扫清数学基础弱和编程实践难的两大障碍

机器学习基础
原理、算法与实践

- 本书精挑细选机器学习的常用算法
- 帮助初学者快速入门，逐步掌握机器学习的基本原理和技能
- 专设QQ群，提供学习辅导和资源下载

袁梅宇 著

清华大学出版社

机器学习基础
原理、算法与实践

袁梅宇 著

清華大學出版社
北京

内 容 简 介

本书讲述机器学习的基本原理，使用 MATLAB 实现涉及的各种机器学习算法。通过理论学习和实践操作，使读者了解并掌握机器学习的原理和技能，拉近理论与实践的距离。全书共分 12 章，主要内容包括：机器学习介绍、线性回归、逻辑回归、贝叶斯分类器、模型评估与选择、K-均值和 EM 算法、决策树、神经网络、HMM、支持向量机、推荐系统、主成分分析。全书源码全部在 MATLAB R2015b 上调试通过，每章都附有习题和习题参考答案，供读者参考。

全书系统讲解了机器学习的原理、算法和应用，内容全面、实例丰富、可操作性强，做到理论与实践相结合。本书适合机器学习爱好者作为入门和提高的技术参考书使用，也适合用作计算机专业高年级本科生和研究生的教材或教学参考书。

图书在版编目(CIP)数据

机器学习基础——原理、算法与实践/袁梅宇著. —北京：清华大学出版社，2018
ISBN 978-7-302-50014-8

Ⅰ. ①机…　Ⅱ. ①袁…　Ⅲ. ①机器学习　Ⅳ. ①TP181

中国版本图书馆 CIP 数据核字(2018)第 081988 号

责任编辑：魏　莹　刘秀青
装帧设计：杨玉兰
责任校对：周剑云
责任印制：丛怀宇
出版发行：清华大学出版社
网　　址：http://www.tup.com.cn, http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969, c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015, zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn, 010-62791865
印 装 者：北京嘉实印刷有限公司
经　　销：全国新华书店
开　　本：185mm×230mm　　印　张：19　　字　数：460 千字
版　　次：2018 年 8 月第 1 版　　印　次：2018 年 8 月第 1 次印刷
定　　价：69.00 元

产品编号：072387-01

前言

机器学习无疑是当今最炙手可热的领域，机器学习工程师、数据科学家和大数据工程师逐渐成为最为热门的新兴职业，各行各业的公司都在寻求具备这些技能的人才。技术职位的爆炸式增长吸引了很多在校大学生、社会 IT 人员将机器学习职位纳入自己的职业规划中。由于具备机器学习相关技能才更有可能在上述新兴职业中获得成功，因此一本容易上手的入门书籍肯定会对初学者有着莫大的帮助，本书就是为初学者精心编写的。

初学者学习机器学习课程一般会面临两大障碍：第一大障碍是数学基础，机器学习要求学习者具备数学基础，书籍中大量的公式是初学者最大的噩梦，尤其是对于已经离开大学走向工作岗位的爱好者，要从头开始去学习和理解数据分布及模型背后的数学原理需要花费很长的时间和精力，学习周期非常漫长；第二大障碍是编程实践，并不是所有人都擅长编代码，学习者只有自己亲手用代码实现机器学习的各种算法，亲眼见到算法解决了实际问题，才能更深入地理解算法。除非想做高精尖的前沿研究，理论研究和公式推导并非是大多数人的专长，如果只是想更合理地应用机器学习来解决实际问题，必需的数学知识就可以降低到大多数人都可以理解的程度，使用 MATLAB 编程实现机器学习算法也比使用 C++或 Java 等语言容易得多。

本书就是为了让初学者顺利入门而设计的。首先，本书只讲述机器学习常用算法的基本原理，并不追求各种算法大而全但简略的罗列，学习并深入理解这些精挑细选的算法后，能够了解基本的机器学习算法，使用适合的算法来解决实际问题。其次，本书使用 MATLAB R2015b 实现了常用的机器学习算法，读者能亲眼看见算法的工作过程和结果，加深对抽象公式和算法的理解，逐步掌握机器学习的原理和技能，拉近理论与实践的距离。再次，每章都附有习题和习题参考答案，其中，一部分习题是为了理解正文内容而设置的，另一部分习题是为了降低正文中的数学要求，将一些必要但枯燥的公式推导放在习题中，供读者有选择性地学习。最后，本书专门设有读者 QQ 群，群号为 278724996，欢迎读者加群，下载书中源代码，与作者直接对话探讨书中技术问题。

