Kernel function theory and signal processing

核函数理论与信号处理

蒋 刚 著

科学出版社

核函数理论与信号处理

蒋　刚　著

科学出版社

北　京

内 容 简 介

本书主要对机器学习问题、核函数方法、支持向量机等内容进行探讨和分析，构建不同的支持向量机模型，将其应用于海色遥感、时间序列预测、交通流、点云建模等领域，理论分析和实验结果验证了可行性和有效性，表明其特别适合于小样本、非线性、受干扰的信号处理。

本书可供机械电子、信号处理等相关专业的科研人员和工程技术人员参考。

图书在版编目(CIP)数据

核函数理论与信号处理/蒋刚著. —北京：科学出版社，2013.1
ISBN 978-7-03-036562-0

Ⅰ. ①核… Ⅱ. ①蒋… Ⅲ. ①核函数 ②信号处理 Ⅳ. ①O174.1 ②TN911.7

中国版本图书馆 CIP 数据核字(2013)第 017946 号

责任编辑：余 江 张丽花 / 责任校对：韩 杨
责任印制：闫 磊 / 封面设计：迷底书装

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

北京通州皇家印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2013 年 1 月第 一 版 开本：720×1000 B5
2013 年 1 月第一次印刷 印张：10 1/2
字数：206 000

定价：45.00 元

（如有印装质量问题，我社负责调换）

前　言

非线性非平稳信号普遍存在于工业、金融、军事、交通运输、航空航天等领域，由于它们本身的复杂性以及各种不明扰动的存在，对它们的处理往往需要更多的技巧和技术。基于传统方法的学习机器在样本趋于无穷时，的确表现出非常不错的性能，但在样本数有限时却表现出很差的推广能力。基于核函数理论的支持向量机最近几年得到了机器学习界的广泛关注，它以统计学习理论为基础，通过 VC 维控制学习机的规模，可以保证收敛到全局最优解，即使在小样本条件下也体现出良好的推广能力。本书主要针对机器学习、核机器方法、小波核机器、模糊小波核机器等内容进行探讨和分析，构建不同的支持向量机模型，将其应用于海色遥感、时间序列预测、交通流、点云建模等领域，以理论计算与分析、数据实测与仿真等方法对所提理论与方法进行验证，有一定的参考价值。

在编写本书的过程中，作者把初稿送给一些初学者、有一定基础的研究人员审阅，并请他们提出意见，在综合各方意见的基础上，对书中内容做了多次调整和修改。“传道、授业、解惑”的思想贯穿于本书始终。不管你是初学，还是已经具有一定的基础，都能从本书获益！

万丈高楼平地起，本书从最基本的知识入手，逐步提高程度和等级。本书尽可能做到语言规范简练、用词恰当合理、表达生动形象、例证丰富实用。

本书的出版得到国家自然科学基金(11176027)和博士点建设基金资助。

由于时间仓促，作者知识水平有限，书中不当之处在所难免，敬请广大读者和各位同行不吝赐教！

蒋　刚

2012 年 8 月于西南科技大学

目　录

第1章 机器学习

1.1 机器学习的起源与发展

核函数虽然是数学中一个早已存在的概念，但是基于核的学习算法却是在最近几年中才得到了机器学习界的广泛关注。核机器方法建立在统计学习理论和核函数理论基础之上，统计学习理论属于机器学习的范畴。机器学习的本质是：设计某种方法，使之能够通过对已知数据的学习，找到这些数据的内在依赖关系。根据这种依赖关系，不但可以较好地解释已知的实例，并且能够对未知的现象或无法观测的现象做出正确的预测和判断。

统计学在解决机器学习问题中起着基础性的作用，传统的统计学研究的主要是渐近理论，即样本数趋于无穷多时的统计性质。但是在实际应用中，我们面对的却通常是有限样本的情况，多数时候的样本甚至是非常有限的。这样就产生了如下问题：基于传统方法的学习机器在样本趋于无穷时，的确表现出非常不错的性能，但在样本数有限时却表现出很差的推广能力。因此，深入研究有限样本情况下的机器学习问题具有极为重要的现实意义。

