AdaBoost 算法,这样组合的线性决策规则比任何一个弱规则的性能都好得多。

① 所谓弱规则,指能够至少比一个随机猜测略好地把测试数据分类的指示函数。

后来得到证明,实际上,AdaBoost 是(用一种贪婪的优化步骤)最小化某个泛函,而这个泛函的最小点定义了 logistic 回归(Friedman, Hestie and Tibshirany, 1998)。并且,人们还证明了,在 AdaBoost 所选择的弱(指示器)规则之上建立的最优超平面,其性能往往高于 AdaBoost 所得的解。

因此,在 AdaBoost 算法中,我们把它分为两个部分来考虑:

(1) 从给定的指示器特征集合中选择 N 个适当的特征;

(2) 用所选择的特征构造一个分类超平面。

在这一节里,我们将介绍构造一个 SVM 组合的两阶段方法。在其第一阶段中,利用给定的训练数据,我们寻找 N 个指示函数(特征),它们一方面是给定的模式识别问题的 SVM 解,另一方面是 AdaBoost 算法中对同一泛函进行贪婪优化的结果。

在第二阶段中,利用给定的训练数据,我们在所得的特征之上构造 SVM 决策规则。因此,我们将对同一模式识别问题构造 N 个不同的 SVM 解,然后把它们组合到一个决策规则中。

5.10.1 AdaBoost 方法

在 5.9.1 节中,我们介绍了风险泛函(5-37)式,它的最小点确定了 logistic 回归的参数。下面我们考虑另一个风险泛函:

$$R(\alpha) = Ee^{-yf(x,\alpha)}, \tag{5-55}$$

它是定义在函数 $f(x,\alpha)$ 的集合上的,该函数集包含了函数

$$f(x,\alpha_0) = \frac{1}{2}\ln\frac{P(y=1|x)}{P(y=-1|x)}. \tag{5-56}$$

容易看到,函数 $f(x,\alpha_0)$ 给出了泛函(5-55)式的最小点。

事实上,方程(5-56)等价于下面两个方程:

$$P(y=1|x) = \frac{e^{2f(x,\alpha_0)}}{1+e^{2f(x,\alpha_0)}} = \frac{e^{f(x,\alpha_0)}}{e^{-f(x,\alpha_0)}+e^{f(x,\alpha_0)}},$$

$$P(y=-1|x) = \frac{1}{1+e^{2f(x,\alpha_0)}} = \frac{e^{-f(x,\alpha_0)}}{e^{-f(x,\alpha_0)}+e^{f(x,\alpha_0)}}. \tag{5-57}$$

因为

$$E(e^{-yf(x,\alpha)}|x) = P(y=1|x)e^{-f(x,\alpha)} + P(y=-1|x)e^{f(x,\alpha)},$$

我们有

$$\frac{E(e^{-yf(x,\alpha)}|x)}{f(x,\alpha)} = -P(y=1|x)e^{-f(x,\alpha)} + P(y=-1|x)e^{f(x,\alpha)}. \tag{5-58}$$

只要把(5-57)式代入,就可以看到,在 α_0 点上导数(5-58)式等于 0。

让我们用经验风险泛函

$$R_{emp}(\alpha) = \sum_{i=1}^{l} e^{-y_i f(x_i,\alpha)} \tag{5-59}$$

来代替(5-55)式,并用下面的贪婪优化步骤来迭代地对它最小化。

贪婪优化步骤:

(1) 我们迭代地最小化泛函(5-59)式, 在第 k 次迭代时, 构造一个如下形式的函数:

$$f(x, \beta) = \sum_{r=1}^{k} d_r \, {}_r(x), \quad d_1 = 1,$$

其中, ${}_r(x)$, r= 1, …, N 属于一个给定的指示函数集(可能是无限的), k 是迭代的次数, $\beta = (d_1, \cdots, d_k)$ 是一个 k 维向量。

在第一次迭代中, 我们选择使训练错误数最小的特征 ${}_1(x)$。

(2) 假设在第 k 次迭代时我们得到了下面的经验风险值:

$$R_{emp}(\beta) = \sum_{i=1}^{l} e^{-y_i f_k(x_i, \beta_k)}.$$

在接下来的第(k+ 1)次迭代时, 我们继续在单参数函数的集合

$$f(x, \beta_{(k+1)}) = f(x, \beta) + d_{(k+1)} \, {}_{(k+1)}(x) \tag{5-60}$$

中最小化经验风险泛函。对(5-60)式, 我们得到下面的经验风险值:

$$R_{emp}(\beta_{(k+1)}) = \sum_{i=1}^{l} e^{-y_i f(x_i, \beta_{k+1})} = \sum_{i=1}^{l} c_i^{k+1} e^{-d_{(k+1)} y_{k(k+1)}(x_i)} \tag{5-61}$$

其中, 我们记

$$c_i^{k+1} = e^{-y_k f_k(x_i, \beta_k)}.$$

假设在第(k+ 1)次迭代时, 我们选择了指示函数 ${}_{(k+1)}(x)$(后面我们将定义如何选择这一函数)。那么, 为了最小化经验风险(5-61)式, 我们必须选择下面的参数值:

$$d_{(k+1)} = \frac{1}{2} \ln \frac{C_+^{k+1}}{C_-^{k+1}}, \tag{5-62}$$

其中我们使用了记号

$$C_+^{k+1} = \sum_{\{i:\, y_{i(k+1)}(x_i) = 1\}} c_i^{k+1},$$

$$C_-^{k+1} = \sum_{\{i:\, y_{i(k+1)}(x_i) = -1\}} c_i^{k+1}.$$

这是由下面的事实所决定的: $y_{i(k+1)}(x_i) \in \{1, -1\}$, 并且, 在最优点 $d_{(k+1)}$ 上, 经验风险(5-61)式对 d 的导数必须等于 0:

$$\frac{\partial}{\partial d} \sum_{i=1}^{l} e^{-y_i [f(x_i, \beta_k) + d_{(k+1)}(x_i)]}$$

$$= -\sum_{i=1}^{l} c_i^{k+1} y_{i(k+1)}(x_i) e^{-d y_{i(k+1)}(x_i)} = 0. \tag{5-63}$$

(3) 要为第(k+ 1)次迭代选择适当的函数 ${}_{(k+1)}(x)$, 注意到在第 k 次迭代之后, 根据(5-63)式, 有下面的等式成立:

$$-\sum_{i=1}^{l} c_i^k y_{ik}(x_i) e^{-d_k y_{ik}(x_i)} = -\sum_{i=1}^{l} c_i^{k+1} y_{ik}(x_i) = 0.$$

假设系数 c_i^{k+1} 被归一化到 1, 即

$$c_i^{k+1} \quad \frac{c_i^{k+1}}{\sum_{i=1}^{l} c_i^{k+1}},$$

这并不会影响结果。然而,归一化使得我们可以为方程(5-63)提出一个很好的统计学解释,就是:归一化系数 c_i^{k+1}, $i=1,\cdots,l$ 可以看作是在第$(k+1)$次迭代时赋予给定训练数据的一个概率测度,指示函数 $_k(x)$ 可以被看作是对赋予这一概率测度训练数据的最差解(即对这一概率测度,规则 $_k(x)$ 有 50% 的错误率)。也就是,在每一次迭代之后,算法为训练数据赋予一个新的概率测度,它对前一个弱规则来说是最困难的。因此,对下一次迭代,即第$(k+1)$次迭代,我们选择对于这个新赋予的概率测度最小化错误率的函数 $_{(k+1)}(x)$。即我们选择函数 $_{(k+1)}(x)$,使之最小化泛函

$$R(\ ) = -\sum_{i=1}^{l} c_i^{k+1} y_i\ (x_i). \tag{5-64}$$

(4) 用上面描述的贪婪最小化步骤得到的指示函数

$$\Phi(x) = \operatorname{sgn}\sum_{k=1}^{N} d_k\ _k(x) \tag{5-65}$$

就是 AdaBoost 决策规则。

5.10.2 SVM 的组合

让我们用上面描述的贪婪优化的思想来构造 SVM 的组合。我们从弱特征是线性决策规则的情况开始,即

$$_k(x) = \operatorname{sgn}\{(x\ ¡¤w_k) + b_k\}.$$

我们的目标是找到 N 个最优超平面,它们以贪婪的方式最小化泛函

$$R(w,b) = \sum_{i=1}^{l} \exp\ -\ y_i \sum_{k=1}^{N} d_k \operatorname{sgn}[(x_i\ ¡¤w_k) + b_k]\ , \tag{5-66}$$

然后用这些线性决策规则作为特征构造所求的组合。

构造特征

要构造 N 个特征,我们只需要按照上一节描述的一般原理确定在线性决策函数集合中最小化泛函(5-64)式的方法,这个线性决策函数集合是

$$_k(x) = \operatorname{sgn}\{(x\ ¡¤w_k) + b_k\}$$

(是由最优超平面定义的)。

像以前一样,我们用下面的问题来代替这个问题:最小化泛函

$$R(w_k) = \frac{1}{2}(w_k\ ¡¤w_k) + C\sum_{i=1}^{l} c_i^k \xi, \qquad c_i^1 = 1, \tag{5-67}$$

约束条件是

$$y_i((w_k\ ¡¤x_i) + b_k) \geq 1 - \xi, \qquad \xi \geq 0. \tag{5-68}$$

与第 5.5 节中描述的构造软间隔超平面问题相比,这里的构造超平面问题的唯一区别是,在软间隔超平面情况下,所有系数 c_i^k 都等于 1,而现在,(5-67)式中的第二项是一个加权和。

我们用与拉格朗日乘子同样的技术来求解这个优化问题。可以得到下面的解:

$$w_k = \sum_{i=1}^{l} y_i \alpha_i^k x_i,$$

其中的系数 α^k 最大化泛函

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j), \tag{5-69}$$

约束条件是

$$0 \leq \alpha_i \leq C c_i^k, \tag{5-70}$$

以及

$$\sum_{i=1}^{l} y_i \alpha_i c_i^k = 0. \tag{5-71}$$

系数 b_k 可以用 Kuhn-Tucker 条件

$$\alpha_i (y_i (w_k \cdot x_i) + b_k - 1 + \xi_i) = 0$$

求得。

因此,决策规则的不同是由系数 c_i^k 决定的。这些系数是迭代计算的,就像前面在贪婪优化步骤中介绍的那样(第 5.10.1 小节):

$$c_i^1 = 1, \quad i = 1, \cdots, l,$$

$$c_i^{(k+1)} = \exp - y_i \sum_{r=1}^{k} d_r r(x_i) = c_i^k \exp\{- y_i d_k k(x_i)\}, \tag{5-72}$$

其中,

$$d_k = \frac{1}{2} \ln \frac{\sum_{\{i \cdot y_i k(x_i) = 1\}} c_i^k}{\sum_{\{i \cdot y_i k(x_i) = -1\}} c_i^k}. \tag{5-73}$$

讨论

注意,如果训练数据是可分的,那么等式(5-73)的分母等于零,因此,根据(5-72)式,对所有的 $k > 1$, $c_i^k = 0$, $i = 1, \cdots, l$。也就是,特征集合只有一个决策规则。为了避免这种情况,我们可以选择充分小的 C 值(大的正则化参数)。然而,如果对充分小的 C,训练数据仍是可分的,那么所得到的超平面具有好的推广能力。

常数 C 的选择在构造一个 SVM 组合中起着十分重要的作用。

构造决策规则

为了得到决策规则,我们在 N 维二值空间

$$z = (\ _1(x), \cdots, \ _N(x))$$

中构造最优超平面。利用给定的训练数据集合,我们得到一个新的训练数据集合

$$(y_1, z_1), \cdots, (y_l, z_l), \tag{5-74}$$

(其中 $z_i = (\ _1(x_i), \cdots, \ _N(x_i))$),我们基于这个新的集合来构造最优超平面。

SVM 的组合

像以前一样, 我们可以利用一般类型的 SVM, 用核函数得到特征。我们可以采用下面形式的特征:

$$_k(x) = \operatorname{sgn}\left(\sum_{i=1}^{l} y_i \alpha K(x, x_i)\right),$$

其中的系数 α 是下面优化问题的解: 最大化泛函

$$W(\alpha) = \sum_{i=1}^{l} \alpha - \frac{1}{2}\sum_{i,j=1}^{l} \alpha\alpha y_i y_j K(x_i, x_j)$$

约束条件是

$$0 \leq \alpha \leq Cc_i^k,$$

以及

$$\sum_{i=1}^{l} y_i \alpha c_i^k = 0.$$

利用所得到的定义一个二值空间 Z 的 N 个特征 $_k(x)$, k= 1, …, N, 我们构造出训练集(5-74)式。在这个训练集的基础上, 利用在 Z 空间中定义的一个核函数 $K^*(z, z_i)$, 我们构造 SVM 解:

$$r(x) = \operatorname{sgn}\left(\sum_{i=1}^{l} y_i \beta K^*(z(x), z(x_i))\right).$$

非正式推导和评述——5

5.11 工程技巧与正式的推理

神经网络的存在可以看作是对理论学家的挑战。

从严格的角度看,我们无法保证神经网络能够较好地推广,因为根据理论,为了控制推广能力,我们需要控制两个因素:经验风险值和置信范围值。然而,神经网络不能控制两者中的任何一个。

事实上,要最小化经验风险,神经网络必须最小化一个泛函,而这个泛函有很多局部极小点。有关理论并没有提供能防止算法终止于不可接受的局部极小点的构造性方法。而另一方面,要控制置信范围,我们必须首先构造神经网络所实现的函数集的一个结构,然后利用这个结构来进行容量控制。对神经网络来说,没有能够做到这一点的准确方法。

因此,从形式化的角度看,对于在解决实际问题时应采用什么类型的机器这一问题,答案似乎是明确的。

然而,现实并不是这样简单。神经网络的设计者们用高超的工程技巧弥补了数学上的缺陷。这就是,他们综合利用各种启发式算法,使得用较少的计算取得一个合理的局部极小点成为可能。

而且,对于给定的问题,他们创造出了一些特殊的神经网络结构,它们既有适当的容量,又包含了对求解问题‘有用的’函数。利用这些启发式方法,神经网络显示出了出人意料的好结果。

在第五章中介绍用美国邮政服务数据库进行数字识别的问题时,介绍了通过构造全局(而非局部)决策规则所得到的最好结果,在那里我们给出了下面两个数字:

- 神经网络 LeNet 1(由 Y. LeCun 设计)的错误率是 5.1%,
- 多项式 SV 机的错误率是 4.0%。

我们还提到了下面两个最好的结果:

- 局部学习方法的错误率是 3.3%,

· 对由训练集给出的模板进行切距匹配可使错误率达 2.7%,这是最好的记录。

在 1993 年,根据学术界对评价基准的需求,美国国家标准与技术研究所(NIST)提供了一个手写数字的数据库,其中包含60 000个训练数据和10 000个测试数据,这些数字都是用 20× 20= 400 的像素空间中的向量来描述的。

针对这个数据库,有人设计了一种特殊的神经网络(LeNet 4)。下面是在报告有关基准研究的文章(Léon Bottou et al, 1994)中关于 LeNet 4 的介绍:

"很长时间以来,LeNet 1 的水平被看作是最先进的。后来发展了局部分类器、SV 分类器和切距分类器,以在 LeNet 1 基础上进一步改进——这一目的达到了。但是这些方法又反过来引发了对改进神经网络构造的研究。这一研究部分程度上是由对各种学习机器容量的估计引导的,这种估计是通过把(在 NIST 的大数据库上的)训练和测试错误作为训练样本数目的一个函数进行度量而得到的[①]。我们发现需要更大的容量。通过对神经网络构造的一系列实验,并结合对误识样本特点的分析,精心制造出了 LeNet 4。"

在这些基准研究中,两种构造整个决策规则的方法达到了同样好的性能: 1.1% 的测试错误率[②],这两种方法是:

(1) LeNet 4;

(2) 多项式 SV 机(多项式阶数是 4)。

局部学习方法和对 60 000 个模板的切距匹配也得到了同样的性能: 1.1% 的测试错误率。

回顾对一个小的数据库(美国邮政数据库),最好的结果(比其他结果好很多)是由切距匹配方法得到的,其中利用了对问题的先验信息(结合在切距的概念中)。而当样本数增加到 60 000 个时,先验知识的优势减小了。局部学习方法的优势也随着观测数目的增多而减小。

与 LeNet 1 相比,针对 NIST 数据库精心制造的 LeNet 4 表现出了显著的性能改进(LeNet 1 对 NIST 数据库有 1.7% 的测试错误率[③])。

标准的多项式 SV 机也表现很好。下面我们继续引用文献(Léon Bottou et al, 1994)中的论述:

"SV 机有出色的精度,是最值得注意的,因为与其他高性能的分类器不同,它没有包含关于问题的几何结构的知识。事实上,如果图像的像素被加密了,比如进行了某个固定的随机排序,这个分类器仍将做得同样好。"

然而,这些学习机器所得到的性能并不是对 NIST 数据库的最好记录。利用字符的模型(与构造切距使用的模型相同)和60 000个训练数据样本,H. Drucker, R. Schapire, P. Simard(1993)产生出了多于1 000 000个样本,他们用这些样本训练三个 LeNet 4 神经

---

① Vapnik V, Levin E, and LeCun Y. 1994. Measuring the VC dimension of a learning machine. Neural Computation, 6(5): 851～876

② 遗憾的是我们不能把这些结果与第五章中介绍的结果进行比较。NIST 数据库中的数字与美国邮政数据库中的数字相比更"容易"识别一些。

③ 注意 LeNet 4 对 60 000 个训练样本的大(NIST)数据库有优势。对包含 7 000 个训练样本的小(美国邮政)数据库,容量较小的网络 LeNet 1 更好。

网络,三个神经网络在一个特殊的“自举系统”中结合起来,最后得到了 0.7% 的错误率。

现在 SV 机面对着一个挑战——它要跨越这个差距(从 1.1% 到 0.7%)。也许仅靠采用蛮力的 SV 机和60 000个训练样本不足以跨越这个差距。也许我们必须结合关于所面对问题的某些先验信息。

这样做可以有几种办法。最简单的是利用同样的(从60 000个 NIST 原型构造出来的)1 000 000个样本数据。但是,令人更感兴趣的是,如何找到一种途径,它可以把在构造新样本时利用的不变性直接集成到 SV 机中来。例如,对于多项式机器,我们可以通过采用形式为$(x^{T}Ax^{*})^{d}$的内积函数来集成关于不变性的先验信息,其中 x 和 x^{*} 是输入向量,A 是一个反映模型不变性的对称正定矩阵①。

我们也可以用下面的方法集成另一种(几何)类型的先验信息,就是只采用特征(单项式)$x_i x_j x_k$,它们是由相互靠近的像素形成的。(这样做反映了我们对问题的几何性质的认识:重要的特征是由相互连接的像素形成的,而不是由互相远离的像素形成的。)这样做可以大大降低特征空间的维数(降低幅度以百万计)。

总之,尽管支持向量机的理论基础看上去要比神经网络的理论基础更坚实,但是这种新的学习机器的实际优势仍需进一步证实②。

5.12 统计模型的高明所在

在这一章里,我们介绍了支持向量机,它通过以下步骤实现了结构风险最小化归纳原则:

(1) 用非线性变换把输入向量映射到一个高维特征空间中。

(2) 在这个空间中,在线性决策规则集合上按照正规超平面权值的模构造一个结构。

(3) 选择结构中最好的元素以及这个元素中最好的函数,以达到最小化错误率的界目标。

但是,在用本章描述的算法实现上述步骤时,却有一个地方违反了 SRM 原则。为了定义线性函数集上的结构,我们采用了对训练数据中的向量 x 构造的正规超平面集合。根据 SRM 原则,这个结构必须在训练数据出现之前事先定义。

在试图全面履行 SRM 原则的努力中,我们得到了学习问题的一种新的表述,它形成了一种新的推理。为了简单起见,我们只考虑对于模式识别问题的这种新模型。

设对一个实现在特征空间中的线性函数集的学习机器,给定如下 $l+k$ 个向量:

---

① B. Scholkopf 考虑了一种中间路线:构造一个 SV 机,通过转换支持向量图像(在 4 个主方向上进行转换)产生新样本,然后用支持向量和新样本重新训练。这样做,把在美国邮政数据库上的性能从 4.0% 提高到了 3.2%,把在 NIST 数据库上的性能从 1.1% 提高到了 0.8%。

② 联系到结合在神经网络中的启发式知识,我回忆起了 R. Feynman 所做的下面的评论:“我们必须从一开始就澄清一个观点,就是如果某事不是科学,它并不一定不好。比如说,爱情就不是科学。因此,如果我们说某事不是科学,并不是说它有什么不对;而只是说它不是科学。”—— 见 The Feynman Lectures on Physics. Addison-Wesley, 3-1, 1975。

$$x_1, \cdots, x_{l+k}, \tag{5-75}$$

它们是按照某个分布函数随机独立地抽取出来的。

现在假设这 $l+k$ 个向量被随机地分成两个子集:一个子集是

$$x_1, \cdots, x_l,$$

它们的分类是给定的(即训练集),类别标号由串

$$y_1, \cdots, y_l, \quad y \in \{-1, +1\}$$

给出。而另一个子集是

$$x_{l+1}, \cdots, x_{l+k},$$

它们的类别标号应由机器得到(即测试集)。学习机器的目标是找到一个规则,使得它能够给出在测试集上错误数目最少的类别标号串。

与本书中考虑的函数估计模型不同,上面这个模型寻找的是这样一个规则,它使得在给定的测试集上的错误数目最小,而不是寻求使得在允许的测试集上的错误概率最小的规则。我们把这个问题叫做在给定点上对函数取值的估计。对在给定点上估计函数值的问题,如果我们定义的正规超平面是关于所有 $l+k$ 个向量(5-75)式的,那么 SV 机就全面地履行了 SRM 原则。(我们可以把(5-75)式的数据看作是先验信息,而后验信息是关于把这个集合分为两个子集的任何信息。)

估计给定点上函数值的问题的解及其求解方法,都与基于对未知函数估计的情况不同。

考虑识别 5 位邮政编码的例子[①]。现有的基于函数估计的方法是独立地识别 5 位数字 $x_1, \cdots, x_5$:首先用在学习过程中构造的规则识别数字 x_1,然后用同一个规则来识别数字 x_2,依此类推。

估计函数值的方法建议对 5 位数字进行联合识别:对一位数字(比如 x_1)的识别不但依赖于训练数据和向量 x_1,也依赖于 $x_2, \cdots, x_5$。在这种方法中,我们使用的规则是在一种特殊的方式上针对一个给定的特定任务的。可以证明,这种方法能够给出更精确的解[②]。

应该注意到,是由于人们试图用 SRM 原则验证在正规超平面集上定义的一个结构,才第一次发现了对学习问题的这种观点。

5.13 从数字识别实验中我们学到了什么?

在本章描述的实验中,观察到 3 个值得进行讨论的现象,它们是:

(1) 在特征空间中构造的结构较好地反映了现实中的问题。

---

① 为了简单起见,我们不考虑分割问题,假设全部 5 位数字都是被分割好的。

② 注意到,在第 4.5 节中描述的局部学习方法可以看作是在函数估计和对兴趣点上函数值的估计之间的一个中间模型。回顾前面已介绍过的,对于小数据库(美国邮政数据库),局部学习方法给出更好的结果(错误率 3.3%),明显好于基于对整个函数估计的方法所得到的最好结果(LeNet 1 得到的是 5.1%,多项式 SV 机是 4.0%)。

(2) 所得决策规则的质量并不很严重地依赖于 SV 机的类型(多项式机器、RBF 机器或者两层神经网络)。但是,的确非常依赖于 VC 维控制(容量控制)的精确度。

(3) 不同类型的机器采用训练数据中同样的样本作为支持向量。

5.13.1 结构类型与容量控制精度的影响

传统的估计高维函数依赖关系的方法是建立在下面的信念基础上的:

实际问题中总存在较少数目的一些"强特征",用它们的简单函数(比如线性组合)就能较好地逼近未知函数。因此,需要仔细地选择一个低维的特征空间,在这个空间中用常规的统计技术来求解一个逼近。

这种方法强调这样一种观点:在特征选择阶段要精心(这是一个非形式的过程),然后就可以用常规的统计方法。

我们的新方法是基于一种不同的信念的:

实际问题中存在较大数目的一些"弱特征",它们"巧妙的"线性组合可以较好地逼近未知的依赖关系。因此,采用什么样的"弱特征"并不十分重要,而形成"巧妙的"线性组合更为重要。

这种方法强调这样一种观点:可以选择任何合理的"弱特征空间"(这是一个非形式的过程),但是在求解"巧妙的"线性组合时要非常精心。从 SV 机的角度看,"巧妙的"线性组合就是容量控制的方法。

一些理论学者和实验学者都多次表达了关于现实问题结构的这种信念。

1940 年,Church 提出一种主张,这就是下面的 Turing-Church 命题[①]:

所有(充分复杂的)计算机都实现同一函数族。

在我们的特殊情况中,我们提出一种更强的观点,就是对在不同内积回旋对应的各种特征空间中的线性组合,如果它们有相同的容量,那么它们就逼近相同的函数集。

Church 是在纯理论分析的基础上提出了他的观点的,然而,当计算机实验变得广泛之后,研究者们很快就意外地遇到了可以用 Church 的观点来描述的情况。

在 70 和 80 年代,对于求解各种导致不适定问题的算子方程,尤其是对密度估计问题,人们进行了很多实验研究。其中一个普遍观察到的现象就是,对(4-32)式中的正则化因子 $\Omega(f)$ 类型的选择,不如对正则化系数 $\gamma(\delta)$ 的正确选择那样重要(前者决定了结构的类型,而后者决定了容量的控制)。

特别地,在用 Parzen 窗法估计密度

$$p(x) = \frac{1}{l}\sum_{i=1}^{l}\frac{1}{\gamma^n}K\left(\frac{x - x_i}{\gamma}\right)$$

时,一个普遍的现象是:如果观测数目不是"非常少",那么估计子中核函数 $K(u)$ 的类型

① 注意,这一命题并不反映某一得到证明的事实,而是反映了对存在某种难以证明的(或者难以用精确的形式表达的)规律的信念。

并不像常数 γ的取值那样重要。(回顾在 Parzen 估计子中的核 K(u)是由泛函 Ω(f)确定的,而 γ是由正则化常数确定的。)

在回归估计问题中也有同样的现象,在人们试图用不同序列的展开来估计回归函数时:如果观测数目不是“非常少”,那么所采用的序列类型就不像逼近中采用的项的数目那样重要。所有这些现象都是在解决低维问题(多数是一维问题)中观察到的。

在本章介绍的实验中,我们在非常高维的空间中观察到了同样的现象。

5.13.2 SRM 原则和特征构造问题

在 SV 机中采用的大量特征的“巧妙”线性组合,有一个重要的结构:支持向量集合。我们可以如下描述这一结构:在弱特征的集合(弱特征空间)中,存在一个与支持向量相对应的复合特征的集合。我们把这个空间记作

$$u = (K(x, x_1), \cdots, K(x, x_N)) \in U,$$

其中

$$x_1, \cdots, x_N$$

是支持向量。我们在这个复合特征的空间 U 中构造线性决策规则。注意到,在定理 5.2 得到的界中,复合特征数目的期望在问题的维数中起了重要的作用。因此我们可以用下面的话来描述支持向量方法与传统方法的区别:

要较好地实现传统方法,需要人工选择(构造)一些数目相对较少的“巧妙的特征”,而支持向量方法则是自动地选择(构造)一些数目较少的“巧妙的特征”。

注意,SV 机是在空间 Z(弱特征的空间)中构造最优超平面,而不是在复合特征的空间中构造。但是(在选择出复合特征后),在空间 U 中寻找最优超平面的系数是容易的。而且,我们可以在 U 空间中构造一个新的 SV 机(用同样的训练数据)。因此,我们可以构造两层(或多层)的 SV 机。换句话说,我们可以采用多级来选择“巧妙的特征”。然而,正如我们在第 4.10 节中指出的,特征选择的问题是十分微妙的(读者可以回顾构造稀疏代数多项式和稀疏三角多项式之间的区别)。

5.13.3 支持向量集合是否是数据的一个鲁棒的特性?