本书共分 12 章。第 1 章介绍机器学习的基本概念、MATLAB 的数据格式和示例数据集；第 2 章介绍线性回归，主要内容包括线性回归的模型定义及模型假设和评估、最小二

前言

乘法、梯度下降、多变量线性回归、随机梯度下降、正规方程、多项式回归和正则化；第3章介绍逻辑回归，主要内容包括逻辑回归的假设函数、决策边界、梯度下降算法、MATLAB优化函数、多项式逻辑回归、多元分类、Softmax回归；第4章介绍贝叶斯分类器，主要内容包括判别模型和生成模型的概念、极大似然估计、高斯判别分析、朴素贝叶斯和文本分类；第5章介绍模型评估与选择，主要内容包括训练集验证集测试集划分、交叉验证、性能度量，以及偏差与方差折中；第6章介绍K-均值和EM算法，主要内容包括聚类分析的基本概念、K-means算法应用、EM算法，以及混合高斯模型；第7章介绍决策树，主要内容包括决策树的基本概念、ID3算法、C4.5算法，以及CART算法的原理与实现；第8章介绍神经网络，主要内容包括神经元、神经网络结构、反向传播算法原理与实现；第9章介绍隐马尔科夫模型，主要内容包括HMM的基本概念、HMM的组成和序列生成、求解HMM三个基本问题的算法，以及MATLAB代码实现；第10章介绍支持向量机，主要内容包括支持向量机的基本概念、最大间隔超平面、对偶算法、非线性支持向量机、软间隔支持向量机、SMO算法和LibSVM库的使用；第11章介绍推荐系统，主要内容包括推荐系统的基本概念、基于用户的协同过滤算法、基于物品的协同过滤算法和基于内容的协同过滤算法；第12章介绍主成分分析，主要内容包括主成分分析的基本概念、本征值分解和奇异值分解、PCA算法的计算步骤、如何从压缩表示中重建、如何选取主成分的数量以及PCA实现。

本书的编写异常艰难，从选题到付梓花费约两年时间。和大多数人一样，笔者的脑袋也是单任务处理系统，不善于同时处理多个任务，因此经常迷失在算法、代码、习题、绘图和文字的沼泽中不能自拔，多亏朋友和家人的支持才能坚持到最后。尽管在写作中付出很多艰辛的劳动，限于笔者的学识、能力和精力，书中难免会存在一些缺陷，甚至错误，敬请各位读者批评指正。感谢提供宝贵建议的贡献者，昆明理工大学计算机系吴霖老师经常与笔者讨论机器学习问题，并为本书的内容选取提出了很多建设性建议，感谢吴霖老师的贡献。另外，还要感谢昆明理工大学提供的宽松的研究环境。感谢清华大学出版社的编辑老师在出版方面提出的建设性意见和给予的无私帮助。感谢购买本书的朋友，欢迎批评指正，你们的批评建议都会受到重视，并在将来再版中改进。

