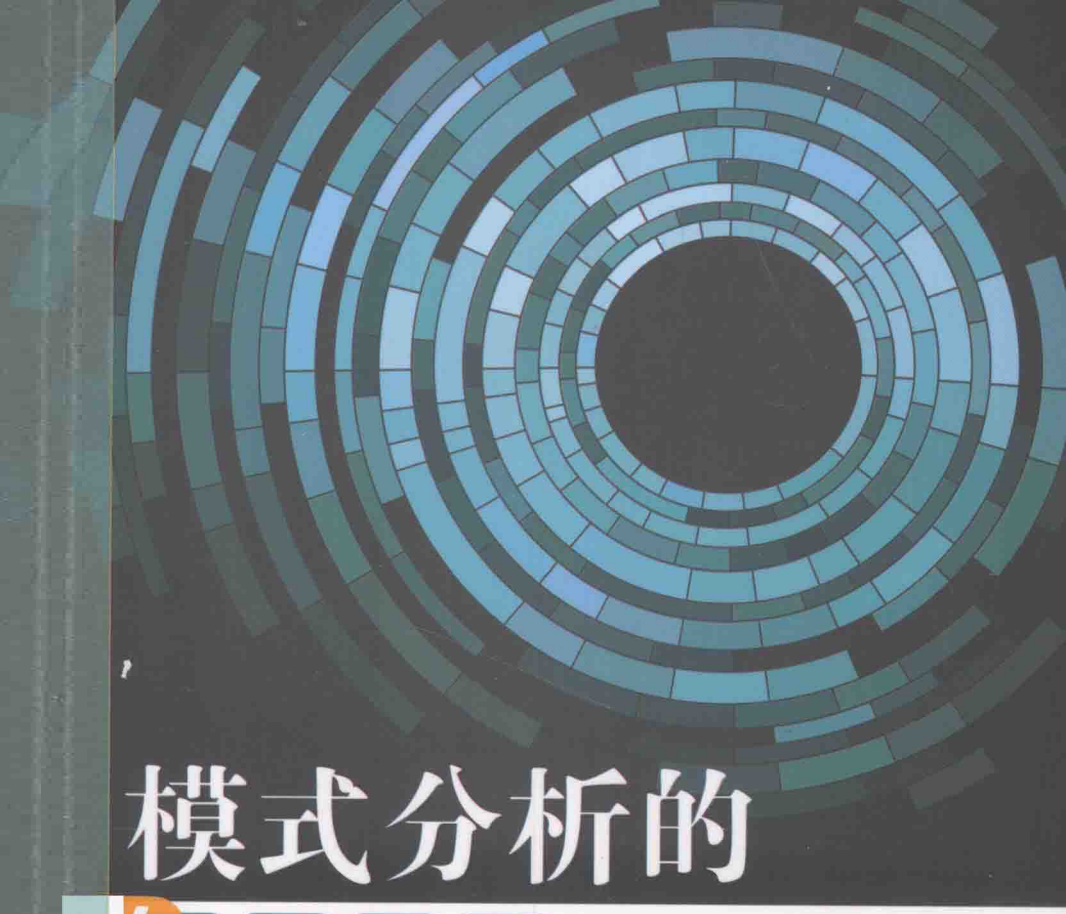


模式分析的多核方法及其应用

汪洪桥 蔡艳宁 王仕成 付光远 孙富春 著

国防工业出版社
National Defense Industry Press

模式分析的多核方法及其应用

汪洪桥　蔡艳宁
王仕成　付光远　孙富春　著

国防工业出版社
·北京·

图书在版编目(CIP)数据

模式分析的多核方法及其应用/汪洪桥等著. —北京:国防工业出版社,2014.3
ISBN 978-7-118-09286-8

Ⅰ.①模… Ⅱ.①汪… Ⅲ.①并行程序—程序设计
Ⅳ.①TP311.11

中国版本图书馆 CIP 数据核字(2014)第 036123 号

※

国防工业出版社出版发行
(北京市海淀区紫竹院南路 23 号　邮政编码 100048)
国防工业出版社印刷厂印刷
新华书店经售

*

开本 880×1230　1/32　**印张** 5⅝　**字数** 250 千字
2014 年 3 月第 1 版第 1 次印刷　**印数** 1—2000 册　**定价** 48.00 元

