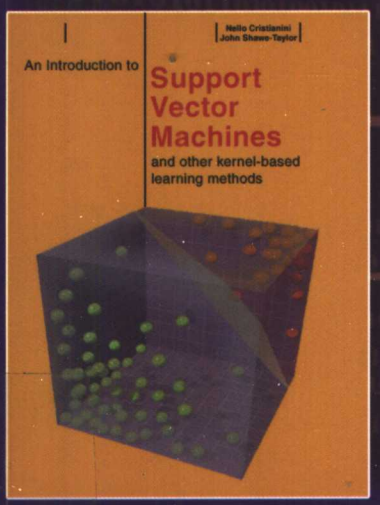
国外计算机科学教材系列

支持向量机导论

An Introduction to Support Vector Machines and Other Kernel-based Learning Methods



[美] Nello Cristianini
John Shawe-Taylor 著

李国正 王 猛 曾华军 译

電子工業出版社
Publishing House of Electronics Industry
http://www.phei.com.cn

国外计算机科学教材系列

支持向量机导论

An Introduction to Support Vector Machines
and Other Kernel-based Learning Methods

[英] Nello Cristianini
John Shawe-Taylor 著

李国正 王 猛 曾华军 译

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

内容简介

支持向量机（SVM）是在统计学习理论的基础上发展起来的新一代学习算法，该算法在文本分类、手写识别、图像分类、生物信息学等领域中获得了较好的应用。本书是第一本支持向量机方面的导论型读物。它从机器学习算法的基本问题开始，循序渐进地介绍相关的背景知识，包括线性分类器、核函数特征空间、推广性理论和优化理论，从而很自然地引出了支持向量机的算法。书的末尾还详细讨论了一系列支持向量机的重要应用以及实现的技巧。该书提供的大量相关文献以及网站链接为进一步学习提供了有效线索，有助于读者及时跟踪该领域的最新信息。本书可作为计算机、自动化、机电工程、应用数学等专业的研究生教材，也可作为神经网络、机器学习、数据挖掘等课程的参考教材，同时还是相关领域的教师和研究人员的参考书。

版权贸易合同登记号　图字：01-2002-4550

图书在版编目（CIP）数据

支持向量机导论/（英）克里斯特安尼（Cristianini, N.）等著；李国正，王猛，曾华军译.
-北京：电子工业出版社，2004.3
（国外计算机科学教材系列）
书名原文：An Introduction to Support Vector Machines and Other Kernel-based Learning Methods
ISBN 7-5053-9336-7

Ⅰ.支…　Ⅱ.①克…②李…③王…④曾…　Ⅲ.电子计算机-算法理论-教材　Ⅳ.TP301.6

中国版本图书馆CIP数据核字（2004）第009806号

责任编辑：史　平
印　　刷：北京兴华印刷厂
出版发行：电子工业出版社
　　北京市海淀区万寿路173信箱　邮编：100036
经　　销：各地新华书店
开　　本：787×980　1/16　　印张：11　字数：215千字
印　　次：2004年3月第1次印刷
定　　价：25.00元

凡购买电子工业出版社的图书，如有缺损问题，请向购买书店调换；若书店售缺，请与本社发行部联系。联系电话：（010）68279077。质量投诉请发邮件至zlts@phei.com.cn，盗版侵权举报请发邮件至dbqq@phei.com.cn。

出 版 说 明

21世纪初的5至10年是我国国民经济和社会发展的重要时期，也是信息产业快速发展的关键时期。在我国加入WTO后的今天，培养一支适应国际化竞争的一流IT人才队伍是我国高等教育的重要任务之一。信息科学和技术方面人才的优劣与多寡，是我国面对国际竞争时成败的关键因素。

当前，正值我国高等教育特别是信息科学领域的教育调整、变革的重大时期，为使我国教育体制与国际化接轨，有条件的高等院校正在为某些信息学科和技术课程使用国外优秀教材和优秀原版教材，以使我国在计算机教学上尽快赶上国际先进水平。