袁梅宇
于昆明理工大学

目录

第 1 章　机器学习介绍 1

1.1　机器学习简介 2

1.1.1　什么是机器学习 2

1.1.2　机器学习与日常生活 3

1.1.3　如何学习机器学习 4

1.1.4　MATLAB 优势 5

1.2　基本概念 5

1.2.1　机器学习的种类 6

1.2.2　有监督学习 6

1.2.3　无监督学习 7

1.2.4　机器学习术语 7

1.2.5　预处理 9

1.3　MATLAB 数据格式 10

1.3.1　标称数据 10

1.3.2　序数数据 11

1.3.3　分类数据 11

1.4　示例数据集 12

1.4.1　天气问题 12

1.4.2　鸢尾花 15

1.4.3　其他数据集 16

1.5　了解你的数据 16

习题 20

第 2 章　线性回归 21

2.1　从一个实际例子说起 22

2.1.1　模型定义 23

2.1.2　模型假设 23

2.1.3　模型评估 24

2.2　最小二乘法 24

2.2.1　最小二乘法求解参数 25

2.2.2　用最小二乘法来拟合奥运会数据 26

2.2.3　预测比赛结果 27

2.3　梯度下降 27

2.3.1　基本思路 28

2.3.2　梯度下降算法 29

2.3.3　梯度下降求解线性回归问题 30

2.4　多变量线性回归 32

2.4.1　多变量线性回归问题 33

2.4.2　多变量梯度下降 34

2.4.3　随机梯度下降 38

2.4.4　正规方程 40

2.5　多项式回归 42

2.5.1　多项式回归算法 42

2.5.2　正则化 45

习题 47

第 3 章　逻辑回归 49

3.1　逻辑回归介绍 50

3.1.1　线性回归用于分类 50

3.1.2　假设函数 51

3.1.3　决策边界 52

3.2　逻辑回归算法 53

3.2.1　代价函数 53

3.2.2　梯度下降算法 54

3.2.3　MATLAB 优化函数 56

3.2.4　多项式逻辑回归 58

3.3　多元分类 60

目录

3.3.1 一对多.....60
3.3.2 一对一.....62
3.3.3 Softmax 回归.....64
习题.....66
第 4 章 贝叶斯分类器.....67
4.1 简介.....68
4.1.1 概述.....68
4.1.2 判别模型和生成模型.....68
4.1.3 极大似然估计.....69
4.2 高斯判别分析.....72
4.2.1 多元高斯分布.....72
4.2.2 高斯判别模型.....73
4.3 朴素贝叶斯.....75
4.3.1 朴素贝叶斯算法.....76
4.3.2 文本分类.....81
习题.....86
第 5 章 模型评估与选择.....87
5.1 简介.....88
5.1.1 训练误差与泛化误差.....88
5.1.2 偏差和方差.....89
5.2 评估方法.....90
5.2.1 训练集、验证集和测试集的划分.....91
5.2.2 交叉验证.....92
5.3 性能度量.....95
5.3.1 常用性能度量.....95
5.3.2 查准率和查全率.....96
5.3.3 ROC 和 AUC.....98
5.4 偏差与方差折中.....100
5.4.1 偏差与方差诊断.....101
5.4.2 正则化与偏差方差.....102
5.4.3 学习曲线.....103
习题.....104
第 6 章 K-均值算法和 EM 算法.....107
6.1 聚类分析.....108
6.1.1 K-means 算法描述.....108
6.1.2 K-means 算法应用.....112
6.1.3 注意事项.....113
6.2 EM 算法.....114
6.2.1 基本 EM 算法.....114
6.2.2 EM 算法的一般形式.....115
6.2.3 混合高斯模型.....118
习题.....123
第 7 章 决策树.....125
7.1 决策树介绍.....126
7.2 ID3 算法.....127
7.2.1 信息熵.....127
7.2.2 信息增益计算示例.....127
7.2.3 ID3 算法描述.....132
7.2.4 ID3 算法实现.....134
7.3 C4.5 算法.....134
7.3.1 基本概念.....135
7.3.2 剪枝处理.....139
7.3.3 C4.5 算法描述.....140
7.3.4 C4.5 算法实现.....142
7.4 CART 算法.....144

目录

7.4.1 CART 算法介绍 144
7.4.2 CART 算法描述 147
7.4.3 CART 算法实现 149
习题 150

第 8 章 神经网络 151

8.1 神经网络介绍 152
8.1.1 从一个实例说起 152
8.1.2 神经元 153
8.1.3 神经网络结构 154
8.1.4 简化的神经网络模型 157
8.1.5 细节说明 160
8.2 神经网络学习 161
8.2.1 代价函数 161
8.2.2 BP 算法 162
8.2.3 BP 算法实现 166
8.2.4 平方代价函数的情形 171
习题 171

第 9 章 隐马尔科夫模型 173

9.1 隐马尔科夫模型基本概念 174
9.1.1 离散马尔科夫过程 174
9.1.2 扩展至隐马尔科夫模型 176
9.1.3 HMM 的组成和序列生成 179
9.1.4 三个基本问题 181
9.2 求解 HMM 三个基本问题 182
9.2.1 评估问题 183
9.2.2 解码问题 187
9.2.3 学习问题 190
习题 196

