HZ BOOKS
华章IT

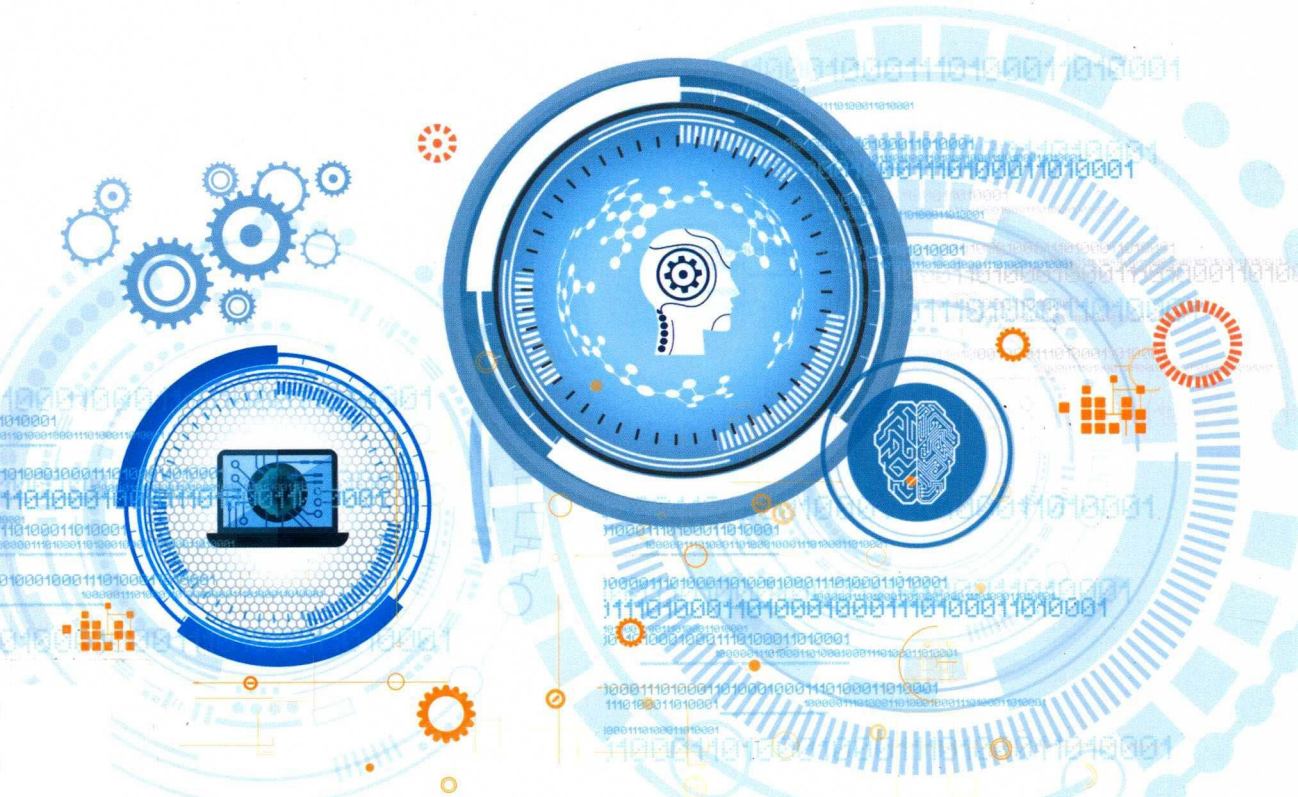
智能系统与技术丛书

把握机器学习和Python语言两个热门领域，用案例驱动方式讲解15个经典的机器学习算法的知识点，以Python语言作为开发语言实现算法的相关步骤和描述。注重理论与实践相结合，在理解机器学习算法原理的同时，能够迅速上手进行实践操作。

Python Machine Learning

Python机器学习

赵涓涓 强彦 主编



机械工业出版社
China Machine Press

Python Machine Learning

Python机器学习

赵涓涓 强彦 主编

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Python 机器学习 / 赵涓涓，强彦主编 . —北京：机械工业出版社，2019.6
（智能系统与技术丛书）

ISBN 978-7-111-63052-4

I. P… II. ①赵… ②强… III. 软件工具 – 程序设计 IV. TP311.561

中国版本图书馆 CIP 数据核字（2019）第 127574 号

本书以案例驱动的方式讲解机器学习算法的知识点，并以 Python 语言作为基础开发语言实现算法，包括目前机器学习主流算法的原理、算法流程图、算法的详细设计步骤、算法实例、算法应用、算法的改进与优化等环节。

