HZ BOOKS
华章科技

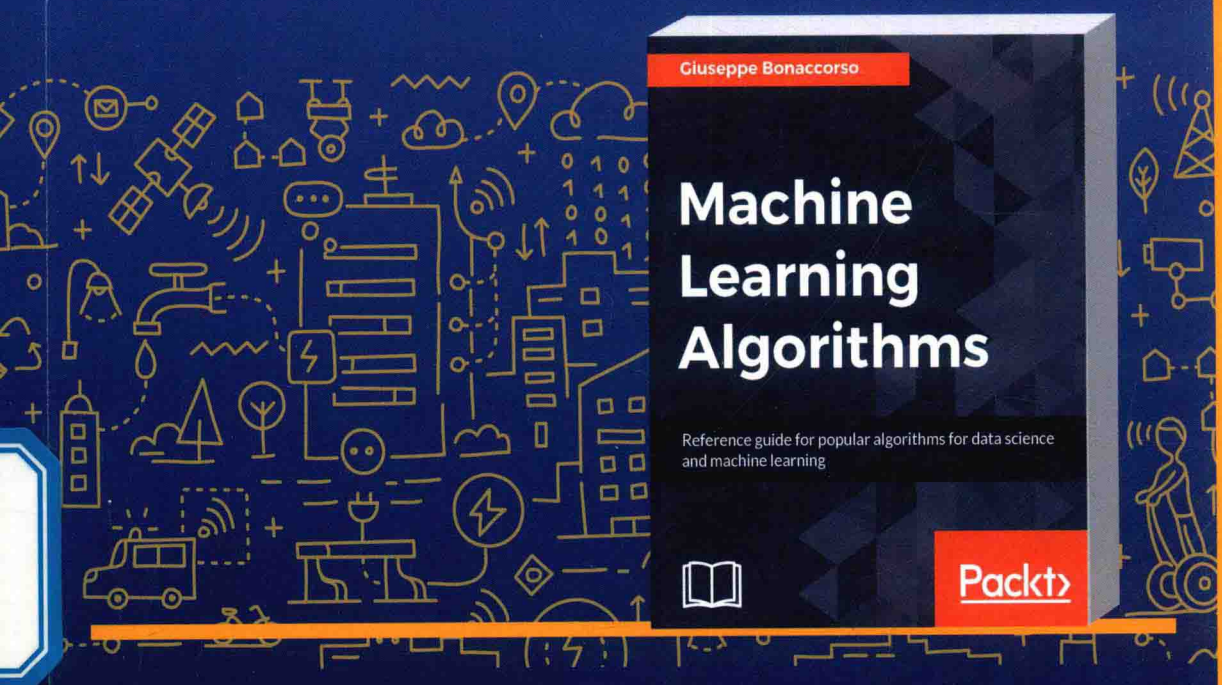
[PACKT]
PUBLISHING

智能科学与技术丛书

Machine Learning Algorithms

机器学习算法

[意] 朱塞佩·博纳科尔索 (Giuseppe Bonaccorso) 著
罗娜 等译



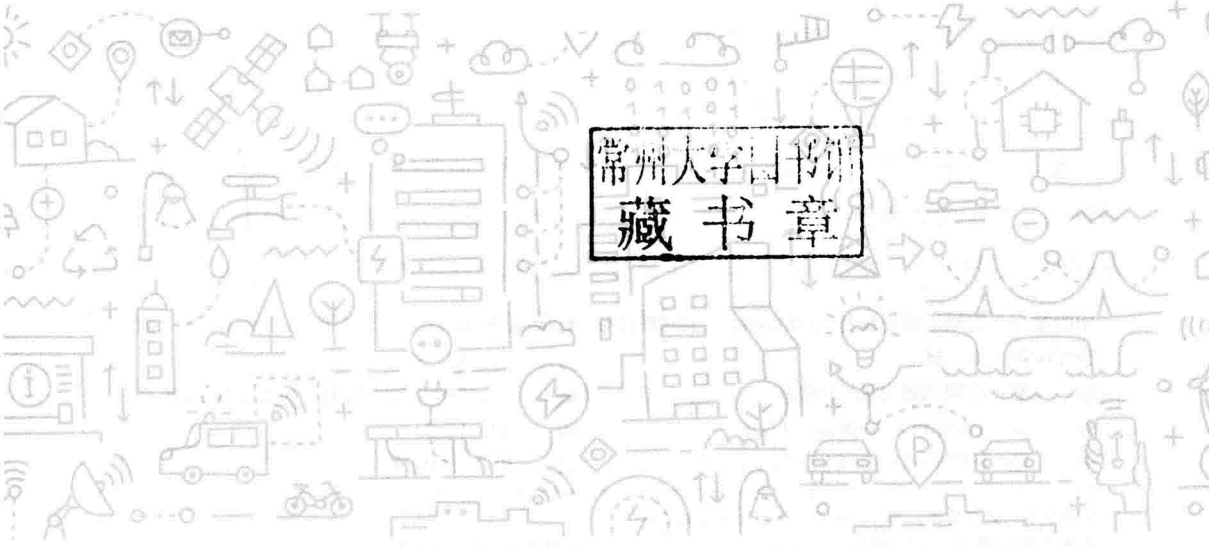
机械工业出版社
China Machine Press

智能科学与技术丛书

Machine Learning Algorithms

机器学习算法

[意] 朱塞佩·博纳科尔索（Giuseppe Bonaccorso）著
罗娜　等译



机械工业出版社
China Machine Press

图书在版编目（CIP）数据

机器学习算法 /（意）朱塞佩·博纳科尔索著；罗娜等译．—北京：机械工业出版社，2018.4
（智能科学与技术丛书）
书名原文：Machine Learning Algorithms

ISBN 978-7-111-59513-7

I. 机… II. ①朱… ②罗… III. 机器学习－算法 IV. TP181

中国版本图书馆 CIP 数据核字（2018）第 059957 号

本书版权登记号：图字 01-2017-7507

本书采用理论与实践相结合的方式，在简明扼要地阐明机器学习原理的基础上，通过大量实例介绍了不同场景下机器学习算法在 scikit-learn 中的实现及应用。书中有大量的代码示例及图例，便于读者理解和学习并实际上手操作。另一方面，书中还有很多的延伸阅读指导，方便读者系统性地了解机器学习领域的现有技术及其发展状况。

