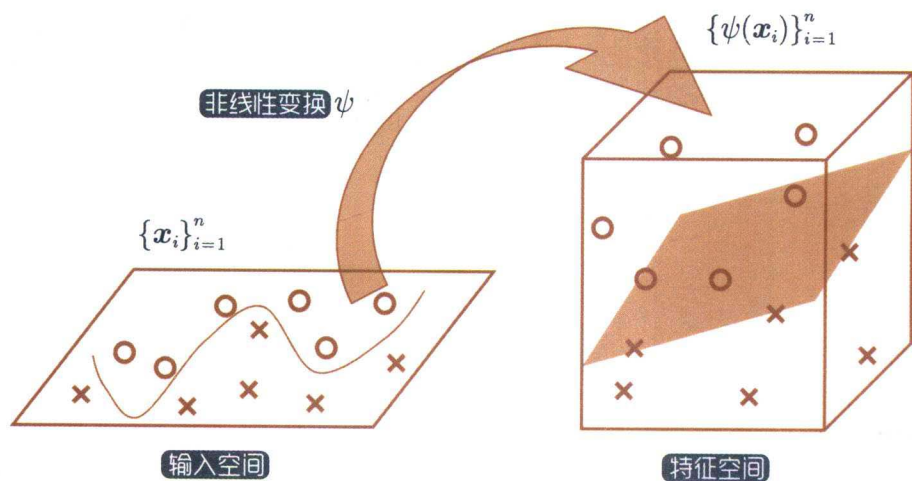


图解机器学习

[日] 杉山将 著

许永伟 译



187张图解轻松入门

提供可执行的MATLAB程序代码
覆盖机器学习中最经典、用途最广的算法



中国工信出版集团

人民邮电出版社
POSTS & TELECOM PRESS

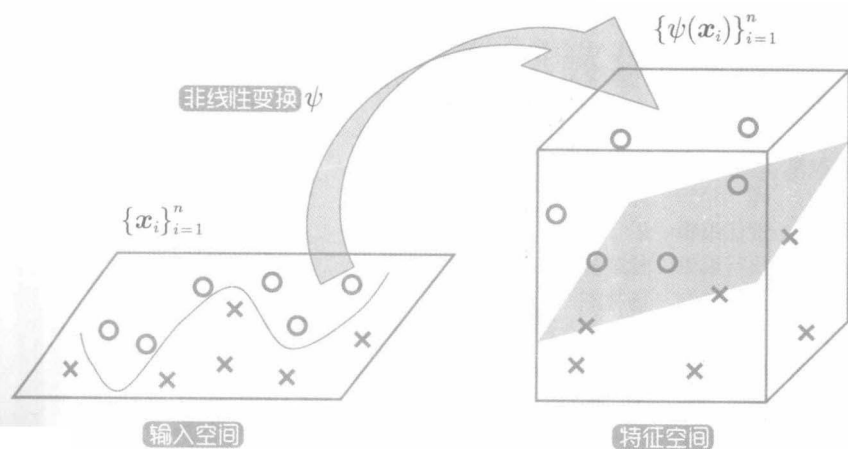
TURING

图灵程序
设计丛书

图解机器学习

[日] 杉山将 著

许永伟 译



人民邮电出版社
北 京

图书在版编目 (C I P) 数据

图解机器学习 / (日) 杉山将著 ; 许永伟译. — 北京 : 人民邮电出版社, 2015. 4
(图灵程序设计丛书)
ISBN 978-7-115-38802-5

I. ①图… II. ①杉… ②许… III. ①机器学习—图解 IV. ①TP181-64

中国版本图书馆CIP数据核字 (2015) 第052954号

内 容 提 要

本书用丰富的图示, 从最小二乘法出发, 对基于最小二乘法实现的各种机器学习算法进行了详细的介绍。第 I 部分介绍了机器学习领域的概况; 第 II 部分和第 III 部分分别介绍了各种有监督的回归算法和分类算法; 第 IV 部分介绍了各种无监督学习算法; 第 V 部分介绍了机器学习领域中的新兴算法。书中大部分算法都有相应的 MATLAB 程序源代码, 可以用来进行简单的测试。