在工业、金融、军事、交通运输、航空航天等领域，非线性非平稳信号是普遍存在的。由于它们本身的复杂性以及各种不明扰动的存在，为了得到更好的效果，对它们的处理往往需要更多的技巧和技术。本书中，我们尝试将小波技术和模糊逻辑引入核机器方法中，并将其统称为核机器方法。

1.1.1 机器学习概述

机器学习最早的研究工作可以追溯到20世纪30年代的Fisher判据，但由于Fisher判据没有考虑归纳推断的问题，因此通常不把它视作机器学习的开端。从机器学习的研究历程来看，大致可以分为以下四个阶段。

1. 20世纪60年代：Rosenblatt感知机

在20世纪60年代，F. Rosenblatt提出了第一个学习机器的模型，称为感知机，在一些文献中也称为感知器，它利用了最简单的神经元模型的自适应特性，标志着人们对机器学习过程进行数学研究的真正开始。

从概念上讲，感知机的思想并不是新的，它已经在神经生理学领域被讨论了很多年。但 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，并通过简单的实验说明该模型能够被推广。Rosenblatt 感知机率先在模式识别问题上取得了成功的应用，最简单的情况就是用于构建两类分类情况下的分类规则。

Rosenblatt 研究了 n 个输入一个输出的多神经元多层模型，前一层的输出作为下一层的输入，最后一层只有一个输出。考虑到算法的实用性，构建一种确定神经元系数的算法规则是必要的，Rosenblatt 提出了一种迭代算法，通过在任意初始权重向量上增加误分的正样本或减去误分的负样本来工作。该算法使得感知机在一些简单的问题上表现出较强的推广能力。

在 Rosenblatt 提出感知机模型后，Novikoff 证明了感知机的第一个定理：给定样本集 $\{x_i, i = 1,2,\cdots,n\}$，如果它能以间隔 γ 分开，并且一阶范数有界，即 $\max \| x_i \| = R < \infty, i = 1,2,\cdots,n$，则最多经过 $N \leqslant [(R/\gamma)^2]$ 次修改，就可以构建出将所有样本分开的超平面，$[\cdot]$ 表示向上取整运算。

2. 20 世纪 60～70 年代：学习理论的创立

在这段时间里，统计学习理论得到很大的发展，两个核心概念：VC(Vapnik & Chervonenkis)维和 VC 熵也是在这段时间提出的，它们直接促使泛函空间的大数定律得以发现，从而得到关于收敛速率的非渐近界的主要结论。关于界的一些结论为结构风险最小化(Structural Risk Minimization，SRM)原则奠定了基础。

正则化理论也在这一时期提出，用于解决不适定问题。事实上，早在 20 世纪初，Hadamard 就已经观察到不适定现象，遗憾的是他认为这只是一个纯数学问题。研究人员发现，通过最小化正则化泛函 $R^*(f) = \| Af - F_\delta \|^2 + \gamma(\delta)\Omega(f)$ 代替最小化泛函 $R(f) = \| Af - F_\delta \|^2$，可以得到一个解的序列，当 $\delta \to 0$ 时能收敛到期望解，从而建立了解决不适定问题的理论。以该理论为基础，Vapnik 和 Stefanyuk 在 20 世纪 70 年代中期进一步发现了创建此类算法的通用方法，将其应用于概率密度估计，提出了密度估计的非参数方法。

此外，作为统计学和信息论最重要的一环，算法复杂度的思想也在这一时期萌芽，在此基础上，最小长度描述(Minimum Description Length，MDL)原则于 1978 年提出。它们为学习理论的创立奠定了坚实的基础。

3. 20 世纪 80 年代：人工神经网络

在 1986 年，LeCun、Rumelhart、Hilton 和 Williams 独立地提出了构造感知器所有神经元的向量系数的后向传播算法(Back Propagation，BP)。它修改了 Rosenblatt