全书共分 17 章，前两章介绍机器学习与 Python 语言的相关基础知识，后面各章以案例的方式分别介绍线性回归算法、逻辑回归算法、K 最近邻算法、PCA 降维算法、k–means 算法、支持向量机算法、AdaBoost 算法、决策树算法、高斯混合模型算法、随机森林算法、朴素贝叶斯算法、隐马尔可夫模型算法、BP 神经网络算法、卷积神经网络算法、递归神经网络算法。

本书适合作为高等院校人工智能、大数据、计算机科学、软件工程等相关专业本科生和研究生有关课程的教材，也适用于各种计算机编程、人工智能学习认证体系，还可供广大人工智能领域技术人员参考。

Python 机器学习

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码：100037）

责任编辑：郎亚妹　　责任校对：殷　虹

印　　刷：北京瑞德印刷有限公司　　版　　次：2019 年 7 月第 1 版第 1 次印刷

开　　本：186mm × 240mm　1/16　　印　　张：15

书　　号：ISBN 978-7-111-63052-4　　定　　价：69.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

客服热线：（010）88379426　88361066　　投稿热线：（010）88379604

购书热线：（010）68326294　　读者信箱：hzit@hzbook.com

PREFACE

前　言

2018 年 12 月，DeepMind 设计的基于 Transformer 神经网络和深度学习的人工智能程序 AlphaStar，在《星际争霸 2》游戏中以 5：0 的成绩分别战胜两位职业选手，这是继 AlphaGo 打败世界围棋冠军李世石以来，机器学习领域又一次震惊世界的壮举，为机器学习的发展历程又增添了一抹浓厚的色彩。

机器学习（Machine Learning，ML）是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多学科。它专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，或者重新组织已有的知识结构使之不断改善自身的性能。它是人工智能的核心，是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域。

Python 语言凭借语法简单、优雅、面向对象、可扩展性等优点，一经面世就受到广大开发者的追捧，这使得 Python 语言不仅提供了丰富的数据结构，还具有诸如 NumPy、SciPy、Matplotlib 等丰富的数据科学计算库，为机器学习的开发带来了极大的便利。因此，本书用 Python 语言来编写机器学习算法。

书中对每一种机器学习算法都按照下列几个方面进行总结和描述。第一，简要介绍算法的原理，通过通俗易懂的语言描述和示例使读者对算法有一个大致的了解；第二，给出标准的算法流程图；第三，具体介绍算法的详细设计步骤，使读者对算法的理解更为深入；第四，为了加深读者对算法的熟练程度，针对每个算法举出示例；第五，将每个算法回归到日常生活的应用中，以提高读者对算法的灵活掌握程度；第六，结合当前的最新研究成果，对经典的机器学习算法提出改进与优化建议，为读者进一步研究算法提供新思路；第七，每一章的最后都对全章的内容进行总结，帮读者梳理整章知识；第八，课后习题的设置旨在帮助读者巩固算法的学习。

全书共分 17 章，第 1 和 2 章介绍机器学习与 Python 语言的相关概念与基础知识，第 3 ～ 17 章分别介绍了线性回归算法、逻辑回归算法、K 最近邻算法、PCA 降维算法、k-means 算法、支持向量机算法、AdaBoost 算法、决策树算法、高斯混合模型算法、随机森林算法、朴素贝叶斯算法、隐马尔可夫模型算法、BP 神经网络算法、卷积神经网络算法、递归神经网络算法。

本书由多人合作完成，其中第 1 ～ 4 章由太原理工大学赵涓涓编写，第 5 ～ 7 章由太

原理工大学强彦编写，第 8 和 9 章由太原理工大学王华编写，第 10 和 11 章由太原科技大学蔡星娟编写，第 12 和 13 章由太原理工大学降爱莲编写，第 14 和 15 章由太原理工大学田玉玲编写，第 16 和 17 章由太原理工大学马建芬编写。全书由赵涓涓审阅。

在本书撰写过程中，车征、王磐、王佳文、史国华、魏淳武、周凯、王梦南、王艳飞、吴俊霞、武仪佳、张振庆等项目组成员做了大量的资料准备、文档整理和代码调试工作，在此一并表示衷心的感谢！

由于作者水平有限，不当之处在所难免，恳请读者及同仁赐教指正。

编　者

2019 年 5 月

CONTENTS

目　录

CHAPTER 1

第 1 章

机器学习基础

1.1 引论

在本书开篇之前，读者首先需要明白一个问题：机器学习有什么重要性，以至于需要学习这本书呢?