电子工业出版社秉承多年来引进国外优秀图书的经验，翻译出版了“国外计算机科学教材系列”丛书，这套教材覆盖学科范围广、领域宽、层次多，既有本科专业课程教材，也有研究生课程教材，以适应不同院系、不同专业、不同层次的师生对教材的需求，广大师生可自由选择和自由组合使用。这些教材涉及的学科方向包括网络与通信、操作系统、计算机组织与结构、算法与数据结构、数据库与信息处理、编程语言、图形图像与多媒体、软件工程等。同时，我们也适当引进了一些优秀英文原版教材，本着翻译版本和英文原版并重的原则，对重点图书既提供英文原版又提供相应的翻译版本。

在图书选题上，我们大都选择国外著名出版公司出版的高校教材，如Pearson Education培生教育出版集团、麦格劳－希尔教育出版集团、麻省理工学院出版社、剑桥大学出版社等。撰写教材的许多作者都是蜚声世界的教授、学者，如道格拉斯·科默（Douglas E. Comer）、威廉·斯托林斯（William Stallings）、哈维·戴特尔（Harvey M. Deitel）、尤利斯·布莱克（Uyless Black）等。

为确保教材的选题质量和翻译质量，我们约请了清华大学、北京大学、北京航空航天大学、复旦大学、上海交通大学、南京大学、浙江大学、哈尔滨工业大学、华中科技大学、西安交通大学、国防科学技术大学、解放军理工大学等著名高校的教授和骨干教师参与了本系列教材的选题、翻译和审校工作。他们中既有讲授同类教材的骨干教师、博士，也有积累了几十年教学经验的老教授和博士生导师。

在该系列教材的选题、翻译和编辑加工过程中，为提高教材质量，我们做了大量细致的工作，包括对所选教材进行全面论证；选择编辑时力求达到专业对口；对排版、印制质量进行严格把关。对于英文教材中出现的错误，我们通过与作者联络和网上下载勘误表等方式，逐一进行了修订。

此外，我们还将与国外著名出版公司合作，提供一些教材的教学支持资料，希望能为授课老师提供帮助。今后，我们将继续加强与各高校教师的密切联系，为广大师生引进更多的国外优秀教材和参考书，为我国计算机科学教学体系与国际教学体系的接轨做出努力。

电子工业出版社

教材出版委员会

主　任	杨芙清	北京大学教授 中国科学院院士 北京大学信息与工程学部主任 北京大学软件工程研究所所长
委　员	王　珊	中国人民大学信息学院院长、教授
	胡道元	清华大学计算机科学与技术系教授 国际信息处理联合会通信系统中国代表
	钟玉琢	清华大学计算机科学与技术系教授 中国计算机学会多媒体专业委员会主任
	谢希仁	中国人民解放军理工大学教授 全军网络技术研究中心主任、博士生导师
	尤晋元	上海交通大学计算机科学与工程系教授 上海分布计算技术中心主任
	施伯乐	上海国际数据库研究中心主任、复旦大学教授 中国计算机学会常务理事、上海市计算机学会理事长
	邹　鹏	国防科学技术大学计算机学院教授、博士生导师 教育部计算机基础课程教学指导委员会副主任委员
	张昆藏	青岛大学信息工程学院教授

译 者 序

支持向量机由 Vapnik 及其合作者发明，在 1992 年计算学习理论的会议上介绍进入机器学习领域，之后受到了广泛的关注。在 20 世纪 90 年代中后期得到了全面深入的发展，现已成为机器学习和数据挖掘领域的标准工具。

支持向量机是机器学习领域若干标准技术的集大成者。它集成了最大间隔超平面、Mercer 核、凸二次规划、稀疏解和松弛变量等多项技术。在若干挑战性的应用中，获得了目前为止最好的性能。在美国科学杂志上，支持向量机以及核学习方法被认为是“机器学习领域非常流行的方法和成功的例子，并是一个十分令人瞩目的发展方向”。

《支持向量机导论》是一本综合性介绍支持向量机各项标准技术的著作，书中从学习方法到超平面、核函数、泛化性理论、最优化理论，最后总结到支持向量机理论，并介绍了其实现技术和应用。本书的叙述循序渐进，内容深入浅出，既严谨又易于理解，得到了很多支持向量机研究者的认可。本书的原著已经是第二版第四次印刷。其一大特色是作者建立了一个网站 http://www.support-vector.net/，提供了在线的参考文献和一些软件的链接。这被评为“独特并具有现代化色彩”。目前国内还很缺乏这样系统性的著作，所以译者很乐意将本书翻译过来，供国内支持向量机、神经网络、机器学习、模式识别、数据挖掘乃至统计数学等领域的研究者参考。