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：刘 锋	责任校对：殷 虹
印 刷：中国电影出版社印刷厂	版 次：2018 年 5 月第 1 版第 1 次印刷
开 本：185mm × 260mm 1/16	印 张：15.5
书 号：ISBN 978-7-111-59513-7	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）88378991 88361066
购书热线：（010）68326294 88379649 68995259
投稿热线：（010）88379604
读者信箱：hzjsj@hzbook.com

作为一门多领域交叉学科，人工智能近年来引起了越来越多的关注，也日益广泛地应用于工业及日常生活中。作为未来能真正意义上实现人工智能的方向之一，机器学习涵盖了包括概率、统计、代数、优化等在内的大量基础知识以及数量繁多的算法，形成了庞大的理论知识体系。对机器学习方法进行实现并封装的软件包，使得人们可以在了解基础理论的基础上，快速地将机器学习的现有技术应用于所关注的领域。

作为一种解释型语言，Python 简单易学，开发生态成熟，有很多非常有用的库可以调用，无论是初学者还是专业的程序员都可以利用它方便地编写出所需要的程序。同时，Python 可以方便地调用很多第三方库，从而给程序员带来了很大的便捷性。基于此，Python 拥有庞大的用户群体以及繁荣的社区，这反过来也促进了 Python 的进一步发展。