前　言

研究人员对核方法的关注，得益于支持向量机(Support Vector Machine，SVM)理论的发展和应用。核函数的采用使得线性的SVM很容易推广到非线性情形，是解决非线性模式分析问题的一种有效方法，也是当前机器学习(Machine Learning，ML)领域研究的焦点。但在一些复杂情形下，由单个核函数构成的核机器并不能满足诸如数据异构或不规则、样本规模巨大、样本不平坦分布等实际的应用需求，因此将多个核函数进行融合，以获得更好的结果是一种必然选择，这就是当前机器学习的热点：多核学习方法(Multiple Kernel Learning，MKL)。

本书以多核方法与尺度分析的融合为主线，全面系统地介绍了多核机器的构造、学习以及相关应用的国内外的最新进展。全书共分为7章。第1章为绪论，在回顾统计学习理论与支持向量机的基础上引出多核学习的研究现状及难点；第2章首先从前期融合的角度，将目标多尺度局部特征与整体特征进行融合，并基于核分类器介绍了有效的低维鲁棒组合特征提取和目标识别方法；第3章针对具有不同尺度效应核的融合，深入研究了基于合成核的无偏LSSVR模型，并通过采用基于Cholesky的无偏扩展核矩阵快速分解方法，导出了该模型的快速在线学习算法；第4章基于多尺度特征提取与多尺度核方法，系统研究了基于局部多分辨分解特征提取与多尺度核有机融合的自动目标识别方法；第5章面向多类别、大数据量的图像检索问题，利用合成核方法对图像的多种特征进行有效融合，研究了以分类概率估计为度量依据的图像检索新方法；第6章基于多尺度核的融合，探讨了通用的多尺度核方法及其自适应序列学习问题；第7章还探讨了多核方法的其他应用及提升核机器学习效率的途径。

本书是作者近年来潜心学习和研究国内外多核学习方法和应用

成果的一个总结。很多研究成果是在清华大学计算机科学与技术系孙富春老师的指导和帮助下完成的，这为本书的写作打下了坚实的基础。同时，清华大学自动化系邹红星老师给了作者大量为人师表、做人、处事、做学问的道理和要求，第二炮兵工程大学信息工程系的领导以及学校机关也一直关心和支持作者的课题研究，在此谨向他们表示衷心的感谢！

另外，借此机会特别感谢第二炮兵工程大学信息工程系和清华大学计算机科学与技术系的李琳琳、刘华平、李洪波、陈宁、丁林阁、裴得利、苏娟、封富君、周志杰、于为中、郭金库、杨晓君、王榕、胡来红、闵海波、杨东方、魏振华、伍明、孙振生、席建祥、黄拳章等老师和同学的帮助。

本书的出版受到了国家自然科学基金项目(61202332)、中国博士后科学基金(2012M521905)和第二炮兵工程大学创新性探索项目的资助。

作　者

2013 年 11 月

目　录

第 1 章　核机器学习与多核学习方法

1.1　核方法基础

人们对核方法[1-5]的关注，得益于支持向量机（Support Vector Machine，SVM）[6,7]理论的发展和应用，核函数的采用使得线性的SVM很容易推广到非线性情形。其核心在于利用相对简单得多的核函数运算，避免了特征空间中复杂的内积计算，又避免了特征空间（学习机器）本身的设计[8-10]，是当前机器学习（Machine Learning，ML）[11]领域研究的焦点。

机器学习的目的是根据给定的训练样本，对某系统输入/输出之间依赖关系进行估计，使它（这种关系）能够对未知输出做出尽可能准确地预测。作为人工智能（Artificial Intelligence，AI）的一个重要研究领域，ML的研究工作主要围绕学习机理、学习方法和面向任务这三个基本方面进行研究。典型的ML问题，包括函数拟合（回归估计）、模式分类和概率密度估计等。