感知机的 McCulloch－Pitts 神经元模型，把非连续函数 $\mathrm{sgn}\{\langle w \cdot x\rangle - b\}$ 替换为 sigmoid函数 $y = f\{\langle w \cdot x\rangle - b\}$，该函数满足

$$\begin{cases} f(-\infty) = -1 \\ f(+\infty) = +1 \end{cases} \tag{1.1}$$

这样，新神经元将合成一个连续函数，对任意固定的 x 都存在对应于所有神经元的所有系数的梯度，利用计算出的梯度值，可实现函数的无限逼近，但它只能保证收敛到局部极值点，这也是基于梯度的方法的共同特点。在这段时间里，感知器被改称为神经网络，在数字识别等问题上率先体现出优越性，之后被陆续引入到其他研究领域并取得了一些较好的效果。

神经网络法本身有一些固有缺陷，如欠学习、过学习、局部极值等，此外，由于它是基于启发式的方法进行寻优，因此在缺乏先验信息的情况下，网络结构的选择是个难以有效解决的问题，在实际应用中，多数是通过实验的方式来“凑”出较好的结构。在解决不适定问题时，人们利用正则化技术，可从一定程度上避免过学习。神经网络以经验风险最小化（Empirical Risk Minimization，ERM）作为优化目标，在训练的时候可以收敛到极小的误差甚至零误差，但是对未经训练新数据，却往往表现出很差的推广能力。

4. 20 世纪 90 年代：统计学习理论

考虑到神经网络的一些不足之处，尤其是在小样本条件下的不良表现，人们试图从更本质的层次上寻求一种更好的学习机器。在这段时间里，神经网络重新被称为多层感知机模型，以结构风险最小化原则和最小描述长度原则为理论基础，对照传统统计方法的渐近理论，针对小样本条件的统计学习理论开始成为研究的热点，并得到逐步完善。在完成学习过程的一般理论性分析后，对最优算法的研究开始启动。在 20 世纪 90 年代中期，Vapnik 等人成功地提出了基于统计学习理论的支持向量机（Support Vector Machine，SVM）方法，它是用统计学习理论解决实际应用问题的通用学习机器，与神经网络相比，它具有更坚实的数学理论基础，可有效地解决有限样本条件下的高维模型构建问题，并具有更强的推广能力。国内的边肇祺、张学工等也在这段时间开始了相关研究。

在统计学习理论的激发下，学习方法领域产生了一些新思想。原有的思想主要是试图构建生物模型，希望通过仿生学方法得到好的学习算法；新的学习思想摆脱了这一思路的束缚，它主要基于最小化错误率的界限理论，具有良好的性能，在解的唯一性、样本规模较大时的优化处理等方面表现出色。

1.1.2 核机器方法

核机器方法以统计学习理论和核函数方法为基础，是一种基于核函数的机器学习算法。它包括两部分内容：一是核函数的选择与构建，通常还包括对核的容许条件的证明，即核的构建问题，有的研究人员将它称为核技术；另一部分是用选择的核构造学习机器，即载体的构建问题，通过载体让核技术得以实现。本书采用支持向量机作为核函数的载体。

核函数的数学结果产生于 1909 年，把它应用于机器学习领域用于构造学习算法始于 1992 年，成为构建内积算法的非线性推广的一般性方法。核函数方法的基本思想是：给定原空间 S 中的向量 x，通过非线性映射 $\Phi(\cdot)$ 将 S 映射到特征空间 H，然后在 H 里构建线性算法，它对应 S 非线性问题的求解。如果 x 各坐标分量间的相互作用仅限于内积，就可以使用满足 Mercer 条件的核函数来代替内积运算，从而越过本来需要计算的映射，避免维数灾难。

为了避免维数灾难，许多学习算法都是通过“降维”的方式，将高维原始空间变换到较低维的特征空间，这容易损失一些有用的特征，导致学习机器性能的下降；而基于核的方法却恰好相反，它将低维向高维映射，却不需要过多地考虑维数对学习机器性能的影响。核函数是支持向量机的重要组成部分，根据 Hilbert-Schmidt 定理，只要变换 Φ 满足 Mercer 条件，就可用于构建核函数。Mercer 条件如下：给定对称函数 $K(x,x')$ 和任意函数 $\varphi(x)\neq 0$，满足约束

$$\begin{cases}\int_{-\infty}^{\infty}\varphi^2(x)\mathrm{d}x<\infty\\ \iint_{-\infty}^{\infty}K(x,x')\varphi(x)\varphi(x')\mathrm{d}x\mathrm{d}x'>0\end{cases}\tag{1.2}$$