与机器学习领域很多偏重于理论的图书相比，本书在简明扼要地阐明基本原理的基础上，侧重于介绍如何在 Python 环境下使用机器学习方法库，并通过大量实例清晰形象地展示了不同场景下机器学习方法的应用。从这个角度来说，本书是一本使机器学习算法通过 Python 实现真正“落地”的书，而这无疑将给想要或致力于机器学习应用的读者带来方法理解和实现上的巨大裨益。

感谢华东理工大学信息科学与工程学院的部分研究生，包括金宇尘、何翔、陈鹏、李福杰、余刚、骆楠、戚鹏程等，他们在本书的翻译过程中做了一些辅助性的工作。感谢崔冬等软件开发人员，他们在本书的审阅过程中提出了建设性意见和建议。这里还要特别感谢机械工业出版社张梦玲编辑的大力引荐和无私帮助。

限于本人水平，对本书中部分内容的理解或中文语言的表达难免存在不当之处，敬请读者批评指正，以便能够不断改进。

罗娜

2017 年 12 月 12 日于上海

前　言

Machine Learning Algorithms

本书是对机器学习领域的介绍。机器学习不仅对于 IT 专业人员和分析师，而且对于所有希望利用预测分析、分类、聚类和自然语言处理等技术的科研人员和工程师，都变得越来越重要。当然，本书不可能覆盖所有细节内容，而是只对有些主题进行了简单的描述，给用户更多机会在关注基本概念的基础上通过参考文献深入研究感兴趣的内容。对于本书中可能出现的任何不准确的表达或错误深表歉意，同时感谢所有 Packt 编辑为本书所付出的辛勤劳动。

谨以此书献给我的父母，在他们的信任和鼓励下，我才得以对这个非凡的主题一直保持着巨大的热情。

本书涵盖的内容

第 1 章　对机器学习领域进行简单的介绍，解释了生成智能应用的重要方法的相关基本概念。

第 2 章　解释了关于最常见的机器学习问题的数学概念，包括可学习性的概念和信息论的一些内容。

第 3 章　介绍了数据集预处理、如何选择信息量最大的特征以及进行降维的重要技术。

第 4 章　描述了连续型变量的线性模型，重点介绍了线性回归算法，介绍了 Ridge、Lasso 和 ElasticNet 优化以及其他高级技术。

第 5 章　介绍了线性分类的概念，重点介绍了逻辑回归和随机梯度下降算法，以及几个重要的评估指标。

第 6 章　解释了贝叶斯概率理论，并描述了朴素贝叶斯分类器的结构。

第 7 章　引入了支持向量机算法，着重介绍了线性和非线性分类问题。

第 8 章　解释了层次决策过程的概念，并描述了决策树分类、Bootstrap 和袋装树以及

投票分类器的概念。

第 9 章　介绍了聚类的概念，描述了 k 均值算法和确定聚类最佳数量的多种方法，还介绍了 DBSCAN 和谱聚类等其他聚类算法。

第 10 章　继续第 9 章聚类的内容，介绍了凝聚聚类。

第 11 章　解释了推荐系统中最常用的算法：基于内容和基于用户的策略、协同过滤和交替最小二乘法。

第 12 章　解释了词袋的概念，并介绍了有效处理自然语言数据集所需的最重要技术。

第 13 章　介绍了主题建模的概念，并描述了最重要的算法，如潜在语义分析和潜在狄利克雷分配。同时，还涵盖了情感分析问题，解释了最常用的解决问题的方法。

第 14 章　介绍了深度学习领域的内容，解释了神经网络和计算图的概念，对 TensorFlow 和 Keras 框架的主要概念进行了简要的介绍并列举了几个实例。

第 15 章　介绍了如何定义一个完整的机器学习管道，重点介绍了每一步的特点和缺点。

阅读本书须知

阅读本书不需要特别的数学基础知识。但是，为充分理解所有的算法，需要有线性代数、概率论和微积分的基本知识。

本书中的例子采用 Python 编写，使用了 scikit-learn 机器学习框架、自然语言工具包（NLTK）、Crab、langdetect、Spark、gensim 和 TensorFlow（深度学习框架），环境为 Linux、Mac OS X 或 Windows 平台的 Python 2.7 或 3.3 + 版本。当一个特定的框架被用于特定的任务时，会提供详细的指导和参考内容。

scikit-learn、NLTK 和 TensorFlow 可以按照以下网站提供的说明进行安装：http://scikit-learn.org、http://www.nltk.org 和 https://www.tensorflow.org。