以一个监督的机器学习问题为例

$$(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_l, y_l) \in \boldsymbol{X} \times \boldsymbol{Y}$$

式中：输入空间 $\boldsymbol{X} \subseteq \mathbb{R}^n$；输出空间 $\boldsymbol{Y} \subseteq \mathbb{R}$（回归问题）或 $\boldsymbol{Y} = \{-1, +1\}$（两类分类问题）。

可以通过一个非线性映射

$$\begin{aligned} \Phi: \boldsymbol{X} &\to \boldsymbol{F} \\ \boldsymbol{x} &\mapsto \Phi(\boldsymbol{x}) \end{aligned} \tag{1-1}$$

把输入数据映射到一个新的特征空间 $\boldsymbol{F} = \{\Phi(\boldsymbol{x}) \mid \boldsymbol{x} \in \boldsymbol{X}\}$，其中 $\boldsymbol{F} \subseteq \mathbb{R}^N$。然后，我们利用新的数据表示方法考虑原来的学习问题：

$$(\Phi(\boldsymbol{x}_1), y_1), (\Phi(\boldsymbol{x}_2), y_2), \cdots, (\Phi(\boldsymbol{x}_l), y_l) \in \boldsymbol{F} \times \boldsymbol{Y} \tag{1-2}$$

通过非线性映射，一个很难的非线性问题可能变成一个很简单的

线性问题。例如，在二维空间中，一个两类分类的问题[12]，如图 1-1 所示。在输入空间 X，需要一个复杂的非线性决策面才能将两类分开。我们可以通过一个非线性映射

$$\Phi:\mathbb{R}^2\to\mathbb{R}^3$$

$$(x_1,x_2)\mapsto(f_1,f_2,f_3)=(x_1^2,\sqrt{2}x_1x_2,x_2^2)$$

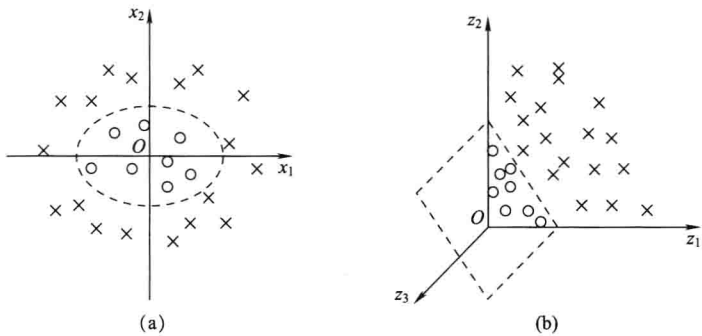


图 1-1　二维分类问题

(a)在输入空间，两类被一个非线性椭圆决策边界划分；

(b) 在特征空间，两类仅需一个线性超平面就可以划分。

把数据映射到特征空间 F，然后只需要一个线性超平面就可以将两类分开。在这个简单的例子中，特征空间只有 3 维，可以先显式地做非线性映射，然后在特征空间里用线性学习机学习。数据原始的量 x_1、x_2 经常被称为属性，为了描述数据而新引入的量 f_1、f_2、f_3 通常称为特征。

输入空间和特征空间采用二次单项式 x_1^2、$\sqrt{2}x_1x_2$、x_2^2 作为特征时的分类结果。

核函数把非线性映射和特征空间中两个向量的内积这两步结合起来，使得非线性映射隐式地进行，而且线性超平面可以由所有训练样本与一个测试样本在特征空间中的内积项的线性组合表示，从而避免了维数灾难。那么什么样的函数才能作为核函数使用呢？或者说如何构造有效的核函数呢？这里存在一个 Mercer 条件。

Mercer 条件[13,14]：设 $\boldsymbol{X}$ 是 $\mathbb{R}^n$ 的一个紧子集，$k:\boldsymbol{X}\times\boldsymbol{X}\to\mathbb{R}$ 是一个