第 10 章 支持向量机 197

10.1 支持向量机介绍 198
10.2 最大间隔超平面 198
10.2.1 SVM 问题的形式化描述 199
10.2.2 函数间隔和几何间隔 199
10.2.3 最优间隔分类器 201
10.2.4 使用优化软件求解 SVM 203
10.3 对偶算法 204
10.3.1 SVM 对偶问题 204
10.3.2 使用优化软件求解对偶 SVM 206
10.4 非线性支持向量机 208
10.4.1 核技巧 208
10.4.2 常用核函数 210
10.5 软间隔支持向量机 213
10.5.1 动机及原问题 213
10.5.2 对偶问题 214
10.5.3 使用优化软件求解软间隔对偶 SVM 215
10.6 SMO 算法 218
10.6.1 SMO 算法描述 218
10.6.2 简化 SMO 算法实现 221
10.7 LibSVM 226
10.7.1 LibSVM 的安装 226
10.7.2 LibSVM 函数 228
10.7.3 LibSVM 实践指南 230
习题 232

第 11 章 推荐系统 233

11.1 推荐系统介绍 234

目录

11.1.1 什么是推荐系统.....234
11.1.2 数据集描述.....235
11.1.3 推荐系统符号.....236
11.2 基于用户的协同过滤.....236
11.2.1 相似性度量.....237
11.2.2 算法描述.....239
11.2.3 算法实现.....240
11.3 基于物品的协同过滤.....241
11.3.1 调整余弦相似度和预测.....241
11.3.2 Slope One 算法描述与实现.....243
11.4 基于内容的协同过滤算法与实现.....247
11.4.1 算法描述.....247
11.4.2 算法实现.....250
习题.....251
第 12 章 主成分分析.....253
12.1 主成分分析介绍.....254
12.2 本征值与奇异值分解.....255
12.2.1 本征值分解.....255
12.2.2 奇异值分解.....256
12.3 PCA 算法描述.....256
12.3.1 PCA 算法.....257
12.3.2 从压缩表示中重建.....258
12.3.3 确定主成分数量.....258
12.4 PCA 实现.....260
12.4.1 假想实例.....260
12.4.2 MNIST 实例.....264
习题.....265
习题参考答案.....267
符号表.....294
参考文献.....295

第1章

机器学习介绍

机器学习试图让机器像人类那样去理解数据，从大量的数据中发现规律和提取知识，不断地完善自我。机器学习是人工智能的一个重要研究方向，研究如何从数据中提取一些潜在的有用模式的算法。

本章首先介绍机器学习的基本概念，然后介绍 MATLAB 的数据格式和示例数据集，最后介绍如何使用各种统计度量来描述数据分布特征。

1.1 机器学习简介

机器学习是人工智能研究领域中极其重要的研究方向，也是发展最快的分支。每过一段时间，我们都能听到一些新的应用在各个领域大展宏图的消息，如谷歌 DeepMind 团队研发的人工智能程序 AlphaGo 战胜世界围棋名将李世石、最强最新 AlphaGo Zero 的横空出世、无人驾驶公交客车正式上路，等等。相比这些新闻，我们也许更关心其背后的支撑技术，机器学习就是 AlphaGo 和无人驾驶等背后的重要技术。

1.1.1 什么是机器学习

机器学习是一门多领域交叉学科，其涉及概率论、统计学、优化理论、算法复杂度理论等多门学科，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并使之不断改善自身的性能。

至今，还没有统一的机器学习定义，而且也很难给出一个公认和准确的定义。一种经常引用的英文定义来自 Tom Mitchell 的《机器学习》一书，原文是：A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its'performance at tasks in T, as measured by P, improves with experience E. 对应的中文译文是：如果用 P 来衡量计算机程序在任务 T 上的性能，根据经验 E 在任务 T 上获得性能改善，那么我们称该程序从经验 E 中学习。

不同于通过编程告诉计算机如何计算来完成特定任务，机器学习是一种数据驱动方法(data-driven approach)，意味着方法的核心是数据。也许读者对此有疑问，让我们举例进行说明。

普通意义上的学习是通过观察获得技能的过程，学习过程如图 1-1 所示。例如，某天大人告诉小孩子前面那只深棕色的小动物是猫，小孩子通过观察认识猫的颜色和形态。另一天大人告诉小孩子前面那只白色的小动物也是猫，小孩子观察到尽管毛色不同，但猫的形态一样，学习到辨识猫的技能是不管毛色，只重形态。因此，下次如果遇到一只黑猫，小孩子也能准确地叫出猫。