支持向量机首先在有限样本的模式识别问题上体现优越的性能，之后，统计学习理论逐渐成为研究热点之一。支持向量机采用结构风险最小化(SRM)准则，认为实际风险由经验风险和置信间隔组成，下式至少以概率 $1-\eta$ 成立：

$$R(\omega)\leqslant R_{\mathrm{emp}}(\omega)+\sqrt{\frac{h[\ln(2n/h)+1]-\ln(\eta/4)}{n}}\tag{1.3}$$

右边第一项表示经验风险，第二项被称做置信间隔。其中，n 是样本数，h 是 VC 维。式(1.3)表明实际风险和预测函数集的 VC 维有很强的依赖关系，可以通过控制 VC 维来控制学习机器的推广能力，我们总是可以通过设计具有较小 VC 维的函数集来提高学习机器的推广能力，这是神经网络不具备的优势。

支持向量机最初是针对线性可分情况下的模式识别问题提出的，随着 ε 不敏感损失函数的引入，逐步推广到了线性回归、非线性回归和概率密度估计等领域。各种改进算法也相继提出，如 Scholkoph 等提出的 v-SVM，Suykens 等提出的 LS-

SVM、加权支持向量机 W-SVM 以及基于线性规划的 SVM 等。这些方法从较大程度上改善了支持向量机的性能，但在某些方面也略有不足。例如，LS-SVM 没有解的稀疏性；W-SVM 主要应用于时间序列预测；线性规划 SVM 的约束数目与 w 的维数有关，因此对于非线性样本集难以求解。

针对大数据条件下的优化问题，研究人员也提出了一些解决方案，如分解法(Decomposition Method)、序贯最优算法(Sequential Minimal Optimization，SMO)和增量学习法(Incremental Learning Method)等。它们各有优势，但都存在收敛速度慢的缺点，样本集越大这个缺点越突出。

目前，支持向量机方法已在金融时间序列预测、非线性系统参数辨识、建模、控制等领域有了成功应用的案例。陈念贻、陆文聪、武海顺等在氮化铝薄膜生长过程控制中应用了支持向量机算法；潘继斌介绍了一种回归函数的支持向量机估计方法；李良敏、屈梁生提出了一种基于遗传编程和支持向量机的故障诊断模型；刘洪、赵金洲、胡永全等将支持向量机应用于天然气工业的井层优选和施工方案优化，结果表明支持向量机在重复压裂研究领域具有良好的应用前景。

以上列举的只是部分研究成果，除此之外，还有很多的研究人员做了大量的研究工作，这里限于篇幅，不一一列举。

1.1.3 非平稳信号与小波技术

在处理科学与工程问题时，信号的统计量起着极其重要的作用。在这些统计量中，最常用的是均值，它是一阶统计量。此外，属于二阶统计量的相关函数和功率谱密度也用得非常多。相对来说，三、四阶的高阶矩，高阶累积量和高阶谱的使用频率要低一些。关于信号的平稳性，目前的主流定义方法如下。

定义 1.1 若 $\{x(t_1),\cdots,x(t_n)\}$ 的联合分布函数与 $\{x(t_1+\tau),\cdots,x(t_n+\tau)\}$ 的联合分布函数对所有的 $t_1,\cdots,t_n$ 和 $\tau \in T$ 都相同，则称随机过程 $\{x(t),t \in T\}$ 为严格平稳随机信号，一些文献中也称之为狭义平稳信号。

定义 1.2 若随机信号 $\{x(t),t \in T\}$ 满足

$$\begin{cases} E\{x(t)\} = m = \text{const} \\ E\{|\ x(t)\ |^2\} < \infty \\ E\{[x(t)-m][^{x}(s)-m]^*\} = R_x(t-s) - |\ m\ |^2 \end{cases} \tag{1.4}$$