那么接下来的两张图片，希望可以帮助大家解决这个首要问题。

图 1-1 所展示的三位学者是当今机器学习界的执牛耳者。中间是 Geoffrey Hinton，加拿大多伦多大学教授，如今被聘为“ Google 大脑”的负责人。右边是 Yann LeCun，纽约大学教授，如今是 Facebook 人工智能实验室的主任。而左边这位相信大家都很熟悉，是 Andrew Ng，中文名吴恩达，斯坦福大学副教授，曾于 2014 年加入百度，担任百度公司首席科学家，负责百度研究院的领导工作，尤其是 Baidu Brain 计划。

图 1-1 机器学习界的执牛耳和互联网界大鳄的联姻

这三位都是目前业界炙手可热的大牛，深受互联网界大鳄的欢迎，这足以证明他们的重要性。而他们的研究方向无一例外都是机器学习的子类学科——深度学习。

图 1-2 所描述的是 Windows Phone 上的语音助手 Cortana，它的名字来源于科幻游戏《光环》中士官长的助手。相比其他竞争对手，微软很迟才推出这个服务。Cortana 背后的核心技术是什么？为什么它能够听懂人的语音？事实上，这个技术正是机器学习。机器学

习是所有语音助手产品（包括 Apple 的 Siri 与 Google 的 Now）能够跟人交互的关键技术。

图 1-2 语音助手产品

通过以上两张图片，相信各位读者可以看出机器学习似乎是一项很重要的、有很多未知特性的技术。

1.2 何谓机器学习

1.2.1 概述

机器学习（Machine Learning，ML）是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。它专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。它是人工智能的核心，是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域，它主要使用归纳、综合而不是演绎。

传统上，如果想让计算机工作，我们给它一串指令，然后它遵照这个指令一步步执行下去，有因有果，非常明确。但这样的方式在机器学习中却行不通。机器学习所接受的不是我们输入的指令，而是我们输入的数据。也就是说，机器学习是一种让计算机利用数据而不是指令来执行各种工作的方法。这听起来非常不可思议，但实际上却是非常可行的。

1.2.2 引例

为了加深机器学习在读者心中的印象，我们使用“等人问题”的例子，来进一步介绍机器学习的概念。

1. 提出问题

小 Y 不是一个守时的人，最常见的表现是他经常迟到。小 A 与他相约 3 点钟在某地见

面，在出门的那一刻小 A 突然想到一个问题：我现在出发合适吗？我会不会到了地点后，又要花上 30 分钟去等他？

2. 解决思路

小 A 把以往跟小 Y 相约的经历在脑海中重现一遍，统计了一下跟他约会的次数中，迟到占了多大的比例，并以此为依据来预测小 Y 这次约会迟到的可能性。如果这个值超出了小 A 心里的某个界限，那么小 A 选择等一会再出发。

假设小 A 跟小 Y 相约过 5 次，小 Y 迟到的次数是 1，那么小 Y 按时到的比例为 80%。如果小 A 心中的阈值为 70%，那么就认为这次小 Y 应该不会迟到，因此小 A 按时出门；如果小 Y 迟到过 4 次，也就是他按时到达的比例仅为 20%，由于这个值低于小 A 心中的阈值，则小 A 选择推迟出门的时间。

这个方法从其利用层面来看，又称为经验法。在经验法的思考过程中，我们事实上利用了以往所有约会的数据，因此也可以称之为依据数据所做的判断。依据数据所做的判断跟机器学习的思想根本上是一致的。

3. 建立模型

上述小 A 的思考过程只考虑“频次”这种属性，而一般的机器学习模型至少考虑两个量：一个是因变量，即希望预测的结果，在上述例子中就是小 Y 迟到与否的判断；另一个是自变量，也就是用来预测小 Y 是否迟到的量。

假设将时间作为自变量，譬如根据以往经验发现小 Y 所有迟到的日子基本都是星期五，而在非星期五情况下他基本不迟到。于是可以建立一个模型，来模拟小 Y 迟到跟日期是否是星期五的概率，如图 1-3 所示。