图 1-1 普通学习过程

机器学习是通过数据来获取模式的过程，模式可以视为对象的组成成分或影响因素间存在的规律性关系，简单地说，模式相当于事物的规律，机器学习过程如图 1-2 所示。机器学习能够自动识别数据中的模式，然后使用已发现的模式去预测未来的数据，或者在不确定条件下进行某种决策。

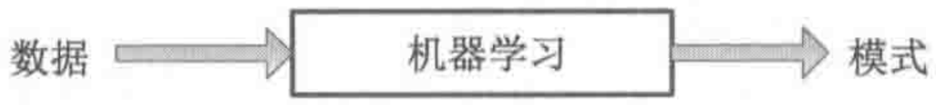

图 1-2　机器学习过程

我们已经知道，使用计算机语言编程能够做很多事情，但是，如果要求编程实现在一堆照片中识别并标记出猫或狗，我们却不知道该怎样做。技术难点在于我们不知道该怎样对猫和狗的照片进行建模，也就是说，一些模式我们无法通过数据直接进行归纳总结。机器学习恰好能解决这类问题，我们将一些标记为猫和狗的照片让某个分类器(如神经网络)进行学习，分类器自动识别照片中猫和狗的模式，经过训练后，分类器分别得到猫和狗的模型，然后使用模型来识别未标记照片中是否有猫或狗。

机器学习的主要内容是研究如何从数据中构建模型的学习算法。有了学习算法之后，将已有数据(称为训练数据集)提供给它，算法就能根据这些数据构建模型，从而使用模型进行预测。因此，机器学习的一个核心内容就是研究学习算法。

1.1.2 机器学习与日常生活

日常生活离不开机器学习。人们或许每天都在不知不觉中使用了机器学习算法，网民经常使用谷歌、必应、百度等搜索引擎来搜索需要的内容，谷歌等公司使用网页排名(PageRank)算法来衡量特定网页的重要程度。网页排名的核心就是机器学习。

人们经常阅读电子邮件，你也许不知道，垃圾邮件过滤器会帮助你过滤掉大量的垃圾邮件。垃圾邮件一度非常猖獗，使用机器学习技术协助识别垃圾邮件并进行过滤后，用户收件箱中的垃圾邮件越来越少。垃圾邮件过滤也是机器学习。

人们在日常生活中经常使用数码相机。你也许不知道，数码相机上的人脸检测技术也是基于机器学习技术。

手机、手写板等手写字符识别使用机器学习；电子商务个性化推荐系统使用的协同过滤算法是机器学习；跳棋、AlphaGo 是机器学习；无人驾驶汽车也是机器学习……在我们的工作与生活中，机器学习的例子不胜枚举，很多智能技术都以机器学习为核心技术。

机器学习的应用领域非常广泛，以下列举一些常见的应用。

- 文本分类，如上面提到的垃圾邮件分类；

- 光学字符识别(optical character recognition，OCR)，如手写识别、车牌识别；
- 计算机视觉，如图像识别、人脸检测；
- 自然语言处理，如词法分析、词性标注、统计句法分析和实体名识别；
- 欺诈检测，如信用卡欺诈、网络入侵检测；
- 推荐系统、搜索引擎、信息提取系统；
- 医疗诊断，如人工智能医生、大数据驱动的个性化诊断；
- 语音识别、语音合成、说话者验证；
- 游戏，如 AlphaGo。

1.1.3 如何学习机器学习

机器学习是一门理论与实践相结合的课程。学习机器学习课程有两种不同的方法：从理论角度出发和从技术角度出发，这两种方法都有其显而易见的优点和缺点。

从理论角度出发是传统的学习机器学习的途径，一般顺序为：掌握必要的数学(微积分、概率论、统计学、线性代数、优化理论等)背景知识，学习机器学习理论，使用编程语言实现算法，使用各种机器学习算法解决实际问题。从理论角度出发的优点是，能够从理论高度抽象出机器学习本质问题的深入理解。缺点是这种方法的学习过程特别长，是为学科前沿的学者设计的，不适合只是想利用机器学习技术的实践者，对一般数学功底较差的爱好者来说学习难度较大。另一个缺点是辛苦学到的理论、公式不够实用，难以直接应用到手边的项目上。

从技术角度出发可以直接学习各类开源软件，如 TensorFlow、Scikit-Learn、Caffe、WEKA、Apache Mahout 等，能够快速上手解决实际问题。缺点是只见树木不见森林，不知道如何从众多可选技术中选择自己所需要的技术，不了解工作原理难以得心应手地使用工具。