本书适合所有对机器学习有兴趣的初学者阅读。

-
- ◆ 著 [日] 杉山将
译 许永伟
责任编辑 乐 馨
执行编辑 杜晓静
责任印制 杨林杰
- ◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路 11 号
邮编 100164 电子邮件 315@ptpress.com.cn
网址 <http://www.ptpress.com.cn>
北京天宇星印刷厂印刷
- ◆ 开本: 880 × 1230 1/32
印张: 7.5
字数: 209 千字 2015 年 4 月第 1 版
印数: 1-4 000 册 2015 年 4 月北京第 1 次印刷
著作权合同登记号 图字: 01-2014-3345 号
-

定价: 49.00 元

读者服务热线: (010)51095186 转 600 印装质量热线: (010)81055316

反盗版热线: (010)81055315

广告经营许可证: 京崇工商广字第 0021 号

站在巨人的肩上

Standing on Shoulders of Giants



iTuring.cn

版 权 声 明

IRASUTO DE MANABU KIKAI GAKUSYUU

SAISYOUNIJYOUHOU NI YORU SHIKIBETSU MODERU GAKUSYUU O TYUUSHIN NI

© Masashi Sugiyama 2013

All rights reserved.

Original Japanese edition published by KODANSHA LTD.

Publication rights for Simplified Chinese character edition arranged with
KODANSHA LTD. through

KODANSHA BEIJING CULTURE LTD. Beijing, China.

本书由日本讲谈社授权人民邮电出版社发行简体字中文版，版权所有，未经书面同意，不得以任何方式做全面或局部翻印、仿制或转载。

译者序

机器学习领域是深不可测的吗？人工智能是天方夜谭吗？时至今日，机器学习研究的重要性与可行性已得到广泛承认，并在模式识别、通信、控制、金融、机器人、生物信息学等许多领域都有着广泛的应用。

如何自动归类筛选邮件和网页？如何向大家推荐你可能感兴趣的人？如何预测整体市场行情的好坏？如何从统计学的角度对照片进行归类？本书就介绍了这样一些算法。

如果想得到最通俗、简洁的讲解，本书最为合适。

如果想立即知道算法的性能，并期望有可运行的源代码，本书最为方便。

很多人都是看着日本的动画长大的。殊不知，大部分日本人都具有熟练的绘画能力。他们总可以把复杂、枯燥的事物用惟妙惟肖的漫画生动地表达出来。广告、网页、海报，甚至政府公告都图文并茂。市面上也有不计其数的“图解……”“图说……”一类的书籍。本书就是其中一例，这也是本书的最大特点。

杉山将博士今年赴任东京大学教授，他在机器学习领域颇有建树。他的研究室吸引了来自世界各地的机器学习研究者。本书承袭了日本特有的绘画特色，依靠作者丰富的机器学习经验，用最精简的文字，将原本复杂抽象的数学原理，用形象的漫画与数据图形进行了清晰的说明。作者也将最前沿和最核心的研究成果汇集到了本书之中。

本书的侧重点不在于机器学习原理的相关推导，而在于结论的分析和应用。读者朋友可以更快地掌握各种算法的特点和使用方法，提纲挈领地消化应用，而不必拘泥于算法的细节不能自拔。另外，本书

旁征博引，图文并茂，结构清晰，范例实用丰富，深入浅出地说明了机器学习中最典型和用途最广泛的算法。

本书内容覆盖面广，不但与市面上众多的机器学习书籍并无重复，更可与其互为补充。大部分算法都有简洁、现成的MATLAB源代码，读者朋友可以轻松地进行验证。以此为原型，再稍加修改扩充，即可做出为自己所用的项目代码。