在实验中我们观察到一个重要的现象,即不同类型的参数最优的 SV 机采用几乎相同的支持向量:在训练数据中有一个小的子集(在我们的实验中只占数据的 3%～5%),对于构造最佳决策规则的问题来说,这个子集等价于整个训练数据集合,而且这个训练数据子集对不同类型的最优 SV 机是几乎相同的(这些最优 SV 机分别是多项式阶数最优的多项式机器、参数 γ最优的 RBF 机器和参数 b 最优的神经网络机器)。

重要的问题是,是不是对一个很广泛的实际问题的集合来说都是这样的。一些间接的理论证据表明这是很可能的。我们可以看到,如果基于各种支持向量机的一种多数投票机制不能使性能进一步改进,那么这些机器中相同的支持向量的比例肯定比较大。

现在讨论 SV 机的特性还为时尚早,因为关于这些特性的分析现在才刚刚开始[①]。因此,我愿意用下面的话来结束本章的评述:

SV 机是一个十分适合理论分析的研究对象。它集以下多种概念模型于一身:

(1) SRM 模型。(SV 机最初就是从这里得到的。见定理 5.1。)

(2) 数据压缩模型。(定理 5.2 的界可以从压缩系数的角度来描述。)

(3) 构造复合特征的一个通用模型。(在希尔伯特空间中的内积回旋可以看作是构造特征的一种标准途径。)

(4) 对实际数据的一种模型。(一个小的支持向量集合可能足以对不同的机器代表整个训练集。)

再经过几年的时间,我们就将认清这些模型的统一是否反映了学习机制的某种本质的特性,还是只是一个新的死胡同[②]。

---

① 在本书完成后,C. Burges 说明了对所得的决策规则

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{N} \alpha_i K(x, x_i) + \alpha_0\right),$$

可以用下面的更简单的决策规则来近似:

$$f^*(x) = \operatorname{sgn}\left(\sum_{i=1}^{M} \beta_i K(x, T_i) + \beta_0\right), \quad M \ll N,$$

其中利用了所谓的广义支持向量 $T_1, \cdots, T_M$(一个特别构造的序列集合)。

对在第 5.7 节中介绍的数字识别问题,只需用基于每类 $M = 11$ 个广义支持向量的近似,就可以达到与每个分类器有 $N = 270$ 个(开始得到的)支持向量同样的性能。

这意味着对支持向量机,存在一种规则的办法,可以合成具有最优复杂度的决策规则。

② 从我在 1995 年说这些话到现在已经过去了 4 年(指本书第二版完稿时的 1999 年——译者注)。从那时起我们得到了很多证据,包括实验证据(比如参见第 5.7 节),表明 SV 方法是解决高维空间中各种函数估计问题的一个一般方法。

第六章

函数估计的方法

在本章中,我们将把在估计指示函数(对模式识别问题)中得到的结果推广到估计实函数(回归)中。我们将引入一种新的损失函数(称作 ε不敏感损失函数),它不但使我们的估计具有鲁棒性,而且使它是稀疏的。正如我们在本章及下一章将要看到的,解的稀疏性对在高维空间中用大量数据估计依赖性关系来说是非常重要的。

6.1 ε不敏感损失函数

在第一章 1.7 节中,为了描述在实值函数集 $\{f(x,\alpha),\ \alpha\in\Lambda\}$ 中估计训练器规则 $F(y|x)$ 的问题,我们考虑了一个二次损失函数

$$L(y,f(x,\alpha))=(y-f(x,\alpha))^2 \tag{6-1}$$

y 是在正态加性噪声下对一个回归函数的度量结果,在这样的条件下,ERM 原则(对于上述损失函数)给出对回归 $f(x,\alpha)$ 的一个有效的(最佳无偏的)估计。

然而我们知道,如果加性噪声是由其他分布产生的,则对回归函数逼近更好(对于 ERM 原则)的估计器是基于其他损失函数的(与这些分布有关),即

$$L(y,f(x,\alpha))=L(|y-f(x,\alpha)|). \tag{6-2}$$

(对于对称密度函数 $p(\xi)$,$L(\xi)=-\ln p(\xi)$。)

在 1964 年,Huber 提出了一个理论,它使得我们可以在只知道关于噪声模型的一般信息的情况下,找到选择损失函数的最佳策略。尤其是,他说明了,如果我们只知道描述噪声的密度是一个对称函数,那么最好的对回归的最小最大逼近(在最坏可能的噪声模型 $p(x)$ 下最好的 L_2 逼近)给出了下面的损失函数:

$$L(y,f(x,\alpha))=|y-f(x,\alpha)| \tag{6-3}$$

在这个损失函数下最小化经验风险被称作最小模方法,它属于所谓的鲁棒回归(robust regression)方法。然而,这是我们知道关于未知密度的最少信息的极端情况。

Huber 还考虑了另一种模型，其中噪声是某种固定的噪声（下面我们将考虑正态噪声）与另一有对称连续密度函数的任意噪声的混合。他说明了，对这种类型的噪声，最优解（依最小最大策略）是在采用下面的损失函数时得到的：

$$L(y-f(x,\alpha))=\begin{cases}c|y-f(x,\alpha)|-\dfrac{c^2}{2} & \text{对 } |y-f(x,\alpha)|>c,\\ \dfrac{1}{2}(y-f(x,\alpha))^2 & \text{对 } |y-f(x,\alpha)|\le c.\end{cases} \tag{6-4}$$

其中的常数 c 是由混合的比例定义的。

为了对实值函数构造 SVM，我们采用一种新的损失函数类型，即 ε不敏感损失函数：

$$L(y,f(x,\alpha))=L(|y-f(x,\alpha)|_\varepsilon), \tag{6-5}$$

其中，我们记

$$|y-f(x,\alpha)|_\varepsilon=\begin{cases}0 & \text{若 } |y-f(x,\alpha)|\le\varepsilon,\\ |y-f(x,\alpha)|-\varepsilon & \text{其他.}\end{cases} \tag{6-6}$$

这种损失函数描述这样一种 ε不敏感模型，即如果预测值和实际值之间的差别小于 ε则损失等于 0。当 $\varepsilon=0$ 时它与 Huber 的损失函数相同，当 c 比较小时与(6-4)式的损失函数接近。

下面我们考虑 3 种损失函数：

(1) 线性 ε不敏感损失函数：

$$L(y,f(x,\alpha))=|y-f(x,\alpha)|_\varepsilon \tag{6-7}$$

（如果 $\varepsilon=0$，则它与鲁棒损失函数(6-3)式相同。）

(2) 二次 ε不敏感损失函数：

$$L(y,f(x,\alpha))=|y-f(x,\alpha)|_\varepsilon^2. \tag{6-8}$$

（如果 $\varepsilon=0$，则它与二次损失函数(6-1)式相同。）

(3) Huber 损失函数：

$$L(y-f(x,\alpha))=\begin{cases}c|y-f(x,\alpha)|-\dfrac{c^2}{2} & \text{对 } |y-f(x,\alpha)|>c,\\ \dfrac{1}{2}(y-f(x,\alpha))^2 & \text{对 } |y-f(x,\alpha)|\le c.\end{cases} \tag{6-9}$$

利用同样的技术，我们可以考虑任意凸损失函数 L(u)。但是以上 3 种损失函数有其独特之处：它们可以引出与我们在模式识别问题中用到的相同的简单的优化问题。

图 6.1　损失函数的比较

6.2 用于回归函数估计的 SVM

如果我们:

(1) 在线性函数集合

$$f(x, \alpha) = (w \cdot x) + b$$

中估计回归函数;

(2) 把回归估计的问题定义为对一个(6-8)式的 ε不敏感损失函数($\varepsilon \geq 0$)进行风险最小化的问题;

(3) 用 SRM 原则进行风险最小化,其中结构 S_n 的元素由不等式

$$(w \cdot w) \leq c_n \tag{6-10}$$

定义,那么就产生了对回归的支持向量估计。

1. 对结构中固定元素的解

假设我们给定了训练数据

$$(x_1, y_1), \cdots, (x_l, y_l),$$

那么,寻找 w_l 和 b_l 使得在约束(6-10)式条件下经验风险[①]

$$R_{emp}(w, b) = \frac{1}{l}\sum_{i=1}^{l} |y_i - (w \cdot x) - b|_\varepsilon$$

最小化的问题就等价于下面的问题:寻找 w, b 对,使得它们最小化由松弛变量 ξ, ξ^*, $i = 1, \cdots, l$定义的下述量:

$$F(\xi, \xi^*) = \sum_{i=1}^{l} \xi_i^* + \sum_{i=1}^{l} \xi_i \tag{6-11}$$

约束条件是

$$\begin{aligned} y_i - (w \cdot x_i) - b &\leq \varepsilon + \xi_i^*, & i &= 1, \cdots, l \\ (w \cdot x_i) + b - y_i &\leq \varepsilon + \xi_i, & i &= 1, \cdots, l \\ \xi_i^* &\geq 0, & i &= 1, \cdots, l \\ \xi_i &\geq 0, & i &= 1, \cdots, l \end{aligned} \tag{6-12}$$

以及约束(6-10)式。

像前面一样,要求解带有不等式类型约束的优化问题,我们必须找到下面的拉格朗日泛函的鞍点:

$$\begin{aligned} &L(w, \xi, \xi^*, \alpha, C^*, \gamma, \gamma^*) \\ &= \sum_{i=1}^{l} (\xi_i^* + \xi_i) - \sum_{i=1}^{l} \alpha_i [y_i - (w \cdot x_i) - b + \varepsilon + \xi_i] \end{aligned}$$

---

① 这里实际上是采用前面提到的线性 ε不敏感损失函数(6-7)式。——译者

$$- \sum_{i=1}^{l} \alpha_i^* [(w \cdot x_i) + b - y_i + \varepsilon + \xi_i^*] - \frac{C^*}{2} (c_n - (w \cdot w))$$

$$- \sum_{i=1}^{l} (\gamma_i^* \xi_i^* + \gamma_i \xi_i). \tag{6-13}$$

(这个鞍点是对元素 w, b, ξ_i^* 和 ξ_i 的极小点和对拉格朗日乘子 $C^* \geq 0$, $\alpha_i^* \geq 0$, $\alpha_i \geq 0$, $\gamma_i^* \geq 0$ 和 $\gamma_i \geq 0$, $i = 1, \cdots, l$ 的极大点。)

对 w, b 和 ξ_i^*, ξ_i 最小化意味着下面 3 个条件:

$$w = \sum_{i=1}^{l} \frac{\alpha_i^* - \alpha_i}{C^*} x_i, \tag{6-14}$$

$$\sum_{i=1}^{l} \alpha_i^* = \sum_{i=1}^{l} \alpha_i, \tag{6-15}$$

$$\begin{aligned} 0 \leq \alpha_i^* \leq 1, \quad i = 1, \cdots, l \\ 0 \leq \alpha_i \leq 1, \quad i = 1, \cdots, l \end{aligned} \tag{6-16}$$

把(6-14)式和(6-15)式代入到(6-13)式中, 可以得到, 要解这个优化问题, 我们需要找到下面的凸泛函的最大值:

$$W(\alpha, \alpha^*, C^*) = - \varepsilon \sum_{i=1}^{l} (\alpha_i^* + \alpha_i) + \sum_{i=1}^{l} y_i (\alpha_i^* - \alpha_i)$$

$$- \frac{1}{2C^*} \sum_{i,j=1}^{l} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(x_i \cdot x_j) - \frac{c_n C^*}{2}, \tag{6-17}$$

约束条件是(6-15)式、(6-16)式以及

$$C^* \geq 0.$$

就像在模式识别中一样, 这里, 在展开式(6-14)中只有部分参数

$$\beta_i = \frac{\alpha_i^* - \alpha_i}{C^*}, \quad i = 1, \cdots, l$$

不为零。它们定义了问题中的支持向量。

2. 基本解

如果我们不是在约束(6-12)式和(6-10)式下最小化泛函(6-11)式, 而是变成在约束条件(6-12)式下(以给定的 C 值)最小化

$$\Phi(w, \xi^*, \xi) = \frac{1}{2}(w \cdot w) + C \sum_{i=1}^{l} \xi_i^* + \sum_{i=1}^{l} \xi_i,$$

那么我们就可以把上面的凸优化问题简化为对一个二次优化问题寻找向量 w 的问题。在这种情况下, 要找到所求的向量

$$w = \sum_{i=1}^{l} (\alpha_i^* - \alpha_i) x_i,$$

我们必须找到最大化二次型

$$W(\alpha, \alpha^*) = - \varepsilon \sum_{i=1}^{l} (\alpha_i^* + \alpha_i) + \sum_{i=1}^{l} y_i (\alpha_i^* - \alpha_i)$$

$$-\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)(x_i\cdot x_j) \tag{6-18}$$

的参数 α_i^*, α_i, $i=1,\cdots,l$, 约束条件是:

$$\sum_{i=1}^{l}\alpha_i^*=\sum_{i=1}^{l}\alpha_i$$

$$0\le\alpha_i^*\le C\quad i=1,\cdots,l,$$

$$0\le\alpha_i\le C\quad i=1,\cdots,l,$$

正如在模式识别情况中一样, 对上述两个问题的解在 $C=C^*$ 时是重合的。

可以证明, 对任何 $i=1,\cdots,l$, 等式

$$\alpha_i^*\times\alpha_i=0$$

成立。因此, 对 $\varepsilon=1-\delta$ (δ很小) 且 $y_i\in\{-1,1\}$ 的情况, 这里考虑的优化问题与在前面对模式识别讨论的问题是等同的。

为了推导 SVM 推广性的界, 假设分布 $F(x,y)=F(y|x)F(x)$ 是这样的: 对任意固定的 w,b, 对应的随机变量 $|y-(w\cdot x)-b|$ 的分布是"轻尾的"(参见 3.4 节), 即

$$\sup_{w,b}\frac{(E|y-(w\cdot x)-b|^p)^{1/p}}{E|y-(w\cdot x)-b|}\le\tau\quad p>2,$$

那么根据(3-30)式, 我们可以断定, 优化问题的解 w_l, b_l 提供了这样一个风险(相对于所选的损失函数), 它以至少 $1-\eta$ 的概率使得界

$$R(w_l,b_l)\le\varepsilon+\frac{R_{emp}(w_l,b_l)-\varepsilon}{(1-a(p)\tau\sqrt{E})_+}$$

成立, 其中

$$a(p)=\sqrt[p]{\frac{1}{2}\left(\frac{p-1}{p-2}\right)^{p-1}},$$

$$E=4\frac{h_n\ln\frac{2l}{h_n}+1-\ln\frac{\eta}{4}}{l}.$$

这里的 h_n 是函数集

$$S_n=\{|y-(w\cdot x)-b|,\ (w\cdot w)\le c_n\}$$

的 VC 维。

6.2.1 采用回旋内积的 SV 机

在第五章中考虑模式识别问题时, 我们将输入向量映射到了高维空间, 在这里, 采用同样的做法, 我们可以构造形式为

$$f(x;v,\beta)=\sum_{i=1}^{N}\beta_iK(x,v_i)+b \tag{6-19}$$

的最优逼近问题, 其中 β_i, $i=1,\cdots,l$ 是标量[①], v_i, $i=1,\cdots,N$ 是向量, $K(\cdot,\cdot)$ 是满足

① 原著误写为 $i=1,\cdots,N$. —— 译者

Mercer 条件的一个给定函数[①]。

1. 对结构固定元素的解

采用凸优化的方法, (6-19)式中的系数 β_i, $i = 1, \cdots, l$ 用下式计算:

$$\beta = \frac{\alpha^* - \alpha}{C^*}, \quad i = 1, \cdots, l,$$

其中的 α^*, α, C 是最大化下列函数的参数:

$$W = -\varepsilon\sum_{i=1}^{l}(\alpha_i^* + \alpha_i) + \sum_{i=1}^{l} y_i(\alpha_i^* - \alpha_i)$$

$$-\frac{1}{2C^*}\sum_{i,j=1}^{l}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(x_i, x_j) - \frac{c_n C^*}{2},$$

约束条件是

$$\sum_{i=1}^{l}\alpha_i^* = \sum_{i=1}^{l}\alpha_i$$

和

$$0 \leq \alpha_i^* \leq 1, \quad i = 1, \cdots, l$$
$$0 \leq \alpha_i \leq 1, \quad i = 1, \cdots, l$$

以及

$$C^* \geq 0.$$

2. 基本解

采用二次优化的方法, 式(6-19)中的系数 β_i, $i = 1, \cdots, l$ 用下式计算:

$$\beta_i = \alpha_i^* - \alpha_i, \quad i = 1, \cdots, l,$$

其中的 α^*, α 是最大化下列函数的参数:

$$W = -\varepsilon\sum_{i=1}^{l}(\alpha_i^* + \alpha_i) + \sum_{i=1}^{l} y_i(\alpha_i^* - \alpha_i)$$

$$-\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(x_i, x_j),$$

约束条件是

$$\sum_{i=1}^{l}\alpha_i^* = \sum_{i=1}^{l}\alpha_i$$

和

$$0 \leq \alpha_i^* \leq C \quad i = 1, \cdots, l,$$
$$0 \leq \alpha_i \leq C \quad i = 1, \cdots, l.$$

通过在二次优化方法中控制 C 和 ε 两个参数, 我们可以控制(即使在高维空间中) SVM 的推广能力。

---

① 本节原文中有多处细小的公式印刷错误和表述错误, 在翻译时根据上下文并参照有关文献进行了适当的更正, 不再一一注明。—— 译者

6.2.2 对非线性损失函数的解

除了线性损失函数外，我们也可以得到凸损失函数 $L(\xi^*)$ 和 $L(\xi)$ 的解。

一般来说，当 $L(\xi)$ 是凸函数[①]时，我们可以用相应的优化技术找到其解。但是，对于二次损失函数 $L(\xi)=\xi^2$ 或者 Huber 的损失函数，我们可以用简单的二次优化技术求解。

1. 二次损失函数[②]

要得到解(在支持向量上超平面的展开系数 α^*，α)，我们必须最大化二次型

$$W(\alpha,\alpha^*)=-\sum_{i=1}^{l}\varepsilon(\alpha_i^*+\alpha_i)+\sum_{i=1}^{l}y_i(\alpha_i^*-\alpha_i)-\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)K(x_i,x_j)+\frac{1}{C}\sum_{i=1}^{l}(\alpha_i^*)^2+\frac{1}{C}\sum_{i=1}^{l}(\alpha_i)^2,$$

约束条件是

$$\sum_{i=1}^{l}\alpha_i^*=\sum_{i=1}^{l}\alpha_i$$

$$0\le\alpha_i^*\quad i=1,\cdots,l,$$

$$0\le\alpha_i\quad i=1,\cdots,l.$$

当 $\varepsilon>0$ 且

$$K(x_i,x_j)=\mathrm{cov}\{f(x_i),f(x_j)\}$$

是随机过程的协方差函数，且

$$Ef(x)=0$$

时，所得的解与在地质统计学中提出的克里金(Kreiging)方法相符(关于地质统计学的内容可参见文献(Matheron, 1987))。

2. Huber 损失函数的解

最后我们考虑对 Huber 损失函数[③]

$$L(\xi)=\begin{cases}c|\xi|-\dfrac{c^2}{2} & \text{对 } |\xi|>c,\\[2mm] \dfrac{1}{2}\xi^2 & \text{对 } |\xi|\le c\end{cases}$$

的 SVM。

对这个损失函数，要找到期望的解

$$f(x)=\sum_{i=1}^{l}(\alpha_i^*-\alpha_i)K(x_i,x)+b,$$

我们必须找到系数 α^*，α，它们最大化二次型

---

① 原著误写为凹函数。—— 译者

② 这里实际上是采用前面提到的二次 ε不敏感损失函数(6-8)式，而不是一般的二次损失函数。—— 译者

③ 下式中 对 $|\xi|>c$，原著中误写为 $|\xi|\le c$；
对 $|\xi|\le c$，原著中误写为 $|\xi|>c$。—— 译者

$$W(\alpha,\alpha^*) = \sum_{i=1}^{l} y_i(\alpha_i^* - \alpha_i) - \frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(x_i,x_j) + \frac{\varepsilon}{C}\sum_{i=1}^{l}(\alpha_i^* + \alpha_i) + \frac{1}{2C}\sum_{i=1}^{l}(\alpha_i^* + \alpha_i)^2,$$

约束条件是

$$\sum_{i=1}^{l}\alpha_i^* = \sum_{i=1}^{l}\alpha_i$$

$$0 \le \alpha_i^* \le C \quad i = 1,\cdots,l,$$

$$0 \le \alpha_i \le C \quad i = 1,\cdots,l.$$

当 $\varepsilon = 0$ 时，对 Huber 损失函数得到的解与对 ε 不敏感损失函数得到的解接近。但是，对 ε 不敏感损失函数得到的解的展开式使用更少的支持向量。

3. 损失函数的样条逼近

如果 $F(\xi)$ 是凹函数，并关于零对称，那么我们可以用线性样条

$$F(\xi) = \sum_{k=1}^{n} c_k(\xi - a_k)_+, \quad 0 < a_1 = \varepsilon < a_2 < \cdots < a_n$$

以任意的精度对它进行逼近。在这种情况下，用在模式识别中对 SV logistic 回归逼近所采用的同样技术，我们可以得到基于二次优化技术的解。

6.2.3 线性优化方法

正如在模式识别中一样，我们可以把这个优化问题进一步简化为一个线性优化任务[①]。假设给定数据是

$$(y_i, x_i), \cdots, (y_l, x_l),$$

我们用集合

$$y(x) = \sum_{i=1}^{l}\beta_i K(x_i, x) + b$$

中的函数进行函数逼近，其中 β 是实数值，x_i 是训练集中的向量，$K(x_i, x)$ 是核函数。我们把训练集中对应非零 β 值的向量称作支持向量。我们把 β 重写为

$$\beta = \alpha_i^* - \alpha_i$$

其中 $\alpha_i^* > 0$，$\alpha_i > 0$。

我们可以用下面优化问题的解来逼近所求的函数，即最小化泛函[②]

$$W(\alpha,\alpha^*,\xi,\xi) = \sum_{i=1}^{l}\alpha_i + \sum_{i=1}^{l}\alpha_i^* + C\sum_{i=1}^{l}\xi_i + C\sum_{i=1}^{l}\xi_i,$$

① 原著中在关于用于模式识别的支持向量机的讨论中并没有提及线性优化的问题。作者在此处这样说，可能是由于本书第二版写作时与第一版已经时隔 4 年，误以为书中前面涉及到了有关内容。就译者掌握的资料，在某些文献中讨论过把用于模式识别的支持向量机简化为线性优化任务的问题，比如 Bradley P S and Mangasarian O L. Massive Data Discriminationvia Linear Support Vector Machines. Mathematical Programming Technical Report 98-05, Univ. of Wisconsin Madison—— 译者

② 原著中误写为 $W(\alpha,\xi)$。—— 译者

约束条件是

$$\alpha \geq 0, \quad \alpha^* \geq 0, \quad i = 1, \cdots, l$$

$$\xi \geq 0, \quad \xi^* \geq 0$$

$$y_i - \sum_{j=1}^{l} (\alpha_j^* - \alpha_j) K(x_i, x_j) - b \leq \varepsilon - \xi_i^*$$

$$\sum_{j=1}^{l} (\alpha_j^* - \alpha_j) K(x_i, x_j) + b - y_i \leq \varepsilon - \xi_i,$$

这一问题的求解只需要用到线性优化技术。

6.3 构造估计实值函数的核

要构造不同类型的 SVM, 我们需要选择满足 Mercer 条件的不同的核 $K(x, x_i)$。

特别地, 我们可以采用与在指示函数逼近中同样的核:

(1) 生成多项式的核

$$K(x, x_i) = [(x \cdot x_i) + 1]^d,$$

(2) 生成径向基函数的核

$$K(x, x_i) = K(\|x - x_i\|),$$

比如

$$K(x, x_i) = \exp\{-\gamma \|x - x_i\|^2\},$$

(3) 生成两层神经网络的核

$$K(x, x_i) = S(v(x \cdot x_i) + c),$$

在这些核的基础上, 我们可以用前面讨论的优化技术得到逼近

$$f(x, \alpha) = \sum_{i=1}^{l} \beta_i K(x, x_i) + b. \tag{6-20}$$

这些核函数意味着逼近函数就是在模式识别问题中采用的符号判别 sgn 中的函数, 即在模式识别中考虑的是 $\mathrm{sgn}[f(x, \alpha)]$。

然而, 与指示函数逼近相比, 实值函数的逼近问题更复杂(由于在函数 $f(x, \alpha)$ 前面没有了 $\mathrm{sgn}(\cdot)$, 逼近问题大大地改变了)。

不同的实值函数估计问题需要不同的逼近函数集。因此, 构造能反映逼近函数特性的特殊的核是十分重要的。

要构造重要的核, 我们采用了两种主要的技术:

(1) 构造逼近一维函数的核,

(2) 用一维核来构成多维核。

6.3.1 生成正交多项式展开的核

要构造这样的核, 使它生成由正交多项式 $P_i(x)$, $i = 1, \cdots, N$ (如 Chebyshev,

Legendre, Hermite 多项式等等) 前 N 项组成的一维函数的展开, 我们可以利用 Christoffel-Darboux 公式

$$K_n(x,y)=\sum_{k=1}^{n}P_k(x)P_k(y)=a_n\frac{P_{n+1}(x)P_n(y)-P_n(x)P_{n+1}(y)}{x-y},$$

$$K_n(x,x)=\sum_{k=1}^{n}P_k^2(x)=a_n[P'_{n+1}(x)P_n(x)-P'_n(x)P_{n+1}(x)], \tag{6-21}$$

其中 a_n 是一个常数, 它依赖于多项式的类型和正交基中元素的序号 n。

然而很明显, 随着 n 的增加, 核 $K(x,y)$ 趋近于 δ函数。但是, 我们可以修正生成核函数, 使之再造一个正则化的函数。考虑核

$$K(x,y)=\sum_{i=1}^{\infty}r_i\psi_i(x)\psi_i(y) \tag{6-22}$$

其中 r_i 随着 i 的增加而趋近于零。这个核定义了正则化的多项式展开。

我们可以选择适当的 r_i 值, 使得它们能改善序列(6-22) 式的收敛特性。比如可以选择 $r_i=q^i, 0\leq q\leq 1$。

例　考虑(一维的) Hermite 多项式

$$H_k(x)=\mu P_k(x)e^{-x^2}, \tag{6-23}$$

其中

$$P_k(x)=(-1)^k e^{x^2}\frac{d}{dx}{}^k e^{-x^2},$$

μ 是归一化常数。

对这些多项式, 我们可以得到下面的核(Mikhlin, 1964):

$$K(x,y)=\sum_{i=0}^{\infty}q^iH_i(x)H_i(y)$$

$$=\frac{1}{\sqrt{\pi(1-q^2)}}\exp\left(\frac{2xyq}{1+q}-\frac{(x-y)^2q^2}{1-q^2}\right). \tag{6-24}$$

从(6-24) 式我们可以看到, q 越接近 1, 核 $K(x,y)$ 越接近 δ函数。

为了构造核, 我们甚至不一定使用正交基。在下一节中构造用于样条逼近的核时, 我们将使用线性无关但不正交的基。

这种一般性(有任意平滑参数的任意线性无关系统) 为构造 SVM 的核提供了广泛的机会。

6.3.2　构造多维核

然而, 我们的目的是构造用于逼近定义在向量空间 $X\subset R^n$ 上的多维函数的核, 向量 $x=(x^1,\cdots,x^n)$ 的所有坐标都定义在相同的有限或无限区间 I 上。

现在假设对任意坐标 x^k 都给出了完备正交基 $b_{i_k}(x^k)$, $i=1,2,\cdots$。考虑 n 维空间中下面的基函数集合

$$b_{i_1,i_2,\cdots,i_n}(x^1,\cdots,x^n)=b_{i_1}(x^1)b_{i_2}(x^2)\cdots b_{i_n}(x^n). \tag{6-25}$$

这些函数是由每个坐标的基函数通过它们的直积(张量积) 构造的, 其中所有索引 i_k 都取

从 0 到∞的所有可能值。可以知道,函数集(6-25)式是一个 $X \quad R^n$ 中的完备正交基。

现在我们来考虑更一般的情况,即一个(有限或无限的)单坐标基函数的集合,它并不一定是正交的。考虑把每个坐标的基的张量积作为 n 维空间中的一个基。

对于这种多维空间结构,有下面的定理成立。

定理 6.1　设一个多维函数集合是由作为单位坐标基函数之张量积的基函数定义的,那么定义了这个 n 维基上内积的核是一维核的积。

续例　现在我们来对 n 维 Hermite 多项式的正则化展开构造一个核。在上面讨论的例子中,我们对一维 Hermite 多项式构造了一个核。根据定理 6.1,如果我们把一维基函数的张量积作为 n 维空间的一个基,那么生成 n 维展开的核是 n 个一维核的积

$$K(x,y)=\prod_{i=1}^{n}\frac{1}{\pi(1-q^2)}\exp\left\{\frac{2x^iy^iq}{1+q}-\frac{(x^i-y^i)^2q^2}{1-q^2}\right\}$$

$$=\frac{1}{(1-q^2)^{n/2}}\exp\left\{\frac{2(x\cdot y)q}{1+q}-\frac{\|x-y\|^2q^2}{1-q^2}\right\}. \tag{6-26}$$

这样,我们得到了一个用于构造半局部逼近的核:

$$K(x,y)=C\exp\{2(x\cdot y)\delta\}\exp\{-|x-y|^2\sigma^2\},\quad \delta,\sigma>0. \tag{6-27}$$

其中,由于 Gaussian 函数定义了逼近的近邻,带有两个向量的内积乘子定义了"全局的"逼近。

6.4　生成样条的核

下面我们介绍可以用来构造高维函数的样条逼近的核。我们将构造有固定数目结点的样条和有无穷多结点的样条。在两种情况下,解的运算复杂度都取决于以 ε精度逼近待求函数所需的支持向量数目,而不是取决于空间的维数或者结点的数目。

图 6.2　用函数 $1, x, (x-t_1)_+, \cdots, (x-t_m)_+$ 的展开,我们可以构造对一个函数的分段线性逼近。类似地,函数 $1, x, \cdots, x^d, (x-t_1)_+^d, \cdots, (x-t_m)_+^d$ 的展开提供了分段多项式逼近

6.4.1　d 阶有限结点的样条

首先讨论用有 m 个结点的 $d\geq 0$ 阶样条逼近区间 $[0,a]$ 上的一维函数的核,这 m 个结

点是

$$(t_1, \cdots, t_m), \quad t_i = \frac{i}{m}, \quad i = 1, \cdots, m.$$

根据其定义,样条逼近有如下的形式:

$$f(x) = \sum_{r=0}^{d} a_r^* x^r + \sum_{i=1}^{m} a_i (x - t_i)_+^d. \tag{6-28}$$

考虑下面的从一维变量 x 到 m+ d+ 1 维向量 u 的映射:

$$x \to u = (1, x, \cdots, x^d, (x - t_1)_+^d, \cdots, (x - t_m)_+^d),$$

其中我们记

$$(x - t_k)_+^d = \begin{cases} 0 & \text{若 } x \leq t_k \\ (x - t_k)^d & \text{若 } x > t_k. \end{cases}$$

因为(6-28)式的样条逼近可以看作是两个向量的内积

$$f(x) = (a \cdot u)$$

(其中 $a = (a_0, \cdots, a_{m+d})$),我们可以定义生成特征空间中内积的核如下①:

$$K(x, x_t) = (u \cdot u_t) = \sum_{r=0}^{d} x^r x_t^r + \sum_{i=1}^{m} (x - t_i)_+^d (x_t - t_i)_+^d. \tag{6-29}$$

采用生成核(6-29)式,SVM 构造出下面的函数:

$$f(x, \beta) = \sum_{i=1}^{l} \beta K(x, x_i) + b,$$

即在 m 个结点上的 d 阶样条。

要构造生成 n 维空间中样条的核,需注意到 n 维空间被定义作一些基函数的展开,而这些基函数是一维基函数的张量积。因此,根据定理 6.1,生成 n 维样条的核是 n 个一维核的积:

$$K(x, x_i) = \prod_{k=1}^{n} K(x^k, x_i^k),$$

其中我们记 $x = (x^1, \cdots, x^k)$。

6.4.2 生成有无穷多结点的样条的核

在 SVM 的应用中,结点的数目并不起十分重要的作用(更重要的是 ε 值)。因此,为了简化计算,我们采用有无穷多结点的样条,它是定义在区间(0, a), $0 < a < \infty$ 上的展开

$$f(x) = \sum_{i=0}^{d} a_i x^i + \int_0^a a(t)(x - t)_+ \, dt,$$

---

① 原文中式(6-29)为

$$K(x, x_t) = (u^* u_t) = \sum_{r=0}^{d} x^r x_i^r + \sum_{i=1}^{m} (x - t_i)_+^d (x_t - t_i)_+^d,$$

译者认为有误,第一个求和项中第二个元素的下标应为 t。另外,在本章中的多处公式中,原著采用符号"*"表示内积运算,这与本书前面的表达不一致,因此我们在翻译时都改为用"·"表示而不一一注明。——译者

其中, a_i, $i = 0, \cdots, d$ 是未知值, $a(t)$ 是定义展开式的未知函数。我们可以把这个展开式看作内积。因此, 可以构造下面的核来生成有无穷多结点的 d 阶样条, 然后在这个空间中采用下面的内积:

$$\begin{aligned} K(x_j, x_i) &= \int_0^a (x_j - t)_+^d (x_i - t)_+^d \, dt + \sum_{r=0}^{d} x_j^r x_i^r \\ &= \int_0^{(x_j \wedge x_i)} (x_j - t)_+^d (x_i - t)_+^d \, dt + \sum_{r=0}^{d} x_j^r x_i^r \\ &= \int_0^{(x_j \wedge x_i)} u^d (u + |x_j - x_i|)^d \, du + \sum_{r=0}^{d} x_j^r x_i^r \text{①} \\ &= \sum_{r=0}^{d} \frac{C_d^r}{2d - r + 1} (x_j \wedge x_i)^{2d - r + 1} |x_j - x_i|^r + \sum_{r=0}^{d} x_j^r x_i^r, \end{aligned} \tag{6-30}$$

其中, 我们记 $(x \wedge x_i) = \min(x, x_i)$。特别地, 对于线性样条 $(d = 1)$, 我们有

$$K_1(x_j, x_i) = 1 + x_j x_i + \frac{1}{2} |x_j - x_i| (x_j \wedge x_i)^2 + \frac{(x_j \wedge x_i)^3}{3}.$$

同样, 有无穷多结点的 n 维空间样条的核是 n 个一维样条的核的乘积。

在这个核的基础上, 我们可以构造形式为

$$f(x, \beta) = \sum_{i=1}^{l} \beta K(x, x_i)$$

的样条逼近(用上一节介绍的技术)。

6.5 生成傅里叶展开的核

傅里叶展开在信号处理中起着很重要的作用。在这一节里, 我们将构造在多维空间中 SV 傅里叶展开的核。像前面一样, 我们仍从一维情况开始讨论。

假设我们要用傅里叶级数展开来分析一个一维信号。

我们把输入变量 x 映射成下面的 $2N + 1$ 维向量:

$$u = (1/\sqrt{2}, \sin x, \cdots, \sin(Nx), \cos x, \cdots, \cos(Nx)),$$

那么, 对任意固定的 x, 其傅里叶展开可以考虑为在这个 $2N + 1$ 维空间中的内积, 即

$$f(x) = (a \cdot u) = \frac{a}{2} + \sum_{k=1}^{N} (a_k \sin(kx) + b_k^* \cos(kx)). \tag{6-31}$$

因此, 在这个空间中, 两个向量的内积有如下的形式:

$$K_N(x, x_i) = \frac{1}{2} + \sum_{k=1}^{N} (\sin(kx) \sin(kx_i) + \cos(kx) \cos(kx_i)).$$

经过显而易见的变换, 并考虑到 Dirichlet 函数后, 我们得到

---

① 原著这一步中的求和是从 $r = 1$ 开始, 译者认为是印刷错误。—— 译者

$$K_N(x, x_i) = \frac{1}{2} + \sum_{k=1}^{N} \cos(k(x - x_i)) = \frac{\sin\frac{2N+1}{2}(x - x_i)}{\sin\frac{x - x_i}{2}}.$$