读者对象

本书主要面向希望进入数据科学领域但对机器学习非常陌生的 IT 专业人员，最好熟悉

Python 语言。此外，需要基本的数学知识(线性代数、微积分和概率论)，以充分理解大部分章节的内容。

排版约定

在本书中，你将找到许多区分不同类型信息的文本样式。下面是这些样式的一些例子以及含义：

任何命令行输入或输出如下所示：

```
>>> nn = NearestNeighbors(n_neighbors=10, radius=5.0, metric='hamming')
>>> nn.fit(items)
```

警告或重要内容。

提示和技巧。

示例代码及彩图下载

本书的代码包可以在 GitHub 上找到，网址为 https://github.com/PacktPublishing/Machine-Learning-Algorithms。读者也可以访问华章图书官网 www.hzbook.com，通过注册并登录个人账号，下载本书的源代码和彩图。

Giuseppe Bonaccorso 是一位拥有 12 年经验的机器学习和大数据方面的专家。他拥有意大利卡塔尼亚大学电子工程专业工程学硕士学位，并在意大利罗马第二大学、英国埃塞克斯大学深造过。在他的职业生涯中，担任过公共管理、军事、公用事业、医疗保健、诊断学和广告等多个业务领域的 IT 工程师，使用 Java、Python、Hadoop、Spark、Theano 和 TensorFlow 等多种技术进行过项目开发与管理。他的主要研究兴趣包括人工智能、机器学习、数据科学和精神哲学。

Manuel Amunategui 是 SpringML 公司数据科学项目副总裁。SpringML 是一家初创公司，提供 Google Cloud、TensorFlow 和 Salesforce 企业解决方案。在此之前，他曾在华尔街担任量化开发人员，为一家大型股票期权交易商工作，之后担任微软的软件开发人员。他拥有预测分析和国际管理硕士学位。

他是数据科学爱好者、博主（http://amunategui. github. io），担任 Udemy. com 和 O'Reilly Media 的培训师，以及 Packt 出版社的技术审校人员。

Doug Ortiz 是 ByteCubed 的一名高级大数据架构师，他在整个职业生涯中一直从事企业解决方案方面的架构、开发和集成工作。他帮助企业通过一些现有的和新兴的技术，诸如 Microsoft BI Stack、Hadoop、NoSQL 数据库、SharePoint 以及相关工具和技术，重新发现和利用未充分利用的数据。他也是 Illustris 公司的创始人，可通过 ougortiz@ illustris. org 与他联系。

在专业领域，他有多平台和产品集成、大数据、数据科学、R 和 Python 方面的丰富经验。Doug 还帮助企业深入了解并重视对数据和现有资源的投资，将其转化为有用的信息来源。他利用独特和创新的技术改进、拯救并架构了多个项目。他的爱好是瑜伽和潜水。

Lukasz Tracewski 是一名软件开发人员和科学家，专攻机器学习、数字信号处理和云计算。作为开源社区的积极成员，他也是众多研究类出版物的作者。他曾在荷兰一家高科技产业作为软件科学家工作了 6 年，先后在光刻和电子显微镜方面帮助构建达到生产量与物理精度极限的算法及机器。目前，他在金融行业领导着一支数据科学团队。