机器学习领域日新月异，书中所涉及的概念和术语数目繁多，且有许多概念和术语目前尚无公认的中文译法。如果有不合读者朋友习惯的术语出现，请参考译者注，确认其原始词意。

本译稿得到了图灵公司编辑的悉心指导，她们为保证本书的质量做了大量的补译、校正及编辑工作，在此表示深深的谢意。

许永伟

2014年12月于东京

序

本书是关于机器学习的入门图书。说到“机器”，可能很多人都会想到机械表或车床等大型机器设备，但是机器学习里的“机器”指的是计算机。机器学习，是指让计算机具有人那样的学习、思考能力的技术的总称。近年来，随着计算机软硬件技术的发展，机器学习领域也得到了巨大的进步。本书就是介绍这一蓬勃发展中的机器学习算法的一本书。

在机器学习领域，借助高级的数学方法，各种新型算法层出不穷。因此对于初涉这一领域的研究人员、技术工作者和学生来说，要理解这些最前沿的技术往往有很多困难。然而大部分这些最新的机器学习算法，都是在最经典的算法——最小二乘法的基础上发展起来的。本书就是立足于这样的视点，对基于最小二乘法实现的各种机器学习算法做简单的介绍，并给出许多具体的实例。因此，只要理解了最小二乘法的基本原理，即可掌握能够处理中等数据规模的大多数高级算法。

本书由以下几部分构成。

第Ⅰ部分介绍了本书所涉及的机器学习领域的概况。首先在第1章，对监督学习、无监督学习和强化学习等基本概念进行了说明；第2章介绍了机器学习里需要使用到的各种各样的模型。

第Ⅱ部分介绍了与连续函数的近似问题相对应的各种回归算法。具体来说，首先在第3章引入了回归算法的基础，即最小二乘学习法；第4章介绍了能够避免过拟合问题的条件约束的最小二乘学习法。第5章介绍了通过把大部分参数置为0来大幅提高学习效率、计算精度的稀疏算法。第6章介绍了不易受到异常值影响的鲁棒学习法。

第Ⅲ部分介绍了各种分类算法。第7章介绍了回归问题中直接使用最小二乘学习法进行分类的算法。第8章引入了基于间隔最大化原

理的支持向量机分类器的算法，并且明确了支持向量机分类器和最小二乘学习法之间的关系，还介绍了把支持向量机分类器向鲁棒学习扩展的方法。第9章引入了把多个性能稍弱的分类器组合在一起生成高性能的分类器的集成学习法，介绍了Bagging和Boosting算法。第10章介绍了把各个模式以概率进行分类的Logistic回归的分类算法，以及最小二乘学习版的最小二乘概率分类器。第11章介绍了能够处理字符串那样的序列数据的模式分类的条件随机场。

第Ⅳ部分介绍了各种无监督学习算法。第12章介绍了除去数据中的异常值的方法。第13章介绍了把高维数据降到低维后进行学习的降维算法。第14章介绍了把数据集合化的聚类算法。

第Ⅴ部分介绍了机器学习领域中的新兴算法。第15章介绍了把训练样本逐次输入的逐次学习算法。第16章介绍了在输入输出成对出现的训练样本集的基础上，灵活应用只有输入的训练样本集的半监督学习算法。第17章介绍了有监督的降维算法。第18章介绍了灵活应用其他学习任务的信息，来提高当前学习任务的学习精度的迁移学习法。第19章介绍了在多个学习任务之间实现信息共享，然后同时进行求解的高性能的多任务学习算法。

在第Ⅵ部分的第20章，主要论述了机器学习领域今后的发展。

如图1所示，第Ⅱ部分、第Ⅲ部分和第Ⅳ部分是相对独立的章节。但是第Ⅱ部分的第5章和第6章，以及第Ⅲ部分的第8章、第9章和第11章，包含稍有难度的数学内容，初学者在最开始的时候可以跳过这些内容。

对于书中的大部分算法，本书同时提供了能够进行简单的数值计算的MATLAB程序源代码。各个程序都浓缩在一页的范围内，读者朋友可以轻松录入，以对书中的各种学习算法进行简单的测试。另外，在各个程序行首添加如下代码：

```
rand('state', 0); randn('state', 0);
```