要用傅里叶展开来定义信号, SVM 采用下面的表达:

$$f(x, \beta) = \sum_{i=1}^{l} \beta K_N(x, x_i).$$

同样地, 要构造 n 维向量空间 $x = (x^1, \cdots, x^n)$ 中的 SVM, 只要用一维核的乘积

$$K(x, x_i) = \prod_{k=1}^{n} K(x^k, x_i^k)$$

作为生成核即可。

正则傅里叶展开的核

然而我们知道, 傅里叶展开并不具有好的逼近特性。因此, 下面我们介绍两种正则化核, 用它们来实现 SVM 对多维函数的逼近。

考虑下面的正则傅里叶展开

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} q^k (a_k \cos(kx) + b_k \sin(kx)), \quad 0 < q < 1,$$

其中, a_k, b_k 是傅里叶展开的系数。这个展开与(6-31)式的展开不同之处在于多了一个乘子 q^k, 它起到正则化作用。与这一正则化展开相应的核是

$$K(x_i, x_j) = \frac{1}{2} + \sum_{k=1}^{\infty} q^k (\sin(kx_i)\sin(kx_j) + \cos(kx_i)\cos(kx_j))$$

$$= \frac{1}{2} + \sum_{k=1}^{\infty} q^k \cos(k(x_i - x_j)) = \frac{1 - q^2}{2(1 - 2q\cos(x_i - x_j) + q^2)}. \quad (6\text{-}32)$$

(公式的最后一个等式参见文献(Gradshtein and Ryzhik, 1980))。我们得到的另外一种正则化方法是下面的正则化傅里叶展开

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} \frac{a_k \cos(kx) + b_k \sin(kx)}{1 + \gamma^2 k^2},$$

其中 a_k, b_k 是傅里叶展开系数。对这种正则傅里叶展开, 我们有下面的核:

$$K(x_i, x_j) = \frac{1}{2} + \sum_{k=1}^{\infty} \frac{\cos(kx_i)\cos(kx_j) + \sin(kx_i)\sin(kx_j)}{1 + \gamma^2 k^2}$$

$$= \frac{\pi}{2\gamma} \frac{\mathrm{ch}\dfrac{\pi - |x_i - x_j|}{\gamma}}{\mathrm{sh}\dfrac{\pi}{\gamma}}, \quad 0 \le |x_i - x_j| \le 2\pi \quad (6\text{-}33)$$

(公式的最后一个等式参见文献(Gradshtein and Ryzhik, 1980))。

不同 q 值和不同 γ值的正则化的核见图 6.3、图 6.4。

同样, 多维傅里叶展开的核是一维傅里叶展开核的乘积。

图 6.3　不同 q 值的强方式正则化的核

图 6.4　不同 γ值的弱方式正则化的核

6.6　用于函数逼近和回归估计的支持向量 ANOVA 分解(SVAD)

在前面几节中定义的核既可以被用来逼近多维函数,也可以被用来估计多维回归。它们可以定义的函数集过于丰富了,因此,为了控制推广性,我们需要构造这个函数集上的一个结构,以便从这个结构的某个适当的元素中选择函数。还要注意,当输入空间的维数很高(比如 100 维)时,n 维核(它是 n 个一维核的乘积)的值可能达到 q^n 的数量级。这些值不论对 q> 1 还是 q< 1 都不合适。

传统统计学考虑了源自 L_2 的多维函数集上的下面的结构,即所谓 ANOVA 分解(ANOVA 是方差分析(analysis of variances)的缩写)。

假设 n 维函数 $f(x)= f(x^1, \cdots, x^n)$ 是定义在集合 $I\times I\times \cdots\times I$ 上的,其中 I 是一个有限或无限区间。

函数 f(x)的 ANOVA 分解展开为

$$f(x^1, \cdots, x^n)= F_0+ F_1(x^1, \cdots, x^n)+ F_2(x^1, \cdots, x^n)+ \cdots+ F_n(x^1, \cdots, x^n)$$

其中,

$$F_0= C,$$

$$F_1(x^1, \cdots, x^n)= \sum_{1\leq k\leq n}{}_k(x^k),$$

$$F_2(x^1, \cdots, x^n)= \sum_{1\leq k_1<k_2\leq n}{}_{k_1,k_2}(x^{k_1}, x^{k_2}),$$

$$\cdots$$

$$F_r(x^1, \cdots, x^n) = \sum_{1 \le k_1 < k_2 < \cdots < k_r \le n} {}_{k_1, \cdots, k_r}(x^{k_1}, x^{k_2}, \cdots, x^{k_r}),$$

$$F_n(x^1, \cdots, x^n) = {}_{k_1, \cdots, k_n}(x^1, \cdots, x^n).$$

ANOVA 分解的传统方法有一个问题,就是随着逼近阶数的增加,求和项的数目将出现指数爆炸。在支持向量技术中,我们没有这一问题。要用一系列一维核 $K(x^i, x_r^i)$, $i = 1, \cdots, n$ 的乘积

$$K_p(x, x_r) = \sum_{1 \le i_1 < i_2 < \cdots < i_p \le n} K(x^{i_1}, x_r^{i_1}) \times \cdots \times K(x^{i_p}, x_r^{i_p})$$

之和来构造用于 p 阶 ANOVA 分解的核,我们可以引入一个循环过程来计算 $K_p(x, x_r)$, $p = 1, \cdots, n$。

我们记

$$K^s(x, x_r) = \sum_{i=1}^{n} K^s(x^i, x_r^i).$$

容易验证,核 $K_p(x, x_r)$, $p = 1, \cdots, n$ 可以用下面的循环过程定义:

$$K_0(x, x_r) = 1,$$

$$K_1(x, x_r) = \sum_{1 \le i \le n} K(x^i, x_r^i) = K^1(x, x_r),$$

$$K_2(x, x_r) = \sum_{1 \le i_1 < i_2 \le n} K(x^{i_1}, x_r^{i_1}) K(x^{i_2}, x_r^{i_2})$$

$$= \frac{1}{2}[K_1(x, x_r) K^1(x, x_r) - K^2(x, x_r)],$$

$$K_3(x, x_r) = \sum_{1 \le i_1 < i_2 < r_3 \le n} K_1(x^{i_1}, x_r^{i_1}) K_2(x^{i_2}, x_r^{i_2}) K(x^{i_3}, x_r^{i_3})$$

$$= \frac{1}{3}[K_2(x, x_r) K^1(x, x_r) - K_1(x, x_r) K^2(x, x_r) + K^3(x, x_r)],$$

一般情况下,我们有[①]:

$$K_p(x, x_r) = \frac{1}{p} \sum_{s=1}^{p} (-1)^{s+1} K_{p-s}(x, x_r) K^s(x, x_r).$$

利用这样的核,并采用 L_2 损失函数的 SVM,我们可以得到任意阶次的逼近。

6.7 求解线性算子方程的 SVM

在这一节里,我们用 SVM 来求解线性算子方程

$$Af(t) = F(x), \tag{6-34}$$

其中,算子 A 实现了一个从希尔伯特空间 E_1 到希尔伯特空间 E_2 的一对一映射。

我们将在下面的情况下求解这个方程,即不知道方程(6-34)右边的函数 $F(x)$,而是

---

① 见 Burges C 和 Vapnik V. A New Method for Constructing Artificial Neural Networks. Interim Technical Report ONR Contract N00014-94-C-0186 Data Item A002. May 1, 1995

给出了对这个函数的一些观测(一般情况下是带有误差的观测):

$$(x_1, F_1), \cdots, (x_l, F_l). \tag{6-35}$$

我们需要从数据(6-35)式估计方程(6-34)的解。

下面我们将说明,支持向量技术实现了求解不适定问题的传统思想,其中对核的选择等价于选择正则化泛函。利用这一技术,我们可以求解高维空间中的算子方程。

支持向量方法

在下一章里,我们将讨论求解算子方程的正则化方法,在这些方法中,要求解算子方程(6-34),需最小化泛函

$$R_\gamma(f, F) = \rho^2(Af, F) + \gamma W(f),$$

其中,解属于某个紧的 $W(f) \leq C$(C 是未知常数)。当用数据(6-35)式求解算子方程(6-34)时,我们考虑泛函

$$R_\gamma(f, F) = \frac{1}{l}\sum_{i=1}^{l} L(Af(t)\big|_{t_i} - F_i) + \gamma(Pf \cdot Pf),$$

其中采用某种损失函数 $L(Af - F)$ 和形式为

$$W(f) = (Pf \cdot Pf)$$

的正则化项,它是由某个非生成算子 P 定义的。令

$$\varphi(t), \cdots, \varphi(t), \cdots$$
$$\lambda, \cdots, \lambda, \cdots$$

是自共轭算子 $P \cdot P$ 的本征函数和本征值,即

$$P \cdot P\varphi = \lambda\varphi$$

把方程(6-34)的解看作下面的展开:

$$f(t) = \sum_{k=1}^{\infty} \frac{w_k}{\lambda}\varphi(t).$$

把这个展开代入到泛函 $R_\gamma(f, F)$ 中,我们得到

$$R_\gamma(f, F) = \frac{1}{l}\sum_{i=1}^{l} L\left(A\sum_{k=1}^{\infty} \frac{w_k}{\lambda}\varphi(t)\right|_{x_i} - F_i + \gamma\sum_{k=1}^{\infty} w_k^2.$$

记

$$_k(t) = \frac{\varphi(t)}{\lambda},$$

我们可以用我们熟悉的形式把问题重新表示如下:在函数集

$$f(t, w) = \sum_{r=1}^{\infty} W_r\ _r(t) = (w \cdot \Phi(t)) \tag{6-36}$$

中最小化泛函

$$R_\gamma(f, F) = \frac{1}{l}\sum_{i=1}^{l} L\left|A(w \cdot \Phi(t))\right|_{x=x_i} - F_i\Big| + \gamma(w \cdot w),$$

其中我们记

$$w = (w_1, \cdots, w_N, \cdots),$$

$$\Phi(t) = (\ _1(t), \cdots, \ _N(t), \cdots). \tag{6-37}$$

算子 A 把函数集(6-36)式映射到函数集

$$F(x, w) = Af(t, w) = \sum_{r=1}^{\infty} w_r A\ _r(t) = \sum_{r=1}^{\infty} w_r \psi(x) = (w \cdot \Psi(x)), \tag{6-38}$$

它在另一个特征空间

$$\Psi(t) = (\psi(t), \cdots, \psi(t), \cdots)$$

中是线性的,其中,

$$\psi(x) = A\ _r(t).$$

要在函数集 $f(t, w)$ 中找到方程(6-34)的解(找到系数向量 w),我们可以在函数 $F(x, w)$ 的空间中(即在像空间中)最小化泛函

$$D(F) = C\sum_{i=1}^{l} (F(x_i, w) - F_i)^k + (w \cdot w), \quad k = 1, 2,$$

然后用所得到的参数 w 定义解(6-36)式(在原像空间中)。为了实现这一思想,我们与核函数一起使用所谓的交叉核函数(cross-kernel function)。让我们定义像空间中的生成核为

$$K(x_i, x_j) = \sum_{r=1}^{\infty} \psi(x_i)\psi(x_j) \tag{6-39}$$

(这里我们假设该式右边对任意固定的 x_i 和 x_j 都收敛),定义交叉核函数为

$$K(x_i, t) = \sum_{r=1}^{\infty} \psi(x_i)\ _r(t) \tag{6-40}$$

(这里我们假设算子 A 使得该式右边对任意固定的 x_i 和 t 都收敛)。

注意,在所考虑的情况中,寻找算子方程的解(寻找对应的系数向量 w)的问题与在像空间中利用测量值(6-35)式寻找线性回归函数(6-38)式的向量 w 的问题是等价的。

让我们用二次优化支持向量技术来求解这个回归问题。这就是,用核(6-39)式求出支持向量 x_i, $i = 1, \cdots, N$ 和相应的系数 $\alpha^* - \alpha$,它们定义了支持向量回归逼近的向量 w:

$$w = \sum_{i=1}^{N} (\alpha^* - \alpha)\Psi(x_i)$$

(这一步只需要利用标准的二次优化支持向量技术就足够了)。因为同样的系数 w 定义了对算子方程的解的逼近,我们可以把这些系数代入到表达式(6-36)中,从而得到

$$f(t, \alpha, \alpha^*) = \sum_{i=1}^{N} (\alpha^* - \alpha)(\Psi(x_i) \cdot \Phi(t)) = \sum_{i=1}^{N} (\alpha^* - \alpha)K(x_i, t).$$

也就是,我们利用交叉核函数在支持向量上的展开,找到了所研究的求解算子方程问题的解决方案。

因此,为了用支持向量方法求解线性算子方程,我们需要进行以下步骤:

(1) 定义在像空间中相应的回归问题。

(2) 构造用支持向量方法求解这个回归问题的核函数 $K(x_i, x_j)$。

(3) 构造对应的交叉核函数 $K(x_i, t)$。

(4) 采用支持向量方法,用核函数 $K(x_i, x_j)$ 求解回归问题(即找到支持向量 x_i, $i = 1, \cdots, N$ 和相应的系数 $\beta = \alpha^* - \alpha$, $i = 1, \cdots, N$)。

(5) 用这些支持向量和相应的系数确定解

$$f(t) = \sum_{i=1}^{N} \beta K(x_i, t). \tag{6-41}$$

在这 5 个步骤中,前 3 个步骤(构造回归问题、构造像空间中的核函数以及构造对应的交叉核函数)考虑了我们所面对的问题中的奇异性(它们是依赖于算子 A 的),而后两个步骤(用 SVM 求解回归问题和构造对所求问题的解)是常规的。

用支持向量技术求解算子方程,主要问题在于,对给定的算子方程,要得到在像空间中核函数的显式表达和对应的交叉核函数的显式表达。对于很多问题来说,比如密度估计问题或求解 Radon 方程的问题,这些方程是容易找到的。

6.8 用 SVM 进行函数逼近

下面我们考虑用 SVM 求解函数逼近问题的一些例子。在要求的精度水平 ε 下,我们要逼近由在均匀网格 $x_i = ia/l$ 上的值

$$(y_1, x_1), \cdots, (y_l, x_l)$$

定义的一维和二维函数。我们的目的是要通过实验说明,构造 SV 逼近所使用的支持向量的数目取决于所要求的精度 ε 逼近精度越低则需要的支持向量越少。

在这一节里,为了逼近实值函数,我们采用有无穷多结点的线性样条。

首先,我们将介绍逼近一维 sinc 函数

$$f(x) = \frac{\sin(x - 10)}{x - 10} \tag{6-42}$$

的例子,它是由区间 $0 \leq x \leq 200$ 上的 100 个均匀网格点定义的。

然后,我们逼近定义在 $0 \leq x \leq 20$, $0 \leq y \leq 20$ 内的均匀网格点上的二维 sinc 函数

$$f(x) = \frac{\sin\sqrt{(x - 10)^2 + (y - 10)^2}}{\sqrt{(x - 10)^2 + (y - 10)^2}}. \tag{6-43}$$

为了构造一维线性样条逼近,我们采用了第 6.4[①]节中定义的核

$$K_1(x, x_i) = 1 + x_i x + \frac{1}{2}|x - x_i|(x \wedge x_i)^2 + \frac{(x \wedge x_i)^3}{3}.$$

得到的逼近形式为

$$y = \sum_{i=1}^{N} (\alpha^* - \alpha) K_1(x, x_i) + b,$$

其中,系数 α^*, α 是求解二次优化问题的结果。

图 6.5 显示了用不同的精度逼近函数(6-42)式的结果。图中,圆圈表示支持向量,小

① 原书错印为 6.3 节。—— 译者.

黑点表示非支持向量。从这些结果我们可以看出,随着逼近精度要求的降低,支持向量的数目减少。

图 6.5　不同精度要求的逼近需要不同数目的支持向量

(图中“39SV/100 total”表示总共 100 个样本中有 39 个支持向量,余类推)

为逼近(6-43)式的二维 sinc 函数,我们用的核是

$$K(x,y;x_i,y_i)=K(x,x_i)K(y,y_i)$$

$$=\left(1+xx_i+\frac{1}{2}|x-x_i|(x\wedge x_i)^2+\frac{(x\wedge x_i)^3}{3}\right)$$

$$\times\left(1+yy_i+\frac{1}{2}|y-y_i|(y\wedge y_i)^2+\frac{(y\wedge y_i)^3}{3}\right),$$

它是由两个一维核的乘积定义的。

我们得到的逼近形式为

$$y=\sum_{i=1}^{N}(\alpha_i^*-\alpha_i)K(x,x_i)K(y,y_i)+b,$$

其中,系数 α_i^*, α_i 是通过求解与一维情况下相同的二次优化问题定义的。

图 6.6 显示了在要求精度 $\varepsilon=0.03$ 下对二次 sinc 函数逼近的结果,它们分别是用不同数目的网格点进行的:图(a)中用 400 个网络点构造的逼近需要 153 个支持向量,图(b)中用2 025个网络点构造的逼近需要 234 个支持向量,图(c)用7 921个网络点构造的逼近

需要 285 个支持向量。可以看到,图(c)的网络点约等于图(a)网络点的 20 倍,而图(c)的支持向量数不到图(a)支持向量数的 2 倍。

图 6.6　在同一精度($\varepsilon = 0.03$)下对二次 sinc 函数逼近的结果
(图中"153SV/400 total"表示总共 400 个样本中有 153 个支持向量,余类推)

为什么 ε值控制支持向量数目?

下面的模型描述了用 ε不敏感损失函数的 SV 机为函数逼近选择支持向量的机制。这一机制解释了为什么 ε的选择控制了支持向量的数目。

假设我们要以精度 ε逼近函数 $f(x)$,即用另一个函数 $f^*(x)$来描述函数 $f(x)$,使得函数 $f(x)$处在 $f^*(x)$的 ε管道内。要构造这样一个函数,我们取一个弹性 ε管道(管道总是趋向于平坦)并把函数 $f(x)$放到这个 ε管道中。因为弹性管道总趋向于变平,因此它会碰到 $f(x)$的一些点。我们在这些点上把管道固定住。于是这个管道的轴线就定义了函数 $f(x)$的 ε逼近 $f^*(x)$,管道碰到函数 $f(x)$处的点的坐标定义了支持向量。核 $K(x_i, x_j)$则描述了管道的弹性规律。

的确,既然函数 $f(x)$在这个 ε管道内,函数中就不会有点与轴线的距离大于 ε 因此这个轴线描述了所求的逼近。

要证明那些接触点定义了支持向量,只需注意到下面一点即可:我们是通过求解在 6.2 节中定义的一个优化问题来得到逼近的,对这个优化问题 Kuhn-Tucker 条件成立。根据定义,支持向量就是那些在 Kuhn-Tucker 条件中拉格朗日乘子不为零,因而第二项

乘子必须为零的向量。这个乘子定义了在一个不等式类型的优化问题中边界点,亦即函数 f(x)接触到 ε管道处的坐标。ε管道越宽,接触点就越少。

这个模型对任意维的函数逼近问题都是有效的。它解释了为什么随着 ε不敏感性的增加,支持向量的数目减少。

图 6.7 显示出了一个 ε管道逼近,它对应以精度 ε= 0.2 逼近一维 sinc 函数的情况。读者可以把此图与图 6.5(d)进行比较。

图 6.7 函数逼近的 ε管道模型(ε= 0.2)

6.9 用于回归估计的 SVM

在这一节中,我们首先讨论一些简单的回归估计任务的例子,其中的回归函数是由一维和二维的 sinc 函数来定义的。之后,我们考虑用 SVM 估计多维线性回归函数的问题。我们将构造一个十分适合采用特征选择方法的线性回归任务,并将前向特征选择方法所得结果与 SVM 所得结果进行比较。最后,我们将就一些问题比较支持向量回归方法与一些新的非线性技术,这些问题包括 J. Friedman 提出的三个人造的多维问题和一个多维的实际问题(波士顿住房问题)(这些问题往往被用来作为研究不同回归估计方法的基准)。

6.9.1 数据平滑的问题

设数据集

$$(y_1, x_1), \cdots, (y_l, x_l)$$

是由区间[-10, 10]上的一维 sinc 函数定义的;y_i 的值受到正态分布噪声的影响:

$$y_i = \frac{\sin x}{x} + \xi, \quad E\xi = 0, \quad E\xi^2 = \sigma^2.$$

问题是,要从区间[-10, 10]内均匀网格上的 100 个这样的观测数据出发估计回归函数

$$y = \frac{\sin x}{x}.$$

图 6.8 和图 6.9 显示了 SV 回归估计实验的结果,它们是从受到不同程度噪声污染的

数据得到的。逼近是用有无穷多结点的线性样条得到的。注意到, 图 6.8 中两种情况下逼近所用的支持向量数目大致相同; 图 6.9 中不同 ε值情况下逼近所用的支持向量数不同。

图 6.8　回归函数和对它的逼近

(图中“15 SV/100 total”表示总共 100 个样本中有 15 个支持向量, 余类推)

图 6.9　回归函数和对它的逼近

(图中“14 SV/100 total”表示总共 100 个样本中有 14 个支持向量, 余类推)

图 6.10～图 6.12 显示了对定义在区间[-5, 5]×[-5, 5]内的均匀网格上的二维回归函数 $y=\frac{\sin\sqrt{x^2+y^2}}{\sqrt{x^2+y^2}}$ 的逼近结果。逼近是用有无穷多结点的二维线性样条得到的。

6.9.2　线性回归函数估计

下面我们描述用 SVM 进行线性回归函数估计的实验(Drucker et al, 1997)。

我们把 SVM 与两种不同的线性回归函数估计方法进行比较, 这两种方法是普通的最小二乘方法(OLS)和前向单步特征选择(FSFS)方法。

OLS 方法是通过最小化泛函

图 6.10　对回归函数的逼近

(σ= 0.1, ε= 0.15, 107 SV/ 400 total)

图 6.11　对回归函数的逼近

(σ= 0.1, ε= 0.25, 159 SV/ 3969 total)

图 6.12　对回归函数的逼近

(σ= 0.1, ε= 0.15, 649 SV/ 3969 total)

$$R(a) = \sum_{i=1}^{l} (y_i - (a \cdot x_i))^2$$

来估计线性回归函数的系数的。而 FSFS 是首先选择向量中给出对数据最佳逼近的一个坐标,然后固定这个坐标,加入第二个坐标,使得这两个坐标对数据有最佳的逼近,依此类推。人们采用某些技术来选择适当的坐标数目。

我们考虑在 30 维向量空间 $x=(x^{(1)}, \cdots, x^{(30)})$ 中根据数据

$$(y_1, x_1), \cdots, (y_l, x_l)$$

进行线性回归估计的问题,其中的回归函数只依赖于 3 个坐标

$$y(x) = 2x^{(1)} + x^{(2)} + x^{(3)} + 0\sum_{k=4}^{30} x^{(k)},$$

且数据是从随机选择的 x 点上对这个函数进行的测量。测量是带有加性噪声的:

$$y = y(x_i) + \xi,$$

噪声与 x_i 独立。

表 6.1 描述了在不同的信噪比、不同的噪声模型下,用 60 个观测数据,用上述三种方法对这个回归函数进行估计的实验结果。表中的数据是 100 次实验的平均。这个表说明了,对大噪声(小信噪比)情况,SV 回归给出的结果与 FSFS 方法接近(这个模型特别适合 FSFS 方法),远好于 OLS 方法。

用模型

$$y_i = \sum_{k=4}^{30} x_i^{(k)} + \xi$$

所做的实验则表明,SV 技术在表 6.1 中列出的各种信噪比水平下都有优势。

表 6.1 OLS, FSFS 和 SV 方法的结果对比

信噪比	正态分布			拉普拉斯分布			均匀分布		
	OLS	FSFS	SV	OLS	FSFS	SV	OLS	FSFS	SV
0.8	45.8	28.0	29.3	40.8	24.5	25.4	39.7	24.1	28.1
1.2	20.0	12.8	14.9	18.1	11.0	12.5	17.6	11.7	12.8
2.5	4.6	3.1	3.9	4.2	2.5	3.2	4.1	2.8	3.6
5.0	1.2	0.77	1.3	1.0	0.60	0.52	1.0	0.62	1.0

6.9.3 非线性回归函数估计

在非线性回归函数估计的实验中,我们选择了 J. Friedman 提出的回归函数,这些回归函数被用在很多基准研究中。

(1) Friedman 的 1 号目标函数是一个有 10 个名义变量的函数

$$y = 10\sin(\pi x^{(1)} x^{(2)}) + 20(x^{(3)} - 0.5)^2 + 10x^{(4)} + 5x^{(5)} + \xi \qquad (6\text{-}44)$$

但它只依赖于其中的 5 个变量。在这个模型中,10 个变量是在[0, 1]中均匀分布的,噪声是正态的且参数为 N(0, 1)。

(2) Friedman 的 2 号目标函数是

$$y=\sqrt{(x^{(1)})^2+[x^{(2)}x^{(3)}-1/(x^{(2)}x^{(3)})]^2},$$

它有 4 个独立变量,它们均匀分布在下面的区间内:

$$\begin{aligned}0 &\le x^{(1)} \le 100\\ 40\pi &\le x^{(2)} \le 560\pi\\ 0 &\le x^{(3)} \le 1\\ 0 &\le x^{(4)} \le 11\end{aligned} \tag{6-45}$$

噪声被调节为信噪比 3∶1。

(3) Friedman 的 3 号目标函数也有 4 个独立变量:

$$y=\tan^{-1}\frac{x^{(2)}x^{(3)}-1/(x^{(2)}x^{(4)})}{x^{(1)}}+\xi \tag{6-46}$$

4 个变量均匀分布在与(6-45)式相同的区间内,噪声水平调节为 3∶1 的信噪比。

下面我们来对两种先进的回归技术与支持向量回归机器进行比较,这两种先进的回归技术是称作 bagging(装袋)(L. Brieman, 1996)和 AdaBoost(自举)的技术[①],它们通过组合由回归树得到的解构造出不同类型的投票机器[②]。

实验是按照与文献(Drucker, 1997)和(Drucker et al, 1997)中相同的方式进行的。

表 6.2 给出了用 bagging、自举和多项式 SVM(d=2)估计 Friedman 函数的实验结果。实验是用 240 个训练样本进行的。表 6.2 给出了模型误差的平均(对 10 次实验的平均),模型误差是用真实目标函数与所得逼近之间的均方偏差定义的。

表 6.2 对 3 种人造数据集的回归比较

	Bagging	Boosting	SV
Friedman # 1	2.2	1.65	0.67
Friedman # 2	11.463	11.684	5.402
Friedman # 3	0.0312	0.0218	0.026

表 6.3 给出了对所谓波士顿住房数据集得到的结果,其中有 506 个 13 维实际数据样本,在实验中,有 401 个随机抽选的样本被作为训练集,80 个作为确认集,25 个作为测试集。表 6.3 给出的是 100 次实验的平均结果。其中,SV 机构造了基于确认集选择的多项式(多数是 4 或 5 阶)。对这些波士顿住房数据,表中的性能指标是测试集上的实际 y 值与预测值之间的均方误差。

表 6.3 对波士顿住房数据用不同方法的性能比较

Bagging	Boosting	SV
12.4	10.7	7.2

① AdaBoost 算法是在模式识别问题中提出的(见第 5.10 节),被 H. Drucker(1997)用到回归估计中来。

② 这里原著表述比较混乱,译者参考作者以前的手稿进行了适当的调整。——译者

非正式推导和评述——6

6.10 回归估计问题中的损失函数

对基于经验数据估计函数依赖关系的方法的研究已经有很长的历史了。这些研究是由两个伟大的数学家开始的: 他们是高斯(Gauss, 1777—1855)和拉普拉斯(Laplace, 1749—1827), 他们提出了从天文学和物理学中的观测结果估计依赖关系的两种不同方法。

高斯提出了最小二乘方法(least squared method, LSM), 而拉普拉斯提出了最小模方法(least modulo method, LMM)。从那时起就有了下面的问题: 哪种方法更好呢? 在19世纪和20世纪初, 人们更倾向于最小二乘方法: 这种方法对线性方程有闭合形式的解。而且人们也证明了, 在线性无偏估计中, LSM是最好的。

后来, 人们注意到, 在很多情况下, 线性无偏估计的集合往往过于局限, 无法保证在这个集合中的最佳估计真正是好的(很可能整个集合中只包含"坏"的估计)。

在20年代, R. Fisher发现了最大似然(ML)方法, 并且提出了带有加性噪声的观测模型。根据这一模型, 在任意点 x^* 上对函数 $f(x, \alpha)$ 的观测都受到了加性噪声的干扰(噪声服从已知的对称密度 $p_0(\xi)$, 且 ξ 与 x^* 无关):

$$y^* = f(x, \alpha) + \xi$$

因为

$$\xi = y - f(x, \alpha),$$

要用最大似然方法从数据

$$x_1, \cdots, x_l$$

估计密度 $p_0(\xi)$(未知函数 $f(x, \alpha)$)的参数 α, 我们必须最大化泛函

$$R_l(\alpha) = \frac{1}{l}\sum_{i=1}^{l} \ln p(y - f(x_i, \alpha)).$$

在 1953 年, L. Le Cam 定义了 ML 方法一致的条件。他发现了(在集合 $\alpha \in \Lambda$ 上)一致收敛的一些充分条件, 在这些条件下, 经验泛函 $R_l(\alpha)$ 收敛到泛函

$$R(\alpha) = \int \ln p(y - f(x, \alpha)) dP(y, x).$$

(这些条件是第二章讨论的充分必要条件的一个特例); 这可以立即引出下面的推断, 即

$$-\int \ln \frac{p(y - f(x, \alpha_l))}{p(y - f(x, \alpha))} \xrightarrow{p} 0, \quad l \to \infty$$

也就是说, ML 解在 Kulbac-Leibler 距离下是一致的。而且, 在无偏估计器集合中(不一定是线性), ML 方法具有最小的方差(具有最小方差的无偏估计被称作是有效的)。

这就意味着, 如果噪声是服从高斯(正态)规律的, 则最小二乘方法给出最好的结果。但是, 如果噪声服从拉普拉斯规律

$$p(x, \Delta) = \frac{1}{2\Delta} \exp\left(-\frac{|x|}{\Delta}\right),$$

则最优解就是最小模估计。从这些结果中也可以得出, 最优(有效)估计中的损失函数是由噪声的分布定义的。

在实际中, (即使测量的加性噪声模型是适用的)噪声的形式往往是未知的。在 60 年代, Tukey 说明了在现实情况中, 噪声的形式与高斯或拉普拉斯规律都相去甚远。

因此, 设计在现实情况中(在噪声形式未知时)估计函数的最佳策略就变得十分重要。P. Huber 提出了这样一种策略, 他创造了所谓鲁棒估计的概念。

6.11 鲁棒估计的损失函数

考虑下面的情况。假设我们的目标是估计随机变量 ξ 的期望 m, 利用的是独立同分布数据

$$\xi_1, \cdots, \xi_l.$$

另外假设相应的未知密度 $p_0(\xi - m_0)$ 是一个平滑函数, 关于位置 m_0 对称、且有二阶矩。

我们知道, 在这种情况下, 最大化

$$L(m) = \sum_{i=1}^{l} \ln p_0(\xi_i - m)$$

的最大似然估计

$$m = \hat{m}(\xi_1, \cdots, \xi_l | p_0)$$

是一个有效估计。这意味着, 在所有可能的无偏估计[①]中, 这个估计取得最小的方差, 或者换句话说, 估计子 $\hat{m}(\xi_1, \cdots, \xi_l | p_0)$ 最小化泛函

$$V(\hat{m}) = \int (\hat{m}(\xi_1, \cdots, \xi_l) - m)^2 dp_0(\xi_1 - m) \cdots dp_0(\xi_l - m). \tag{6-47}$$

下面假设尽管密度 $p_0(\xi - m)$ 未知, 但知道它属于某个容许密度集 $p_0(\xi - m) \in \Omega$。这

① 如果估计子 $\hat{m}(\xi_1, \cdots, \xi_l)$ 满足

$$E\{\hat{m}(\xi_1, \cdots, \xi_l)\} = m$$

则称之为无偏的。

种情况下怎样选择估计子呢? 设未知密度是 $p_0(\xi - m)$。但是,我们构造的估计是对密度 $p_1(\xi - m) \in$ 最优的,即我们定义估计子 $(\xi_1, \cdots, \xi_l)$,使它最大化泛函

$$L_1(m) = \sum_{i=1}^{l} \ln p_1(\xi_i - m). \tag{6-48}$$

现在这个估计的质量取决于两个密度:实际密度 $p_0(\xi - m)$ 和用来构造估计子(6-48)式的密度 $p_1(\xi - m)$[①],评价这个估计子质量的泛函是:

$$V(p_0, p_1) = \int (\quad (\xi_1, \cdots, \xi_l) - m)^2 dp_0(\xi_1 - m) \cdots dp_0(\xi_l - m).$$