4 年来，Lukasz 一直在自然保护领域利用他的专业技能提供无偿服务，如从录音或卫星图像分析中进行鸟类分类等。他在业余时间从事濒危物种的保护工作。

译者序
前言
作者简介
审校人员简介

第1章 机器学习简介 …… 1

1.1 经典机器和自适应机器简介 …… 1
1.2 机器学习的分类 …… 2
1.2.1 监督学习 …… 3
1.2.2 无监督学习 …… 5
1.2.3 强化学习 …… 7
1.3 超越机器学习——深度学习和仿生自适应系统 …… 8
1.4 机器学习和大数据 …… 9
延伸阅读 …… 10
本章小结 …… 10

第2章 机器学习的重要元素 …… 11

2.1 数据格式 …… 11
2.2 可学习性 …… 13
2.2.1 欠拟合和过拟合 …… 15
2.2.2 误差度量 …… 16
2.2.3 PAC 学习 …… 18
2.3 统计学习方法 …… 19
2.3.1 最大后验概率学习 …… 20
2.3.2 最大似然学习 …… 20
2.4 信息论的要素 …… 24
参考文献 …… 26
本章小结 …… 26

第3章 特征选择与特征工程 …… 28

3.1 scikit-learn 练习数据集 …… 28
3.2 创建训练集和测试集 …… 29
3.3 管理分类数据 …… 30
3.4 管理缺失特征 …… 33
3.5 数据缩放和归一化 …… 33
3.6 特征选择和过滤 …… 35
3.7 主成分分析 …… 37
3.7.1 非负矩阵分解 …… 42
3.7.2 稀疏 PCA …… 42
3.7.3 核 PCA …… 43
3.8 原子提取和字典学习 …… 45
参考文献 …… 47
本章小结 …… 47

第4章 线性回归 …… 48

4.1 线性模型 …… 48
4.2 一个二维的例子 …… 48
4.3 基于 scikit-learn 的线性回归和更高维 …… 50
4.4 Ridge、Lasso 和 ElasticNet …… 53
4.5 随机采样一致的鲁棒回归 …… 57
4.6 多项式回归 …… 58
4.7 保序回归 …… 60
参考文献 …… 62
本章小结 …… 62

第5章 逻辑回归 …… 64

5.1 线性分类 …… 64

5.2 逻辑回归 …… 65
5.3 实现和优化 …… 67
5.4 随机梯度下降算法 …… 69
5.5 通过网格搜索找到最优超参数 …… 71
5.6 评估分类的指标 …… 73
5.7 ROC 曲线 …… 77
本章小结 …… 79

第 6 章 朴素贝叶斯 …… 81

6.1 贝叶斯定理 …… 81
6.2 朴素贝叶斯分类器 …… 82
6.3 scikit-learn 中的朴素贝叶斯 …… 83
6.3.1 伯努利朴素贝叶斯 …… 83
6.3.2 多项式朴素贝叶斯 …… 85
6.3.3 高斯朴素贝叶斯 …… 86
参考文献 …… 89
本章小结 …… 89

第 7 章 支持向量机 …… 90

7.1 线性支持向量机 …… 90
7.2 scikit-learn 实现 …… 93
7.2.1 线性分类 …… 94
7.2.2 基于内核的分类 …… 95
7.2.3 非线性例子 …… 97
7.3 受控支持向量机 …… 101
7.4 支持向量回归 …… 103
参考文献 …… 104
本章小结 …… 104

第 8 章 决策树和集成学习 …… 105

8.1 二元决策树 …… 105
8.1.1 二元决策 …… 106
8.1.2 不纯度的衡量 …… 107
8.1.3 特征重要度 …… 109
8.2 基于 scikit-learn 的决策树分类 …… 109
8.3 集成学习 …… 113
8.3.1 随机森林 …… 114
8.3.2 AdaBoost …… 116
8.3.3 梯度树提升 …… 118
8.3.4 投票分类器 …… 120
参考文献 …… 122
本章小结 …… 122

第 9 章 聚类基础 …… 124

9.1 聚类简介 …… 124
9.1.1 k 均值聚类 …… 125
9.1.2 DBSCAN …… 136
9.1.3 光谱聚类 …… 138
9.2 基于实证的评价方法 …… 139
9.2.1 同质性 …… 140
9.2.2 完整性 …… 140
9.2.3 修正兰德指数 …… 141
参考文献 …… 142
本章小结 …… 142