式中，$R_x(t-s)=E\{x(t)x^*(s)\}$，$x^*(t)$ 表示 $x(t)$ 的复共轭。则称 $\{x(t),t\in T\}$ 为广义平稳过程。广义平稳也称弱平稳、协方差平稳或二阶平稳。

定义 1.3 如果一个信号不是广义平稳，则称它为非平稳信号。最常见的非平稳信号是自相关函数或功率谱密度随时间变化的信号。

对非平稳信号的分析和处理有很多种方法，最主要的是傅里叶变换和小波变

换。在经典方法的基础上，研究人员提出了很多改良方法，如短时 Fourier 变换、时频分析、小波分析、Gabor 展开、Radon-Wigner 变换、分数阶 Fourier 变换、线调频小波分析方法(Chirplet Transform)等。

Fourier 变换包括两部分：Fourier 变换和反变换，它们构成时域、频域的连接桥梁，使我们能在时域和频域观测同一个信号。Fourier 变换能在整体上将信号分解为不同的频率分量，但不能告诉我们某个频率分量具体发生在哪些时间段内，而对非平稳信号的分析和处理来说，这又是非常重要也是非常关键的。

事实上，从 $s(t)\rightarrow S(f)$ 的 Fourier 变换过程可以看出：频谱 $S(f)$ 等于原信号和基于无穷区间的正弦波基函数的内积，即

$$S(f)=\int_{-\infty}^{\infty}s(t)\mathrm{e}^{-\mathrm{j}2\pi ft}\,\mathrm{d}t=\langle s(t),\mathrm{e}^{\mathrm{j}2\pi ft}\rangle \tag{1.5}$$

容易看到，非平稳信号 $s(t)$ 的局域性是不可能通过无穷区间的平稳基函数表现出来的。这是经典 Fourier 方法在处理非平稳信号时的一个重大不足，如果一定要用 Fourier 方法处理非平稳信号，需要使用具有局域性的基函数，由此诞生了短时 Fourier 变换。短时 Fourier 变换也称谱图方法，它的基本思想是：假定非平稳信号在分析窗函数 $g(t)$ 的一个短的时间间隔内是平稳或伪平稳的，移动分析窗函数，使得 $s(u)g(u-t)$ 在不同的有限时间宽度内为不同的伪平稳信号，就可以计算出各个不同时刻的功率谱。

从本质上讲，短时 Fourier 变换是一种单一分辨率的信号分析方法，因为它使用一个固定的短时窗函数。后来，一些研究人员拓展了该方法，采用自适应的方法对不同的信号段选择长度不一的窗函数，从而具有了多分辨率分析的能力。短时 Fourier 变换的主要缺陷是：对应一定的时刻，只是对其附近窗口内的信号做分析，若选择的 $g(t)$ 窄，则时间分辨率高，而频率的分辨率低；如果为了提高频率分辨率而使 $g(t)$ 变宽，那么伪平稳假设的近似程度便会变差。对实际信号而言，谱分量的变化极快，并且难以用某个规律来描述，因此实际上很难找到合适的短时窗函数 $g(t)$，使得信号在某时间段内既能较好地满足平稳性假设，又能保证窗的一定宽度。

小波是为了弥补 Fourier 变换的不足而发展起来的数学学科，最早可追溯到 1910 年 Haar 提出的小波规范基。它的真正发展起源于 20 世纪 80 年代：1984 年法国地球物理学家 Morlet 在分析地震波的性质时，发现传统 Fourier 方法难以达到要求，于是用小波对信号进行分解；随后物理学家 Grosses 对该信号按一个确定函数的伸缩、平移系展开的可行性进行了分析；Meyer 意外地发现了存在小波正交系。1987 年，Mallat 巧妙地将计算机视觉多尺度分析思想引入到小波方法中，提出了小波多尺度分析的概念。1988 年，Daubechies 构造了具有紧支撑的正交小

波基。1990年,C. K. Chui 和 J. Z. Wang 提出了基于样条函数的单正交小波函数;1994年,Geronimo 等开始研究多小波变换。此后,许多研究人员从理论和应用上对小波做了大量的研究工作,取得了很多成果,限于篇幅,这里不一一列举。至此,小波逐渐得到完善,并形成了一个较为完整的理论体系。