即可完全再现本书中介绍的所有实例的计算结果。

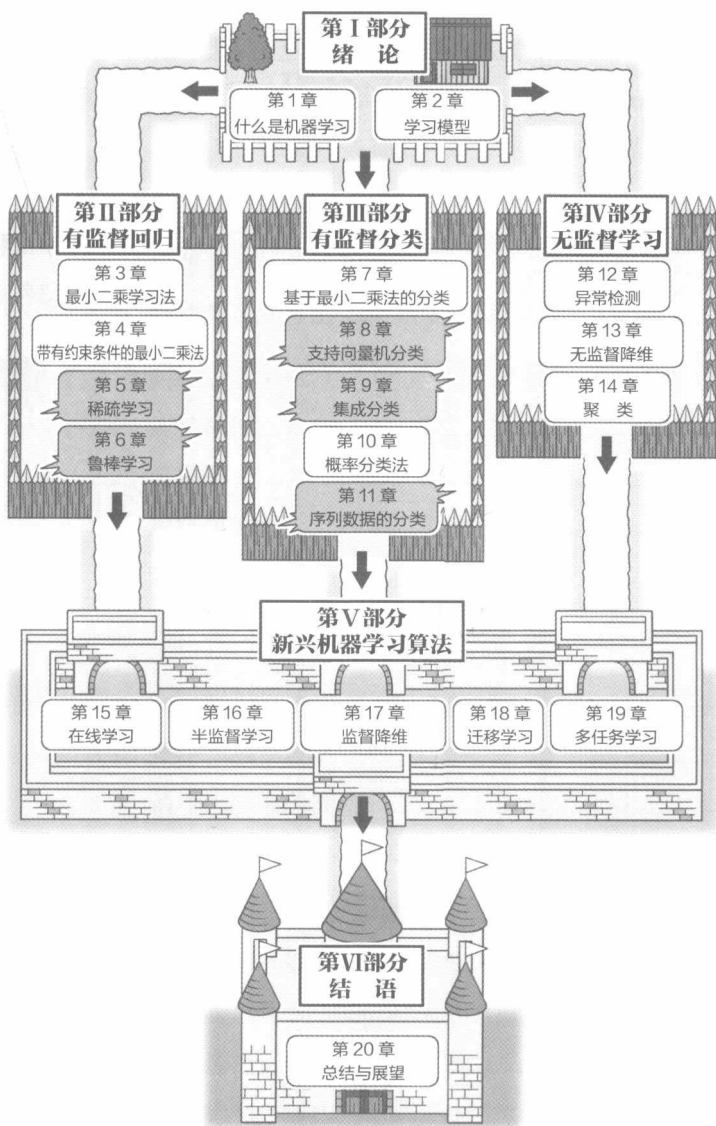


图1 本书的构成

最后，在本书执笔过程中，名古屋大学的金森敬文副教授、名古屋工业大学竹内一郎副教授、NTT Communications科学技术研究所的山田诚博士、东京大学的鹿岛久嗣副教授、东京大学的武田朗子副教授、东京工业大学的山根一航先生、讲谈社的横山真吾先生、绘制插图的Horiguchi Hiroshi先生，给予了笔者巨大的支持和鼓励，在此一并表示真诚的感谢。

杉山将

2013年6月

目 录

第I部分 绪 论

第1章	什么是机器学习	2
1.1	学习的种类	2
1.2	机器学习任务的例子	4
1.3	机器学习的方法	8
第2章	学习模型	12
2.1	线性模型	12
2.2	核模型	15
2.3	层级模型	17

第II部分 有监督回归

第3章	最小二乘学习法	22
3.1	最小二乘学习法	22
3.2	最小二乘解的性质	25
3.3	大规模数据的学习算法	27
第4章	带有约束条件的最小二乘法	31
4.1	部分空间约束的最小二乘学习法	31
4.2	ℓ_2 约束的最小二乘学习法	33
4.3	模型选择	37
第5章	稀疏学习	43
5.1	ℓ_1 约束的最小二乘学习法	43
5.2	ℓ_1 约束的最小二乘学习的求解方法	45
5.3	通过稀疏学习进行特征选择	50