本书试图把理论和实践结合在一起。机器学习中有一些基本的哲学思想、关键理论和核心技术，是每一个机器学习爱好者需要了解的。但要完整而系统地学习所有的机器学习知识是不必要的，疯狂英语创始人李阳先生曾说过：“系统全面地学习语法只会系统全面地忘记，而且非常辛苦”，机器学习面临和英语一样的困境：系统地学习还是零敲碎打地学习。对于大部分机器学习爱好者来说，也许零敲碎打地学习更符合他们的实际情况。鉴于此，本书的指导思想是：精心挑选出最常用的经典算法，详细讲解原理和实现方法，兼顾理论和实践，使读者能够快速入门。然后以点带面，逐步形成机器学习的整体概念，为进一步深造打下基础。

机器学习不可避免地要涉及一些数学推导过程，为了避免淹没在公式的海洋中，本书

将必要的公式推导从正文中剔除，放到作业中，供读者有选择性地学习。

1.1.4 MATLAB 优势

本书所有示例都使用 MATLAB 编程环境。它是美国 MathWorks 公司出品的商业数学软件，是用于算法开发、数据可视化、数据分析以及数值计算的高级技术计算语言和交互式环境。MATLAB 将数值分析、矩阵计算、科学数据可视化以及非线性动态系统的建模和仿真等诸多强大功能集成在一个易于使用的窗口环境中，为科学研究、工程设计以及必须进行有效数值计算的众多科学领域提供了一种全面的解决方案。MATLAB 可以进行矩阵运算、绘制函数和数据、实现算法、创建用户界面、连接其他编程语言的程序等，MATLAB 专门为机器学习提供一个工具箱，称为 Statistics and Machine Learning Toolbox(统计和机器学习工具箱)。

本书使用 MATLAB 而不使用 Python、R 等其他计算机程序设计语言是有充分理由的。MATLAB、Python 和 R 都有非常广泛的用户群，使用都很方便，不需要非常高深的编程能力，适合算法开发，且已有大量的成熟包可供使用。MATLAB 作为高级语言，在缩短代码编程量的同时也带来运行效率不高的问题，如果非常看重运行效率，C++和 Java 是更好的选择。尽管 Python 近年发展得很不错，主要原因似乎是 Python 开源且免费，但我们已经见证过很多好的开源项目最终不开源也不免费的情形。选择 MATLAB 的主要原因是它更有用户基础，很多大学都开设相关课程。另一个原因是，相对于其他解释性语言，MATLAB 在矩阵运算上的运行效率很高，还可以通过矢量化来尽量避免使用循环结构，通过编译器将 m 源文件转换为可执行文件来提高运行效率。当然，MATLAB 也有缺点，最大的缺点是它既不免费也不开源，如果对版权方面的因素和开销很介意，不妨使用免费开源的 GNU Octave。

为了更好地讲解算法，本书提供了若干算法的 MATLAB 实现，目的是让读者更好地了解算法。要特别说明的是，本书代码只是演示性质，只是为了让读者能够深入理解算法原理，没有对代码进行优化，也没有考虑让代码能够应用到所有可能的数据集。因此不建议也不鼓励直接将书中代码用于实际项目，除非项目实施者有很强的开发能力，能很好地整合代码。

1.2 基本概念

本节讲述机器学习的基本概念，包括机器学习的种类、有监督学习和无监督学习、常见术语和预处理等。

1.2.1 机器学习的种类

机器学习可分为两种主要类型。第一种机器学习类型称为有监督学习或预测学习，其目标是在给定一系列输入 $\mathbf{x}$ 和输出 y 实例所构成的数据集的条件下，学习输入 $\mathbf{x}$ 到输出 y 的映射关系。这里的数据集称为训练集，实例的个数 N 称为训练样本数。第二种机器学习类型称为无监督学习或描述学习，其目标是在给定一系列仅由输入实例 $\mathbf{x}$ 构成的数据集的条件下，发现数据中的有趣模式。无监督学习有时候也称为知识发现，这类问题并没有明确定义，因为我们不知道需要寻找什么样的模式，也没有明显的误差度量可供使用。而对于给定的 $\mathbf{x}$，有监督学习可以对所观察到的值 y 与预测的值 $\hat{y}$ 进行比较，得到明确的误差值。