译者在翻译过程中力求忠于原著。专业术语尽量遵循各学科的标准。由于水平和时间有限，对于原著的理解可能会有偏差，书中错误和不妥之处在所难免，恳请读者批评指正。

本书第 1 章和第 5 章初稿分别由曾华军和王猛翻译，其余章节由李国正翻译，全书由李国正整理完成。上海大学陈念贻教授校对了部分内容，并对一些翻译用语提出了中肯的建议；南京大学周志华教授就前六章细致地修改了一些错误的地方，并对专业术语的使用提出了中肯的意见；天津大学郑建华同学仔细阅读了译稿并提出了若干修改意见；清华大学孙建涛博士，北京的邹涛研究员，上海的叶晨周博士，交通大学的王永刚博士和青岛朗讯的刘玉海研究员分别对部分章节的翻译提出了一些宝贵的修改意见。南京航空航天大学陈松灿教授也对本书的翻译予以支持。译者谨在此对这些老师和同学表示感谢！李国正和王猛还要感谢导师杨杰教授提供的宽松舒适的学术环境以及在学习研究过程中所给予的指导。

前　言

近年来，支持向量机（SVM）的理论已经取得重大进展，其算法实现策略以及实际应用也发展迅速。可以确信，该技术的研究已发展成为机器学习中一个独立的子领域，在理论和实践两方面都有着光明的前景。尽管如此，目前还很缺乏对这个主题较系统的介绍材料。SVM 的理论体系涵盖的对象极为广泛，包括对偶表示、特征空间、学习理论、优化理论和算法等。

虽然关于这些主题仍在进行大量的研究工作，但其基础理论已经发展成熟，足以构建 SVM 的整个框架。本书从这些基础理论出发，尝试以 SVM 为主题进行深入浅出的介绍。

本书提供了严谨但易懂的阅读材料，可以作为学生或研究者进入机器学习这一新领域的指导读物。本书组织成教材形式，既可作为 SVM 课程的核心材料，也可作为神经网络、机器学习或模式识别等课程的附加读物。因为采取了教材形式，本书的内容尽可能做到自包含，以使非机器学习或者没有计算机背景的读者易于掌握。这样其他领域的读者就可以很轻松地应用 SVM 到自己的实际问题中去。作者还尝试通过严谨的推导来提供清晰的学习路线，因此书中提供了定理简要的证明过程以增强读者对概念的理解。对详细的证明过程感兴趣的读者可以参考原始文献。

每章后面附有习题，以及相关的参考文献或软件。由于在线资源可能会更新，所以有些引用会指向某一专门网站，其中相关的链接也会相应更新，从而保证读者找到相应的在线资源。

尽管有些引用是间接的，本书还是尽量标明所引用素材的作者。希望各位作者不会因为这些间接引用而感到不快。每章的结尾有补充读物和高级主题，其作用有二：一是将所有引用包含在这一节以使正文尽可能整洁，这里要再次请求所引用素材的作者宽恕这种延后的引用说明；二是希望为读者提供一个出发点以便能够进一步学习该章所讨论的主题。这些参考材料也包含在网站中，并保持更新。将引用说明移到正文之外的另一考虑是该领域已经到达一个成熟阶段，从而能使用前后一致的表述。但在这个方面有两个例外，一是某些定理通常可由其命名得知其作者，比如 Mercer 定理；另一例外是在第 8 章，其中描述了研究领域的一些特定实验。

本书写作中遵循的基本原则是对于不喜欢复杂的证明和概念的学生和研究者而言易于理解。相信通过直观且严谨的内容安排，SVM 的出现会显得简单又自然。本

书还尽量用简单的例子来介绍概念，之后才说明在复杂情形下的用法。本书是自包含的，附录中提供了基本线性代数和概率论之外的一些数学工具，从而更适合于跨学科的读者。

书中大部分材料在一个 5 小时的 SVM 和大间隔推广能力的讲座中展示过，该讲座是 1999 年在 California 大学 Santa Cruz 校区举办的。其中的大多数反馈意见也包含在书中。本书的部分章节是 Nello 在 California 大学 Santa Cruz 校区访问时写的，感谢这里的东道主和优美的校园环境。写作过程中，Nello 多次长时间访问 London 大学 Royal Holloway 校区。他很感谢 Lynda 及其家人的三次款待。Nello 和 John 还想感谢技术管理员 Alex Gammerman 以及 Royal Holloway 的计算机系同仁，他们提供了宽松舒适的环境，使得作者安心写作。