5.4	ℓ_p 约束的最小二乘学习法	51
5.5	$\ell_1 + \ell_2$ 约束的最小二乘学习法	52
第6章	鲁棒学习	55
6.1	ℓ_1 损失最小化学习	56
6.2	Huber 损失最小化学习	58
6.3	图基损失最小化学习	63
6.4	ℓ_1 约束的 Huber 损失最小化学习	65

第 III 部分 有监督分类

第7章	基于最小二乘法的分类	70
7.1	最小二乘分类	70
7.2	0/1 损失和间隔	73
7.3	多类别的情形	76
第8章	支持向量机分类	80
8.1	间隔最大化分类	80
8.2	支持向量机分类器的求解方法	83
8.3	稀疏性	86
8.4	使用核映射的非线性模型	88
8.5	使用 Hinge 损失最小化学习来解释	90
8.6	使用 Ramp 损失的鲁棒学习	93
第9章	集成分类	98
9.1	剪枝分类	98
9.2	Bagging 学习法	101
9.3	Boosting 学习法	105
第10章	概率分类法	112
10.1	Logistic 回归	112
10.2	最小二乘概率分类	116

第11章	序列数据的分类	121
11.1	序列数据的模型化	122
11.2	条件随机场模型的学习	125
11.3	利用条件随机场模型对标签序列进行预测	128

第IV部分 无监督学习

第12章	异常检测	132
12.1	局部异常因子	132
12.2	支持向量机异常检测	135
12.3	基于密度比的异常检测	137
第13章	无监督降维	143
13.1	线性降维的原理	144
13.2	主成分分析	146
13.3	局部保持投影	148
13.4	核函数主成分分析	152
13.5	拉普拉斯特征映射	155
第14章	聚类	158
14.1	K均值聚类	158
14.2	核K均值聚类	160
14.3	谱聚类	161
14.4	调整参数的自动选取	163

第V部分 新兴机器学习算法

第15章	在线学习	170
15.1	被动攻击学习	170
15.2	适应正则化学习	176

第16章	半监督学习	181
16.1	灵活应用输入数据的流形构造	182
16.2	拉普拉斯正则化最小二乘学习的求解方法	183
16.3	拉普拉斯正则化的解释	186
第17章	监督降维	188
17.1	与分类问题相对应的判别分析	188
17.2	充分降维	195
第18章	迁移学习	197
18.1	协变量移位下的迁移学习	197
18.2	类别平衡变化下的迁移学习	204
第19章	多任务学习	212
19.1	使用最小二乘回归的多任务学习	212
19.2	使用最小二乘概率分类器的多任务学习	215
19.3	多次维输出函数的学习	216

第VI部分 结 语

第20章	总结与展望	222
参考文献		225

第

部分

绪论



什么是机器学习

近些年来,得益于互联网的普及,我们可以非常轻松地获取大量文本、音乐、图片、视频等各种各样的数据。机器学习,就是让计算机具有像人一样的学习能力的技术,是从堆积如山的数据(也称为大数据)中寻找出有用知识的数据挖掘技术。通过运用机器学习技术,从视频数据库中寻找出自己喜欢的视频资料,或者根据用户的购买记录向用户推荐其他相关产品等成为了现实(图 1.1)。本章将从宏观角度对什么是机器学习做相应的介绍,并对机器学习的基本概念进行说明。

1.1 学习的种类

计算机的学习,根据所处理的数据种类的不同,可以分为监督学习、无监督学习和强化学习等几种类型。

监督学习,是指有求知欲的学生从老师那里获取知识、信息,老师提供对错指示、告知最终答案的学习过程(图 1.2)。在机器学习里,学生对应于计算机,老师则对应于周围的环境。根据在学习过程中所获得的经验、技能,对没有学习过的问题也可以做出正确解答,使计



图 1.1 机器学习