Huber 证明,对一个很宽的容许密度集 ,泛函 $V(p_0, p_1)$ 存在一个鞍点。也就是,对任意的容许密度集,存在这样一个密度 $p_r(\xi - m)$,使得不等式

$$V(p, p_r) \leq V(p_r, p_r) \leq V(p_r, p) \tag{6-49}$$

对任意函数 $p(\xi - m) \in$ 都成立。

不等式(6-49)确定了对任意容许密度集,存在一个最小最大密度,即所谓鲁棒密度,它在最坏的情况下保证最小损失。

利用这个鲁棒密度,我们可以构造所谓的鲁棒回归估计。鲁棒回归估计子就是使泛函

$$R_h(w) = - \sum_{i=1}^{l} \ln p_r(y_i - f(x_i, \theta))$$

最小化的估计子。

下面我们给出 Huber 定理的正式表达,这个定理是鲁棒估计理论的一个基础。

考虑密度类 H,它们是由两个密度的混合形成的:

$$p(\xi) = (1 - \varepsilon)g(\xi) + \varepsilon h(\xi),$$

其中 $g(\xi)$ 是某个固定的密度,$h(\xi)$ 是一个任意密度,它们都关于原点对称。混合的权值分别是 $(1-\varepsilon)$ 和 ε 对这种类型的密度,下面的定理成立。

定理(Huber) 设 $-\lg(\xi)$ 是一个二次连续可微函数,那么类 H 中包含下面的鲁棒密度:

$$p_r(\xi) = \begin{cases} (1-\varepsilon)g(\xi)\exp\{-c(\xi - \xi_0)\}, & \text{对 } \xi < \xi_0 \\ (1-\varepsilon)g(\xi), & \text{对 } \xi_0 \leq \xi < \xi_1 \\ (1-\varepsilon)g(\xi)\exp\{-c(\xi - \xi_1)\}, & \text{对 } \xi \geq \xi_1 \end{cases} \tag{6-50}$$

其中,ξ_0 和 ξ_1 是区间 $[\xi_0, \xi_1]$ 的端点,在这个区间上,单调(单调性是由于 $-\lg(\xi)$ 是凸函数)函数

$$-\frac{d\lg(\xi)}{d\xi} = -\frac{g'(\xi)}{g(\xi)}$$

的绝对值以一个常数 c 为界,这个常数是由归一化条件

$$1 = (1-\varepsilon)\int_{\xi_0}^{\xi_1} g(\xi)d\xi + \frac{g(\xi_0) + g(\xi_1)}{c}$$

决定的。

这一定理使得我们可以构造各种鲁棒密度。特别是,如果我们选择 $g(\xi)$ 为正态密度

① 原文此处表达略有混乱,在翻译时根据译者的理解进行了一定的调整。——译者

$$g(\xi)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\xi^2}{2\sigma^2}\right)$$

并考虑密度类 H

$$p(\xi)=\frac{1-\varepsilon}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\xi^2}{2\sigma^2}\right)+h(\xi),$$

那么根据这一定理,密度

$$p_r(\xi)=\begin{cases}\dfrac{1-\varepsilon}{\sqrt{2\pi}\sigma}\exp\left(\dfrac{c^2}{2}-\dfrac{c}{\sigma}|\xi|\right) & \text{对 } |\xi|>c\sigma\\[2ex] \dfrac{1-\varepsilon}{\sqrt{2\pi}\sigma}\exp\left(-\dfrac{\xi^2}{2\sigma^2}\right) & \text{对 } |\xi|\le c\sigma\end{cases}\tag{6-51}$$

在这个类中将是鲁棒的,其中 c 是由下面的归一化条件确定的:

$$1=\frac{1-\varepsilon}{\sqrt{2\pi}\sigma}\int_{-c\sigma}^{c\sigma}\exp\left(-\frac{\xi^2}{2}\right)d\xi+\frac{2\exp\left(-\dfrac{c^2}{2}\right)}{c}.$$

从这一鲁棒密度导出的损失函数是

$$L(\xi)=-\ln p(\xi)=\begin{cases}c|\xi|-\dfrac{c^2}{2} & \text{对 } |\xi|>c,\\[2ex] \dfrac{\xi^2}{2} & \text{对 } |\xi|\le c.\end{cases}\tag{6-52}$$

这个损失函数平滑地将两个函数结合起来:一个是二次函数,一个是线性函数。在一种极端的情况下(当 c 趋向于无穷大时),它定义了最小二乘方法;在另一种极端情况下(当 c 趋向于零时),它定义了最小模方法。在一般情况下,鲁棒回归的损失函数是两个函数的组合,一个是 $f(u)=|u|$,而另一个函数对 u 的偏差更不敏感得多(函数 f(u) 的非线性部分的导数小于线性部分的导数)。

6.12 支持向量回归机器

我们创建用于回归的 SVM 是基于 ε不敏感损失函数的。这个损失函数与鲁棒损失函数有相同的结构:它也是两个函数的组合,一个是 $f(u)=|u|$,另一个是常数函数[①] $f(u)=\text{const}$(我们考虑 const= 0 的情况)。

ε不敏感性质带来了 SVM 解的一些新特性,这就是解的稀疏性。通过变化(增加) ε值,我们可以控制(增加)SVM 解的稀疏性。

但是,鲁棒性方法与 SVM 方法之间的区别也反映出 SVM 回归的损失函数比鲁棒回归的损失函数更复杂。对线性函数,它有下面的形式[②]:

① 严格地说它不属于 Huber 鲁棒估计子族,因为均匀分布函数没有平滑的导数。

② 在这一章的主要部分中我们采用了这一泛函的一种等价形式。

$$L(\alpha) = \frac{1}{C}(w, w) + \sum_{i=1}^{l} \xi_i - (w, x)\xi_i$$

其中,(w, w) 是正则化泛函,$1/C$ 是正则化参数(我们将在下一章讨论正则化技术)。

把正则化项加入到泛函中使情形发生了大大的改变:一方面,它把 SVM 回归与为解决不适定问题引入的正则化技术联系了起来;另一方面,它增加了自由参数的数目。

现在,为了估计回归函数我们必须指明三个自由参数:ε不敏感性值、正则化参数 C 值以及核参数(多项式核的阶数、径向基核的宽度参数、样条生成核的样条阶数等等)。

在下一章里我们将说明,利用在传统统计学中发展出的一些一般思想和在不适定问题理论中发展出的解决不适定问题的一般原则,我们将不但能够确定这些参数应该如何相互联系,从而得到最优估计;而且我们还将能够描述,评估对于解决统计学习理论的主要问题的最好可能参数的有效算法,这些主要问题是:估计密度函数、条件概率(这是比前面讨论的模式识别问题更一般一些的解)和回归函数。ε不敏感估计将在这些算法中起到决定性的作用。

第七章

统计学习理论中的直接方法

在这一章里,我们将介绍解决统计学习理论中主要问题的一种新方法,这些问题是:模式识别、回归估计和密度估计。

我们将介绍所谓的直接方法,它需要求解定义了所求的函数的算子方程。这些方程的求解是建立在求解随机不适定问题的基础上的。为了有效地求解这些问题,我们把源于3种不同数学分支的思想结合起来,它们是:不适定问题的理论、传统的非参数统计学以及统计学习理论。本书的大部分内容中并没有考虑在前两个分支中得到的结果(只是在各章的非正式推导和评述中有简要的讨论)。

在本章中,我们将介绍这些分支中一些必要的结果,并将相关的技术结合起来得到一种新的算法。

7.1 密度、条件概率和条件密度的估计问题

7.1.1 密度估计的问题:直接表示

我们从密度估计问题开始本章的讨论。设 ξ 是一个随机变量,随机事件

$$F(x) = P\{\xi < x\}$$

的概率称作随机变量 ξ 的概率分布函数。随机向量 ξ 是随机变量概念的一个推广,而函数

$$F(x) = P\{\xi < x\}$$

被称作随机向量 ξ 的概率分布函数,其中的比较大小符号代表对向量的每一维坐标进行比较。如果存在一个非负函数 $p(x)$,使得对所有 x,等式

$$F(x) = \int_{-\infty}^{x} p(x')dx'$$

都成立,那么我们说随机变量 ξ 随机向量 ξ 存在一个密度。

函数 $p(x)$ 被叫做随机变量(随机向量)的概率密度。这样,依据定义,要从数据估计一

个概率密度,我们需要在一个给定的密度集 $p(x,\alpha)$, $\alpha\in\Lambda$上得到积分方程①

$$\int_{-\infty}^{x} p(x',\alpha)\,dx' = F(x) \tag{7-1}$$

的解,求解的条件是分布函数 $F(x)$未知,但是给出了依照 $F(x)$得到的随机独立样本

$$x_1,\cdots,x_l. \tag{7-2}$$

我们可以用数据(7-2)式构造对分布函数 $F(x)$的逼近,即所谓经验分布函数:

$$F_l(x) = \frac{1}{l}\sum_{i=1}^{l}\theta(x-x_i), \tag{7-3}$$

其中,我们对向量② u 定义阶跃函数如下:

$$\theta(u)=\begin{cases}1 & \text{若向量 } u \text{ 的所有分量都为正,}\\ 0 & \text{其他.}\end{cases}$$

在下一节里我们将说明,经验分布函数 $F_l(x)$是实际分布函数 $F(x)$的一个好的逼近。

这样,密度估计的问题就成为,如果概率分布函数未知,但可以定义它的一个逼近,寻找积分方程(7-1)的解的一个逼近。

我们把密度估计问题的这种表示称作直接表示,因为它是基于密度的定义的。在下面的几节里,我们将讨论求解带有近似的右边和近似的算子的积分方程问题。现在我们先来看看条件概率 $P(\omega|x)$估计问题的直接表示,条件概率 $P(\omega|x)$是给定向量 x 下类 ω的概率。

7.1.2 条件概率估计问题

考虑变量对(ω,x),其中 x 是一个向量,ω是一个标量,它只取 k 个值$\{0,1,\cdots,k-1\}$中的一个。根据定义,条件概率 $P(\omega|x)$是下面积分方程的解:

$$\int_{-\infty}^{x} P(\omega|x')\,dF(x') = F(\omega,x), \tag{7-4}$$

其中,$F(x)$是随机向量 x 的分布函数,$F(\omega,x)$是变量对(ω,x)的联合分布函数。事实上,由于 $dF(x)=p(x)dx$(我们假设密度存在)且

$$P(\omega|x') = \frac{p(\omega,x')}{p(x')},$$

方程(7-4)的解定义了条件概率。

在函数集 $P_\alpha(\omega|x)$, $\alpha\in\Lambda$中估计条件概率的问题就是得到对积分方程(7-4)的解的逼近,方程中的分布函数 $F(x)$和 $F(\omega,x)$都是未知的,但给出了数据

$$(\omega,x_1),\cdots,(\omega,x_l).$$

就像在密度估计中一样,我们可以用经验分布函数(7-3)式和函数

---

① 当 $x=(x^1,\cdots,x^n)$是向量时,我们用这里的写法定义对向量每一维的积分,即

$$\int_{-\infty}^{x} p(x,\alpha)\,dx = \int_{-\infty}^{x^1}\cdots\int_{-\infty}^{x^n} p(x^1,\cdots,x^n;\alpha)\,dx^1\cdots dx^n.$$

② 也包括作为一维向量的标量。

$$F_l(\omega, x) = \frac{1}{l}\sum_{i=1}^{l} \theta(x - x_i)\,\delta(\omega, x_i)$$

来逼近未知分布函数 $F(x)$ 和 $F(\omega, x)$，其中

$$\delta(\omega, x) = \begin{cases} 1 & \text{若向量 } x \text{ 属于 } \omega \text{ 类,} \\ 0 & \text{其他.} \end{cases}$$

这样，问题就成为在概率分布函数 $F(x)$ 和 $F(\omega, x)$ 都未知，但给出了其逼近 $F_l(x)$ 和 $F_l(\omega, x)$ 的情况下，在函数集 $P_\alpha(\omega|x)$，$\alpha \in \Lambda$ 中求得积分方程(7-4)解的逼近。

注意，与第一章讨论的方法相比，对条件概率函数 $P(\omega|x)$ 的估计是对模式识别问题的一个更强的解。在第一章中，目标是从给定的决策规则集合中找到最佳的决策规则，至于这个集合中是否包含对训练器决策规则的一个好的逼近并没有关系。在这里，目标是找到对训练器决策规则的最佳逼近(根据问题的表示，训练器的决策规则就是条件概率函数，参见第一章)。当然，如果对训练器算子 $P(\omega|x)$ 的逼近是已知的，那么我们可以容易地构造最优决策构造。对 $\omega \in \{0, 1\}$ 且类先验概率相等的情况，最优决策函数的形式是

$$f(x) = \theta\left(P(\omega = 1|x) - \frac{1}{2}\right).$$

这就是所谓的贝叶斯规则：如果向量 x 属于第一类的概率大于 1/2，则决策规则把向量 x 赋予类别 1，否则赋予类别 0。然而，对条件概率的知识不但给出了模式识别问题的最优解，而且提供了对任意特定向量 x 的错误率的估计。

7.1.3 条件密度估计问题

最后，我们来考虑条件密度估计的问题。在变量对 (y, x)，设变量 y 是一个标量，x 是一个向量。考虑等式

$$\int_{-\infty}^{y}\int_{-\infty}^{x} p(y'|x')\,dF(x')\,dy' = F(y, x), \tag{7-5}$$

其中 $F(x)$ 是一个概率分布函数，其密度存在，$F(y, x)$ 是定义在变量对 (y, x) 上的联合概率分布函数①。

像前面一样，我们要通过在给定函数集上求解积分方程(7-5)式得到对条件密度 $p(y|x)$ 的逼近，方程求解是在分布函数 $F(x)$ 和 $F(y, x)$ 未知、但给出了随机独立同分布数据对

$$(y_1, x_1), \cdots, (y_l, x_l) \tag{7-6}$$

的情况下进行的。同样，我们可以用经验分布函数(7-3)式逼近 $F(x)$，用经验分布函数

$$F_l(y, x) = \frac{1}{l}\sum_{i=1}^{l} \theta(y - y_i)\,\theta(x - x_i)$$

逼近 $F(y, x)$。这样，我们的问题就成为，在概率分布函数未知、但可以用(7-6)式的数据构

① 事实上，这个方程的解就是条件密度的定义。假设 $p(x)$ 和 $p(y, x)$ 是与概率分布函数 $F(x)$ 和 $F(y, x)$ 对应的密度，那么等式(7-5)等价于等式

$$p(y|x)\,p(x) = p(y, x).$$

造逼近 $F_l(x)$ 和 $F_l(y, x)$ 的情况下，在函数集 $p_\alpha(y|x)$，$\alpha \in \Lambda$ 中得到对积分方程(7-5)的解的逼近。

注意，与回归函数相比，条件密度 $p(y|x)$ 中包含了关于给定 x 下随机值 y 的行为更多的信息。从条件密度可以容易地得到回归函数。根据定义，回归函数就是

$$r(x) = \int yp(y|x)dy.$$

7.2 求解近似确定的积分方程的问题

估计随机依赖关系的 3 个问题都可以用下面的一般方法描述。即需要求解线性算子方程

$$Af = F, \quad f \in \mathfrak{F}. \tag{7-7}$$

形成方程的某些函数是未知的，而给出的是数据。用这些数据可以得到对未知函数的逼近。

在密度估计问题与条件概率和条件密度估计问题之间有一点不同。在密度估计问题中，对方程右边给出的是它的逼近。我们要从关系

$$Af \approx F_l, \quad f \in \mathfrak{F}$$

中得到对方程(7-7)的解的逼近。在条件概率和条件密度估计问题中，不但知道方程(7-7)右边的逼近，而且也近似地知道算子 A(在积分方程(7-4)和(7-5)的左边，我们不是用分布函数，而是用它们的逼近)。因此我们的问题是从关系

$$A_l f \approx F_l, \quad f \in \mathfrak{F}$$

中得到方程(7-7)的解的逼近，其中 A_l 是算子 A 的一个逼近。

求解这些问题既有有利的一方面，也有不利的一方面。有利的方面是，经验分布函数构成了对未知分布函数的一个好的逼近。在下一节里我们将说明，随着观测数目趋近于无穷大，经验分布函数以快的速度 $1/\sqrt{l}$ 收敛到所要的分布函数。在一维情况下，对定义经验分布函数与真实分布函数之间距离的不同度量，人们已经知道了其收敛速度的渐近精确的描述。

特别地，对一维情况，所要的函数与其逼近之间的距离(在一致度量 C 下)的 Kolmogorov-Smirnov 分布是已经知道的。在多维情况下，我们可以计算这个分布的任意分位数(Paramasamy, 1992)。

不利的一方面是，求解算子方程(7-7)属于所谓的不适定问题。在第 7.4 节中，我们将定义“不适定”问题的概念，并讨论我们求解不适定问题时遇到的困难。我们将描述传统不适定问题求解理论的主要结果，以及在随机不适定问题上的推广。解决随机不适定问题的理论将被用来求解我们的积分方程。

7.3 Glivenko-Cantelli 定理

我们已经提到过，在 30 年代，Glivenko 和 Cantelli 证明了统计学中最重要的定理之一。他们证明，当观测数目趋近于无穷大时，经验分布函数 $F_l(x)$ 收敛到实际分布函数

$F(x)$。这一定理在理论统计学的基础中起了十分重要的作用。

定理(Glivenko-Cantelli) 收敛

$$\sup_x \left| F_l(x) - F(x) \right| \xrightarrow{p} 0, \quad l \to \infty$$

成立。

在这一表述中,Glivenko-Cantelli 定理断定了经验分布函数 $F_l(x)$ 以概率收敛[①] (在一致度量下)到实际分布函数 $F(x)$。

我们可以从第二章讨论的一致收敛的角度来表述这一定理。事实上,考虑下面的事件集合

$$e(\alpha) = \theta(\alpha - x), \quad \alpha \in \Lambda, \tag{7-8}$$

对任意固定的 α 它定义了小于 α 的 x 的集合。现在,设在这个 x 的集合上定义了一个概率测度。那么,作为 α 的一个函数,期望

$$R(\alpha) = E\theta(\alpha - x)$$

就定义了一个概率分布函数,而从独立同分布数据 $x_1, \cdots, x_l$ 计算出的经验泛函

$$R_l(\alpha) = \frac{1}{l}\sum_{i=1}^{l} \theta(\alpha - x_i) \quad \alpha \in R^1$$

就定义了一个经验分布函数。因此,实际上,Glivenko-Cantelli 定理是对定义在 R^1 上的一个特殊的事件集合(7-8)式的一致收敛理论。

在 n 维情况下,$\alpha = (\alpha^1, \cdots, \alpha^n)$, $x = (x^1, \cdots, x^n)$, Glivenko-Cantelli 定理描述了在事件集合

$$e(\alpha) = \prod_{k=1}^{n} \theta(x^k - \alpha^k), \quad \alpha \in R^n \tag{7-9}$$

上频率到其概率的一致收敛。

在第三章中,我们分析了在任意给定的事件集合(不一定是(7-9)式的集合)上一致收敛的条件。因此,在统计学习理论中发展的一致收敛理论中,Glivenko-Cantelli 定理是一个特例。

Kolmogorov-Smirnov 分布

在 Glivenko-Cantelli 定理得到证明后不久,马上就出现了 $F_l(x)$ 收敛到 $F(x)$ 的速度的问题。

对一维连续函数 $F(x)$,人们对 $F_l(x)$ 收敛到 $F(x)$ 的速度的研究,导致建立了下面这一重要的统计学规律:

Kolmogorov-Smirnov 分布

随机变量

$$\xi = \sqrt{l}\sup_x \left| F_l(x) - F(x) \right|$$

① 此收敛也几乎必然发生。

有下面的极限概率分布(Kolmogorov):

$$\lim_{l\to\infty} P\left\{\sqrt{l}\sup_x |F(x)-F_l(x)| \geq \varepsilon\right\} = 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2\varepsilon^2 k^2}. \tag{7-10}$$

随机变量

$$\xi = \sqrt{l}\sup_x (F(x)-F_l(x)),$$

$$\xi = \sqrt{l}\sup_x (F_l(x)-F(x)),$$

有下面的极限概率分布(Smirnov):

$$\lim_{l\to\infty} P\left\{\sqrt{l}\sup_x (F(x)-F_l(x)) \geq \varepsilon\right\} = e^{-2\varepsilon^2},$$

$$\lim_{l\to\infty} P\left\{\sqrt{l}\sup_x (F_l(x)-F(x)) \geq \varepsilon\right\} = e^{-2\varepsilon^2}. \tag{7-11}$$

正如我们在上一节提到的, Glivenko-Cantelli 定理(原来是在一维情况提出的)是统计学习理论的一个特例。在第三章中,我们讨论了一致收敛的界,它们是对任意特定的 l 和任意维空间中的事件集合都成立的。

特别地,这一理论可以应用到事件集(7-9)式上。因为在 R^n 中定义的这个集合的 VC 维等于 n(空间的维数),我们也可以得到在事件集合(7-9)式上一致收敛的界。因此,用统计学习理论中的结果,我们可以得到不等式形式的非渐近界。

然而,对事件(7-9)式的一致收敛分析中有些东西在针对一般事件类型的统计学习理论中没有得到。对事件集合(7-9)式,存在对其一致收敛速度的一个精确描述,它不依赖于概率测度(通用分布)。对一维情况,这个精确分布是由 Kolmogorov 和 Smirnov(对足够大的 l)得到的。在多维情况下,这种分布尚未知道,但是已经知道这样一个分布是存在的①

在 7.5 节中,我们将看到,对我们的估计问题来说,拥有这种分布的通用等式型特性是多么重要。尽管对多维情况和/或有限观测数目情况,这种分布的解析表达尚不得而知,但我们可以容易地创建一个表,在表中对任意观测数目 l 和任意适度的维数 n(比如 $n<100$)计算这一分布的任何分位数。在第 7.8、7.9 和 7.10 节中,我们将用这样一个表来估计我们算法的最佳参数。

7.4 不适定问题

设算子方程

$$Af(t) = F(x) \tag{7-12}$$

是由连续算子 A 定义的,它把度量空间 E_1 中的元素 f 一对一地映射到度量空间 E_2 的元素 F。

对算子方程(7-12),如果方程右边 $F(x)\in F(x,\alpha)$ 的一个小的变化导致解的一个小的变化,即,如果对任意 $\varepsilon>0$,存在一个 $\delta(\varepsilon)$,使得只要不等式

$$\rho_2(F(x,\alpha),F(x,\alpha)) \leq \delta(\varepsilon)$$

① 描述拥有通用的(不依赖于概率测度)一致收敛速度且精确分布的事件集合是非常有意思的。

成立,则不等式

$$\rho_{z_1}(f(t,\alpha),f(t,\alpha)) \leq \varepsilon$$

也成立,那么我们说算子方程(7-12)是稳定的。

对算子方程(7-12),如果方程的解

· 存在,

· 唯一,

· 稳定,

则我们说方程(7-12)是在 Hadamard 意义下适定的。而如果算子方程的解不满足上述条件中的任何一个,则求解这个算子方程的问题就被看作是不适定的。下面我们考虑算子方程的解存在、唯一、但不稳定的不适定问题。我们将考虑由第一类 Fredholm 积分方程定义的不适定问题,即

$$\int_a^b K(t,x)f(t)\,dt = F(x).$$

但所得到的所有结果对于由其他线性连续算子定义的方程也成立。

考虑 Fredholm 的第一类积分方程:

$$\int_a^b K(t,x)f(t)\,dt = F(x), \tag{7-13}$$

它是由核 $K(t,x)$ 定义的,核 $K(t,x)$ 在 $a\leq t\leq b, a\leq x\leq b$ 上几乎处处连续。这个核把在 $[a,b]$ 上连续的函数集$\{f(t)\}$映射到也在$[a,b]$上连续的函数集$\{F(x)\}$。

容易说明,求解方程(7-13)的问题是一个不适定问题。为此,我们注意到,用核 $K(t,x)$ 形成的连续函数 $G_\nu(x)$:

$$G_\nu(x) = \int_a^b K(t,x)\sin(\nu t)\,dt$$

拥有下面的特性:

$$\lim_{\nu\to\infty} G_\nu(x) = 0.$$

考虑积分方程

$$\int_a^b K(t,x)f^*(t)\,dt = F(x) + G_\nu(x),$$

因为 Fredholm 方程是线性的,这个方程的解有下面的形式:

$$f^*(t) = f(t) + \sin(\nu t),$$

其中 $f(t)$ 是方程(7-13)的解。对于充分大的 ν,这个方程的右边与方程(7-13)的右边只相差一个小量 $G_\nu(x)$,而它的解却与方程(7-13)的解相差 $\sin(\nu t)$。

注意,方程(7-1)、(7-4)和(7-5)也属于第一类 Fredholm 方程。我们可以把它们重写如下[①]:

$$\int_I \theta(x-x')p(x')\,dx' = F(x),$$

① 下面公式的等号右边是 $F(x)$,原著误写为 $F(x')$。——译者

$$\int_I \theta(x - x')P(\omega; x')\,dF(x') = F(\omega, x),$$

$$\iint_I \theta(y - y')\theta(x - x')p(y'; x')\,dF(x')dy' = F(y, x).$$

为了简单起见，我们这里假定 x（变量对(x, y)）属于单位立方体 I。

7.5 解决不适定问题的三种方法

在 60 年代，人们提出了解决不适定问题的 3 种方法。它们的基础都是引入所谓的正则化泛函 $\Omega(f)$。

正则化泛函 $\Omega(f)$ 是一个半连续的正泛函，且 $\Omega(f) \leq c$, $c > 0$ 是一个紧统（在函数 f 的空间中）。它是定义在函数集 $f \in F$ 上的，函数集 $f \in F$ 是方程的解域。

下面，为了保证解的惟一性，我们考虑具有以下性质的正则化泛函：

(1) $\Omega(f)$ 是一个非负凸泛函。也就是，对任意的 $0 \leq \lambda \leq 1$，不等式

$$\Omega(\lambda f_1 + (1 - \lambda)f_2) \leq \lambda\Omega(f_1) + (1 - \lambda)\Omega(f_2), \quad f_1, f_2 \in F$$

成立。

(2) 等式

$$\Omega(0) = 0$$

成立。

(3) 对每个固定的 f，函数

$$r(\gamma) = \Omega(\gamma f)$$

是 γ 的严格增函数。

在正则化泛函的基础上，人们提出了下面 3 种方法：

1. Tikhonov（1963）的变分方法（方法 T）

最小化泛函

$$W_T(f) = \| Af - F \|^2_{E_2} + \gamma\Omega(f),$$

其中 $\gamma > 0$ 是某个预定义的常数。

2. Phillips（1962）的残差方法（方法 P）

在约束条件

$$\| Af - F \|_{E_2} \leq \sigma$$

下最小化泛函

$$W_P(f) = \Omega(f),$$

其中 $\sigma > 0$[①] 是某个预定义的常数。

---

① 原文中此处的参数为 μ，在下文中又在同样的地方多次使用了 σ，故我们在翻译时把这里也改为 σ。—— 译者

3. Ivanov(1962)的拟解方法(方法 I)

在约束条件

$$\Omega(f) \leq C$$

下最小化泛函

$$W_I(f) = \| Af - F \|_{E_2}$$

其中 C> 0 是某个预定义的常数。

已经证明(Vasin, 1970),这些方法在一定意义上是等价的,即如果一种方法(比如方法 T)对某个给定的参数(比如 γ^*)得到解 f^*,那么在其他两种方法中存在相应的参数值,使它们得到同样的解。

残差原则

解决不适定问题的 3 种方法都含有一个自由参数(对方法 T 是参数 γ,对方法 P 是参数 σ,对方法 I 是参数 C)。选择适当的参数对于得到一个不适定问题的好的解是十分关键的。

在解决不适定问题的理论中,有一个选择这种参数的一般原则,即所谓的残差原则(Morozov, 1983)。

假设我们知道用函数 F_δ 估计方程(7-12)右边 F 的精度,即我们知道使得下面的等式成立的 σ值:

$$\| F - F_\delta \|_{E_2} = \sigma,$$

那么,残差原则建议选择这样的参数(方法 T 的参数 γ 或方法 I 的参数 C_l),使得所得的解满足等式

$$\| Af_\delta - F_\delta \|_{E_2} = \sigma \tag{7-14}$$

(对方法 P 则是选择在参数 σ下严格满足条件(7-14)式的解)。

通常,要得到方程右边与它的一个给定的逼近之间差异的一个准确估计是不容易的。

幸运的是,在我们估计密度、条件概率和条件密度的问题中,情况并不是这样。对这些问题,存在值 $\sigma = \sigma$ 的精确估计,它依赖于样本数目 l 和空间维数 n。

注意,我们 3 种问题中的共同之处是,方程的右边都是概率分布函数。在求解中,用的不是实际的分布函数,而是经验分布函数。正如我们在第 7.3 节中讨论的,对任意固定的观测数目 l 和任意固定的维数 n,存在一个差异

$$\xi = \sqrt{l}\ \sup_x |F_l(x) - F_l(x)|$$

的通用分布函数。让我们取这个分布的一个适当的分位数 q^*(比如 50% 分位数),并选择

$$\sigma = \sigma = \frac{q^*}{\sqrt{l}}. \tag{7-15}$$

下面,我们将选择在约束(7-15)式下满足残差原则的解。

7.6 不适定问题理论的主要论断

在这一节里,我们将描述 Tikhonov 方法的主要定理。因为所有方法是等价的,对其他两种方法也有类似的结论。

7.6.1 确定性不适定问题

假设对方程

$$Af = F,$$

给出的不是其准确的右边,而是一个估计 F_δ,使得

$$\| F_\delta - F \|_{E_2} \leq \delta \tag{7-16}$$

我们的目标是,确定值 $\delta > 0$ 和正则化参数 $\gamma_\delta > 0$ 之间的关系,以使得一旦 δ 收敛到零时,我们的正则化方法的解就收敛到所求的解。

下面的定理就建立了这样的关系(Tikhonov and Arsenin, 1977)。

定理 7.1 设 E_1 和 E_2 是度量空间,并假定对 $F \in E_2$,存在一个方程(7-12)的解 $f \in E_1$。设方程(7-12)右边准确的 F 没有给出,而给出的是逼近 $F_\delta \in E_2$,且 $\rho_2(F, F_\delta) \leq \delta$。假设参数 $\gamma(\delta)$ 的值是这样选择的:

$$\text{对 } \delta \to 0 \quad \gamma(\delta) \to 0,$$

$$\lim_{\delta \to 0} \frac{\delta}{\gamma(\delta)} \leq r < \infty, \tag{7-17}$$

那么当 $\delta \to 0$ 时,在 E_1 上最小化泛函 $W_T(f)$ 的元素 $f_\delta^{\gamma(\delta)}$ 收敛于方程的准确解。

在希尔伯特空间中,有下面的定理。

定理 7.2 设 E_1 是一个希尔伯特空间,且 $\Omega(f) = \| f \|^2$。那么对满足下面的关系的 $\gamma(\delta)$:

$$\text{对 } \delta \to 0 \quad \gamma(\delta) \to 0,$$

$$\lim_{\delta \to 0} \frac{\delta^2}{\gamma(\delta)} = 0, \tag{7-18}$$

最小化泛函

$$W_T^*(f) = \rho_2^2(Af, F_\delta) + \gamma\Omega(f) \tag{7-19}$$

的函数 $f_\delta^{\gamma(\delta)}$ 当 $\delta \to 0$ 时收敛于在空间 E_1 的度量下的准确解 f。

7.6.2 随机不适定问题

现在考虑这样的情况,即对方程

$$Af = F, \tag{7-20}$$

其右边 F 没有给出,而是给出一个随机函数序列 F_l,它以概率收敛于 F。也就是,给出的

是一个序列 $F_1, \cdots, F_l, \cdots$，它使得下面的等式成立：

$$\lim_{l\to\infty} P\{\rho_2(F_l, F) > \varepsilon\} = 0 \quad \forall \varepsilon > 0.$$