1.2.2 有监督学习

有监督学习主要分为分类(classification)与回归(regression)两种形式，是数据挖掘应用领域的重要技术。分类就是在已有数据的基础上学习出一个分类函数或构造出一个分类模型，这就是通常所说的分类器(classifier)。该函数或模型能够把数据集中的样本 $\mathbf{x}$ 映射到某个给定的类别 y，从而用于数据预测。分类和回归是预测的两种形式，分类预测的输出目标是离散值，而回归预测的输出目标是连续值。

在分类之前，要先将数据集划分为训练集和测试集两个部分。分类分为两步：第一步，分析训练集的特点并构建分类模型，常用的分类模型有决策树、贝叶斯分类器、k-最近邻分类等；第二步，使用构建好的分类模型对测试集进行分类，评估分类模型的分类准确度等指标，选择满意的分类模型。

有监督学习的过程如图 1-3 所示。首先使用训练数据对机器学习算法进行训练，得到模型(若干假设)，然后使用构建好的模型对测试数据进行预测，计算预测输出值和真实输出值的误差，从而得到模型的性能评估指标，再反馈给机器学习算法。

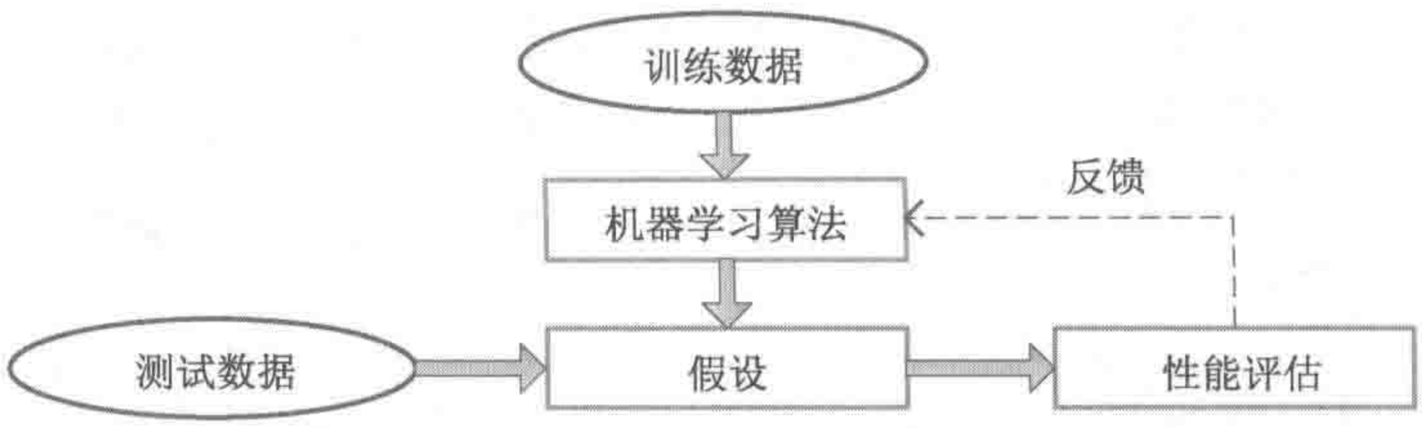

图 1-3　有监督学习

本书涉及的监督学习算法有：线性回归、逻辑回归、高斯判别分析、朴素贝叶斯、决策树、神经网络、支持向量机和协同过滤等。

1.2.3 无监督学习

无监督学习主要分为聚类(clustering)和关联分析(association analysis)。聚类就是将数据集划分为由若干相似实例组成的簇(cluster)的过程，使得同一个簇中实例间的相似度最大化，不同簇中实例间的相似度最小化。也就是说，一个簇就是由彼此相似的一组对象所构成的集合，不同簇中的实例通常不相似或相似度很低。

聚类分析是数据挖掘和机器学习中十分重要的技术，其应用领域极为广泛，如统计学、模式识别、生物学、空间数据库技术、电子商务等。

作为一种重要的数据挖掘技术，聚类主要依据样本间相似性的度量标准将数据集自动划分为几个簇，聚类中的簇不是预先定义的，而是根据实际数据的特征按照数据之间的相似性来定义的。聚类分析算法的输入是一组样本及一个度量样本间相似度的标准，输出是簇的集合。聚类分析的另一个副产品是对每个簇的综合描述，这个结果对于进一步深入分析数据集的特性尤为重要。聚类方法适合用于讨论样本间的相互关联，从而能初步评价其样本结构。