连续的对称函数，如果它在希尔伯特(Hilbert)空间上的积分算子满足积分正定条件

$$\int_{\boldsymbol{X}\times\boldsymbol{X}} k(\boldsymbol{x},\boldsymbol{z})f(\boldsymbol{x})f(\boldsymbol{z})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{z}\geqslant 0,\forall f\in L_2(\boldsymbol{X}) \tag{1-3}$$

那么一定存在一个特征空间 $\boldsymbol{F}$ 和一个映射 $\boldsymbol{\Phi}:\boldsymbol{X}\rightarrow\boldsymbol{F}$，使得

$$k(\boldsymbol{x},\boldsymbol{z})=\boldsymbol{\Phi}(\boldsymbol{x})\cdot\boldsymbol{\Phi}(\boldsymbol{z}) \tag{1-4}$$

因为任何一个满足了积分正定条件式的核函数，可以展开成它的特征函数 $\psi_j\in L_2(\boldsymbol{X})$，$\|\psi_j\|_{L_2}=1$ 和对应的特征值 $\lambda_j>0$ 的形式：

$$k(\boldsymbol{x},\boldsymbol{z})=\sum_{j=1}^{\infty}\lambda_j\psi_j(\boldsymbol{x})\psi_j(\boldsymbol{z}) \tag{1-5}$$

此时，核函数定义的映射可以为

$$\boldsymbol{\Phi}(\boldsymbol{x})=(\sqrt{\lambda_1}\psi_1(\boldsymbol{x}),\sqrt{\lambda_2}\psi_2(\boldsymbol{x}),\cdots) \tag{1-6}$$

针对不同的应用，可以设计不同的核函数。常用的核函数主要有以下几种。

线性核：$k(\boldsymbol{x},\boldsymbol{z})=\boldsymbol{x}^{\mathrm{T}}\boldsymbol{z}$

多项式核：$k(\boldsymbol{x},\boldsymbol{z})=(\boldsymbol{x}\cdot\boldsymbol{z})^d$(齐次)或 $k(\boldsymbol{x},\boldsymbol{z})=(\boldsymbol{x}\cdot\boldsymbol{z}+C)^d$(非齐次)

径向基核(RBF)：$k(\boldsymbol{x},\boldsymbol{z})=\exp\left(-\dfrac{\|\boldsymbol{x}-\boldsymbol{z}\|^2}{2\sigma^2}\right)$

Sigmoid 核：$k(\boldsymbol{x},\boldsymbol{z})=\tanh(\kappa(\boldsymbol{x}\cdot\boldsymbol{z})+\upsilon)$

Fourier 核：$k(\boldsymbol{x},\boldsymbol{z})=\dfrac{1-q^2}{2(1-2q\cos(\boldsymbol{x}-\boldsymbol{z})+q^2)},0<q<1$

核方法的共同特点：①把数据从输入空间非线性映射到特征空间；②在高维的(甚至无穷维)特征空间里采用线性学习；③学得的线性学习机在输入空间则表现为非线性学习机。核方法的学习结果，如分界面函数、拟合函数、特征分量、判别分量、密度函数等，一般都可以写成由训练样本与测试样本组成的核项的线性组合形式，即

$$f(\boldsymbol{x})=\boldsymbol{w}\cdot\boldsymbol{\Phi}(\boldsymbol{x})+b=\sum_{i=1}^{l}\alpha_i k(\boldsymbol{x}_i,\boldsymbol{x})+b \tag{1-7}$$

式中：$\boldsymbol{w}$、$\boldsymbol{\Phi}(\boldsymbol{x})\in\boldsymbol{F}$、$b\in\mathbb{R}$（或 $b=0$）为偏差项；$\alpha_i\in\mathbb{R}$（或 $\alpha_i\geqslant 0$ 且 $\sum_i\alpha_i=1$）为核项系数。

核方法的根本是核函数，当前已经存在了较多的核函数，这些核函数大体可分为两类：局部核（Local Kernel）和全局核（Global Kernel）[15,16]。所谓局部核，就是只有当数据点间的距离很近或者两者的特征相似时，才会对核函数的值产生大的影响；对于全局核，与之相反，当数据点间的距离很远或特征差别很大时，也能对核函数的值产生较大影响。图 1-2 所示的就是在不同参数下，典型的局部核 RBF 核与全局核多项式核的映射特性对比。不同的核函数，体现了其对数据总体或数据局部映射性能的差异。