我们的目标是用序列 $F_1, \cdots, F_l, \cdots$ 来寻找一个方程(7-20)的解的序列，它以概率收敛于真实解。我们把这个问题叫做随机不适定问题，因为我们是用随机函数 $F_l(x)$ 来求解方程。

要解决这些随机不适定问题，我们采用方法 T。对任意的 F_l，我们通过最小化泛函

$$W_T(f) = \rho_2^2(Af, F_l) + \gamma\Omega(f)$$

来找到序列 $f_1, \cdots, f_l, \cdots$。下面我们考虑

$$\text{当 } l \to \infty \text{ 时，} \gamma \to 0$$

的情况。

在这些条件下，有下面的定理成立(Vapnik and Stefnyuk, 1978)，它描述了两个随机变量之间的关系，这两个随机变量是：随机变量 $\rho_2(F, F_l)$ 和随机变量 $\rho_1(f, f_l)$。

定理 7.3　对任意正数 ε 和 μ，存在一个正数 $n(\varepsilon, \mu)$，使得对所有 $l > n(\varepsilon, \mu)$，不等式

$$P\{\rho_1(f_l, f) > \varepsilon\} \le P\left\{\rho_2(F_l, F) > \sqrt{\gamma\mu}\right\} \tag{7-21}$$

成立。

对 E_1 是一个希尔伯特空间的情况，下面的定理成立。

定理 7.4　设 E_1 是一个希尔伯特空间，方程(7-20)中的 A 是一个线性算子，且

$$W(f) = \|f\|^2 = (f, f),$$

那么对任意正数 ε，存在一个数 $n(\varepsilon)$，使得对所有 $l > n(\varepsilon)$，不等式

$$P\{\|f_l - f\|^2 > \varepsilon\} < 2P\left\{\rho_2^2(F_l, F) > \frac{\varepsilon}{2}\gamma\right\}$$

成立。

这些定理是定理 7.1 和定理 7.2 在随机情况下的推广。

推论　根据定理 7.3 和 7.4，如果算子方程(7-20)右边的逼近 F_l 在空间 E_2 的度量下依概率收敛到真实函数 $F(x)$，收敛速度是

$$\rho_2(F(x), F_l(x)) \le r(l),$$

那么，若

$$\lim_{l\to\infty} \frac{r(l)}{\sqrt{\gamma}} = 0$$

且当 $l\to\infty$ 时 γ 收敛于零，则方程(7-20)的解的序列依概率收敛到所求的解。

7.7　密度估计的非参数方法

7.7.1　密度估计问题解的一致性

考虑积分方程

$$\int_{-\infty}^{x} f(x')dx' = F(x),$$

我们不是用实际的分布函数,而是用经验分布函数 $F_1, \cdots, F_l, \cdots$ 求解这个方程。对不同的 l,我们最小化泛函

$$W_l(f) = \rho_2^2(Af, F_l) + \gamma\Omega(f),$$

其中我们选择度量 $\rho_2(Af, F_l)$,使得

$$\rho_2(Af, F_l) \le \sup_x |(Af)x - F_l(x)|, \tag{7-22}$$

假设

$$f_1, \cdots, f_l, \cdots$$

是所得解的序列。

根据定理 7.3,对充分大的 l,对任意的 ε 和对任意的 μ,不等式

$$P\{\rho_1(f_l, f) > \varepsilon\} \le P\{\sup_x |F_l(x) - F(x)| > \varepsilon\sqrt{\gamma}\mu\}$$

成立。

既然事件集合(7-9)式的 VC 维是有界的(界等于空间的维数),对充分大的 l,不等式

$$P\{\sup_x |F_l(x) - F(x)| > \varepsilon\} \le C\exp\{-\varepsilon^2 l\}$$

成立(参见(3-3)式和(3-23)式的界)。因此,存在一个 $l(\varepsilon, \mu)$,使得对 $l > l(\varepsilon, \mu)$,不等式

$$P\{\rho_1(f_l, f) > \varepsilon\} \le C\exp\{-\gamma\mu\} \tag{7-23}$$

成立。

如果 $f(x) \in L_2$,那么根据定理 7.4 和 VC 维的界,对充分大的 l,不等式

$$P\left\{\int (f_l(x) - f(x))^2 dx > \varepsilon\right\} \le C\exp\{-\gamma\mu\} \tag{7-24}$$

成立。

不等式(7-23)和(7-24)式意味着,如果

$$\gamma \underset{l\to\infty}{\to} 0,$$

$$\frac{l}{\ln l}\gamma \underset{l\to\infty}{\to} \infty, \tag{7-25}$$

则解 f_l 依概率收敛于所求的解(在度量 $\rho_1(f_l, f)$ 下)(在这种情况下,不等式(7-23)和(7-24)式的右边趋向于零)。

还可以说明(用 Borel-Cantelli 引理),如果

$$\gamma \underset{l\to\infty}{\to} 0,$$

$$l\gamma \underset{l\to\infty}{\to} \infty,$$

则解以概率 1 收敛于真实解。注意,这一论断对任意正则化泛函 $\Omega(f)$,以及任意满足(7-22)式的度量 $\rho_1(f, f_l)$ 都成立。选择特定的泛函 $\Omega(f)$ 和特定的满足条件

$$\rho_2(F, F_l) \le \sup_x |F_l(x) - F_2(x)|$$

的度量 $\rho_2(F, F_l)$,我们可以构造特定的密度估计器。

7.7.2 Parzen 估计

让我们定义这样的度量 $\rho_2(F, F_l)$ 和泛函 $\Omega(f)$，使得最小化泛函

$$W(f) = \rho_2^2(Af, F_l) + \gamma\Omega(f) \tag{7-26}$$

的方法 T 得到的是 Parzen 估计。

考虑函数集 F 上的 L_2 度量

$$\rho_2(F, F_l) = \sqrt{\int_{-\infty}^{\infty}(F(x) - F_l(x))^2 dx}$$

以及下面的正则化泛函

$$\Omega(f) = \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} R(z - x)f(x)dx\right]^2 dz.$$

这里，$R(z-x)$ 是定义了线性算子

$$Bf = \int_{-\infty}^{\infty} R(z - x)f(x)dx$$

的核。特别地，如果 $R(z-x) = \delta(z-x)$，则算子

$$Bf = \int_{-\infty}^{\infty} \delta(z - x)f(x)dx = f^{(p)}(x)$$

定义了函数 $f(x)$ 的 p 阶导数。

在这些因素下，我们就有泛函

$$W_T(f) = \int_{-\infty}^{\infty}\left[\int_{-\infty}^{x} f(t)dt - F_l(x)\right]^2 dx + \gamma\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} R(z - x)f(x)dx\right]^2 dz. \tag{7-27}$$

下面我们说明，最小化这个泛函的估计子 f_γ 就是 Parzen 估计

$$f_l(x) = \frac{1}{l}\sum_{i=1}^{l} G_\gamma(x - x_i),$$

其中的核函数 $G_\gamma(u)$ 是由核函数 $R(u)$ 定义的。

我们记函数 $f(t)$ 的傅里叶变换为 $\bar{f}(\omega)$，记函数 $R(x)$ 的傅里叶变换为 $R(\omega)$，于是我们就可以计算函数 $F(x)$ 的傅里叶变换：

$$F(\omega) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} F(x)e^{-i\omega x}dx$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-i\omega x}\int_{-\infty}^{x} f(t)dtdx = \frac{\bar{f}(\omega)}{i\omega}$$

以及函数 $F_l(x)$ 的傅里叶变换：

$$F_l(\omega) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} F_l(x)e^{-i\omega x}dx$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\frac{1}{l}\sum_{j=1}^{l}\theta(x - x_j)e^{-i\omega x}dx = \frac{1}{l}\sum_{j=1}^{l}\frac{e^{-i\omega x_j}}{i\omega}.$$

注意到，两个函数卷积的傅里叶变换等于两个函数的傅里叶变换的乘积。在这里，这意味着有

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}(R(x)*f(x))e^{-i\omega x}dx=\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}R(z-x)f(x)dx\ e^{-i\omega z}dz$$

$$=\overline{R}(\omega)\bar{f}(\omega).$$

最后，根据 Parseval 等式，任何函数 $f(x)$ 的 L_2 模都等于它的傅里叶变换 $\bar{f}(\omega)$ 的 L_2 模（在常数 $1/(2\pi)$ 之内）。因此，我们可以把(7-27)式重写为下面的形式：

$$\overline{W}_T(f)=\left\|\frac{\bar{f}(\omega)-\frac{1}{l}\sum_{j=1}^{l}e^{-i\omega x_j}}{i\omega}\right\|_{L_2}^2+\gamma\|\overline{R}(\omega)\bar{f}_l(\omega)\|_{L_2}^2.$$

这个泛函对 $\bar{f}(\omega)$ 是二次的。

因此，这个泛函最小的条件是

$$\frac{\bar{f}_l(\omega)}{\omega^2}-\frac{1}{l\omega^2}\sum_{j=1}^{l}e^{i\omega x_j}+\gamma\overline{R}(\omega)\overline{R}(-\omega)\bar{f}(\omega)=0. \tag{7-28}$$

对 $\bar{f}_l(\omega)$ 求解这个方程，我们得到

$$\bar{f}_l(\omega)=\frac{1}{1+\gamma\omega^2\overline{R}(\omega)\overline{R}(-\omega)}\frac{1}{l}\sum_{j=1}^{l}e^{-i\omega x_j}.$$

引入记号

$$g_{\gamma_l}(\omega)=\frac{1}{1+\gamma\omega^2\overline{R}(\omega)\overline{R}(-\omega)}$$

和

$$G_{\gamma_l}(x)=\int_{-\infty}^{\infty}g_{\gamma_l}(\omega)e^{i\omega x}d\omega$$

要得到对密度的一个逼近，须计算下面的傅里叶反变换：

$$f_l(x)=\int_{-\infty}^{\infty}\bar{f}_l(\omega)e^{i\omega x}d\omega=\int_{-\infty}^{\infty}g_{\gamma_l}(\omega)\frac{1}{l}\sum_{j=1}^{l}e^{-i\omega x_j}e^{i\omega x}d\omega$$

$$=\frac{1}{l}\sum_{j=1}^{l}\int_{-\infty}^{\infty}g_{\gamma_l}(\omega)e^{i\omega(x-x_j)}d\omega=\frac{1}{l}\sum_{j=1}^{l}G_{\gamma_l}(x-x_j).$$

最后一个表达式就是采用核函数 $G_{\gamma_l}(u)$ 的 Parzen 估计。

7.8 密度估计问题的 SVM 解

下面我们考虑在方程右边不是 $F(x)$ 而是其逼近 $F_l(x)$ 的情况下算子方程（密度估计问题）

$$\int_{-\infty}^{x}p(x')dx'=F(x)$$

的另一种解。

我们将用方法 P 来解决这一问题，其中考虑由一致度量

$$\rho_2(F(x),F_l(x))=\sup_x\|F_l(x)-F_l(x)\|_l \tag{7-29}$$

定义的 $F(x)$ 与 $F_l(x)$ 之间的距离，以及正则化泛函

$$\Omega(f) = (f, f)_H, \tag{7-30}$$

它是由某个再生核希尔伯特空间(Reproducing Kernel Hilbert Space, RKHS)的模定义的。

要定义 RKHS 空间，我们必须定义一个对称正定核 $K(x, y)$ 和希尔伯特空间 H 中的一个内积 $(f, g)_H$，使得

$$(f(x), K(x, y))_H = f(y), \quad \forall f \in H \tag{7-31}$$

(再生特性)。注意，任意对称正定函数 $K(x, y)$ 都可以展开为

$$K(x, y) = \sum_{i=1}^{\infty} \lambda_i \varphi_i(x) \varphi_i(y), \tag{7-32}$$

其中，λ 和 $\varphi_i(x)$ 分别是算子

$$Df = \int K(x, y) f(y) dy$$

的本征值和本征函数。

考虑函数集

$$f(x, c) = \sum_{i=1}^{\infty} c_i \varphi_i(x), \tag{7-33}$$

对这个函数集，我们引入内积

$$(f(x, c^*), f(x, c^{**}))_H = \sum_{i=1}^{\infty} \frac{c_i^* c_i^{**}}{\lambda_i}. \tag{7-34}$$

核(7-32)式、内积(7-34)式和集合(7-33)式定义了一个 RKHS 空间。

这是因为，

$$\begin{aligned}(f(x), K(x, y))_H &= \left(\sum_{i=1}^{\infty} c_i \varphi_i(x), K(x, y)\right)_H \\ &= \left(\sum_{i=1}^{\infty} c_i \varphi_i(x), \sum_{i=1}^{\infty} \lambda_i \varphi_i(x) \varphi_i(y)\right)_H \\ &= \sum_{i=1}^{\infty} \frac{c_i \lambda_i \varphi_i(y)}{\lambda_i} = f(y).\end{aligned}$$

对来自一个 RKHS 空间的函数，泛函(7-34)式有如下的形式：

$$\Omega(f) = \sum_{i=1}^{\infty} \frac{c_i^2}{\lambda_i}, \tag{7-35}$$

其中，λ_i 是核 $K(x, y)$ 的第 i 个本征值。因此，核的选择决定了对解的平滑性要求。

要解决密度估计问题，我们使用方法 P，它采用由(7-30)式定义的泛函和一致度量(7-29)式。在约束条件中选择参数 $\sigma = \sigma_l$，使之满足(7-14)式的残差原则。因此，我们要在约束

$$\sup_x \left| F_l(x) - \int_{-\infty}^{x} f(x') dx' \right| = \sigma$$

下最小化泛函

$$\Omega(f) = (f, f)_H.$$

但是，为了计算上的原因，我们考虑只在训练集的点 x_i 上定义的约束条件：

$$\max_i \left| F_l(x) - \int_{-\infty}^{x} f(x') dx' \right|_{x = x_i} = \sigma_i, \quad 1 \le i \le l.$$

我们要寻找的是方程的一个有如下形式的解:

$$f(x) = \sum_{i=1}^{l} \beta_i K(x_i, x), \tag{7-36}$$

其中,核 $K(x_i, x)$ 与定义 RKHS 空间的核相同。考虑到(7-31)式和(7-36)式,我们把泛函(7-30)式重写如下:

$$\begin{aligned}\Omega(f) &= (f, f)_H \\ &= \left(\sum_{i=1}^{l} \beta_i K(x, x_i), \sum_{i=1}^{l} \beta_i K(x, x_i)\right)_H \\ &= \sum_{i=1}^{l} \beta_i \sum_{j=1}^{l} \beta_j (K(x, x_i), K(x, x_j))_H \\ &= \sum_{i,j=1}^{l} \beta_i \beta_j K(x_i, x_j),\end{aligned} \tag{7-37}$$

这里的最后一个等式是利用再生特性(7-31)式得到的。

因此,为了求解我们的方程,要最小化泛函

$$W(\beta) = \Omega(f, f) = \sum_{i,j=1}^{l} \beta_i \beta_j K(x_i, x_j), \tag{7-38}$$

约束条件是

$$\max_i \left| F_l(x) - \sum_{j=1}^{l} \beta_j \int_{-\infty}^{x} K(x_j, x')\,dx' \right|_{x=x_i} = \sigma, \quad 1 \le i \le l. \tag{7-39}$$

这个最优化问题与带有 σ 不敏感区的 SV 回归问题有紧密的联系。它可以用 SVM 技术解决(参见第六章)。

为了得到一个形式为多个密度的组合解,我们选择一个满足以下条件的非负核 $K(x, x_i)$,把这些条件称作条件 K:

(1) 核的形式为

$$K_\gamma(x, x_i) = a(\gamma) K\left(\frac{x - x_i}{\gamma}\right), \tag{7-40}$$

$$a(\gamma) \int K\left(\frac{x - x_i}{\gamma}\right) dx = 1, \quad K(0) = 1, \tag{7-41}$$

其中 $a(\gamma)$ 是归一化常数。

(2) 参数 γ 的值影响由核定义的本征值 $\lambda_1(\gamma), \cdots, \lambda_k(\gamma), \cdots$。我们考虑这样的核,它们使得比值 $\lambda_{k+1}(\gamma)/\lambda_k(\gamma)$, $k = 1, 2, \cdots$ 随着 γ 的增加而减小。这样的核函数的例子是

$$K_\gamma(x, x_i) = a(\gamma) \exp\left(-\left|\frac{x - x_i}{\gamma}\right|^p\right), \quad 0 < p \le 2. \tag{7-42}$$

另外,为了得到形式为多个密度组合的解,我们增加另外两个约束条件:

$$\beta \ge 0, \quad \sum_{i=1}^{l} \beta_i = 1. \tag{7-43}$$

注意,我们的目标泛函也是依赖于参数 γ 的。

$$W_\gamma(\beta) = \Omega(f) = \sum_{i,j=1}^{l} \beta_i \beta_j K_\gamma(x_i, x_j). \tag{7-44}$$

如果对参数 γ,我们的优化问题存在解(解满足残差原则(7-14)式),则我们把参数 γ

的值称作容许值。

容许参数的集合

$$\gamma_{min} \leq \gamma \leq \gamma_{max}$$

是非空的,因为对 Parzen 方法存在这样一个值(Parzen 方法也有(7-36)式的形式)。回顾前面指出的,核函数中的 γ 值决定着对解的平滑性要求: γ值越大,比值 λ_{k+1}/λ_k 越小,因此泛函(7-35)式将导致更强的平滑性要求。

对任意的容许 γ,SVM 技术给出唯一的解,它是一定数目元素的组合。我们选择一个解,它对应一个容许参数 γ,对于系数 β 和参数 γ使得泛函(7-44)式最小化。这种参数选择控制着解的精度。通过选择一个大的容许参数 γ,可以达到另一个目标:增加满足(7-14)式的解的平滑性要求,并选择有较少组合元素的解[①] (即较少数目的支持向量,参见第 6.7 节)。我们可以继续(通过增加(7-14)式中的 σ)增加稀疏性,在解的稀疏性和精度之间进行折衷。

7.8.1 SVM 密度估计方法:总结

用方法 P 求解密度估计方程的 SVM,解决方案实现了下面的思想:

(1) 优化问题中的目标泛函由 RKHS 空间中的模定义,其核(依赖于一个参数)可以有效地控制解的平滑特性。

(2) 选择方程的解为核函数展开的形式(以非负的加权系数),核函数与定义 RKHS 空间的核相同。

(3) 定义优化约束条件的距离 $\rho_2(Af_l, F_l)$是由一致度量给出的(一致度量使得我们可以有效地利用残差原则)。

(4) 解满足残差原则,残差值是从 Kolmogorov-Smirnov 类型的分布得到的(残差值只依赖于维数和观测数目)。

(5) 核的容许参数 γ被用来控制解的精度和/或解的稀疏性。

7.8.2 Parzen 和 SVM 方法的比较

注意到,Parzen 估计

$$f_P(x) = \frac{1}{l}\sum_{i=1}^{l} G_\gamma(x, x_i) \tag{7-45}$$

和 SVM 估计

$$f_{SVM}(x) = \sum_{i=1}^{l} \beta K_\gamma(x, x_i)$$

---

① 注意,我们对同一泛函有两种不同的描述:函数 $_k(x)$ 空间中的描述(7-35)式和核 $K(x, x_i)$ 空间中的描述(7-44)式。从(7-35)式可知,通过增加 γ,我们增强了对在 $_k$ 空间中展开式的"高频成分"的滤波。已经知道,只有目标密度是平滑的(可以用"低频函数"描述),我们才可以在高维空间中用小数目的观测估计这个密度。因此,在高维空间中,最精确的解往往对应于最大的容许的参数 γ。在我们的实验中还观察到,在容许的参数集之内,不同 γ得到的解的精度差别不大。

有相同的结构。在

$$G_\gamma(x, x_i) = K_\gamma(x, x_i)$$

和

$$\beta = \frac{1}{l}$$

的情况下,SVM 估计与 Parzen 估计相同。然而,(7-45)式的解并不一定是我们优化问题的解。尽管如此,我们仍可以说明,SVM 容许解越不平滑,它就越接近 Parzen 解。这是因为,核函数 $a(\gamma)K\frac{\|x_j - x_i\|^2}{\gamma}$ 中的 γ 越小,则泛函

$$W(\beta) = a(\gamma)\sum_{i=1}^{l}\beta_i^2 \tag{7-46}$$

越好地逼近我们的目标泛函(7-38)式。

Parzen 估计是下面的优化问题对最小容许参数 γ 的解:在约束条件(7-39)式和(7-43)式下最小化泛函(7-46)式(而不是泛函(7-38)式)。

因此,Parzen 估计是这个(修正的)优化问题的较不稀疏的容许 SVM 解。

下面对参数 γ 不同的容许值,比较用 Parzen 方法得到的解和用 SVM 方法得到的解。我们估计一个两维情况中的密度,它是由两个拉普拉斯函数的组合定义的①:

$$p(x,y) = \frac{1}{8}\exp\{-(\|x-1\|+\|y-1\|)\} + \exp\{-(\|x+1\|+\|y+1\|)\}.$$

在两种方法中,我们都用同样的高斯核:

$$G_\gamma(x, y; x', y') = K_\gamma(x, y; x', y')$$

$$= \frac{1}{2\pi\gamma^2}\exp\left(-\frac{(x-x')^2+(y-y')^2}{2\gamma^2}\right),$$

并在 $\sigma = q/\sqrt{l}$ 和 $q = 1.2$ 下用残差原则定义最好的参数 γ。

在两种情况下,密度都是用 200 个观测进行估计的。逼近的精度是用 L_1 度量来衡量的:

$$\Delta = \int |p_l(x,y) - p(x,y)|\,dx\,dy.$$

我们进行了 100 次试验,由此构造了对这些试验所得的 q 值②的分布。这一分布是用框状图③表示的,图中的横线分别指示出了误差分布的 5%、25%、50%、75% 和 95% 分位数。

---

① 原著中此公式误写为

$$p(x,y) = \frac{1}{8}\exp\{-(\|x-1\|+\|y-1\|_2)\} + \exp\{-(\|x+1\|_2+\|y+1\|)\}.$$

—— 译者

② 从上下文看这里指的应该是逼近误差 Δ 值。—— 译者

③ “框状图”原文是 boxplot,是一种常用的表示多次试验结果分布的方式。其思路与股市技术分析中表示一个单位时间内股价变动范围的 K 线图略有些类似。基本方法是:用纵坐标代表要表示的值(比如这里是解的逼近误差),横坐标一般代表几组不同的试验。对同一组试验,将多次实验中的多数结果的分布范围框在一个小框内,通常还在其中用一条横线表示出平均结果的值,比如在图 7.1(a)中,每个框代表一组试验的结果分布,框的下线位置是多次试验中误差值从小到大排列的前 25% 的误差值(即所谓 25% 分位数(quantile)),框的上线是前 75% 的误差值,而框中的横线则代表前 50% 的误差值(即平均误差值),这样,框中的误差值范围就代表了所有实验中 50% 的误差所落在的范围。在框的上下两端,还可以延伸出两条虚线,用它们的端点表示更宽一些的数据分布范围,比如图 7.1(a)中的虚线表示了误差值的前 5% 到 95% 的分布范围,这样,连同框内的取值范围,从虚线底端到顶端的误差值就代表了全部实验的 90% 的结果。某些分布情况下,框状图中的某些线可能会重合。—— 译者

图 7.1 和图 7.2 表示出了精度与稀疏性之间的折衷关系。图 7.1(a) 显示了各种方法的 L_1 误差的分布,图 7.1(b) 显示了各种方法中包含的项数的分布,其中的方法是 Parzen 方法和分别采用 $\gamma=0.9$、$\gamma=1.1$ 和最大容许 γ 的 SVM 方法[①]。图 7.2 显示了在(7-39)式中不是采用最优的 $\sigma=q/\overline{l}$,而是用 $\sigma=mq/\overline{l}$,且分别用 $m=1.0$、$m=1.5$ 和 $m=2.3$的 SVM 方法结果,其中,图 7.2(a) 显示了 L_1 误差分布,图 7.2(b) 显示了结果所包含项数的分布。

图 7.1 用不同方法得到的 L_1 误差和项数分布图

(4 种方法是 $\gamma=\gamma_{max}$,$\gamma=1.1$,$\gamma=0.9$ 的 SVM 方法及 Parzen 方法)

图 7.2 采用 $\gamma=\gamma_{max}$的 SVM 方法得到的 L_1 误差及项数分布图

(3 组结果分别是用 $\sigma=mq/\overline{l}$ 和 $m=1.0$、$m=1.5$ 及 $m=2.3$ 得到的)

7.9 条件概率估计

在这一节中,我们要推广上一节的 SVM 密度估计方法来估计条件概率。用同样的思想,我们求解方程

① 原著图 7.1(b) 中并没有给出 Parzen 法中的项数。——译者

$$\int_{-\infty}^{x} p(\omega, x')\,dF(x') = F(\omega, x), \tag{7-47}$$

而其中的概率分布函数 $F(x)$ 和 $F(x,y)$ 未知，给出的是数据

$$(\omega, x_1), \cdots, (\omega, x_l).$$

下面，我们首先描述在什么条件下可以在方程的右边和算子都是近似定义的情况下得到方程的解，然后描述条件概率估计的 SVM 方法。

7.9.1 近似定义的算子

考虑求解算子方程

$$Af = F$$

的问题，其中不但方程的右边给出的是(随机)逼近，而且算子也是如此。我们假设替代精确的算子 A 的是一个逼近序列 A_l, $l = 1, 2, \cdots$，它是一个随机连续算子序列，依概率收敛于算子 A(我们将在下面定义两个算子之间的距离)。

像前面一样，考虑用方法 T 来求解这个算子方程，即最小化泛函

$$W(f) = \rho_2^2(A_l f, F_l) + \gamma\Omega(f).$$

我们用距离

$$\| A_l - A \| = \sup_f \frac{\rho_2(A_l f, Af)}{\Omega^{1/2}(f)} \tag{7-48}$$

来度量算子 A 和算子 A_l 之间的接近程度。

有下面的定理成立(Stefanyuk，1986)。

定理 7.5 对任意的 $\varepsilon > 0$ 和任意的常数 $C_1, C_2 > 0$，存在一个值 $\gamma_0 > 0$，使得对任意的 $\gamma \le \gamma_0$，不等式

$$P\{\rho_1(f_l, f) > \varepsilon\}$$

$$\le P\{\rho_2(F_l, F) > C_1\sqrt{\gamma}\} + P\{\| A_l - A \| > C_2\sqrt{\gamma}\} \tag{7-49}$$

成立。

推论 根据这一定理，如果算子方程右边的逼近 $F_l(x)$ 在空间 E_2 的度量下依概率收敛于真实函数 $F(x)$，收敛速度为 $r(l)$，且逼近 A_l 在(7-48)式定义的度量下依概率收敛于真实算子 A，收敛速度为 $r_A(l)$，那么存在一个这样的函数

$$r_0(l) = \max\{r(l), r_A(l)\} \underset{l\to\infty}{\to} 0,$$

使得如果

$$\frac{r_0(l)}{\sqrt{\gamma}} \underset{l\to\infty}{\to} 0,$$

且当 $l\to\infty$ 时 γ 收敛于零，则方程解的序列依概率收敛于所求的解。

设向量 x 属于某个有界支集 $x \in C^*$。有下面的定理成立。

定理 7.6 如果泛函 $\Omega(f)$ 满足下面的条件：

$$\Omega(f) \geq C \sup_x \|f_l(x)\|_l + \sup_x \|f_l'(x)\|_l^2 \tag{7-50}$$

且 E_2 中的度量满足条件

$$\rho_2(Af_1, Af) \leq \sup_x \|(Af_1)x - (Af)\|_l \tag{7-51}$$

那么，用方法 T 对条件概率的估计是一致的。

也就是说，如果与度量 $\rho_2(\cdot,\cdot)$ 相比正则化足够强，那么通过求解近似定义的积分方程来估计条件概率[①]的方法是一致的。

事实上，考虑差

$$\|(A_1 f)(x) - (Af)(x)\|_l = \left| \int_{-\infty}^{x} f(x') d(F_1(x') - F(x')) \right|$$

$$= \left| f(x)(F_1(x) - F(x)) - \int_{-\infty}^{x} f'(x')(F_1(x') - F(x'))dx \right|$$

$$\leq \sup_x \|f_l(x)\| F_1(x) - F(x)\|_l + \sup_x \|f_l'(x)\| \int_0^x \|F_1(x) - F(x)\| dx.$$

考虑到(7-51)式、(7-50)式及 x 属于有界支集的事实，我们有

$$\|A_1 f - Af\|_{E_2}$$

$$\leq \sup_x \|f_l(x)\|_l + C^* \sup_x \|f_l'(x)\|_l \sup_x \|F_1(x) - F(x)\|_l$$

$$\leq \Omega^{1/2}(f) \sup_x \|F_1(x) - F(x)\|_l \tag{7-52}$$

从这个不等式和算子的模的定义，我们有

$$\|A_1 - A\| = \sup_x \frac{\|A_1 f - Af\|_{E_2}}{\Omega^{1/2}(f)} \leq \sup_x \|F_1(x) - F(x)\|_l \tag{7-53}$$

根据定理 7.5，基于正则化方法得到的算子方程的解 f_γ 有下面的性质：对任意的 $\varepsilon > 0, C_1 > 0, C_2 > 0$，存在 γ_0，使得对任意的 $\gamma < \gamma_0$，不等式

$$P\{\rho_1(f_1, f) > \varepsilon\}$$

$$\leq P\left\{\rho_2(F_1, F) > C_1 \sqrt{\gamma}\right\} + P\left\{\|A_1 - A\| > C_2 \sqrt{\gamma}\right\}$$

$$\leq P\left\{\sup_x \|F_1(x) - F(x)\|_l > C_1 \sqrt{\gamma}\right\} + P\left\{\sup_x \|F_1(x) - F(x)\|_l > C_2 \sqrt{\gamma}\right\}$$

成立。因此，考虑到 VC 维为 n 的事件集合(7-9)式上一致收敛的界，我们对充分大的 l 得到下面的不等式(参见界(3-3)式和(3-23)式)：

$$P\{\rho_1(f_1, f) > \varepsilon\}$$

$$\leq P\left\{\sup_x \|F_1(x) - F(x)\|_l > C_1 \sqrt{\gamma}\right\} + P\left\{\sup_x \|F_1(x) - F(x)\|_l > C_2 \sqrt{\gamma}\right\}$$

$$\leq C(\exp\{-\gamma l C_1\} + \exp\{-\gamma l C_2\}).$$

从这个不等式，我们发现，条件(7-50)式和(7-51)式意味着解收敛且几乎必然收敛到所求的真实解。

---

① 原著中此处误写为条件密度。——译者

7.9.2 条件概率估计的 SVM 方法

现在我们把求解密度估计方程的方法推广到求解条件概率方程

$$\int_{-\infty}^{x} p(\omega|x')\,dF(x') = F(\omega, x) = p(\omega)F(x|\omega), \tag{7-54}$$

其中,我们用经验分布函数 $F_l(x)$ 和 $F_l(x|\omega)$ 来替代方程中的真实分布函数 $F(x)$ 和 $F(x|\omega)$。

我们按照第 7.8 节中介绍的步骤来求解。

(1) 采用方法 P,其目标泛函是 RKHS 空间中的模,RKHS 空间是由满足条件 K

$$\Omega(f) = (f, f)_H$$

的核 $K_\gamma(x, x')$ 定义的(参见第 7.8 节)。

(2) 寻找形式为

$$f_w(x) = p(\omega|x) = p(\omega)\sum_{i=1}^{l}\beta K_\gamma(x, x_i) \tag{7-55}$$

的解,其中 β 为非负系数。

因此我们须最小化泛函

$$W_\gamma(\beta) = \Omega(f) = \sum_{i,j=1}^{l}\beta_i\beta_j K_\gamma(x_i, x_j)$$