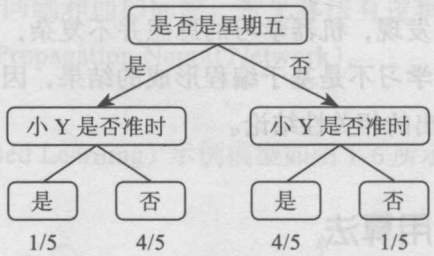


图 1-3　决策树模型

这样的图就是一个最简单的机器学习模型，我们称之为决策树。

当仅仅考虑一个自变量时，情况较为简单。但实际情况较为复杂，还需要考虑一个其他因素，例如小 Y 迟到的部分原因是他开车赶往约定地点时出现的状况（比如，开车比较慢或者路较堵）。考虑到这些信息，就需要建立一个更复杂的模型，这个模型包含两个自变量与一个因变量；而考虑更复杂的情况，小 Y 的迟到跟天气也有一定的关系，例如，下雨

时路比较滑导致路途花费的时间更长，这时需要考虑三个自变量。

如果将所有的自变量和因变量输入计算机，将建模过程交给计算机，由计算机生成一个模型，同时让计算机根据小 A 当前的情况，给出小 A 是否需要晚出门以及需要晚几分钟的建议，那么计算机执行这些辅助决策的过程就是机器学习的过程。

4. 建立对比

通过上述分析，可以看出机器学习与人类思考的经验过程是类似的，并且可以考虑更多的情况，执行更加复杂的计算。事实上，机器学习的一个主要目的就是把人类思考归纳经验的过程转化为计算机通过对数据的处理计算得出模型的过程。经过机器学习，计算机训练出的模型能够以近似于人的方式解决很多灵活而复杂的问题。

将机器学习的过程与人类对历史经验归纳的过程进行比对，如图 1-4 所示。

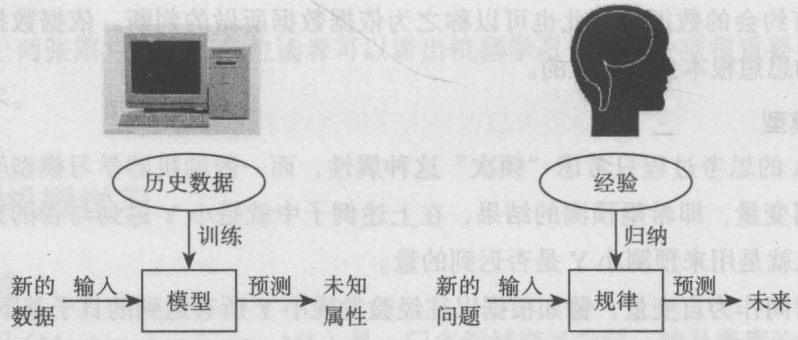


图 1-4　机器学习与人类思考的类比

机器学习中的“训练”与“预测”过程可以对应到人类的“归纳”和“预测”过程。通过这样的对应，我们可以发现，机器学习的思想并不复杂，仅仅是对人类在生活中学习成长的一个模拟。由于机器学习不是基于编程形成的结果，因此它的处理过程不是因果的逻辑，而是通过归纳思想得出的相关性结论。

1.3　机器学习中的常用算法

机器学习有许多算法，这里，我们分别按照两个标准对常用的机器学习算法进行划分：第一个标准是算法的学习方式，第二个标准是算法的相似性。

1.3.1　按照学习方式划分

根据数据类型的不同，对一个问题的建模有多种不同的方式。在机器学习领域，通常将算法按照学习方式进行分类，这样可以在建模和算法选择的时候根据输入数据的类型来

选择最合适的算法以获得最好的结果。按学习方式可分为监督学习、无监督学习、半监督学习和强化学习。

1. 监督学习

监督学习（Supervised Learning）示例模型如图 1-5 所示。

图 1-5　监督学习示例模型

在监督学习中，输入数据称为“训练数据”，每组训练数据有一个明确的标识或结果。如防垃圾邮件系统中“垃圾邮件”“非垃圾邮件”，手写数字识别中的“1”“2”“3”“4”等。在建立预测模型的时候，监督学习建立一个学习过程，将预测结果与“训练数据”的实际结果进行比较，不断地调整预测模型，直到模型的预测结果达到一个预期的准确率。监督学习的常见应用场景有分类问题和回归问题。常见算法有逻辑回归（Logistic Regression）和反向传递神经网络（Back Propagation Neural Network）。