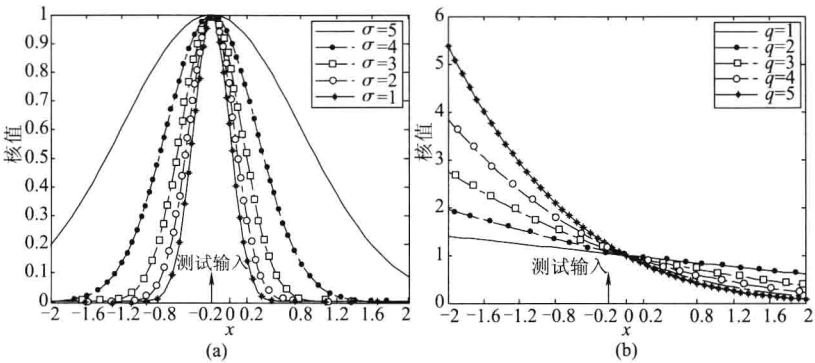


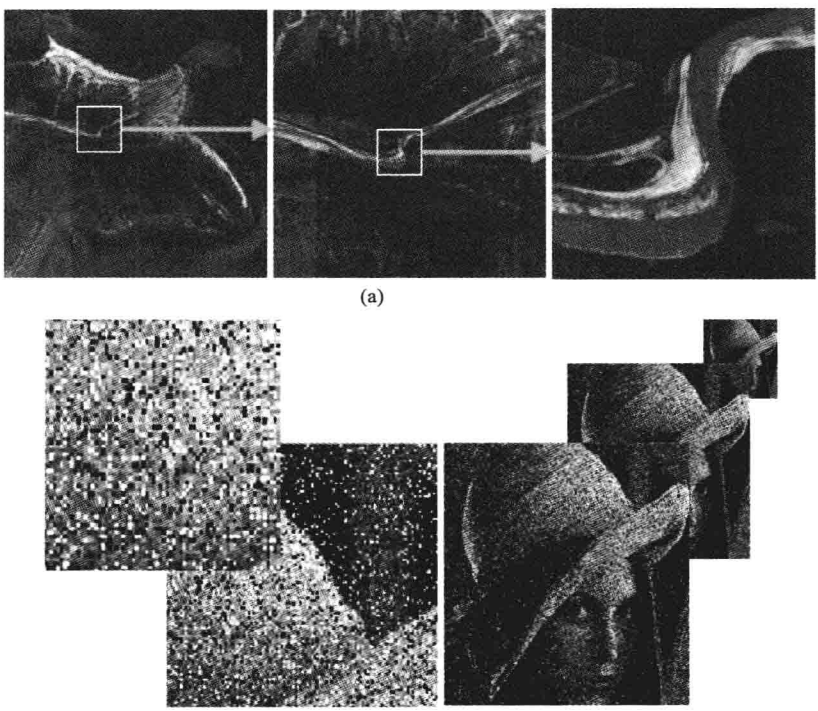
图 1-2　RBF 核与多项式核的映射特性

（a）RBF 核映射特征；（b）多项式核映射特征。

对于任一局部核，通过不同的参数选择，核函数呈现出不同的尺度效应。例如，人在不同的距离下，观测同一图像获得的感受是不一样的，远距离看到的是图像轮廓，近距离下看到的是更多细节，这就是尺度效应。当观察图像中的目标时，由于目标只在某些尺度下才能呈现出来，而在更小的尺度或者更大的尺度下，它们就消失了，如图 1-3 所示。计算机视觉的一个重要的任务就是对目标或者特征进行识别，但是事实上只有在特定的尺度下，这些目标或者特征才会出现，可见尺度选择十分重要。