(参见第 7.8 节)。

(3) 从等式

$$\sup_x |(A_l f)x - F_l(\omega, x)| \le \sigma^* p(\omega)$$

定义优化约束条件,对我们的方程来说,这一等式的形式为

$$\sup_x \left| \int_{-\infty}^{x} p(\omega)\sum_{i=1}^{l}\beta K_\gamma(x, x_i)\,d\,\frac{1}{l}\sum_{j=1}^{l}\theta(x - x_j) - p(\omega)F(x|\omega) \right|$$

$$= \sigma^* p(\omega).$$

经过显而易见的推导后,我们可以得到优化条件

$$\sup_x \left| \sum_{i=1}^{l}\beta\,\frac{1}{l}\sum_{j=1}^{l} K_\gamma(x_j, x_i)\,\theta(x - x_j) - F_l(x|\omega) \right| = \sigma^*.$$

出于计算上的原因,我们只在训练集中的点上检查这一等式。也就是说,我们用下面的等式来代替这一等式:

$$\max_p \left| \sum_{i=1}^{l}\beta\,\frac{1}{l}\sum_{j=1}^{l} K_\gamma(x_j, x_i)\,\theta(x_p - x_j) - F_l(x_p|\omega) \right| = \sigma^*, \quad p = 1, \cdots, l.$$

注意到,有下面的等式成立:

$$\int_{-\infty}^{\infty} p(\omega|x)\,dF(x) = p(\omega).$$

把表达式(7-55)式的 $p(\omega|x)$ 代入到积分中,我们得到

$$\int_{-\infty}^{\infty}\sum_{i=1}^{l}\beta K_\gamma(x, x_i)\,dF(x) = 1,$$

在积分中用 $F_l(x)$ 代替 $F(x)$,得到另一个约束条件:

$$\sum_{i=1}^{l} \beta_i \frac{1}{l} \sum_{j=1}^{l} K_\gamma(x_j, x_i) = 1.$$

(4) 设属于类 ω 的样本向量数目为 $l(\omega)$，那么，对残差原则，我们采用

$$\sigma^*_\omega = \sigma_{l(\omega)} = \frac{q}{\sqrt{l(\omega)}},$$

其中 q 是 Kolmogorov-Smirnov 型分布的适当的分位数。我们对类 ω 向量出现的概率也给出一个估计：

$$p(\omega) = \frac{l(\omega)}{l}.$$

(5) 从容许参数集

$$\gamma_{\min} \le \gamma \le \gamma_{\max}$$

中选择一个 γ 来控制解的精度(通过最小化 $W_\gamma(\beta)$)或/和解的稀疏性(通过选择大的 γ)。

7.9.3 SVM 条件概率估计：总结

SVM 条件概率估计是

$$p(\omega|x) = \frac{l(\omega)}{l} \sum_{i=1}^{l} \beta_i K_\gamma(x, x_i), \quad \beta_i \ge 0,$$

其中系数 β_i 在约束条件

$$\max_p \left| \sum_{i=1}^{l} \beta_i \frac{1}{l} \sum_{j=1}^{l} K_\gamma(x_j, x_i)\theta(x_p - x_j) - F_l(x_p|\omega) \right| = \sigma^*_\omega, \quad p = 1, \cdots, l,$$

和

$$\beta_i \ge 0,$$

$$\sum_{i=1}^{l} \beta_i \frac{1}{l} \sum_{j=1}^{l} K_\gamma(x_j, x_i) = 1$$

下使得泛函

$$W_\gamma(\beta) = \sum_{i,j=1}^{l} \beta_i \beta_j K_\gamma(x_i, x_j)$$

最小化。

我们从容许参数集

$$\gamma_{\min} \le \gamma \le \gamma_{\max}$$

中选择一个 γ，通过最小化 $W_\gamma(\beta)$ 和/或通过选择一个大的 γ 来控制解的特性(精度和/或稀疏性)。

7.10 条件密度和回归的估计

要用方法 P 估计条件密度函数，我们需求解积分方程

$$\int_{-\infty}^{y} \int_{-\infty}^{x} p(y|x) dF(x) dy = F(x, y), \tag{7-56}$$

其中的条件分布函数 $F(y,x)$ 和 $F(x)$ 未知，而给出了数据

$$(x_1,y_1),\cdots,(x_l,y_l).$$

要用逼近

$$F_l(x)=\frac{1}{l}\sum_{i=1}^{l}\theta(x-x_i)$$

$$F_l(y,x)=\frac{1}{l}\sum_{i=1}^{l}\theta(y-y_i)\theta(x-x_i)$$

来求解这一方程，我们采用与求解密度估计和条件概率估计的方程时完全相同的步骤（参见第 7.8 节、7.9 节）。

(1) 选择 RKHS 空间中函数的模作为正则化泛函

$$\Omega(f)=(f(x,y),f(x,y))_H,$$

它是由满足条件 K 的核[①]

$$K((x,y),(x',y'))=K_\gamma(x,x')K_\gamma(y,y')$$

定义的。

(2) 寻找形式为

$$p(y\mid x)=\sum_{i=1}^{l}\beta K_\gamma(x,x_i)K_\gamma(y,y_i),\qquad \beta\geq 0 \tag{7-57}$$

的解。因此，我们的目标泛函是

$$W_\gamma(\beta)=\Omega(f)=\sum_{i,j=1}^{l}\beta_i\beta_j K_\gamma(x_j,x_i)K_\gamma(y_j,y_i) \tag{7-58}$$

（见 7.8 节）。

(3) 用一致度量

$$\rho_2(A_lf,F_l)=\sup_x\left|(A_lf)(x,y)-F_l(x,y)\right|_l=\sigma$$

得到优化约束条件。对我们的方程，有

$$\sup_{x,y}\left|\int_{-\infty}^{y}\int_{-\infty}^{x}\sum_{i=1}^{l}\beta K_\gamma(x',x_i)K_\gamma(y',y_i)\,d\left[\frac{1}{l}\sum_{j=1}^{l}\theta(x'-x_j)\right]dy'-F_l(x,y)\right|=\sigma.$$

经过简单的推导，可得到约束条件

$$\sup_{x,y}\left|\sum_{i=1}^{l}\beta\frac{1}{l}\sum_{j=1}^{l}K_\gamma(x_j,x_i)\theta(x-x_j)\int_{-\infty}^{y}K_\gamma(y',y_i)dy'-F_l(x,y)\right|=\sigma.$$

由于计算上的原因，我们只在训练向量上检查这一约束条件，即

$$\max_p\left|\sum_{i=1}^{l}\beta\frac{1}{l}\sum_{j=1}^{l}K_\gamma(x_j,x_i)\theta(x_p-x_j)\int_{-\infty}^{y_p}K_\gamma(y',y_i)dy'-F_l(x_p,y_p)\right|=\sigma,$$

$$p=1,\cdots,l. \tag{7-59}$$

注意到有下面的等式成立：

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}p(y\mid x)dF(x)dy=1,$$

把表达式(7-57)的 $p(y\mid x)$ 代入到这个积分中，得到

---

① 原著中误写为

$$K_\gamma(x,x_i)K_\gamma(y,y_i)$$ —— 译者

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\sum_{i=1}^{l}\beta K_\gamma(x',x_i)K_\gamma(y',y_i)\,dy'\,dF(x)$$

$$=\int_{-\infty}^{\infty}\sum_{i=1}^{l}\beta K_\gamma(x',x_i)\,dF(x)=1.$$

用 $F_l(x)$ 代替 $F(x)$，我们得到

$$\sum_{i=1}^{l}\frac{1}{l}\sum_{j=1}^{l}\beta K_\gamma(x_i,x_j)=1. \tag{7-60}$$

(4) 应用残差原则，选择从 Kolmogorov-Smirnov 型分布得到的

$$\sigma=\frac{q}{l},$$

并选择一个容许的参数 γ。

(5) 为控制解的性质(精度和/或稀疏性)，我们选择一个最小化目标泛函的容许参数 γ 和/或选择一个大的容许参数 γ。

因此，我们用(7-57)式来逼近条件密度函数，其中的系数 β 是通过求解下面的优化问题得到的：在约束条件(7-59)式和(7-60)式下最小化泛函(7-58)式。从容许参数集中选择 γ 来控制解的特性。

为了估计回归函数

$$r(x)=\int yp(y\mid x)\,dy, \tag{7-61}$$

回顾核 $K_\gamma(y,y_i)$ 是一个对称的(密度)函数，它的积分是 1。对这样一个函数，我们有

$$\int yK_\gamma(y,y_i)\,dy=y_i. \tag{7-62}$$

因此，从(7-57)式、(7-61)式和(7-62)式，我们得到下面的回归函数：

$$r(x)=\sum_{i=1}^{l}y_i\beta K_\gamma(x,x_i).$$

将这个表达式与下面的 Nadaraya-Watson 回归相比较是很有意思的：

$$r(x)=\sum_{i=1}^{l}y_i\frac{K_\gamma(x_i,x)}{\sum_{j=1}^{l}K_\gamma(x_j,x)}, \tag{7-63}$$

这里，括号中表达式是由 Parzen 密度估计定义的(括号中是 Parzen 密度估计的第 i 项与 Parzen 密度估计之比)。

SVM 回归是平滑的，而且有稀疏的表达。

7.11 评注

7.11.1 评注 1. 我们可以利用未知密度的一个好估计

在构造估计密度、条件概率和条件密度的算法时，我们用经验分布函数 $F_l(x)$ 作为真

实分布函数 $F(x)$ 的一个逼近。从 $F_l(x)$，我们得到密度函数的逼近

$$p(x) = \frac{1}{l}\sum_{i=1}^{l}\delta(x - x_i),$$

它是若干 δ 函数的和。事实上，用这一逼近可以得到相应的约束条件。

但是，我们可以用更好的密度逼近，它是基于第 7.8 节中描述的(稀疏的)SVM 估计的。采用密度函数的这个逼近，我们可以得到与这里不同的约束条件(或许更精确)。在第八章中，我们将介绍反映这一思想的一种新的风险最小化原则。

7.11.2 评注 2. 我们可以利用有标号的(训练)数据，也可以利用无标号的(测试)数据

为了估计条件概率函数和条件密度函数，我们可以利用两种数据：训练数据

$$(\omega, x_1), \cdots, (\omega, x_l) \tag{7-64}$$

和无标号的(测试)数据[①]：

$$x_1^*, \cdots, x_k^*.$$

因为根据我们的学习模型，训练数据和测试数据中的向量 x 有同样的分布 $F(x)$，它由产生器 G 产生的(见第一章)，我们可以用联合集合

$$x_1, \cdots, x_l, x_1^*, \cdots x_k^*$$

来估计分布 $F(x)$(或者是密度函数 $p(x)$)。为估计分布函数 $F(x|\omega)$，我们利用(7-64)式中的对应于 $\omega = \omega^*$ 的向量 x 的子集。

7.11.3 评注 3. 得到不适定问题的稀疏解的方法

在密度、条件概率和条件密度估计中使用的方法是有很大的一般性的。可以用这些方法来得到其他算子方程的稀疏解。

要得到这样的稀疏解，我们需要：

- 选择 RKHS 的模作为正则化因子；
- 在 E_2 中选择 L_∞ 度量；
- 采用残差原则；
- 从容许集中选择适当的 γ 值。

① 原著中为 $x^*, \cdots, x_l^*$ ——译者

非正式推导和评述——7

7.12 科学理论的三个要素

根据 Kant 的观点,任何理论都应包含三个要素:

(1) 问题的表示,

(2) 问题的解决,

(3) 证明。

第一眼看到这种说法,人们会觉得这是显然的,然而其意义是很深刻的。这一说法的关键在于它反映了一种思想,即理论的这三种因素在某种意义上是独立的且同等重要。

(1) 对于理解一个问题来说,对问题进行精确表示的重要性并不比解决问题或是证明这一解决的正确性更少。

(2) 对问题的解决并不是从对问题的深入理论分析而来,而是产生于这种分析之前。

(3) 证明并不是为了寻找问题的解,而是为了证明已经提出的解的正确性。

理论的前两个要素反映了对所关心问题的本质和其基本原理的理解,而证明使得一般的(原理性的)模型成为一个科学理论。

7.12.1 密度估计的问题

分析密度估计理论的发展过程,我们可以看到 Kant 的这一论点是多么深刻。传统的密度估计理论,包括参数和非参数估计,都只包括两个要素:对问题的解决和证明。它们不包括对问题的表示。

在参数估计中,Fisher 提出了最大似然方法(对问题的解决),之后,Le Cam(1953)、Ibragimov 和 Hasminski(1981)及其他学者证明,在某些条件下(条件不是很宽,见第 1.7.4 节),最大似然方法是一致的。

同样的情况也发生在问题的非参数解决方法上。首先人们提出了方法: 直方图方法(Rosenblatt, 1956)、Parzen 方法(Parzen, 1962)、投影方法(Chentsov, 1963)等等, 然后才证明了它们的一致性。与参数化方法相对比, 非参数方法在很宽的条件下是一致的。

由于缺少对问题的一般表示, 密度估计的方法看上去就像一个处方清单。同时, 这也导致了对这些方法可能的改进似乎只能是启发式的。这一点导致人们创造了对非参数方法实际应用的一大堆启发式修正。

尝试提出密度估计问题的一般表示是在 1978 年(Vapnik and Stefanyuk, 1978), 在所提出的表示中, 密度估计问题是直接从密度的定义得到的, 它被考虑为一个在方程右边未知、而给出了一些数据的情况下求解积分方程的问题。这种一般表示(因为它是从密度的定义得到的, 所以是一般性的)马上将密度估计理论与不适定问题的求解这一基本理论联系了起来。

7.12.2 不适定问题的理论

不适定问题的理论最早是为了解决数学物理反问题而发展起来的。但后来, 人们发现了这一理论的一般性。人们发现, 每当我们面对一个反演问题, 即需要从已知结果推出未知的原因时, 总要考虑到这一理论的论述。尤其是, 不适定问题理论的结果对统计反演问题是很重要的, 其中包括了密度估计问题、条件概率估计问题和条件密度估计问题。

不适定问题的存在是 Hadamard(1902)发现的。Hadamard 当时认为, 不适定问题是纯数学现象, 现实中的问题是适定的。但是, 很快人们就发现, 在现实中存在的一些重要问题是不适定的。

在 1943 年, A. N. Tikhonov 证明了关于反算子的一个引理, 描述了适定问题的特性, 从而发现了不适定问题正则化的方法。直到又过了 20 年以后, Phillips(1962)、Ivanov(1962)和 Tikhonov(1963)才得到了同样的构造性正则化思想, 只是描述形式略有不同。正则化理论中重要的信息就是, 在求解定义了不适定问题的算子方程

$$Af(t) = F(x)$$

的问题中, 其显而易见的解决方案——最小化泛函

$$R(f) = \| Af - F \|^2$$

并不能得到一个好的解。相反, 我们应该采用并不是显而易见的解决方案, 即最小化“恶化的”(正则化的)泛函

$$R^*(f) = \| Af - F \|^2 + \gamma\Omega(f),$$

在 60 年代初, 这一思想并不是显然的。现在, 人人都自然地接受了这一思想, 这反映了正则化理论对数学科学的不同分支尤其是统计学的深刻影响。

7.13 随机不适定问题

为了构造密度估计的一般理论, 有必要把解决不适定问题的理论推广到随机情况中。

把在确定性情况下提出的解决不适定问题的理论推广到随机不适定问题是很直接的。利用与求解确定性不适定问题相同的正则化技术,以及相同的基于反算子引理的核心论据,我们把关于正则化方法的主要定理推广到了一个随机模型上(V. Vapnik and A. Stefanyuk, 1978)。后来 Stefanyuk(1986)又把这一结果推广到了近似定义的算子的情况。

统计学中的主要问题——从比较宽的函数集中估计函数的问题——是不适定的,这一事实每个人都知道。尽管如此,对解决这一基本统计学问题尤其是密度估计问题的方法,却从没有人对它从正则化理论的形式化角度上进行考虑[①]。

的确,在统计学的传统中,人们首先是提出解决问题的某种方法,证明它的良好特性,然后再引入一些启发式修正,使得这种方法在实际任务中有用(尤其是对多维问题)。

我们从解决随机不适定问题的角度得到新的估计子,这一研究是从对密度估计问题的各种已知算法的分析开始的(Aidu and Vapnik, 1989)。我们发现,几乎所有的传统算法(比如 Parzen 方法和投影方法),都可以在求解随机不适定问题的标准正则化方法的基础上得到,条件是选择经验分布函数作为未知分布函数的一个逼近。

在那时,基于现有数据对未知分布函数构造一个更好的逼近,这一想法引发了构造新算法的尝试。利用这一思想,我们构造了一个新的估计,它证明了人们提出的很多对估计一维密度函数的启发性建议。

80 年代统计学的理论学者和应用学者都非常重视密度估计的非参数方法问题。其中主要的问题是寻找一个如何选择 Parzen 方法中最佳的宽度参数的规律。人们发现了把宽度值与关于实际密度的平滑性、核的特性和观测数目等信息联系起来的渐近原则。

但是,对应用者来说,这些结果并不充分。原因有二:其一,这些结论只对充分大的数据集成立;其二,对一个自由参数的估计是基于某些未知参数的(比如平滑性参数,它是由未知密度所具有的导数的阶数定义的)。

因此,应用者们研究了他们自己的估计宽度参数的方法。在这些方法中,留一法(leave-one-out)估计成为最常用的方法。有大量的文献都是关于对宽度参数的实验分析。

在 80 年代末,我们提出了估计正则化参数(宽度参数)的残差方法(Vapnik, 1988)。已经证明,这种方法几乎是最优的(Vapnik et al, 1992)。同时,在对一个很宽的一维密度集合的实验中,也说明了这种选择宽度参数的方法比很多其他理论的或启发式的方法性能更好(Markovich, 1989)。

遗憾的是,多数关于密度估计的结果都是针对一维情况的,而密度估计问题的主要应用在于多维情况。对这种情况,人们提出了一些特殊的方法。

这些方法中最常用是高斯混合模型(Gaussian mixture model)方法,人们发现它是不一致的(见第 1.7.4 节)。尽管如此,这种方法仍被用在多数密度估计的高维(比如 50 维)问题中(比如在语音识别中)。

---

① 一种可能的解释是,密度估计的非参数方法的理论是在 50 年代开始的,是在解决不适定问题的正则化方法发现之前。在 60 年代末和 70 年代,当不适定问题理论吸引了不同数学分支的很多学者的注意力时,分析密度估计问题的框架已经建立起来了。

然而,我们知道,即使是构造一个好的二维密度估计,也需要新的思想。

然而,真正的挑战在于,要寻找定义在有界支集上多维密度的一个好的估计。

在这一章里,我们提出了多维密度估计的一种新方法。它结合了数学的三个分支中的思想:用残差原则求解积分方程的理论、通用 Kolmogorov-Smirnov 分布(它使我们可以估计残差原则的参数)及统计学习理论中的 SVM 技术(它是设计用来逼近高维空间中的函数的)。

我们还对解决一维密度估计问题检查了这三方面思想中的两个(Vapnik, 1988; Aidu and Vapnik, 1989; Vapnik et al, 1992 及 Markovich, 1989)。

第三个思想是用 RKHS 空间中的模作为正则化泛函,并用 L_∞ 模来度量函数差,它是采用 ε 不敏感损失函数的 SVM 函数逼近方法的直接结果,对此在本书第一版中首次进行了介绍。它部分地在一维密度估计中得到了检查。

本章描述方法的实现是由 Sayan Mukherjee 完成的。他对在一维、二维和六维空间中估计密度的实验显示出所得解具有高的精度和好的稀疏性。本书中给出了其中的两个实验结果。

本章描述的条件概率和条件密度估计问题的直接解决方法,是直接密度估计方法的一个简单推广。这些方法尚未在实验中进行检查。

第八章

邻域风险最小化原则与 SVM

在这一章中,我们介绍最小化期望风险的一种新原则,称作邻域风险最小化(vicinal risk minimization, VRM)原则[①]。用这一原则来解决我们的主要问题:模式识别、回归估计和密度估计。

我们用 SVM 技术来最小化邻域风险泛函,得到的解的形式为核的展开式,而这些核对不同训练点是不同的。

8.1 邻域风险最小化原则

我们再来考虑函数估计问题的标准表示:在函数集 $f(x, \alpha)$, $\alpha \in \Lambda$ 中最小化泛函

$$R(\alpha) = \int L(y - f(x, \alpha)) dP(x, y), \tag{8-1}$$

其中,$L(u)$ 是一个给定的损失函数,概率测度 $P(x, y)$ 未知但是给出了数据

$$(y_1, x_1), \cdots, (y_l, x_l). \tag{8-2}$$

在本书第一章中,为了解决这个问题,我们考虑了经验风险最小化原则,它是要最小化泛函

$$R_{emp}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} L(y_i - f(x_i, \alpha)), \tag{8-3}$$

以取代泛函(8-1)式。

之后,我们介绍了结构风险最小化原则,其中我们在函数集 $f(x, \alpha)$, $\alpha \in \Lambda$ 上定义一个结构:

$$S_1 \subset \cdots \subset S_n,$$

---

① 之所以用这个名字,是为了强调我们的目标是在训练向量 x_i, $i = 1, \cdots, l$ 的邻域 $x \in v(x_j)$ 内最小化风险,不是指最小化只由训练向量定义的经验风险,其中(我们认为)在这些邻域 $x \in v(x_i)$ 中的多数点与训练向量 x_i 保持相同的(或几乎相同的)值 y_i。

然后在这个结构的某个适当选择的元素 S_k 上最小化泛函(8-3)式。

现在,我们考虑一个新的基本泛函来取代经验风险泛函(8-3)式,并将这个泛函用到结构风险最小化过程中。

注意到,引入经验风险泛函的原因是由于:我们的目标是在概率测度未知的情况下最小化期望风险(8-1)式。让我们根据数据来估计密度函数,然后在泛函(8-1)式中用这个估计 $p(x,y)$ 得到目标泛函

$$R_T(\alpha)=\int (L(y-f(x,\alpha))p(x,y)\,dxdy. \tag{8-4}$$

当我们用 δ 函数之和来估计未知密度时,即

$$p(x,y)=\frac{1}{l}\sum_{i=1}^{l}\delta(x-x_i)\,\delta(y-y_i),$$

我们就得到了经验风险泛函。

如果我们认为密度函数和目标函数都是平滑的,那么经验风险泛函可能就不是对期望风险泛函最好的逼近。于是问题就出现了:是否存在反映下面两个假设的对风险泛函更好的逼近呢?这两个假设是:

(1) 未知密度函数任意点 x_i 的一个邻域内是平滑的;

(2) 使得风险泛函最小的函数也是平滑的,且在任意点 x_i 的邻域内是对称的。

下面介绍一个新的目标泛函,我们将用它来代替经验风险泛函。为了引入这个泛函,我们对所有训练向量(用数据)构造向量 x_i 的邻域函数 $v(x_i)$,然后用这些邻域函数构造目标泛函。正如第 4.5 节中那样,我们区分两种邻域函数,即硬邻域函数和软邻域函数。人们也可以根据所面临的实际问题采用其他更合适的邻域函数。

8.1.1 硬邻域函数

(1) 对任意的 x_i, $i=1,\cdots,l$,我们定义集合 $X\in R^n$ 的一个可测子集 $v(x_i)$(即点 x_i 的邻域),其体积是 v_i。

我们可以把这个点的邻域定义为与 $x_i=(x_i^1,\cdots,x_i^n)$ 为 r_i-接近的点的集合(r_i 依赖于点 x_i),即

$$v(x_i)=\{x:\ \|x-x_i\|_E\le r_i\},$$

其中,$\|x-x_i\|_E$ 是空间 E 中的一种度量,比如它可以是 l_1, l_2 或 l_∞ 度量,l_1 度量定义的邻域是集合:

$$v(x_i)=\left\{x:\ \sum_{k=1}^{n}\left|x^k-x_i^k\right|\le r_i\right\},$$

l_2 度量定义的邻域是以点 x_i 为中心、半径为 r_i 的球①:

① 原著中误写为 $v(x_i)=\left\{x:\ \sum_{k=1}^{n}\left|x-x_i\right|^2\le r_i^2\right\}$。——译者

$$v(x_i) = \left\{ x: \sum_{k=1}^{n} \left| x^k - x_i^k \right|^2 \leq r_i^2 \right\},$$

而 l_∞ 度量定义的邻域是一个以点 $x_i = (x_i^1, \cdots, x_i^n)$ 为中心、大小为 $2r_i$ 的立方体:

$$v(x_i) = \{x: x_i^k - r_i \leq x^k \leq x_i^k + r_i, \quad \forall k = 1, \cdots, n\}.$$

(2) 不同训练向量的邻域中没有共同的点。

(3) 像下面这样在向量 x_i 的这个邻域中逼近未知密度函数 $p(x)$, 训练数据的所有 l 个邻域都有相同的概率测度:

$$P(x \in v(x_i)) = \frac{1}{l},$$

在邻域内向量的分布是均匀的

$$P(x \in v(x_i)) = \frac{1}{\nu},$$

其中 ν 是邻域 $v(x_i)$ 的体积。

图 8.1 显示了不同度量下点的邻域: (a) 在 l_1 度量下的邻域, (b) 在 l_2 度量下的邻域, (c) 在 l_∞ 度量下的邻域。

图 8.1 不同度量下点的邻域

考虑下面的泛函

$$V(\theta) = \frac{1}{l} \sum_{i=1}^{l} L(y_i, f(x_i, \theta)) - \int_{v(x_j)} f(x, \theta) dx, \tag{8-5}$$

我们称之为邻域风险泛函。

为了找到对最小化风险泛函(8-1)式的函数的一个逼近, 我们寻找使泛函(8-5)式最小的函数。用最小化泛函(8-5)式来取代最小化泛函(8-1)式, 我们把这种方法称作邻域风险最小化(VRM)原则(方法)[①]。注意, 当 $\nu \to 0$ 时, 邻域风险泛函收敛于经验风险泛函。

因为对不同的训练数据点, 邻域的体积可以不同, 通过引入这个泛函, 我们期望使之最小化的函数在不同点的邻域内有不同的平滑特性。

在某种意义上, VRM 方法结合了两种不同的估计方法: 经验风险最小化方法和最近邻方法(1-近邻方法)。

① 邻域风险最小化原文为 vicinal risk minimization, 因 vicinal(邻近的)一词也有"本地的、局部的"之意, 故也可译为本地风险最小化。——译者

8.1.2 软邻域函数

在对邻域方法的定义中，我们利用从训练数据得到的参数 x_i 和 r_i 构造了一个均匀分布函数，在 VRM 的公式中使用了这一分布函数。

然而，我们可以用这些参数来构造其他的分布函数 $p(x|x_i, r_i)$，在其中用这些参数来定义分布函数的位置和宽度的参数（比如，我们可以用正态分布函数 $p(x|x_i, r_i) = N(x_i, d_i)$）。这样定义的函数称作软邻域函数。对软邻域函数来说，空间的所有点都可以属于向量 x_i 的邻域，但它们有不同的度量。

软邻域函数定义了如下（一般）形式的 VRM：

$$V(\alpha) = \frac{1}{l}\sum_{i=1}^{l} L(y_i - Ef(x, \alpha))$$

$$= \frac{1}{l}\sum_{i=1}^{l} L\left(y_i - \int f(x, \alpha)p(x|x_i, r_i)dx\right)$$

在 8.3.1 小节中，我们将定义一种基于硬邻域函数和基于软邻域函数的 VRM 方法。

8.2 用于模式识别问题的 VRM 方法

在这一节里，我们把 VRM 方法应用到两类 $\{-1, 1\}$ 模式识别问题中。考虑下面的指示函数集：

$$y = g(x, \alpha) = \mathrm{sgn}[f(x, \alpha)], \tag{8-6}$$

其中，$f(x, \alpha)$，$\alpha \in \Lambda$ 是一个实值函数集合。在前面的几章中，我们没有对指示函数的结构（8-6）式给予关注。为了从 $f(x, \alpha)$，$\alpha \in \Lambda$ 中找到最小化风险泛函的函数，我们在损失函数 $|y - f(x, \alpha)|$ 下最小化经验泛函（8-3）式。

现在考虑到指示函数的结构（8-6）式，我们考虑另一种损失函数

$$L(y, f(x, \alpha)) = \theta(-yf(x, \alpha)), \tag{8-7}$$

它定义了风险泛函

$$R(\alpha) = \int \theta[-yf(x, \alpha)]dP(x, y), \tag{8-8}$$

其中 $\theta(u)$ 是阶跃函数。

为最小化这个泛函，VRM 方法提出要最小化泛函

$$V(\alpha) = \frac{1}{l}\sum_{i=1}^{l} \theta\left(-\int_{v(x_i)} f(x, \alpha)p(x|x_i, r_i)dx\right). \tag{8-9}$$

对硬邻域函数，我们得到

$$V(\alpha) = \frac{1}{l}\sum_{i=1}^{l} \theta\left(-\frac{y_i}{v}\int_{v(x_i)} f(x, \alpha)dx\right).$$

就像在第五章中一样，我们把这个问题简化为下面的优化问题：最小化泛函

$$W(f) = C\sum_{i=1}^{l} \xi_i + \Omega(f), \tag{8-10}$$

约束条件是

$$\int_{y_i} f(x, \alpha) p(x - x_i, r_i)\,dx \geq 1 - \xi_i, \tag{8-11}$$

其中 $\Omega(f)$ 是某个正则化泛函,我们将在后面再定义。

假设我们的函数集是这样定义的:把输入向量 x 映射到特征空间中的特征向量 z,在特征空间中构造一个超平面①

$$(w \cdot z) + b = 0,$$

它把数据

$$(y_1, z_1), \cdots, (y_l, z_l)$$

分开,这些数据是我们的训练数据(8-2)式在特征空间中的像。(设核 $K(x, x')$ 定义了这个特征空间中的内积。)

我们的目标是找到函数 $f(x, \alpha)$,它满足下面的条件:

$$\int_{y_i} f(x, \alpha) p(x - x_i, r_i)\,dx \geq 1 - \xi_i, \tag{8-12}$$

且该函数在特征空间中的像是一个线性函数

$$l(z) = (w^* \cdot z) + b,$$

它最小化泛函

$$W(w) = (w \cdot w) + C\sum_{i=1}^{l} \xi_i. \tag{8-13}$$

我们将用 SVM 技术来解决这个问题,并称此解为邻域 SVM 解(vicinal SVM 简称 VSV)。注意到,对输入空间中的线性函数

$$f(x, \alpha) = (w \cdot x) + b, \quad \alpha \in \Lambda$$

和以 x_i 为其质心的邻域

$$x_i = E_{v(x_i)} x,$$

VSV 解与 SVM 解重合。这是因为,在这两种情况下的目标泛函是相同的,而且有

$$\int [(w \cdot x) + b] p(x - x_i, r_i)\,dx = (w \cdot x_i) + b,$$

所以两个问题重合。

ERM 与 VRM 的区别会出现在两种情况下:①若点 x_i 不是邻域 $v(x_i)$ 的质心,②若我们考虑的是非线性函数。

让我们(用核 $K(x, x')$)引入两个新的核:单邻域核

$$(x_i, x) = E_{v(x_i)} K(x, x') = \int K(x, x') p(x - x_i, r_i)\,dx' \tag{8-14}$$

与双邻域核

$$(x_i, x_j) = E_{v(x_i)} E_{v(x_j)} K(x, x')$$

---

① 原著中把 $(w \cdot z)$ 或 $(w^* \cdot z)$ 误写为 (w, z) 或 (w^*, z),下面几式中类似。——译者

$$= \iint K(x, x')p(x \mid x_i, r_i)p(x' \mid x_j, r_j)dxdx' \tag{8-15}$$

有下面的定理成立。

定理 8.1 邻域支持向量解(VSV)有下面的形式:

$$f(x) = \sum_{i=1}^{l} \beta_i K(x, x_i) + b, \tag{8-16}$$

其中,为了定义系数 β,我们必须最大化泛函

$$W(\beta) = \sum_{i=1}^{l} \beta_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \beta_i \beta_j K(x_i, x_j), \tag{8-17}$$

约束条件是

$$\sum_{i=1}^{l} y_i \beta_i = 0, \tag{8-18}$$

$$0 \leq \beta_i \leq C. \tag{8-19}$$

证明 让我们把向量 x 映射到特征向量 z。考虑从点 x_i, $i = 1, \cdots, l$ 的邻域中抽取的 N 个样本点

$$x_{i_1}, \cdots, x_{i_N}, \quad i = 1, \cdots, l.$$

设这些点在特征空间中的像为

$$z_{i_1}, \cdots, z_{i_N}, \quad i = 1, \cdots, l.$$

考虑在特征空间中构造下面的邻域最优超平面的问题:最小化泛函①

$$W^*(w) = \frac{1}{2}(w \cdot w) + C \sum_{i=1}^{l} \xi_i, \tag{8-20}$$

约束条件是

$$y_i \frac{1}{N} \sum_{k=1}^{N} [(w \cdot z_{i_k}) + b] \geq 1 - \xi_i. \tag{8-21}$$