许多人为本书的雏形和主干做出了贡献，包括间接的讨论和对手稿的直接评注。作者感谢 Kristin Bennett, Colin Campbell, Nicolò Cesa-Bianchi, David Haussler, Ralf Herbrich, Ulrich Kockelkorn, John Platt, Tomaso Poggio, Bernhard Schölkopf, Alex Smola, Chris Watkins, Manfred Warmuth, Chris Williams 和 Bob Williamson。

作者还感谢David Tranah和Cambridge大学出版社对本书出版过程中的大力支持和帮助。

Alessio Cristianini 帮助建立了网站。Kostantinos Veropoulos 在 Bristol 大学用他的软件包生成了第 6 章的图片。感谢 John Platt 提供了附录 A 中的 SMO 伪码。

Nello 衷心感谢 EPSRC 的支持，感谢 Colin Campbell 主管的理解和协助。John 感谢欧洲委员会对于 NeuroCOLT2 工作组 EP27150 的支持。

由于第一版中出现了少量错误，作者尽力保证在重印前修改这些错误。欢迎读者指出更多的问题，并通过本书的网站 www.support-vector.net 反馈给作者。

Nello Cristianini 和 John Shawe-Taylor

2000 年 6 月

符 号 说 明

N	特征空间维数
$y \in Y$	输出和输出空间
$\mathbf{x} \in X$	输入和输入空间
F	特征空间
$\mathscr{F}$	实值函数类
$\mathscr{L}$	线性函数类
$\langle \mathbf{x} \cdot \mathbf{z} \rangle$	$\mathbf{x}$ 和 $\mathbf{z}$ 的内积
$\phi : X \to F$	到特征空间的映射
$K(\mathbf{x}, \mathbf{z})$	核$\langle \phi(\mathbf{x}) \cdot \phi(\mathbf{z}) \rangle$
$f(\mathbf{x})$	阈值化前的实值函数
n	输入空间的维数
R	包含数据的球半径
ε-insensitive	ε误差不敏感损失函数
$\mathbf{w}$	权重向量
b	偏置
α	对偶变量或者拉格朗日乘子
L	原拉格朗日函数
W	对偶拉格朗日函数
$\|\cdot\|_p$	p 阶范数
ln	自然对数
e	自然对数的底
log	以 2 为底的对数
$\mathbf{x}'$, $\mathbf{X}'$	向量、矩阵的转置
$\mathbb{N}$, $\mathbb{R}$	自然数、实数
S	训练样本
ℓ	训练样例个数
η	学习率
ε	误差概率

δ	置信范围
γ	间隔
ξ	松弛变量
d	VC 维

目　录

第1章 学习方法

长期以来，构造可以从经验中学习的机器在哲学界和科技界都是研究目标之一。从技术角度来讲，机器学习从电子计算机的发明中获得了强大的原动力。机器具有可观的学习能力，这一点已经得到了证实，然而这种学习能力的界定还远不够清晰。

开发可靠学习系统的重要性体现在有许多实际任务不能用传统的编程技术实现。目前并没有针对此问题的数学模型。比如，即使给予大量的样例，也不知道怎样用计算机程序来识别手写字符。因此很自然地产生疑问，能否训练计算机从样例中学习来识别字符“A”？毕竟这也是人类学习阅读的手段，可以把这种问题求解的途径称为学习方法。

同样的方法可适用于在DNA序列中寻找基因，过滤电子邮件，在机器视觉中检测或识别目标，等等。其中每个问题的解决都将为人类的生活带来革命性的变化，而每一个问题中机器学习都是其解决方案的关键所在。

本章将介绍学习方法的重要组成部分，总结各种不同种类的学习，并讨论其具有理论重要性的原因。在介绍学习方法的框架之后，给出了全书内容和关键问题的预览，并且解释了支持向量机为什么能够解决机器学习系统面临的问题。因此本章是对全书的预览，希望读者在阅读后面章节之前查阅本部分。