第 10 章 层次聚类 …… 143

10.1 分层策略 …… 143
10.2 凝聚聚类 …… 143
10.2.1 树形图 …… 145
10.2.2 scikit-learn 中的凝聚聚类 …… 147
10.2.3 连接限制 …… 149
参考文献 …… 151
本章小结 …… 152

第 11 章　推荐系统简介 …… *153*

11.1　朴素的基于用户的系统 …… *153*
11.2　基于内容的系统 …… *156*
11.3　无模式(或基于内存的)协同过滤 …… *158*
11.4　基于模型的协同过滤 …… *160*
11.4.1　奇异值分解策略 …… *161*
11.4.2　交替最小二乘法策略 … *163*
11.4.3　用 Apache Spark MLlib 实现交替最小二乘法策略 …… *164*
参考文献 …… *167*
本章小结 …… *167*

第 12 章　自然语言处理简介 …… *169*

12.1　NLTK 和内置语料库 …… *169*
12.2　词袋策略 …… *171*
12.2.1　标记 …… *172*
12.2.2　停止词的删除 …… *174*
12.2.3　词干提取 …… *175*
12.2.4　向量化 …… *176*
12.3　基于路透社语料库的文本分类器例子 …… *180*
参考文献 …… *182*
本章小结 …… *182*

第 13 章　自然语言处理中的主题建模与情感分析 …… *183*

13.1　主题建模 …… *183*
13.1.1　潜在语义分析 …… *183*
13.1.2　概率潜在语义分析 …… *188*
13.1.3　潜在狄利克雷分配 …… *193*
13.2　情感分析 …… *198*
参考文献 …… *202*
本章小结 …… *202*

第 14 章　深度学习和 TensorFlow 简介 …… *203*

14.1　深度学习简介 …… *203*
14.1.1　人工神经网络 …… *203*
14.1.2　深层结构 …… *206*
14.2　TensorFlow 简介 …… *208*
14.2.1　计算梯度 …… *210*
14.2.2　逻辑回归 …… *212*
14.2.3　用多层感知器进行分类 …… *215*
14.2.4　图像卷积 …… *218*
14.3　Keras 内部速览 …… *220*
参考文献 …… *225*
本章小结 …… *225*

第 15 章　构建机器学习框架 …… *226*

15.1　机器学习框架 …… *226*
15.1.1　数据收集 …… *227*
15.1.2　归一化 …… *227*
15.1.3　降维 …… *227*
15.1.4　数据扩充 …… *228*
15.1.5　数据转换 …… *228*
15.1.6　建模、网格搜索和交叉验证 …… *229*
15.1.7　可视化 …… *229*
15.2　用于机器学习框架的 scikit-learn 工具 …… *229*
15.2.1　管道 …… *229*
15.2.2　特征联合 …… *232*
参考文献 …… *233*
本章小结 …… *233*

第1章

Machine Learning Algorithms

机器学习简介

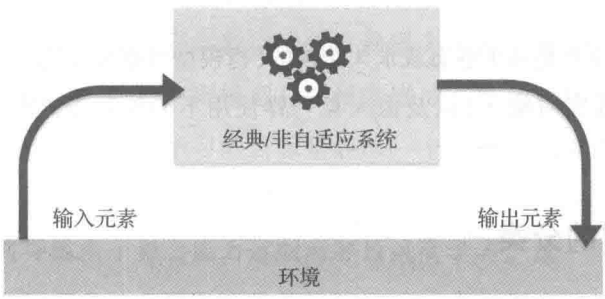
近年来，机器学习已经成为IT和人工智能产业最重要、最多产的分支之一。机器学习相关的应用程序在各个业务部门日益普及，给人们提供了新的更强大的工具和成果。开放源码、产品就绪框架以及每个月发表的数百篇机器学习论文，让机器学习成为IT历史上发展最快的过程之一。但为什么机器学习如此重要和有价值？