注意,(8-21)式在输入空间中的等效表达式是

$$\frac{y_i}{N} \sum_{k=1}^{N} f(x_{i_k}, \alpha) \geq 1 - \xi_i. \tag{8-22}$$

当 $N \to \infty$,表达式(8-22)收敛于(8-12)式。因此,由(8-20)式和(8-21)式定义的优化问题的解收敛于由(8-13)式和(8-12)式定义的优化问题的解。

为了在约束条件(8-21)式下最小化(8-20)式,我们引入拉格朗日函数

$$L(w) = \frac{1}{2}(w \cdot w) + C \sum_{i=1}^{l} \xi_i - \sum_{i=1}^{l} \beta_i \left(y_i \frac{1}{N} \sum_{k=1}^{N} (w \cdot z_{i_k}) + b - 1 + \xi_i \right) + \sum_{i=1}^{l} \eta_i \xi_i. \tag{8-23}$$

我们的优化问题的解是由这个拉格朗日函数的鞍点定义的,这个鞍点对 b, ξ 和 w 使得泛函最小化,对 β 和 η 使得泛函最大化。作为最小化的结果,我们得到

---

① 式中$(w \cdot w)$或$(w \cdot z_{i_k})$在原著中误写为(w, w)或(w, z_{i_k}),下文还有相同情况,不再一一标出。——译者

$$\sum_{i=1}^{l} y_i \beta_i = 0, \tag{8-24}$$

$$0 \leq \beta_i \leq C, \tag{8-25}$$

和

$$w = \sum_{i=1}^{l} \beta_i \frac{1}{N} \sum_{k=1}^{N} z_{i_k}. \tag{8-26}$$

把(8-26)式代入到超平面的表达式中,得到

$$l(z) = (w \cdot z) + b = \sum_{i=1}^{l} y_i \beta_i \frac{1}{N} \sum_{k=1}^{N} (z \cdot z_{i_k}) + b, \tag{8-27}$$

把(8-26)式代回到拉格朗日函数中,得到

$$W(\beta) = \sum_{i=1}^{l} \beta_i - \frac{1}{2} \sum_{i,j=1}^{l} \beta_i \beta_j y_i y_j \frac{1}{N} \sum_{k=1}^{N} \frac{1}{N} \sum_{m=1}^{N} (z_{i_k} \cdot z_{j_m}). \tag{8-28}$$

因为$(z, z') = K(x, x')$,我们可以把(8-27)式和(8-28)式重写为下面的形式:

$$f(x) = \sum_{i=1}^{l} y_i \beta_i \frac{1}{N} \sum_{k=1}^{N} K(x, x_{i_k}) + b, \tag{8-29}$$

其中,系数 β 在约束(8-24)式和(8-25)式下最大化泛函

$$W(\beta) = \sum_{i=1}^{l} \beta_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \beta_i \beta_j \frac{1}{N} \sum_{k=1}^{N} \frac{1}{N} \sum_{m=1}^{N} K(x_{i_k}, x_{j_m}).$$

增大 N,我们得到

$$\lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} K(x, x_{i_k}) = \quad (x, x_i),$$

$$\lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} \frac{1}{N} \sum_{m=1}^{N} K(x_{i_k}, x_{j_m}) = \quad (x_i, x_j).$$

因此,VSV 解是

$$f(x) = \sum_{i=1}^{l} y_i \beta_i \quad (x, x_i) + b, \tag{8-30}$$

其中,为定义系数 β,我们必须在约束条件

$$\sum_{i=1}^{l} y_i \beta_i = 0,$$

$$0 \leq \beta_i \leq C$$

下最大化泛函

$$W(\beta) = \sum_{i=1}^{l} \beta_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \beta_i \beta_j \quad (x_i, x_j). \tag{8-31}$$

8.3 邻域核的例子

在这一节里,我们将给出一对邻域和核 $K(x, y)$的例子,它使得我们可以用解析形式构造单邻域核 (x, x_i)和双邻域核 (x_i, x_j)。在 8.3.1 小节中,我们将介绍对于硬邻域函数的这些核;在 8.3.2 小节中介绍对软邻域函数的这些核。

8.3.1 硬邻域函数

我们用 l_∞ 度量:

$$\| x - x_i \|_\infty = \sup_{1 \le k \le n} | x^k - x_i^k | \tag{8-32}$$

来定义点 x_i, $i = 1, \cdots, l$ 的邻域, 其中 $x = (x^1, \cdots, x^n)$ 是 R^n 中的一个向量。

我们用下面的算法根据训练数据

$$(y_1, x_1), \cdots, (y_l, x_l)$$

定义向量 x_i, $i = 1, \cdots, l$ 的邻域的大小:

(1) 定义三角矩阵①

$$A = [a_{i,j}], \quad i > j,$$

其元素为(l_∞度量下) 训练集中向量两两之间的距离。

(2) 定义矩阵 A 的最小元素(比如说是 a_{ij})。

(3) 把值

$$d_i = K a_{ij}$$

赋给元素 x_i, 把值

$$d_j = K a_{ij}$$

赋给元素 x_j。

这里, $K \le 1/2$ 是控制着邻域大小的参数(通常选择最大可能尺寸 $K = 1/2$ 是合理的)。

(4) 选择矩阵 A 的下一个最小的元素 a_{ms}。如果它对应的两个向量中的一个(比如说 x_m) 已经被赋予某值 d_m, 那么把值

$$d_s = K a_{ms}$$

赋给另外一个向量 x_s; 否则把这个值赋给这两个向量。

(5) 继续上述过程, 直到 d 值被赋给了所有的向量。

利用 d_i 值, 我们定义点 x_i 的邻域

$$v(x_i) = \{ x : x_i^k - d_i \le x^k \le x_i^k + d_i, \quad \forall k = 1, \cdots, n \}$$

和这个邻域的体积

$$V = (2d_i)^n.$$

我们引入记号

$$v(x_i^k) = \{ x^k : x_i^k - d_i \le x^k \le x_i^k + d_i \}.$$

下面, 我们对拉普拉斯类型的核

$$K(x, x') = \exp\left\{ - \frac{\| x - x' \|_{l_1}}{\Delta} \right\} = \prod_{k=1}^{n} \exp\left\{ - \frac{| x^k - (x')^k |}{\Delta} \right\}$$

计算单邻域和双邻域核。

我们得到单邻域核②

---

① 原著误写为 $A = \| a_{i,j} \|$。——译者

② 译文中统一为 $\| x - x' \|$。——译者

$$
\begin{aligned}
(x, x_i) &= \frac{1}{(2d_i)^n}\int_{v(x_i)} \exp\left(-\frac{|x - x'|}{\Delta}\right) dx' \\
&= \frac{1}{2^n d_i^n}\prod_{k=1}^{n}\int_{v(x_i^k)} \exp\left(-\frac{|x^k - (x')^k|}{\Delta}\right) d(x')^k \\
&= \prod_{k=1}^{n} {}^k(x^k, x_i^k).
\end{aligned}
$$

经过基本的运算之后，我们得到

$$
{}^k(x^k, x_i^k) = \frac{1}{2d_i}\int_{v(x_i^k)} \exp\left(-\frac{|x^k - (x')^k|}{\Delta}\right) d(x')^k
$$

$$
= \begin{cases}
\dfrac{\Delta}{2d_i}\left[2 - \exp\left(-\dfrac{(d_i + x^k - x_i^k)}{\Delta}\right) - \exp\left(-\dfrac{(d_i - x^k + x_i^k)}{\Delta}\right)\right] & \text{若} |x_i^k - x^k| \le d_i, \\
\dfrac{\Delta}{2d_i}\exp\left(-\dfrac{|x^k - x_i^k|}{\Delta}\right)\left[\exp\left(\dfrac{d_i}{\Delta}\right) - \exp\left(-\dfrac{d_i}{\Delta}\right)\right] & \text{若} |x_i^k - x^k| \ge d_i.
\end{cases}
$$

n 维双邻域核是一维核的乘积：

$$
(x_i, x_j) = \prod_{k=1}^{n} {}^k(x_i^k, x_j^k).
$$

要计算 ${}^k(x_i^k, x_j^k)$，我们区分下面的两种情况：$i \neq j$ 的情况（比如 $i > j$）和 $i = j$ 的情况。对 $i \neq j$ 情况，（考虑到不同的邻域中没有相同的点）我们得到[①]

$$
{}^k(x_i^k, x_j^k) = \frac{1}{4d_i d_j}\int_{v(x_i^k)}\int_{v(x_j^k)} -\frac{|x^k - (x')^k|}{\Delta} d(x')^k dx^k
$$

$$
= \frac{\Delta^2}{4d_i d_j}\exp\left(-\frac{|x^k - (x')^k|}{\Delta}\right)\left[e^{\frac{d_j}{\Delta}} - e^{-\frac{d_j}{\Delta}}\right]\left[e^{\frac{d_i}{\Delta}} - e^{-\frac{d_i}{\Delta}}\right].
$$

对 $i = j$ 的情况，我们得到[②]

$$
{}^k(x_i^k, x_j^k) = \frac{1}{4d_i^2}\int_{v(x_i^k)}\int_{v(x_j^k)} \exp\left(-\frac{|x^k - (x')^k|}{\Delta}\right) d(x')^k dx^k
$$

$$
= \frac{2}{4d_i^2}\int_{v(x_i^k)} dx^k \int_{x_i^k - d_i}^{x} \exp\left(-\frac{(x^k - (x')^k)}{\Delta}\right) d(x')^k
$$

$$
= \frac{\Delta^2}{2d_i^2}\left[e^{\frac{2d_i}{\Delta}} - 1 - \frac{2d_i}{\Delta}\right],
$$

因此，我们有

$$
{}^k(x_i^k, x_j^k) = \begin{cases}
\dfrac{\Delta^2}{4d_i d_j} e^{-\frac{|x_i^k - x_j^k|}{\Delta}}\left[e^{\frac{d_j}{\Delta}} - e^{-\frac{d_j}{\Delta}}\right]\left[e^{\frac{d_i}{\Delta}} - e^{-\frac{d_i}{\Delta}}\right] & \text{若 } i \neq j, \\
\dfrac{\Delta^2}{2d_i^2}\left[e^{\frac{2d_i}{\Delta}} - 1 - \dfrac{2d_i}{\Delta}\right] & \text{若 } i = j.
\end{cases}
$$

---

① 原著中把 $d(x')^k dx^k$ 误写为 $dx'dx$。——译者

② 原著中把 $4d_i^2, 2d_i^2$ 误写为 $4\delta^2, 2\delta^2$。——译者

注意到,当

$$\frac{d}{\Delta} \quad 0$$

时,我们就得到传统的 SVM 解

$$(x, x_i) \quad K(x, x_i),$$

$$(x_i, x_j) \quad K(x_i, x_j).$$

图 8.2 显示出了从拉普拉斯函数得到的单邻域核,其中参数 $\Delta = 0.25$,采用了 3 种不同的邻域:(a) $d = 0.02$,(b) $d = 0.5$ 和(c) $d = 1$。注意到,点 x_i 的邻域越大,则在这个邻域内的核逼近函数就越平滑。

图 8.2 从拉普拉斯函数得到的单邻域核($\Delta = 0.25$)

8.3.2 软邻域函数

对高斯型的核

$$K(x, x') = \exp - \frac{(x - x')^2}{2\gamma^2}$$

构造单邻域核和双邻域核,我们相应采用以下步骤:

1. 用 l_2 度量定义两点之间的距离。
2. 用与上小节同样的算法,对训练数据的所有点 x_i 定义值 d_i。
3. 用参数为 x_i 和 d_i 的正态分布定义软邻域函数。
4. 计算单邻域函数和双邻域函数[①]:

$$(x, x_i) = \frac{1}{(2\pi)^{\frac{n}{2}} d_i^n} \int \exp - \frac{(x - x')^2}{2\gamma^2} \exp - \frac{(x' - x_i)^2}{2d_i^2} dx'$$

$$= \left(1 + \frac{d_i^2}{\gamma^2}\right)^{-\frac{n}{2}} \exp - \frac{(x - x_i)^2}{2(\gamma^2 + d_i^2)},$$

$$(x_j, x_i) = \frac{1}{(2\pi d_i d_j)^n} \exp - \frac{(x - x')^2}{2\gamma^2} - \frac{(x' - x_i)^2}{2d_i^2} - \frac{(x - x_j)^2}{2d_j^2} dx dx'$$

$$= \left(1 + \frac{d_i^2}{\gamma^2} + \frac{d_j^2}{\gamma^2}\right)^{-\frac{n}{2}} \exp - \frac{(x - x')^2}{2(\gamma^2 + d_i^2 + d_j^2)}.$$

① 原著中此处两个公式有几处错误,在翻译时根据作者后来的说明进行了相应的修改。——译者

8.4 非对称邻域

在上一节中,为了得到邻域函数的解析表达,我们考虑了对称邻域。这种邻域反映了所面对问题的最简单的信息。现在,我们的目标是定义这样的邻域,它使我们能构造出反映某些局部不变量的邻域核。

下面,我们考虑对数字识别问题构造这样的核的例子。在这个例子中介绍的主要思想对于各种函数估计问题也适用。

我们知道,二维图像 x_i 任意的小连续线性变换都可以用 6 个函数(Lie 导数) $x^*_{i,k}$, $k=1,\cdots,6$ 来描述,变换后的图像是

$$x = x_i + \sum_{k=1}^{6} x^*_{i,k} t_k,$$

其中, t_k, $k=1,\cdots,6$ 是适当的小值。因此,图像 x_i 的不同小线性变换是由 x_i 的 6 个 Lie 导数和不同的小向量 $t=(t_1,\cdots,t_6)$ (比如 $\|t\| \le c$) 定义的。

让我们引入 x_i 如下的邻域:

$$v_L(x_i) = \{x: x = x_i + \sum_{k=1}^{6} x^*_{i,k} t_k, \quad \|t\| \le c\}.$$

这个邻域不一定是对称的。注意,如果我们能够构造单邻域和双邻域核:

$$_L(x, x_i) = E_{v_L(x_i)} K(x, x'),$$

$$_L(x_i, x_j) = E_{v_L(x_i)} E_{v_L(x_j)} K(x, x'),$$

那么, VSV 解

$$f_L(x, \alpha) = \sum_{i=1}^{l} y_i \alpha_{\ L}(x, x_i)$$

将考虑对于小 Lie 变换的不变性。

当然,要得到邻域核的解析形式并不容易。但是,我们可以用下面的求和逼近这些核:

$$_L(x, x_i) = \frac{1}{N}\sum_{k=1}^{N} E_{v(x_k(x_i))} K(x, x') = \frac{1}{N}\sum_{k=1}^{N} (x, x_k(x_i)),$$

$$_L(x_i, x_j) = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{N}\sum_{m=1}^{N} E_{v(x_k(x_i))} E_{v(x_m(x_j))} K(x, x')$$

$$= \frac{1}{N^2}\sum_{k=1}^{N}\sum_{m=1}^{N} (x_k(x_i), x_m(x_j)),$$

其中, $x_k(x_i)$, $k=1,\cdots,N$ 是从 x_i 利用小 Lie 变换得到的虚拟样本, $v(x_k(x_i))$ 是从 x_i 得到的第 k 个虚拟样本 $x_k(x_i)$ 的对称邻域。

换句话说,我们可以利用(从样本 x_i 得到的)虚拟样本的对称邻域的组合来逼近样本 x_i 的一个非对称邻域。

注意到,为了在数字识别问题中得到当前最好的性能,有多位学者都采用了虚拟样本来增加训练样本的数目(Y. LeCun et al, 1998; P. Simard et al, 1993 及 B. Scholkopf et

al, 1996)。

在 SVM 方法中, B. Scholkopf 等人考虑的解是在训练数据的扩展集上的展开

$$f(x,\alpha)=\sum_{i=1}^{l}y_i\sum_{k=1}^{N}\alpha_k K(x,x_k(x_i)), \tag{8-33}$$

其中, 扩展集既包括训练数据, 也包括用 Lie 变换从训练数据得到的虚拟样本。

在简化的邻域方法(其中控制邻域 $v(x_i)$ 的系数 K 非常小, 以至于 $(x,x_i)=K(x,x_i)$)中, 我们得到另一种展开:

$$f^*(x,\alpha)=\sum_{i=1}^{l}y_i\alpha\frac{1}{N}\sum_{k=1}^{N}K(x,x_k(x_i)), \tag{8-34}$$

其中的 $x_k(x_i)$ 是从训练数据中的向量 x_i 得到的第 k 个虚拟样本。

解 $f(x,\alpha)$ 与 $f^*(x,\alpha)$ 之间的区别可以描述如下:

- 在 $f(x,\alpha)$ 中, 我们利用的信息是新样本(虚拟样本)与样本 x_i 属于同一类。
- 在 $f^*(x,\alpha)$ 中, 我们利用的信息是新样本(虚拟样本)与 x_i 是相同的样本。

即使是在不能构造虚拟样本的情况下, 仍可以把非对称邻域构造为对称邻域之组合。我们可以把属于相同类的一个(小的)样本聚类看作是同一个组合中的样本。

8.5 对于估计实值函数的推广

在第六章中, 为了从一个给定函数集中估计一个实值函数, 我们采用了 ε不敏感损失函数

$$L(y,f(x,\alpha))=L(|y-f(x,\alpha)|_\varepsilon).$$

对这个损失函数, 我们构造了经验风险泛函①

$$R_{emp}(\alpha)=\frac{1}{l}\sum_{i=1}^{l}L(|y_i-f(x_i,\alpha)|_\varepsilon). \tag{8-35}$$

现在, 我们用下面的邻域风险泛函来代替泛函(8-35)式:

$$V(\alpha)=\frac{1}{l}\sum_{i=1}^{l}L\left(|y_i-\int f(x,\alpha)p(x|x_i,d_i)\,dx|_\varepsilon\right). \tag{8-36}$$

我们可以把最小化(8-36)式的问题重写为下面的形式: 最小化泛函

$$\Phi(\xi)=\sum_{i=1}^{l}L(\xi),\quad \xi\geq 0, \tag{8-37}$$

约束条件是

$$\begin{aligned} y_i-\int f(x,\alpha)p(x|x_i,d_i)\,dx &\geq -\varepsilon-\xi,\\ y_i-\int f(x,\alpha)p(x|x_i,d_i)\,dx &\leq \varepsilon+\xi. \end{aligned} \tag{8-38}$$

然而, 我们希望最小化正则化泛函

① 式(8-35)中$|y_i-f(x_i,\alpha)|_\varepsilon$部分, 原著中误写为$|y-f(x,\alpha)|_\varepsilon$ ——译者

$$B(f) = C\sum_{i=1}^{l} L(\xi_i) + \Omega(f), \tag{8-39}$$

而不是(8-37)式,其中的泛函 $\Omega(f)$ 将在下面定义。

就像在 8.2 节中那样,假设我们的函数集是如下定义的:把输入向量 x 映射到特征向量 z,在特征空间中,构造一个线性函数

$$l(z) = (w \cdot z) + b,$$

它逼近训练数据(8-2)式在特征空间中的像数据

$$(y_1, z_1), \cdots, (y_l, z_l).$$

设定义了特征空间中内积的核是 $K(x, x')$。

我们将定义满足约束条件(8-38)式的函数并最小化泛函

$$\Phi = C\sum_{i=1}^{l} \xi_i + (w \cdot w).$$

考虑 $L(u) = |u|_\varepsilon$ 的情况。

有下面的定理成立。

定理 8.2　邻域支持向量解的形式为

$$f(x) = \sum_{i=1}^{l} (\beta_i - \beta_i^*) K(x_i, x) + b,$$

为定义系数 β 和 β^*,我们须在约束条件

$$\sum_{i=1}^{l} \beta_i = \sum_{i=1}^{l} \beta_i^*$$
$$0 \leq \beta_i \leq C$$
$$0 \leq \beta_i^* \leq C$$

下最大化泛函①

$$W(\beta, \beta^*) = -\varepsilon\sum_{i=1}^{l} (\beta_i + \beta_i^*) + \sum_{i=1}^{l} y_i(\beta_i - \beta_i^*) - \frac{1}{2}\sum_{i,j=1}^{l} (\beta_i - \beta_i^*)(\beta_j - \beta_j^*) K(x_i, x_j),$$

其中,邻域核 $K(x_i, x)$ 和 $K(x_i, x_j)$ 是由(8-14)式和(8-15)式定义的。

这一定理的证明与定理 8.1 的证明相同。

我们可以对不同的损失函数 $L(u) = L(|y - f(x, \alpha)|_\varepsilon)$ 证明类似的定理。尤其是,对 $L = (y - f(x, \alpha))^2$ 的情况,我们可以得到闭合形式的解。

定理8.3　对损失函数

$$L = (y - f(x, \alpha))^2,$$

VSV 解是

$$f(x) = Y^{\mathrm{T}}\left(M + \frac{1}{C}I\right)^{-1} L,$$

---

① 原著中将等式右边第二项误写为 $\sum_{i=1}^{l} y_i(\beta_i - \beta_i^*)\sum_{i=1}^{l} \beta_i$ —— 译者

其中,

$$Y^T = (y_1, \cdots, y_l)$$

是一个由观测的 y 值组成的 $1 \times l$ 矩阵,

$$M = \| \quad (x_i, x_j) \|$$

是一个 $l \times l$ 矩阵, 其元素是由双邻域核定义的,

$$L = \| \quad (x, x_1), \cdots, \quad (x, x_l) \|^T$$

是一个 $l \times 1$ 矩阵, 其元素是由单邻域核 (x_i, x), $i = 1, \cdots, l$ 定义的, I 是一个 $l \times l$ 单位矩阵。

8.6 密度和条件密度估计

8.6.1 估计密度函数

在第七章中, 在用方法 P 求解密度估计问题时, 我们把它简化为下面的优化问题: 最小化泛函

$$\Omega(f) = (f, f)_H, \tag{8-40}$$

约束条件是

$$\sup_x \left| F_l(x) - \int_{-\infty}^{x} f(x')dx' \right| = \sigma. \tag{8-41}$$

但是, 为了计算上的原因, 我们只在由观测数据定义的 l 个点上检查这个约束条件, 即

$$\max_i \left| F_l(x) - \int_{-\infty}^{x} f(x')dx' \right|_{x = x_i} = \sigma, \quad i = 1, \cdots, l. \tag{8-42}$$

我们还把解考虑为是核的一个展开(这个核定义了 RKHS 空间), 即

$$f(x) = \sum_{i=1}^{l} \beta K_\gamma(x, x_i),$$

$$\sum_{i=1}^{l} \beta = 1, \quad \beta \geq 0. \tag{8-43}$$

现在, 让我们寻找下面形式的解:

$$f^*(x) = \sum_{i=1}^{l} \beta \frac{1}{v(x_i)} \int_{v(x_i)} K_\gamma(x, x')dx' = \sum_{i=1}^{l} \beta \quad _\gamma(x, x_i). \tag{8-44}$$

对这样的解, 我们得到下面的优化问题(考虑到核 $K(x, x')$ 的再生特性):

最小化泛函

$$W(\beta) = \Omega(f, f) = \sum_{i,j=1}^{l} \beta\beta \quad _\gamma(x_i, x_j), \tag{8-45}$$

约束条件是(8-43)式和约束

$$\max_i \left| F_l(x) - \sum_{j=1}^{l} \beta_j \int_{-\infty}^{x} K_\gamma(x', x_j)\, dx' \right|_{x=x_i} = \sigma, \quad 1 \le i \le l, \tag{8-46}$$

其中 $K_\gamma(x_j, x')$ 和 $K_\gamma(x_i, x_j)$ 是由(8-14)式和(8-15)式定义的函数，γ 是核的宽度参数

$$K_\gamma(x, x') = a(\gamma) K\left(\frac{x - x'}{\gamma}\right).$$

就像在第七章中那样，我们从容许参数集合中选择 γ 以得到(8-45)式的最小点或/和得到稀疏解。

依据 $v(x_i)$ 的不同，密度函数的这种估计是在不同核上的展开。

8.6.2 估计条件概率函数

为了在条件概率估计中采用 VSV 解，我们考虑与密度估计问题中形式类似的展开

$$p(w|x) = \frac{l(w)}{l} \sum_{i=1}^{l} \beta_i K_\gamma(x, x_i). \tag{8-47}$$

重复与前面相同的推导过程，我们可以说明，为找到系数 β_i，我们须最小化泛函

$$W_\gamma(\beta) = \sum_{i,j=1}^{l} \beta_i \beta_j K_\gamma(x_j, x_i), \tag{8-48}$$

约束条件是

$$\max_p \left| \sum_{i=1}^{l} \beta_i \frac{1}{l} \sum_{j=1}^{l} K_\gamma(x_j, x_i)\, \theta(x_p - x_j) - F_l(x_p|w) \right| = \sigma^*, \quad 1 \le p \le l \tag{8-49}$$

和

$$\sum_{i=1}^{l} \beta_i \frac{1}{l} \sum_{j=1}^{l} K_\gamma(x_j, x_i) = 1 \tag{8-50}$$

$$\beta_i \ge 0. \tag{8-51}$$

我们从容许参数集合

$$\gamma_{\min} \le \gamma \le \gamma_{\max} \tag{8-52}$$

中选择 γ，通过使 $W_\gamma(\beta)$ 最小化和/或选择大的容许 γ 来控制解的特性(精度和/或稀疏性)。

8.6.3 估计条件密度函数

要估计条件密度函数，我们重复同样的推理过程。

我们采用展开

$$p(y|x) = \sum_{i=1}^{l} \beta_i K_\gamma(x, x_i) K_\gamma(y, y_i), \quad \beta_i \ge 0. \tag{8-53}$$

要找到系数 β_i，我们最小化泛函

$$W_\gamma(\beta) = \sum_{i,j=1}^{l} \beta_i \beta_j K_\gamma(x_j, x_i) K_\gamma(y_j, y_i), \tag{8-54}$$

约束条件是

$$\max_p \left| \sum_{i=1}^{l} \beta_i \frac{1}{l} \sum_{j=1}^{l} \nu(x_j, x_i)\theta(x_p - x_j) \int_{-\infty}^{y_p} K_\gamma(y', y_i)\,dy' - F_l(x_p, y_p) \right| = \sigma,$$

$$p = 1, \cdots, l \tag{8-55}$$

以及

$$\sum_{i=1}^{l} \frac{1}{l} \sum_{j=1}^{l} \beta_i \nu(x_i, x_j) = 1, \tag{8-56}$$

$$\beta_i \geq 0. \tag{8-57}$$

为了控制解的特性(精度和/或稀疏性),我们选择容许参数 γ 使得目标泛函最小化和/或选择大的参数。

评注　在估计密度、条件概率和条件密度函数时,我们寻找这样的解:

$$f(x, \beta) = \sum_{i=1}^{l} \beta_i \nu(x, x_i),$$

它有下列特性:

$$\beta_i \geq 0, \quad i = 1, \cdots, l,$$

$$\nu(x, x_i) = E_{v(x_i)} K_\gamma(x, x'),$$

$$\nu(x_i, x_j) = E_{v(x_i)} E_{v(x_j)} K_\gamma(x, x'),$$

其中,

$$K_\gamma(x, x_i) = a(\gamma) K\left(\frac{x - x_i}{\gamma}\right),$$

$a(\gamma)$ 是一个归一化参数(参见 7.8 节)。

既然参数 β 是非负的,我们就可以在核 $K(x, x')$ 的基础上构造这样的解,使之具有轻尾或具有有限支撑。特别地,我们可以采用由正态规律定义的核:

$$K_\Delta(x, x') = \frac{1}{\sqrt{2\pi\gamma}} \exp\left(-\frac{(x - x')^2}{2\gamma^2}\right).$$

对于这个核,我们有

$$\nu(x, x') = [2\pi(\gamma^2 + d_i^2)]^{-\frac{n}{2}} \exp\left(-\frac{(x - x')^2}{2(\gamma^2 + d_i^2)}\right), \tag{8-58}$$

$$\nu(x_j, x_i) = [2\pi(\gamma^2 + d_i^2 + d_j^2)]^{-\frac{n}{2}} \exp\left(-\frac{(x - x')^2}{2(\gamma^2 + d_i^2 + d_j^2)}\right). \tag{8-59}$$

作为定义在有限支撑上的一个核 $K_\gamma(x, x_i)$,我们可以考虑 B_n-样条:

$$B_n(x, x') = \frac{1}{\gamma} \sum_{j=0}^{n+1} \frac{(-1)^j}{(n+1)!} C_j^{n+1} ((x - x') + (n - 1 - j)\gamma)_+^n.$$

我们知道,从 $n = 2$ 开始,B_n-样条可以用高斯函数逼近:

$$B_n(x, x') \approx \sqrt{\frac{6}{\pi\gamma^2(n+1)}} \exp\left(-\frac{6(x - x')^2}{\gamma^2(n+1)}\right). \tag{8-60}$$

因此,对于在一个 B_n-样条定义的函数基础上构造的单邻域和双邻域核,我们既可以直接计算它们,也可以用其逼近(8-60)式及表达式(8-58)和(8-59)来计算。

8.6.4 估计回归函数

估计回归函数

$$r(x) = \int y p(y \mid x) dy, \tag{8-61}$$

回顾核 $K_\gamma(y, y_j)$ **是一个对称(密度)函数,其积分等于** 1。**对这样一个函数,我们有**

$$\int y K_\gamma(y, y_i) dy = y_i, \tag{8-62}$$

因此,从(8-53)**式、**(8-61)**式和**(8-62)**式**[①]**,我们得到下面的回归函数**:

$$r(x) = \sum_{i=1}^{l} y_i \beta_{\gamma}(x, x_i).$$

① 原著中误写为(8-51)式、(8-56)式和(7-57)式。——译者

非正式推导和评述——8

本章中介绍的归纳原则是非常新的。对这一原则还需要做进一步的分析，但开始得到的结果是好的。

Sayan Mukherjee 在基于 VSV 方法解决密度估计问题中采用了这一原则(到目前为止还只是在低维空间中)。他通过与已有方法的比较，说明了这种方法的优势，尤其在样本数目少时更是如此。

在非参数密度估计的文献中已经有与此接近的思想。特别是，为了改进密度估计的 Parzen 方法，人们进行了很多有关的讨论。人们提出了在不同点上采用不同宽度值的方法。可以看到，在某一点上核的宽度应该与这一点上邻域的大小存在一定的联系。

但是，当时实现这一思想的方法过于简单：当时人们提出，选择核的宽度，使之正比于相应的点 x_i 的邻域值 d_i。也就是说，当时提出的方法是采用核 $a(\gamma)K\left(\frac{x-x_i}{d_i\gamma}\right)$。但是，这种方法产生了下面的问题：当邻域的值减小时，这个核收敛于 δ函数

$$\lim_{d_i\to 0} a(d_i\gamma)K\left(\frac{x-x_i}{d_i\gamma}\right) = \delta(x-x_i).$$

为了避免这样，人们考虑了用距第 k 个最近点的距离 d_i^k 来代替 d_i。但这又带来了有关新问题——需要选择 k。

80 年代，在从积分方程的各种解构造密度估计子时，我们发现，传统的方法，比如 Parzen 方法或投影方法，都是由求解这一积分方程的不同条件定义的，积分方程右边采用的是同样的近似——经验分布函数。这里，积分方程定义了其(给定的)右边项的导数，在求解这种积分方程的问题中，用不连续函数来逼近连续函数的思想可能并不是最好的。

在同样的方程中，采用对分布函数的连续逼近，我们得到了非传统的估计子。特别地，用一个连续的分段线性(多边形)逼近，我们(在一维情况下)得到了 Parzen 类型的估计子，其中新的核是如下定义的(Vapnik, 1988)：

$$g_{new}(x;x_i,x_{i+1},\gamma) = \frac{a(\gamma)}{(x_{i+1}-x_i)}\int_{x_i}^{x_{i+1}} K\left(\frac{x-z}{\gamma}\right)dz,$$

其中，x_i, x_{i+1} 是样本的变化序列中的元素，$K_\gamma(u)$ 是 Parzen 核。

当 $(x_{i+1}-x_i)\to 0$ 时，这个核收敛于 Parzen 核：

$$\lim_{(x_{i+1}-x_i)\to 0} g_{new}(x; x_i, x_{i+1}, \gamma) = a(\gamma) K\left(\frac{x-z}{\gamma}\right).$$

在提出了 SVM 方法之后,(稀疏)核逼近开始在求解各种函数估计问题中起重要的作用。就像在 Parzen 密度估计方法中一样,SVM 方法采用的是相同的核(以不同的展开系数和不同的支持向量)。当然,这样就有一个问题,即是否可能对不同的支持向量构造不同的核。采用 VRM 原则,我们在本书考虑的所有问题中都得到了新类型的核。

VRM 原则实际上是我们在研究用不同宽度的核得到解的性质时提出的。

第九章

结论:什么是学习理论中重要的?