机器学习关心聚类算法的如下特性：处理不同类型属性的能力、对大型数据集的可扩展性、处理高维数据的能力、发现任意形状簇的能力、处理孤立点或“噪声”数据的能力、对数据顺序的不敏感性、对先验知识和用户自定义参数的依赖性、聚类结果的可解释性和实用性、基于约束的聚类等。

关联分析方法用于发现隐藏在大型数据集中有意义的联系，这种联系可以用关联规则(association rule)进行表示。本书不涉及关联分析内容。

1.2.4 机器学习术语

根据应用的不同，机器学习的对象可以是各种各样的数据，这些数据能以诸如数据库、数据仓库、数据文件、流数据、多媒体、网页等形式进行存储。这些数据既可以集中存储，也可以分布在网络服务器上。

数据集通常被视为待处理的数据对象的集合，数据对象用于学习或评估。例如，在垃圾邮件分类问题中，数据对象是指用于学习和测试的电子邮件。由于历史原因，数据对象有多个别名，如记录、点、行、向量、案例、样本、观测等。

数据对象也是对象，因此可用刻画对象基本特征的属性来进行描述。属性也有多个别名，如变量、特征、字段、维、列等。例如，在电子邮件中，相关属性可包括邮件长度、

发送者姓名、在正文出现的某些关键字，等等。

数据集一般类似于一个二维的电子表格或数据库表，每行为一个样本。在最简单的情形下，第 i 个训练样本 $\mathbf{x}^{(i)}$ 是一个 D 维的数值向量，表示特定事物的一些特征，如人的身高、体重等。MATLAB 可以将数据集表示为 $N \times D$ 的矩阵。有时 $\mathbf{x}^{(i)}$ 也可以是复杂结构的对象，如图像、电子邮件、时间序列、语句等。

MATLAB 对上述数据类型提供了良好的支持，具体请参见 1.3 节。

大部分数据集都以数据库表和数据文件的形式存在，MATLAB 支持读取数据库表，也支持读取文本文件、数据文件(MAT 文件)、EXCEL 文件等格式的数据文件。

属性可以分为四种类型：标称(nominal)、序数(ordinal)、区间(interval)和比率(ratio)。其中，标称属性的值仅仅是不同的名称，即标称值仅提供区分对象的足够信息，如性别(男、女)、衣服颜色(红、黄、蓝)、天气(阴、晴、雨、多云)等；序数属性的值可以提供确定对象顺序的足够信息，如成绩等级(优、良、中、及格、不及格)、职称(初级、中级、高级)、学生(本科生、硕士生、博士生)等。序数属性相继值之间的度量值未知，即不关心本科生到硕士生、硕士生到博士生之间的差值是否相同；区间属性的值之间的差值是有意义的，即存在测量单位，如温度、日历日期等；比率属性的值之间的差和比值都是有意义的，如绝对温度、年龄、长度、成绩分数等。

标称属性和序数属性统称为分类的(categorical)或定性的(qualitative)属性，它们的取值为集合，即便使用数值来表示，也不具备数的大部分性质，因此，应该像对待符号一样对待它们；区间属性和比率属性统称为定量的(quantitative)或数值的(numeric)属性，定量属性采用数值来表示，具备数的大部分性质，可以使用整数值或连续实数值来表示。

标签是指样本的目标属性。在分类问题中，每个样本都有一个标称型的类别值，如正常邮件还是垃圾邮件。在回归问题中，标签是连续型的数值。

训练样本是用于训练机器学习算法的样本。在垃圾邮件问题中，每个训练样本由一封电子邮件正文、主题及其标签组成。

验证样本用于调节机器学习算法参数。学习算法通常有多个参数，验证样本为这些模型参数选择适当的值，使得学习算法的性能最佳。

测试样本用于评估机器学习算法的性能。测试样本与训练和验证数据分离，在学习阶段不允许“偷看”测试样本。在垃圾邮件问题中，每个测试样本由一封电子邮件组成，学习算法必须根据电子邮件的特征来预测其标签，然后将预测标签与测试样本的真实标签相比较，以评估算法的性能。

损失函数是能够度量预测标签与真实标签之间差异(或损失)的函数，它是一个非负实值