1.1 监督学习

当计算机应用到实际问题中时，通常可以显式地描述出给定一组输入如何推出所需的输出。而系统设计者和最终的程序实现人员的任务仅仅是将其转换为一系列的指令，使计算机能够遵循指令达到期望的结果。

由于计算机应用于更复杂的问题中，有时并不知道如何由给定输入计算出期望的输出，或者这种计算可能代价很高。例如对一个复杂的化学反应进行建模，则无法精确获知不同反应物质的相互作用关系；再如利用DNA序列对于蛋白质类型分类，或者对信用卡申请表的分类，以区分哪些人有能力偿还债务。

这些任务都不能用传统的编程途径来解决，因为系统设计者无法精确指定从输入数据得到输出的方法。解决此类问题的一种策略是让计算机从样例中学习输入到输出的函数对应关系，就像儿童学习辨认赛车的过程，是给他们大量赛车的例子，

而不是告诉他们赛车的精确规格说明。这种使用样例来合成计算机程序的过程称为学习方法，其中当样例是由输入/输出对给出时，称为监督学习。有关输入/输出函数关系的样例称为训练数据。

输入/输出对通常反映了把输入映射到输出的一种函数关系，然而有些情形下并非如此，比如输出中包含了噪声干扰。当输入到输出存在内在函数时，该函数称为目标函数。由学习算法输出的对目标函数的估计称为学习问题的解。对于分类问题，该函数有时称为决策函数。存在一系列候选函数可把输入空间映射到输出域，可以选择其中之一为解。通常可以选择一组或一类候选函数，它们称为假设集合。例如，决策树是通过构造二叉树而产生假设，树的内部节点是简单的决策函数，而叶节点是输出值。因此把假设集合或者假设空间的选择看做学习过程的关键因素，而从训练数据中学习并从假设空间中选择假设的算法是第二个重要因素，它称为学习算法。

在学习区分赛车的例子中，输出为简单的是/否，它可看做是二元输出值。对于识别蛋白质类型的问题，输出值为有限数量的类别之一；对于化学反应的问题输出值为实数值表示的反应化合物的浓度。有二元输出的问题称为二类问题，有多个类别的问题称为多类问题，而实数值输出的问题称为回归问题。本书考察所有这些类型的学习问题，其中二类问题通常作为最简单情形被率先考虑。

其他还有一些学习类型不在本书中考虑。例如非监督学习，其数据不包含输出值，学习的任务是理解数据产生的过程。这种类型的学习包括密度估计、分布类型的学习和聚类等。还有一些学习模型考虑了学习器与其环境复杂的交互过程，其中最简单的情形是学习器可针对一特定输入向系统查询其输出。关于在该过程中如何影响学习器能力的研究称为查询学习。更复杂的交互在强化学习中考虑。这里学习器拥有一组可任意执行的动作，学习器通过执行这些动作尝试达到较高回报的状态。学习方法也可应用于强化学习中，前提是要将最优动作看做是学习器当前状态的一个函数输出。然而这样会产生相当的复杂性，因为输出的质量只能在动作的后果清晰后才能被间接衡量。

学习模型的另一方面的问题是训练数据如何生成及如何输入到学习器。例如，批量学习和在线学习之间存在明显的区别，前者在学习的一开始就把所有的数据提供给学习器；后者则让学习器一次只接收一个样例，并在接收正确输出前给出自己对输出的估计。在线学习中学习器根据每个新样例更新当前假设，学习器的质量由学习期间产生的总错误数量来衡量。

本书着重于在批量学习的设置下，根据有输出值的数据来学习输入/输出映射，即应用监督学习方法到批量训练数据上。

1.2 学习和泛化性

前面讨论了在线学习算法的质量可由其训练阶段的出错总数来衡量。然而，怎样衡量批量学习中所产生假设的质量还不明确。早期的学习算法致力于通过学习产生简单的符号表示。它可由专家来理解和验证。而当前情形下，学习的目标是输出一个假设以正确分类训练数据，早期的学习算法目标也是寻找对数据的精确拟合，这样寻找到的假设称为一致假设。然而生成可验证的一致假设这一目标存在两个问题。

一个问题是待学习的目标函数可能没有简单的表示，因此不能很容易地加以验证，例如在 DNA 序列中定位基因。某些子序列是基因，某些不是。但没有一种简单的方法来区分两者。