9.1 在问题的表示中什么是重要的?

在本书开始时我们提出,学习就是一个基于经验数据的函数估计问题(在那里我们并没有对此说法进行讨论)。为了解决这个问题,我们利用了传统的归纳原则——ERM 原则。但是后来,我们提出了一个新的原则——SRM 原则。尽管如此,对问题的一般理解仍是建立在大样本统计学基础上的:我们的目标是得到一个规则,使之有最低的风险。这个得到"最低风险"的目标反映了大样本统计学的哲学:低风险的规则是好的,因为如果我们对一个大的测试集使用这个规则,那么以高的概率,损失的均值将是小的。

然而,多数情况下,我们面对的是另一个问题。我们同时给出了训练数据((x_i, y_i)对)和测试数据(向量 x_j^*),目标是用函数集为 $f(x, \alpha)$, $\alpha \in \Lambda$的学习机器为给定的测试数据找到其 y_j^* 值。换句话说,我们面对的问题是估计未知函数在给定点上的值的问题。

前面所说的思路实际上是用两个步骤来解决这个问题:首先估计出函数,然后再用估计的函数计算所求的值。为什么估计未知函数在所关心的给定点上的值的问题要通过这两步来解决呢?在这种两步策略中,我们实际上是试图通过解决一个更难得多的问题(估计函数),(以此作为一个中间问题)来解决一个相对简单的问题(估计函数在所关心的给定点上的值)。回顾一下,估计一个函数需要估计函数在其定义域内所有(无穷多)点上的值,其中也包括所关心的点。为什么我们要通过首先估计函数在定义域内所有点上的值来估计函数在所关心的点上的值呢?

可能出现这样的情况,即我们没有足够的信息(训练数据)来很好地估计一个函数,但却有足够的信息来估计函数在给定的有限数目的兴趣点上的值。

进一步,在人的生活中,决策选择问题起着重要的作用。对学习机器,这种决策选择问题可以表述如下:给定训练数据

$$(x_1, y_1), \cdots, (x_l, y_l)$$

函数为 $f(x, \alpha)$, $\alpha \in \Lambda$ 的机器需要在测试数据

$$x_1^*, \cdots, x_k^*$$

中找到一个 x^*, 它以最大的概率属于第一类(模式识别形式的决策问题)[①]。要解决这个问题, 我们甚至不需要估计这个函数在所有给定点上的值; 因此, 在我们没有足够的信息(没有足够的数据)来估计函数在给定点上的值的情况下, 这个问题仍可能解决。

解决这些问题的关键在于下面的观察, 为了简单起见我们在这里只讨论模式识别问题。

同时给学习机器(实现指示函数集 $Q(z, \alpha)$, $\alpha \in \Lambda$)两个串: 来自训练集和测试集的 $l+k$ 个 x 向量的串, 以及来自训练集的 l 个 y 值的串。在模式分类中, 机器的目标是定义包含对测试数据的 k 个 y 值的串。

对估计函数在给定点上值的问题, 学习机器实现的函数集可以被分解为一个等价类的有限集合(如果两个指示函数在串 $x_1, \cdots, x_{l+k}$ 上相同, 则它们落到同一个等价类中)。这些等价类可以用它们的基数(其中包含多少函数)来描述。

等价类的基数是一个重要的概念, 它使得在给定点上估计函数的理论与函数估计的理论区别开来。我们是在 70 年代研究这个概念(以及在给定点上估计函数的理论)的(Vapnik, 1979)。我们发现, 对线性函数集合, 其推广能力的界, 在只在给定点上最小化错误数目的意义上(连同本书中考虑的因素), 也依赖于一个新的因素 —— 等价类的基数。因此, 既然要最小化一个风险, 我们可以对很多因素最小化所得的界, 可以得到一个更小的最小点。现在, 问题就成为构造一个关于在给定点上估计函数的一般理论。这给我们带来了一个新的学习概念。

传统的哲学通常考虑两种类型的推理: 演绎和归纳。演绎描述从一般到特殊的过程, 而归纳描述从特殊到一般的过程。

估计函数在给定的兴趣点上的值的模型描述了一种新的推理概念: 从特殊到特殊。我们把这种类型的推理称作转导推理[②], 参见图 9.1。

注意, 这一概念是在我们希望从有限数量的信息得到最好的结果时产生的。在这种情况下的基本思想已经在第 1.9 节中进行了讨论, 这就是

图 9.1 不同类型的推理。归纳: 从给定数据导出函数。演绎: 导出给定函数在兴趣点上的值。转导: 从给定数据导出未知函数在兴趣点上的值。传统的策略需要用两步来得到未知函数在兴趣点上的值: 先用归纳步骤, 再用演绎步骤, 而不是用一步得到直接的解

① 或者找到以最大概率有最大 y^* 值的点(回归形式的决策选择问题)。

② 演绎的英文为 deduction, 归纳的英文为 induction, 而作者把这里提出的从特殊到特殊的推理称作 transduction, 我们把它译作转导,“转导(transduction)”一词本来是分子生物学中的一个术语, 大意是指某种信息借助一定的控制因素从一处到达另一处的过程), 根据这里的含义, 我们也可以把它译作直推。—— 译者

如果我们受限于有限数量的信息,那么不要通过解决一个更一般的问题来解决所面对的特殊问题。

我们应用这一思想来构造一种估计函数的直接方法。现在我们继续这一思想:不要通过估计整个函数来解决估计函数在给定点上的值的问题,不要通过估计一个函数在给定点上的值来解决一个决策选择问题,等等。

估计一个函数在给定点上的值的问题引出了一个问题,这个问题已经在哲学中被研究了两千多年:

什么是人类智慧的基础:是关于规律(规则)的知识,还是直接感悟真理的修养(直觉,专门的推理)?

关于学习问题的表达存在几种不同的模型,但是从概念角度看,没有一个能够与估计函数在给定点上的值的问题相比。这一模型可以为两千年来关于人类智慧本质的讨论做出最大的贡献。

9.2 在学习过程一致性理论中什么是重要的?

学习过程一致性的理论是比较完善的。对于理解实现 ERM 原则的学习过程的概念模型,这一理论回答了有关的几乎所有问题。惟一尚未解决的问题是收敛速度快的充分必要条件。在第二章中,我们对模式识别情况考虑了用退火熵描述的充分条件:

$$\lim_{l\to\infty}\frac{H^{\Lambda}_{ann}(l)}{l}=0.$$

对于回归估计情况,我们也可以证明,用退火熵 $H^{\Lambda}_{ann}(\varepsilon,l)=\ln EN^{\Lambda}(\varepsilon,z_1,\cdots,z_l)$ 给出的条件

$$\lim_{l\to\infty}\frac{H^{\Lambda}_{ann}(\varepsilon,l)}{l}=0,\quad\forall\varepsilon>0$$

定义了快速收敛的充分条件。

剩下的问题是:

这些不等式也是必要条件吗?如果不是,那么什么才是快速收敛的充分必要条件?

为什么研究描述快速收敛的充分必要条件的概念是这样重要呢?

正如我们前面看到的,这一概念在界的理论中起了关键的作用。在我们的体系中,利用退火熵来得到(非构造性的)不依赖于分布的界和(非构造性的)依赖于分布的界。在退火熵的基础上,我们构造了生长函数和广义生长函数。如果证明了退火熵条件对快速收敛的必要性,就相当于说明了这对于推导学习机器推广能力的界是最好的体系。而如果这个充分必要条件是由另外的函数描述的,那么这种体系就需要重新考虑。

9.3 在界的理论中什么是重要的?

界的理论包括两部分:非构造性界的理论和构造性界的理论。前者是在生长函数和广义生长函数概念的基础上得到的,后者的主要问题是用某种构造性的概念来估计这些函数。

在界的理论中主要问题在其第二部分。我们必须引入某种构造性的概念,使得通过它来估计生长函数或广义生长函数。1968 年我们提出了 VC 维的概念,并发现了生长函数的界(Vapnik and Chervonenkis, 1968, 1971)。我们证明了,值 $N^{\Lambda}(l)$ 要么是 2^{l},要么以多项式为界[①]:

$$N^{\Lambda}(z_1, \cdots, z_l) \leq \left(\frac{el}{h}\right)^{h}.$$

注意,右边的多项式依赖于一个自由参数 h。这个界(它只依赖于一个容量参数)不能被进一步改进(存在使得其等号成立的例子)。

我们面对的挑战是寻找更精细的概念,其中包含多于一个参数(比如两个),它们描述容量的(以及分布函数集合 $F(z) \in$ 的)某些特性,从而可以利用这样一个概念得到更好的界[②]。

这是一个很重要的问题,其答案将直接影响学习机器推广能力的界。

9.4 在控制学习机器推广能力的理论中什么是重要的?

在控制学习机器推广能力的理论中,最重要的问题是找到一个对小样本数的新归纳原则。在 70 年代中期,人们提出了若干技术来改进传统的函数估计方法。这些技术包括多项式回归问题中选择多项式阶数的各种规则、多维回归问题中的各种正则化技术、解决不适定问题的正则化方法等。所有这些技术都是基于同一个思想:为函数集提供一个结构,再在这个结构的元素中最小化风险。在 70 年代发现了容量控制所起的关键作用。我们把这种一般思想称作 SRM,以强调在结构的元素中最小化风险的重要性。

在 SRM 中,我们试图同时控制两个参数:经验风险值和结构元素的容量。

在 70 年代提出了 MDL 原则。运用这一原则,人们可以控制压缩系数。

最重要的问题是:

对于从小样本数估计依赖关系,是否存在一种新的归纳原则?

在对归纳原则的研究中,关键是找到影响风险的界的新概念,并因此用它来最小化这些界。为了利用另外一个概念,我们引入了对学习问题的一种新的表述:局部风险最小化

---

① Sauer(1972)也发表这个界。

② 回顾 MDL 界:即使是像压缩系数这样一个细化了的概念,所得到的界也不如建立在三个(实际上是粗糙的)的概念(比如经验风险值、观测数目和集合中函数的数目)之上得到的界。

问题。在这种表述中,在 SRM 原则的框架下,我们可以控制三个参数:经验风险、容量和局部性。

在估计函数在给定点上的值的问题中,我们利用了一个新的概念——等效类的基数。这有助于控制推广能力:通过在四个参数上最小化界,我们可以得到比在较少参数上最小化界更小的极小点。问题是,要找到影响风险上界的一种新概念。这将立即引出新的学习过程,甚至新的推理类型(就像转导推理那样)。

最后,找到函数集的新结构是重要的。对某些结构,其元素中包含的函数是由很多参数描述的,但是其 VC 维却较低,找到这样的结构是非常有意义的。我们只找到了一种这样的结构,它使得我们得到了 SV 机。这种类型的新结构将可能导致新型的学习机器。

9.5 在构造学习算法的理论中什么是重要的?

学习的算法应该是有良好控制的。这指的是,我们需要控制两个关系到推广能力的参数:经验风险的值和包含所选函数在内的最小结构元素的 VC 维。

如果结构是定义在某个高维特征空间中的线性函数集合上,那么 SV 机可以看作是控制这两个参数的有效工具。这一技术并不仅限于指示函数集(对解决模式识别问题)。在第六章中,我们把 SV 机推广到了解决回归问题。在这种推广的框架下,采用特殊的内积回旋函数,我们可以构造属于具有选定 VC 维的样条子集的高维样条函数。对内积采用不同的回旋函数,我们还可以构造出不同类型的在输入空间中非线性的函数[①]。

而且,SV 技术超出了学习机器的框架。从一般的角度它可以看作是对函数集的一种新的参数化表达方式。

这一点是很重要的,因为不论在计算统计学(比如模式识别、回归、密度估计)中,还是在计算数学(比如寻求对各种多维(算子)方程解的逼近)中,求解函数估计问题的第一步都是对一个函数集的描述(参数化表达),我们在这个函数集中寻找解。

在 20 世纪前半叶,参数化表达的主要思想(在 Weierstrass 定理之后)是多项式序列展开。然而,即使在一维情况下,人们有时也需要数十项来精确地逼近一个函数。对于求解很多问题中出现的此种序列,目前计算机的精度可能是不够的。

因此,在 50 年代中期,人们提出了函数参数化表达的一种新方法,即所谓样条函数(分段多项式函数)。这种参数表达使得我们可以对多数一维(有时是二维)问题得到精确的解。但是,它经常在多维情况下失败,比如四维情况。

函数的 SV 参数化表达可以用在高维空间中(回顾对这种参数化表达,逼近的复杂度依赖于支持向量的数目,而不是空间的维数)。通过控制函数集的"容量",我们可以控制逼

---

① 再次注意到,80 年代统计学中发展出的高级估计技术,比如投影追踪回归(projection pursuit regression)、MARS、铰接超平面(hinging hyperplanes)等等,实际上是考虑在下面的函数集中的某个特殊逼近:

$$y = \sum_{j=1}^{N} \alpha_j K\{(x, w_j)\} + b,$$

其中 $\alpha, \cdots, \alpha_N$ 是标量而 $w_1, \cdots, w_N$ 是向量。

近的“平滑度”特性。

每当我们考虑函数估计(函数逼近)的多维问题时,就应当考虑到这种类型的参数表达。

目前我们只对用 SV 技术解决模式识别问题有经验。但是,从理论上看,在解决统计学的不同领域(比如回归估计、密度估计、条件密度估计)中遇到的依赖关系估计问题中,以及在解决计算数学(比如求解某些多维线性算子方程)中遇到的依赖关系估计问题中,应用这一技术得到同样高的精度并不存在什么障碍。

我们可以把 SV 技术考虑为一种新型的多维函数参数化表达,在很多情况下,它能使我们克服维数灾难问题[①]。

9.6 什么是最重要的?

学习问题属于自然科学的问题:存在一个现象,我们必须对它构造一个模型。在建立这样的模型的努力中,理论学者可以选择两种不同立场中的一种,这取决于他们倾向于黑格尔法则(它描述一般的自然哲学)的哪一部分:

真实的即是合理的,且合理的即是真实的[②]。

这一法则的第一部分可以作如下的解释。某人(比如说一个实验者)知道描述现实的一个模型,理论学者的问题是证明这个模型是合理的(他还需要定义,这里的合理性的含义是什么)。比如,如果某人相信神经网络是真实大脑的一个好模型,并且可以说服理论学者,那么理论学者的目标就是证明这个模型是合理的。

假设模型具有某些很好的渐近特性,则理论学者就认为它是“合理的”。在这种情况下,如果理论学者证明了神经网络中的学习过程渐近地收敛于局部极小点,且一个足够大的神经网络可以很好地逼近任意平滑函数(这些已经得到了证明),那么他或她就成功了。如果可以证明所得的局部极值点接近全局极值点,这一理论的概念部分就完全了。

黑格尔法则的第二种立场对理论学者来说是一个更重的负担:理论学者必须定义什么是一个合理的模型,然后必须找到这个模型,最后必须说服实验者来证明这个模型是真实的(即证明这个模型描述了现实)。

很可能,一个合理的模型不但应该有很好的渐近特性,而且要在对付给定的有限数目观测方面有突出的特性[③]。在这种情况下,小样本数哲学是构造合理模型的一个有用的工具。

合理的模型可能是非常不平常的,可能需要克服常识的偏见才能找到它们。例如,我们已经看到,学习机器的推广能力依赖于函数集的 VC 维,而不是在给定集合中定义函数

---

① 参见本书第 125 页的脚注①。

② 在黑格尔(Hegel)本来的论断中,“真实”和“合理”两词的含义与它们通常的含义并不相同。不过,根据 B. Russell 的评注,认同真实和合理两词通常含义将导致“存在即是合理”的信念。Russel 不接受这一思想(见 Russell B. A History of Western Philosoply)。然而,这里我们就把黑格尔法则解释为“存在的就是正确的,且正确的就存在”。

③ 也许还需要有另外的特性。是什么呢?

的参数个数。因此,我们可以构造高维输入空间中的高阶多项式,使它具有良好的推广能力。没有关于控制推广能力的理论,这样的机会就不会明朗。现在,实验者必须回答这个问题:真实大脑所实现的推广性包含类似于支持向量技术的机制吗[①]?

这就是为什么在对学习过程的研究中,理论的作用可以比在自然科学的很多其他分支中更具构造性的原因。

然而,这一点是依赖于在研究学习现象时对一般立场的选择的。这一立场的选择反映了对下面问题的信念:在自然科学的这一特定邻域中,什么是发现真理的主要途径,是实验还是理论?

① 有的思想认为,推广能力、确定观察现象之重要性的能力及对重要事实的存储,是同一脑机制的不同方面,这种思想很有吸引力。

参考文献及评述

对参考文献的评述

20 世纪最伟大的数学家之一, A. N. Kolmogorov 曾经注意到, 数学科学与历史科学的一个重要区别是, 数学中发现的事实一旦被发现就永远成立, 而在历史中发现的事实却被每一代历史学家重新考虑。

就像在数学中一样, 在统计学习理论中, 已经得到的结果的重要性取决于关于学习现象的新的事实, 不管这些事实揭示了什么, 而不是取决于对已经知道的事实的新的描述。因此, 我尝试列出那些反映了本书所描述的统计学习理论发展过程中的主要事件的文献, 这些主要事件按年份排列是:

1958—1562 构造感知器。

1962—1964 证明关于学习过程的第一个定理。

1958—1963 发现非参数统计学。

1962—1963 发现解决不适定问题的方法。

1960—1965 发现算法复杂度的概念和它与归纳推理的关系。

1968—1971 发现对指示函数空间的大数定律和它与模式识别问题的关系。

1965—1973 创造对随机逼近归纳推理的一个一般的渐近学习理论。

1965—1972 创造对于 ERM 原则模式识别的一个一般的非渐近理论。

1974 形成 SRM 原则。

1978 形成 MDL 原则。

1974—1979 创造基于 ERM 和 SRM 原则上的一般的非渐近学习理论。

1981 把大数定律推广到实值函数空间。

1986 构造基于后向传播方法的神经网络。

1989 发现 ERM 原则、ERM 原则和最大似然方法一致性的充分必要条件。

1989—1993 发现用一系列 sigmoid 函数的叠加进行函数逼近的普遍性。

1992—1995 构造 SV 机。

参考文献

Aizerman M A, Braverman E M and Rozonoer L I. 1964. Theoretical foundation of potential function method in pattern recognition learning. Automation and Remote Control, 25: 821～837

Aizerman M A, Braverman E M and Rozonoer L I. 1965. The Robbince-Monroe process and the method of potential functions. Automation and Remote Control, 28: 1882～1885

Akaike H. 1970. Statistical predictor identification. Annals of the Institute of Statistical Mathematics, 202～217

Amari S. 1967. A theory of adaptive pattern classifiers. IEEE Trans. Elect. Comp., EC-16: 299～307

Anderson T W and Bahadur R R. 1966. Classification into two multivariate normal distributions with different covariance matrices. The Annals of Mathematical Statistics, 133(2)

Barron A R. 1993. Universal approximation bounds for superpositions of a sigmoid function. IEEE Transactions on Information Theory, 39(3): 930～945

Berger J. 1985. Statistical Decision Theory and Bayesian Analysis, Springer

Boser B, Guyon I and Vapnik V N. 1992. A training algorithm for optimal margin classifiers. Fifth Annual Workshop on Computational Learning Theory, Pittsburgh ACM, 144～152

Bottou L, Cortes C, Denker J, Drucker H, Guyon I, Jackel L, LeCun Y, Miller U, S ckinger E, Simard P and Vapnik V. 1994. Comparison of classifier methods: A case study in handwritten digit recognition. Proceeding 12th IAPR International Conference on Pattern Recognition, 2, IEEE Computer Society Press, Los Alamos, California, 77～83

Bottou L and Vapnik V. 1992. Local learning algorithms. Neural Computation, 4(6): 888～901

Breiman L. 1993. Hinging hyperplanes for regression, classification and function approximation. IEEE Transaction on Information Theory, 39(3). 999～1013

Breiman L, Friedman J H, Olshen R A and Stone C J. 1984. Classification and regression trees, Wadsworth, Belmont, CA

Bryson A, Denham W and Dreyfuss S. 1963. Optimal programming problem with inequality constraints, I: Necessary conditions for extremal solutions. AIAA Journal, 1: 25～44

Cantelli F P. 1933. Sulla determinazione empirica della leggi di probalita. Giornale

dell' Institute Italiano degli Attuari, (4)

Chaitin G J. 1966. On the length of programs for computing finite binary sequences. J. Assoc. Comput. Mach., 13: 547～569

Chentsov N N. 1963. Evaluation of an unknown distribution density from observations. Soviet Math., 4: 1559～1562

Cortes C and Vapnik V. 1995. Support Vector Networks. Machine Learning, 20: 1～25

Courant R and Hilbert D. 1953. Methods of Mathematical Physics, J. Wiley, New York

Cybenko G. 1989. Approximation by superpositions of sigmoidal function. Mathematics of Control, Signals, and Systems, 2: 303～314

Devroye L. 1988. Automatic pattern recognition: A study of the probability of error. IEEE Transaction on Pattern Analysis and Machine Intelligence, 10 (4): 530～543

Devroye L and Gy△rfi L. 1985. Nonparametric density estimation in L_1 view, J. Wiley, New York

Drucker H, Schapire R and Simard P. 1993. Boosting performance in neural networks. International Journal in Pattern Recognition and Artificial Intelligence, 7 (4): 705～719

Dudley R M. 1978. Central limit theorems for empirical measures. Ann. Prob., 6 (6): 899～929

Dudley R M. 1984. Course on empirical processes, Lecture Notes in Mathematics, Vol. 1097, 2～142, Springer, New York

Dudley R M. 1987. Universal Donsker classes and metric entropy. Ann. Prob., 15 (4): 1306～1326

Fisher R A. 1952. Contributions to Mathematical Statistics, J. Wiley, New York

Friedman J H, Hastie T and Tibshirany R. 1998. Technical report. Stanford University, Statistic Department. (www.stat.stanford.edu/ghf/# papers)

Friedman J H and Stuetzle W. 1981. Projection pursuit regression. JASA, 76: 817～823

Girosi F and Anzellotti G. 1993. Rate of convergence for radial basis functions and neural networks. Artificial Neural Networks for Speech and Vision, Chapter & Hall, 97～113

Glivenko V I. 1933. Sulla determinazione empirica di probabilita. Giornale dell' Institute Italiano degli Attuari, (4)

Grenander U. 1981. Abstract inference, J. Wiley, New York

Hoerl A E and Kennard R W. 1970. Ridge regression: Biased estimation for non-orthogonal problems. Technometrics, 12: 55～67

Huber P. 1964. Robust estimation of location parameter. Annals of Mathematical Statistics, 35 (1)

Jones L K. 1992. A simple lemma on greedy approximation in Hilbert space and convergence rates for Projection Pursuit Regression. The Annals of Statistics, 20 (1): 608～613

Ibragimov I A and Hasminskii R Z. 1981. Statistical estimation: Asymptotic theory, Springer, New York

Ivanov V V. 1962. On linear problems which are not well-posed. Soviet Math. Docl., 3 (4): 981～983

Ivanov V N. 1976. The theory of approximate methods and their application to the numerical solution of singular integral equations, Leyden, Nordhoff International

Karpinski M and Werther T. 1989. VC dimension and uniform learnability of sparse polynomials and rational functions. SIAM J. Computing. Preprint 8537-CS, Bonn University, 1989

Kolmogoroff A N. 1933. Sulla determinatione empirica di una leggi di distributione. Giornale dell' Institute Italiano degli Attuari, (4)

Kolmogorov A N. 1933. Grundbegriffe der Wahrscheinlichkeitsrech-nung, Springer. (英译本: Kolmogorov A N. 1956. Foundation of the Theory of Probability, Chelsea)

Kolmogorov A N. 1965. Three approaches to the quantitative definitions of information. Problem of Inform. Transmission, 1 (1): 1～7

LeCam Y. 1953. On some asymptotic properties of maximum likelihood estimates and related Bayes estimate. Univ. Calif. Public. Stat., 11

LeCun Y. 1986. Learning processes in an asymmetric threshold network. Disordered systems and biological organizations, Les Houches, France, Springer, 233～240

LeCun Y, Boser B, Denker J S, Henderson D, Howard R E, Hubbard W and Jackel L J. 1990. Handwritten digit recognition with back-propagation network. Advances in Neural Informtation Processing Systems, 2, Morgan Kaufman, 396～404

LeCun Y, Bottou L, Bengio Y and Haffner P. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86: 2278～2324

Lorentz G G. 1966. Approximation of functions, Holt-Rinehart-Winston, New York

Matheron G and Armstrong M (ed.). 1987. Geostatistical case studies (Quantitative geology and geostatistics), D. Reider Publishing Co.

Mhaskar H N. 1993. Approximation properties of a multi-layer feed-forward artificial neural network. Advances in Computational Mathematics, 1: 61～80

Micchelli C A. 1986. Interpolation of scattered data: distance matrices and

conditionally positive definite functions. Constructive Approximation, 2: 11～22

Miller M L. 1990. Subset selection in regression, Chapman and Hall, London

More J J and Toraldo G. 1991. On the solution of large quadratic programming problems with bound constraints. SIAM Optimization, 1(1): 93～113

Novikoff A B J. 1962. On convergence proofs on perceptrons. Proceedings of the Symposium on the Mathematical Theory of Automata, Polytechnic Institute of Brooklyn, XII: 615～622

Paramasamy S. 1992. On multivariant Kolmogorov-Smirnov distribution. Statistics & Probability Letters, 15: 140～155

Parrondo J M and Broeck C Van den. 1993. Vapnik-Chervonenkis bounds for generalization. J. Phys. A., 26: 2211～2223

Parzen E. 1962. On estimation of probability function and mode. Annals of Mathematical Statistics, 33 (3)

Phillips D Z. 1962. A technique for numerical solution of certain integral equation of the first kind. J. Assoc. Comput. Math., 9: 84～96

Poggio T and Girosi F. 1990. Networks for Approximation and Learning. Proceedings of the IEEE, 78(9)

Pollard D. 1984. Convergence of stochastic processes, Springer, New York

Popper K. 1968. The logic of Scientific Discovery. 2nd ed. Harper Torch Book, New York

Powell M J D. 1992. The theory of radial basis functions approximation in 1990. W A Light ed. Advances in Numerical Analysis Volume II: Wavelets, Subdivision algorithms and radial basis functions, Oxford University, 105～210

Rissanen J. 1978. Modeling by shortest data description. Automatica, 14: 465～471

Rissanen J. 1989. Stochastic complexity and statistical inquiry, World Scientific

Robbins H and Monroe H. 1951. A stochastic approximation method. Annals of Mathematical Statistics, 22: 400～407

Rosenblatt F. 1962. Principles of neurodinamics: Perceptron and theory of brain mechanisms, Spartan Books, Washington D. C.

Rosenblatt M. 1956. Remarks on some nonparametric estimation of density functions. Annals of Mathematical Statistics, 27: 642～669

Rumelhart D E, Hinton G E and Williams R J. 1986. Learning internal representations by error propagation. Parallel distributed processing: Explorations in macrostructure of cognition, Vol. I., Badford Books, Cambridge, MA., 318～362

Russell B. 1989. A History of Western Philosophy, Unwin, Lodon

Sauer N. 1972. On the density of families of sets. J. Combinatorial Theory (A), 13: 145～147

Schwartz G. 1978. Estimating the dimension of a model. Annals of Statistics, 6: 461～464

Scholkopf B, Burges C and Vapnik V. 1996. Incorporating invariance in support vector learning machines. in book Malsburg C von der, Seelen W von, Vonbruggen J C and Sendoff S(eds.), Artificial Neural Network-ICANN'96, Springer Lecture Notes in Computer Science, Vol. 1112, Berlin, 47～52

Simard P Y, LeCun Y and Denker J. 1993. Efficient pattern recognition using a new transformation distance. Neural Information Processing Systems, 5: 50～58

Smirnov N V. 1970. Theory of probability and mathematical statistics (Selected works), Nauka, Moscow

Solomonoff R J. 1960. A preliminary report on general theory of inductive inference. Technical Report ZTB-138, Zator Company, Cambridge, MA

Solomonoff R J. 1964. A formal theory of inductive inference. Parts. 1 and 2, Inform. Contr., 7: 1～22, 224～254

Tapia R A and Thompson J R. 1978. Nonparameteric probability density estimation, The John Hopkins University Press, Baltimore

Tikhonov A N. 1963. On solving ill-posed problem and method of regularization. Doklady Akademii Nauk USSR, 153: 501～504

Tikhonov A N and Arsenin V Y. 1977. Solution of ill-posed problems, W. H. Winston, Washington DC

Tsypkin Ya Z. 1971. Adaptation and learning in automatic systems, Academic, New York

Tsypkin Ya Z. 1973. Foundation of the theory of learning systems, Academic, New York

Vapnik V N. 1979. Estimation of dependencies based on empirical data (in Russian), Nauka, Moscow (英译本: Vapnik Vladimir. 1982. Estimation of dependencies based on empirical data, Springer, New York)

Vapnik V N. 1993. Three fundamental concepts of the capacity of learning machines. Physica A, 200: 538～544

Vapnik V N. 1988. Inductive principles of statistics and learning theory. Yearbook of Academy of Sciences of the USSR on Recognition, Classification, and Forecasting, 1, Nauka, Moscow (英译本:. 1995. Inductive principles of statistics and learning theory. in book: Smolensky, Moser, Rumelhart ed., Mathematical perspectives on neural networks, Lawrence Erlboum Associates, Inc.)

Vapnik V N. 1998. Statistical Learning Theory, J. Wiley, New York

Vapnik V N and Bottou L. 1993. Local algorithms for pattern recognition and dependencies estimation. Neural Computation, 5 (6): 893～908

Vapnik V N and Chervonenkis A Ja. 1968. On the uniform convergence of relative frequencies of events to their probabilities. Doklady Akademii Nauk USSR, 181 (4) (英译本: Sov. Math. Dokl.)

Vapnik V N and Chervonenkis A Ja. 1971. On the uniform convergence of relative frequencies of events to their probabilities. Theory Probab. Apl., 16: 264～280

Vapnik V N and Chervonenkis A Ja. 1974. Theory of Pattern Recognition (in Russian), Nauka, Moscow (德译本: Wapnik W N, Tschervonenkis A Ja. 1979. Theorie der Zeichenerkennung, Akademia, Berlin)

Vapnik V N and Chervonenkis A Ja. 1981. Necessary and sufficient conditions for the uniform convergence of the means to their expectations. Theory Probab. Apl., 26: 532～553

Vapnik V N and Chervonenkis A Ja. 1989. The necessary and sufficient conditions for consistency of the method of empirical risk minimization (in Russian). Yearbook of the Academy of Sciences of the USSR on Recognition, Classification, and Forecasting, 2, Nauka, Moscow, 207～249 (英译本: 1991. The necessary and sufficient conditions for consistency of the method of empirical risk minimization. Pattern Recogn. and Image Analysis, 1 (3): 284～305)

Vapnik V N and Stefanyuk A R. 1978. Nonparametric methods for estimating probability densities. Autom. and Remote Contr., (8)

Vasin V V. 1970. Relationship of several variational methods for approximate solutions of ill-posed problems. Math. Notes, 7: 161～166

Wenocur R S and Dudley R M. 1981. Some special Vapnik-Chervonenkis classes. Discrete Math., 33: 313～318

索引[①]

① 本索引按照原著索引中关键词的英文字母顺序排列,它的中文译文在前,英文在后。关键字后的页码是该关键词在本书中第一次出现的页码。