1.2 统计学习理论基础

经典统计学研究的是渐近理论,即样本数趋于无穷大时的性质。但在实际应用中,要获得无穷数量的样本基本上是不可能的,而且,计算机的处理能力有限,不可能处理无穷数量的样本。在这种情况下,一些问题逐渐凸现出来:在样本数无穷或近于无穷时,学习机器确实具有很不错的性能,但在有限数量的样本条件下,往往表现出较差的推广能力。

统计学习理论是专门研究小样本条件下的机器学习问题的一种有效的理论。它以结构风险最小化和最小描述长度原则为理论基础,能有效地控制 VC 维,从而控制学习机器的推广能力,被认为是目前针对小样本的统计估计和预测学习的最佳理论。统计学习理论主要包括以下内容:

(1)经验风险最小化原则下统计学习一致性的条件;

(2)在这些条件下关于统计学习方法推广能力的界的结论;

(3)在界的基础上建立的小样本归纳推理原则;

(4)实现新的归纳原则的算法。

学习问题主要有三种类型:概率密度估计、模式识别和回归估计。本书主要对模式识别和回归估计进行探讨和分析,书中涉及的也主要是与模式识别和回归估计相关的定理和概念,以后不再一一注明。下面将重点介绍这一理论的一些相关概念和重要定理,它们是后续章节的基础。

1.2.1 一致性概念与函数集 VC 维

学习问题的一般表示如下:给定独立同分布样本集 $S = \{(x_i, y_i), i = 1, 2, \cdots, n\}$ 和参数集 P,定义在空间 Z 上的概率测度为 $F(z) = F(x, y)$,考虑一个损失函数的集合 $L(z, \alpha), \alpha \in P$,那么学习的目标就是最小化风险泛函

$$R(\alpha) = \int L(z, \alpha) \mathrm{d}F(z) = \int L[(x, y), \alpha] \mathrm{d}F(x, y), \quad \alpha \in P \tag{1.6}$$

在上述学习问题的一般表达式中,如果把损失函数取成特定的函数,就可以得到相应的特定方法。

1. 非平凡一致性概念

给定独立同分布数据集 $S=\{(x_i,y_i),i=1,2,\cdots,n\}$ 和参数集 P，在空间 Z 上的概率测度为 $F(z)=F(x,y)$，假定损失函数 $L(z_i,\alpha),\alpha\in P$ 是使风险泛函

$$R_{\text{emp}}(\alpha)=\frac{1}{n}\sum_{i=1}^{n}L(z_i,\alpha),\quad \alpha\in P \tag{1.7}$$

最小的函数。做如下定义：

定义 1.4 如果泛函值 $R(\alpha)$ 和 $R_{\text{emp}}(\alpha)$ 构成的序列依概率收敛于同一个极限，即

$$\begin{cases}\lim\limits_{n=\infty}P[R(\alpha_n)]=\inf[R(\alpha)],\quad \alpha\in P\\ \lim\limits_{n=\infty}P[R_{\text{emp}}(\alpha_n)]=\inf[R(\alpha)],\quad \alpha\in P\end{cases} \tag{1.8}$$

我们就说这一归纳原则(ERM 原则)对函数集 $L(z,\alpha),\alpha\in P$ 和概率分布函数 $F(z)$ 是一致的。第 1 式用于保证所能达到的风险收敛于最好的可能值，第 2 式保证可以在经验风险的取值基础上估计出最小可能的风险值。

定义 1.5 给定损失函数集 $L(z,\alpha),\alpha\in P$，P 为参数集，它的一个子集 $L_S(g)$ 具有如下形式：

$$L_S(g)=\left\{\alpha:\int L(z,\alpha)\mathrm{d}F(z)>g,\alpha\in P\right\} \tag{1.9}$$

如果对函数集 $L(z,\alpha),\alpha\in P$ 的任意非空子集 $L_S(g)$，$-\infty<g<+\infty$，都有

$$\lim_{n=\infty}P\{\inf[R_{\text{emp}}(\alpha_n)]\}=\inf[R(\alpha)],\quad \alpha\in P[L_S(g)] \tag{1.10}$$