第二个问题是，通常训练数据是有噪声的。因此不能保证存在一个目标函数能够正确地映射训练数据。很明显信用检测是其中一例，因为偿债能力可能取决于其他一些系统无法获知的因素。另一个例子是网页分类的问题，这也是一个不精确的科学问题。

对于机器学习研究者来说，他们感兴趣的数据将越来越多地属于这种类型。因此使得上面的质量衡量标准难以实现。还有一个更为基本的问题在于，即使能够找到与训练数据一致的假设，它也可能无法对未见数据进行分类。一个假设正确分类训练集之外数据的能力称为泛化性，这正是要优化的属性。

把研究目标转移到泛化性，就不再要求把假设看做真实目标函数的正确表示。如果一个假设能给出正确的输出，它就满足泛化性准则，在这种意义上泛化性成为了一个功能性准则而不是描述性准则。同时该准则不再限制假设的规模以及“含义”，它们今后可以是任意的。

在后面可能看到这种重心改变被削弱的情形，因为可能需要搜索假设的简洁表示（即简短描述），这些可看做是有较好泛化性的属性。但目前，这种改变可看做是从符号表示到亚符号表示的转移。

第 4 章将给出这些概念的精确定义，那时将阐明使用这些模型的动机。

1.3 提高泛化性

泛化性准则对于学习算法附加了另一种约束。这一点可以由一种极端情形下的机械式学习来充分说明。许多经典的机器学习算法能够表示任意函数，并且对于困难的训练数据集会得到一个类似机械式学习器的假设。所谓机械式学习器是指能够

正确分类训练数据，但对所有未见数据会做出根本无关联性的预测。例如，决策树有可能过度增长直至针对每个训练样例有一叶子节点。为了得到一致假设而使假设变得过度复杂称为过拟合。控制此问题的一种方法是限制假设的规模，例如对于决策树可进行修剪操作。奥卡姆（Ockham）剃刀是该类方法中的准则之一，它建议如无必要，不必增加复杂性，或者说更精细的复杂性必须有利于显著提高训练数据的分类正确率。

这些观点自从奥卡姆的建议被采用后已有很长的历史。借此可引发复杂性和精度之间启发式的平衡，而且人们已提议了多种准则在这两者之间选择最优的折中。例如最小描述长度（MDL, Minimum Description Length）准则建议使用这样的假设：其函数描述长度与训练错误列表长度的和最短。

将要采用的方法是为了获得另一种平衡，它涉及泛化误差率上的统计边界，这些边界通常依赖于分类器间隔这样的变量，并引发最优化该变量的算法。该途径的缺点在于此算法不会好于统计结果。另一方面，算法的优点在于统计结果为其提供了一个有充分依据的理论基础，因此能避免基于错误直觉的启发式方法所带来的危险。

算法设计基于统计结果这一点并非意味着忽略解决此优化问题的计算复杂度。所感兴趣的技术需具有可伸缩性，它应能处理从玩具世界的问题到包含上万条记录的真实数据集的问题。只有通过对计算复杂度的原则性分析，才可能避免满足于那些只在小样本上表现良好，却对大训练集完全失效的启发式规则。计算复杂度理论研究了两类问题，第一类问题是是否存在算法能够在输入规模的多项式时间内运行的问题；第二类问题是如果存在这样的算法，任意解能否在多项式时间内检验，也就是能否在多项式时间内求解的问题。后一类问题即为 NP 完全问题，通常认为这些问题不能有效求解。

第 4 章将详细地阐述促使本书算法产生的统计结果。以下两种统计结果应加以区分：一种衡量了在给定训练样例有限时可获得的泛化性能；另一种是渐进统计结果，它研究了当样例数目趋于无穷时的泛化性能。本书要介绍的是前一种类型，该方法由 Vapnik 和 Chervonenkis 开创。

应当强调指出，还存在一些算法，它们与本书描述的不同，可由其他一些学习方法产生。在正文中将引用这些方法的相关文献。现在要稍微详细地介绍一下贝叶斯方法。

贝叶斯分析的出发点是假设集合上的先验分布，它描述了学习器对于数据特定假设的似然性的先验信念。只要能假定这样的先验分布，再加上数据如何被噪声干扰的模型，原则上就有可能在给定训练集合的情况下估计最可能的假设，甚至也可以在可能假设的集合上做加权平均。