随着小波理论和多分辨率分析方法的成熟和不断扩展，尺度空间（Scale Space）和尺度分析（Scale Analysis）在诸多领域，特别是在计算机视觉（Computer Vision）方向，体现出了强大的生命力和极佳的实用

性。2004 年 David Lowe 在《计算机视觉》杂志发表的关于 SIFT 算法[17]的经典论文中，将尺度空间定义为“……It has been shown by Koenderink (1984) and Lindeberg (1994) that under a variety of reasonable assumptions the only possible scale-space kernel is the Gaussian function. **Therefore, the scale space of an image is defined as a function, L(x; y; delta) that is produced from the convolution of a variable-scale Gaussian, G(x; y; delta), with an input image I(x; y)**”。简单理解就是，一个图像的尺度空间 $L(x, y, \text{delta})$，定义为原始图像 $I(x, y)$ 与一个可变尺度的二维高斯函数 $G(x, y, \text{delta})$ 的卷积运算。



(a)

(b)

图 1-3　图像的尺度效应

(a)图像在多尺度下的细节变化；(b)多尺度下图像中的目标。

核函数的选择、核参数的确定对数据的映射以及最终的决策效果

发挥了至关重要的作用。不同核函数和核函数在不同参数下，具有不同的尺度效应。基于这一特性，核函数也可以进行这种类似多分辨分析的尺度分析方法。当采用核方法将数据从输入空间映射到特征空间时，借鉴前述的尺度空间和尺度分析，可以构建一个融合多个特征空间的增广空间，通过参数选择，使这个增广空间可以融合这些特征空间的优点，实现更佳的映射性能。在一些复杂情形下，同时考虑核机器分类、回归性能和泛化能力，将不同核组合使用，将是更合理的选择。简单理解，将多种核函数、或不同尺度的核函数进行融合，就称为多核方法。而多核方法的学习问题，即多核学习[18-26]也成为了机器学习的一项重要任务。

本书针对当前多核机器学习问题的研究现状，从尺度分析的角度出发，将多核学习、支持向量机方法相结合，研究了多核学习的相关理论以及在具体模式分类与回归应用中的关键问题。本书以多尺度鲁棒特征的融合、具有不同尺度效应核函数的合成与基于合成核的在线回归分析、基于视觉认知特性的局部多分辨特征提取、多尺度核分类器与自动目标识别、通用多尺度核方法自适应学习、其他提升多核方法学习效率和核机器性能的途径为主线展开研究，给核机器学习提供了更丰富的设计思路，也为模式分析领域的多核应用，如图像目标的特征提取与目标识别、目标的跟踪与定位、在线回归分析、数据的预处理等，提供了多种快速、有效、实用的解决方案。

1.2 统计学习理论与支持向量机

从数学的角度来考虑，机器学习问题就是已知 n 个独立同分布的观测样本，在同一组预测函数中求一个最优的函数对依赖关系进行估计，使期望风险 $R[f]$ 最小。损失函数是评价预测准确程度的一种度量，它与预测函数 $f(x)$ 密切相关。而 $f(x)$ 的期望风险依赖于概率分布和损失函数：前者是客观存在的；后者是根据具体问题选定的，带有人为(主观)的偏好色彩。期望风险的大小直观上可以理解为：当我们用 $f(x)$ 进行预测时“平均”的损失程度，或“平均”犯错误的程度。

但是，只有样本却无法计算期望风险。因此，传统的学习方法用样

本定义经验风险 $R_{emp}[f]$ 作为对期望风险的估计，并设计学习算法使之最小化，即所谓的经验风险最小化（Empirical Risk Minimization，ERM）归纳原则。经验风险是用损失函数来计算的。对于模式识别问题的损失函数来说，经验风险就是训练样本错误率；对于函数逼近问题的损失函数来说，就是平方训练误差；而对于概率密度估计问题的损失函数来说，ERM 准则就等价于最大似然法。事实上，用 ERM 准则代替期望风险最小化并没有经过充分的理论论证，经验风险最小不一定意味着期望风险最小。其实，只有样本数目趋近于无穷大时，经验风险才有可能趋近于期望风险。但是很多问题中样本数目离无穷大很远，那么在有限样本下 ERM 准则就不一定能使真实风险较小。ERM 准则不成功的一个例子就是神经网络的过学习问题（某些情况下，训练误差过小反而导致推广能力下降，或者说是训练误差过小导致了预测错误率的增加，即真实风险的增加）。