则我们说基于某一归纳原则的方法对函数集 $L(z,\alpha),\alpha\in P$ 和概率分布函数 $F(z)$ 是非平凡一致的。

2. 函数集的 VC 维

VC 维是统计学习理论中的一个非常重要的概念，由 Vapnik 和 Chervonenkis 提出，并以其名字的首字母命名。

1)指示函数集的 VC 维

给定样本集 $S=\{(x_i,y_i),i=1,2,\cdots,h\}$ 和指示函数集 $F=f[(x,y),\alpha]$，$\alpha\in P$，P 为参数集，如果 F 能把 S 中的所有样本按所有可能的 2^h 种方式分成两类，则称 F 能把 S 打散，F 能打散的最大样本数目 h 称为该函数集的 VC 维。

2)实函数集的 VC 维

给定一有界实函数集 $F=f[(x,y),\alpha],\alpha\in P$，满足 $-\infty<A<F<B<\infty$，同时给定一个阶跃函数：

$$\theta(z)=\begin{cases}0, & \text{if } z<0\\ 1, & \text{if } z\geqslant 0\end{cases} \tag{1.11}$$

那么，如果将这个实函数集 F 一起考虑它的指示器集合，则有

$$I[(x,y),\alpha,\beta]=\theta\{f[(x,y),\alpha]-\beta\} \tag{1.12}$$

式中，$\alpha\in P$ 是参数集，$\beta\in(A,B)$ 是阈值，用于刻画函数集所属类别。这样实函数集 F 的 VC 维可定义为相应指示器集合式(1.12)的 VC 维。

1.2.2 四种重要的归纳原则

以下四种归纳原则在机器学习理论中具有非常重要的作用：经验风险最小化、结构风险最小化、随机逼近方法和邻域风险最小化。讨论这些归纳原则的时候，要用到学习机器推广能力的界的一些结论，相关内容参见后面小节，这里直接使用部分结论。

给定一参数 $0<\eta<1$，下面涉及的界的结论至少以概率 $1-\eta$ 成立。在完全有界且为非负函数集的情况下，学习机器的推广能力满足界

$$P\left\{R(\alpha_n)\leqslant R_{\text{emp}}(\alpha_n)+\frac{B\psi}{2}\left[1+\sqrt{1+\frac{4R_{\text{emp}}(\alpha_n)}{B\psi}}\right]\right\}\geqslant 1-\eta \tag{1.13}$$

对无界函数集来说，学习机器的推广能力满足界

$$P\left\{R(\alpha_n)\leqslant R_{\text{emp}}(\alpha_n)\left[1-\tau\cdot\sqrt{\psi}\cdot\sqrt[p]{\frac{1}{2}\left(\frac{p-1}{p-2}\right)^{(p-1)}}\right]_+^{-1}\right\}\geqslant 1-\eta \tag{1.14}$$

如果损失函数集 $L(z,\alpha_i),\alpha_i\in P,i=1,2,\cdots,N$ 包含 N 个元素，则有

$$\psi=[2(\ln N-\ln\eta)]/n \tag{1.15}$$

如果损失函数集 $L(z,\alpha_i),\alpha_i\in P,i=1,2,\cdots,\infty$ 包含无穷个元素，且学习机器的 VC 维 h 有限，则

$$\psi=\frac{4h\left(\ln\frac{2n}{h}+1\right)-4\ln\frac{\eta}{4}}{n} \tag{1.16}$$

1. 经验风险最小化原则

经验风险最小化(Empirical Risk Minimization，ERM)是应用最为广泛的归纳推理原则之一，主要考虑的是大样本数问题。

根据样本集 $S=\{(x_i,y_i),i=1,2,\cdots,n\}$ 和参数集 P，采用经验风险泛函

$$R_{\text{emp}}(\alpha)=\frac{1}{n}\sum_{i=1}^{n}L[(x_i,y_i),\alpha],\quad \alpha\in P \tag{1.17}$$

替代风险泛函 $R(\alpha)$。它定义了一个学习过程，使经验风险式(1.7)最小的函数记