2. 无监督学习

无监督学习（Unsupervised Learning）示例模型如图 1-6 所示。

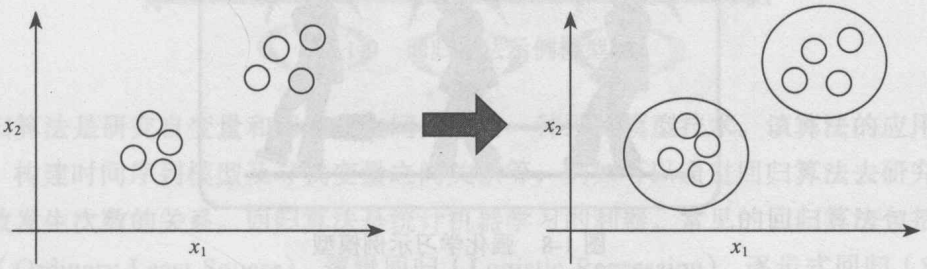


图 1-6　无监督学习示例模型

在无监督学习中，数据并没有被特别标识，学习模型是为了推断出数据的一些内在结构。常见的应用场景包括关联规则的学习以及聚类等。常见算法包括 Apriori 算法和 k-means 算法。

3. 半监督学习

半监督学习（Semi-Supervised Learning）示例模型如图 1-7 所示。

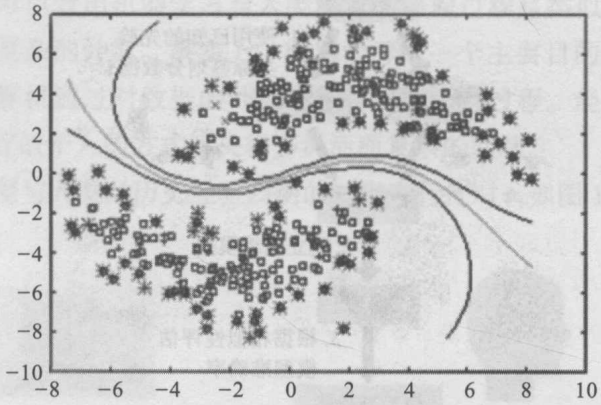


图 1-7　半监督学习示例模型

在半监督学习方式下，输入数据部分被标识，部分没有被标识，这种学习模型可以用来进行预测，但是模型首先需要学习数据的内在结构以便合理地组织数据来进行预测。应用场景包括分类和回归，算法包括一些对常用监督式学习算法的延伸，这些算法首先试图对未标识数据进行建模，在此基础上再对标识的数据进行预测，如图论推理算法（Graph Inference）或者拉普拉斯支持向量机（Laplacian SVM）等。

4. 强化学习

强化学习（Reinforcement Learning）示例模型如图 1-8 所示。

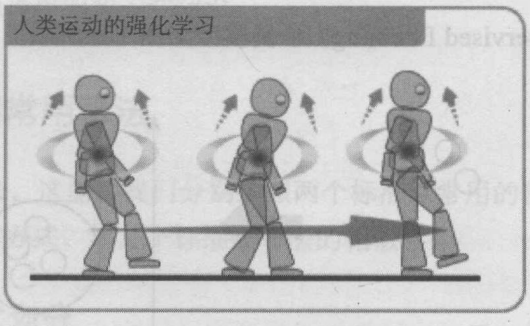


图 1-8　强化学习示例模型

在这种学习模式下，输入数据作为对模型的反馈，不像监督模型那样，输入数据仅

仅是作为一个检查模型对错的方式，在强化学习下，输入数据直接反馈到模型，模型必须对此立刻做出调整。常见的应用场景包括动态系统以及机器人控制等。常见算法包括 Q-Learning 以及时间差学习（Temporal Difference Learning）。

在企业数据应用场景下最常用的是监督学习模型和无监督学习模型；在图像识别等领域，由于存在大量的非标识的数据和少量的可标识数据，故半监督学习是当前一个热门话题；而强化学习更多应用在机器人控制及其他需要进行系统控制的领域。