统计学习理论（Statistical Learning Theory，SLT）和支持向量机建立了一套较好的有限训练样本下机器学习的理论框架和通用方法，既有严格的理论基础，又能较好地解决小样本、非线性、高维数和局部极小点等实际问题，其核心思想就是学习机器（又称为预测函数，或学习函数，或学习模型）要与有限的训练样本相适应。

1.2.1 统计学习理论

基于数据的机器学习的目的是根据给定的训练样本估计输入输出之间的依赖关系，使它能够对未知输出做出尽可能准确的预测。该问题可以描述为：输出 y 与输入 x 之间存在一定的未知依赖关系，即遵循某一未知的联合概率 $F(x,y)$。根据 l 个独立同分布观测样本 $(x_i, y_i)(i=1,2,\cdots,l)$，从候选函数集 $\{f(x,\omega)\}$ 中寻求一个最优函数 $f(x,\omega_0)$ 对依赖关系进行估计，使期望风险最小。

$$R(\omega)=\int L(y,f(x,\omega))\mathrm{d}F(x,y) \tag{1-8}$$

式中：$\omega\in\Lambda$，Λ 为参数集；$\{f(x,\omega)\}$ 表示任何函数集；$L(y,f(x,\omega))$ 为用 $f(x,\omega_0)$ 对 y 预测造成的损失。

有三类基本的学习问题，即模式识别、函数逼近和概率密度估计，

不同类型的学习问题有不同形式的损失函数。

(1)对于模式识别问题,输出 y 是类别标号,在两类情况下 $y=\{0,1\}$ 或 $y=\{-1,1\}$,损失函数可以定义为

$$L(y,f(x,\omega))=\begin{cases}0, y=f(x,\omega)\\1, y\neq f(x,\omega)\end{cases} \tag{1-9}$$

(2)对于函数逼近问题,输出 y 是连续变量,损失函数可以定义为

$$L(y,f(x,\omega))=(y-f(x,\omega))^2 \tag{1-10}$$

(3)对于概率密度估计问题,学习的目的是根据独立同分布的训练样本确定 x 的概率密度。要估计的概率密度函数定义为 $p(x,\omega)$,则损失函数可以定义为

$$L(p(x,\omega))=-\lg p(x,\omega) \tag{1-11}$$

学习问题更一般的表述为:设有定义在空间 Z 上的概率测度 $F(z)$,考虑函数集合 $Q(z,\omega),\omega\in\Lambda$,学习的目标是最小风险泛函

$$R(\omega)=\int Q(z,\omega)\mathrm{d}F(z),\omega\in\Lambda \tag{1-12}$$

式中:概率测度 $F(z)$ 未知,只有独立同分布样本 $z_i(i=1,2,\cdots,l)$;$Q(z,\omega)$ 是特定的损失函数。

在上面的学习问题中,学习的目标在于使期望风险最小化,但由于可以利用的信息只有有限的数据样本,一般不知道样本空间的概率密度,这样式(1-12)的期望风险实际上无法计算,在传统的机器学习中采用经验风险来代替,因此造成小样本条件下模型泛化能力下降。统计学习理论较好的解决了这一问题。该理论着重研究小样本情况下的统计规律及统计决策方法性质。它为小样本统计决策问题建立了一个较好的理论框架,也发展了一种新的通用统计决策学习方法——支持向量机。统计学习理论有三部分核心内容分别为 VC 维[27,28]、推广性的界理论[29] 和结构风险最小化原则(Structural Risk Minimization,SRM)[30],相关内容见相应参考文献,这里不作赘述。

1.2.2 支持向量机

SVM 是一种基于统计的学习方法,是对 SRM 的近似。概括地说,SVM 就是首先通过用内积函数定义的非线性变换将输入空间变换