1.3.2　按照算法相似性划分

根据算法的功能和形式的相似性，我们可以把算法分类，比如分为基于树的算法、基于神经网络的算法等。当然，机器学习的范围非常庞大，有些算法很难明确归类到某一类。而对于有些分类来说，同一分类的算法可以针对不同类型的问题。这里，我们尽可能把常用的算法按照最容易理解的方式进行分类。

1. 回归算法

回归算法示例模型如图 1-9 所示。

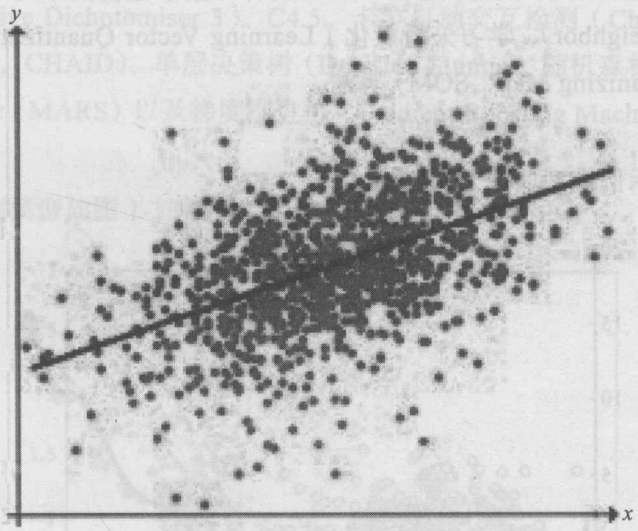


图 1-9　回归算法示例模型

回归算法是研究自变量和因变量之间关系的一种预测模型技术。该算法的应用场景包括预测、构建时间序列模型及寻找变量之间关系等，例如可以通过回归算法去研究超速与交通事故发生次数的关系。回归算法是统计机器学习的利器，常见的回归算法包括：最小二乘法（Ordinary Least Square）、逻辑回归（Logistic Regression）、逐步式回归（Stepwise Regression）、多元自适应回归样条（Multivariate Adaptive Regression Splines）以及本地散

点平滑估计（Locally Estimated Scatterplot Smoothing）。

2. 基于实例的算法

基于实例的算法示例模型如图 1-10 所示。

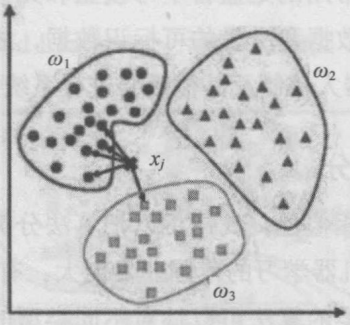


图 1-10 基于实例的算法示例模型

基于实例的算法通常用来对决策问题建立模型，这样的模型常常先选取一批样本数据，然后根据某些相似性把新数据与样本数据进行比较。通过这种方式来寻找最佳匹配。因此，基于实例的算法常常也称为“赢家通吃”学习或者“基于记忆的学习”。常见的算法包括 KNN（K-Nearest Neighbor）、学习矢量量化（Learning Vector Quantization，LVQ），以及自组织映射（Self-Organizing Map，SOM）算法。

3. 正则化方法

正则化方法的示例模型如图 1-11 所示。

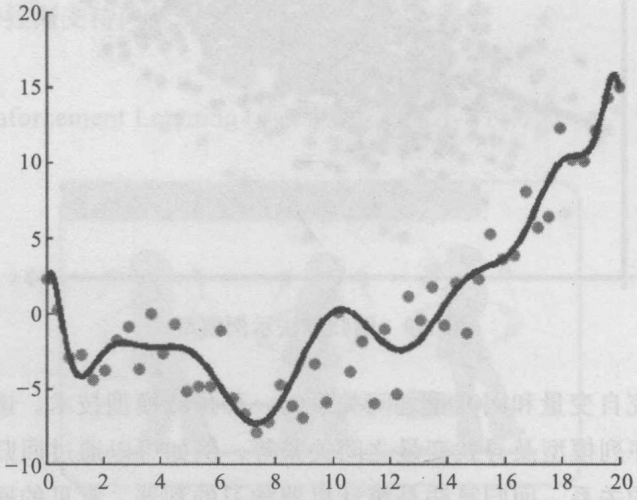


图 1-11 正则化方法示例模型

正则化方法是其他算法（通常是回归算法）的延伸，根据算法的复杂度对算法进行调