1.1 经典机器和自适应机器简介

自古以来，人类通过使用工具和机器来简化工作、减少完成许多不同任务所需的工作量。即使不知道任何物理规律，人们还是发明了杠杆(阿基米德首次正式描述)、仪器和更复杂的机器来执行更长、更复杂的程序。使用简单的技巧可以使锤击钉子变得更容易和更省力，同样，移动重的石头或木头时使用推车可以更加省力。但是，这两个例子有什么区别？可以看出，第二个例子使用了简单的机器，允许单个的人执行移动重物这样的综合任务，而不必考虑其中的步骤。基本的机械定律使得水平力能够有效地对抗重力，即使古人不了解这个原理，仅仅通过观察到的技巧(轮子)就可以改善他们的生活。

因此，一台机器如果不能进行实际应用，就不会有高效或新潮之说。当用户知道用特定的机器可以省力或全自动完成某项任务时，那么这个机器是有用的，而且会不断得到改进。在对机器改进的过程中，开始对齿轮、轮子或轴等部件附加一定的智能，因此出现了(现在我们说编程实现)自动化的机器或程序，并使用能源来实现特定的目标。以风、水磨坊等这种基本工具为例，能够实现以最少的人力控制(与直接活动相比)来完成特定任务。

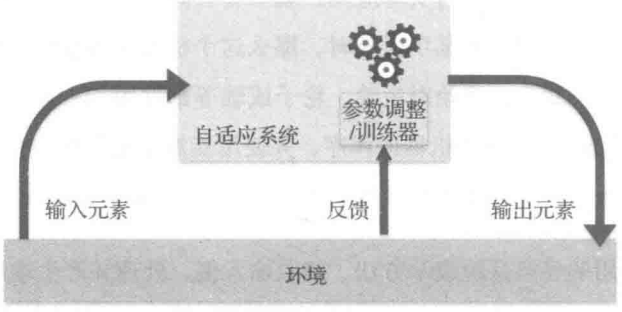
下图是一个通用的经典系统表示方法，接收输入值、处理并产生输出的结果：



但是，水磨坊成功的关键是什么？严格来讲，自从相关技术出现以来，人们一直在试图将智能转移到自己的工具中。河水和风都表现出流动的特性，这种特性具有一定的能量并且是免费提供，但机器需要具有一些意识来使用这些能量。轮子可以绕固定轴转动数百万次，但是需要风推动合适的表面。答案似乎很明显，但对于没有任何知识或经验的人，可能是不自知地，他们也开始采用了一种全新的技术方法。如果对智能这个词定义的话，可以说是从工具开始，先过渡到简单的机器，然后再转到更智能的机器的过程。

进入现代社会后，对于机器，不需要任何中间步骤(但不意味着不重要)，也不需要改变使用的范围。计算机的应用正在变得越来越广泛、灵活和强大，同时网络使得人们能够以最小的代价共享软件应用程序和相关信息。我们现在使用的文字处理软件、电子邮件客户端、网络浏览器以及其他工具都体现了这种灵活性。不可否认的是，IT 革命大大改变了人们的生活和工作，但是如果没有机器学习及其所有的应用程序，计算机将难以完成很多任务。仅仅是垃圾邮件过滤、自然语言处理、通过网络摄像头或智能手机进行视觉跟踪以及预测分析等应用就创造了人机互动并增加了我们的期望。在很多情况下，机器学习将电子工具转变为实际的认知延伸，通过填补人类的感觉、语言、推理、模型和人造工具之间的差距，正在改变人类与许多日常情况的交互方式。

以下是一个自适应系统的示意图：



以上自适应系统不是基于静态或永久结构的(指模型参数和架构)，而是具有能够适应外部信号(数据集或实时输入)以及像人类一样使用不确定和零碎的信息来预测未来的能力。

1.2　机器学习的分类

学习究竟是什么？简单来说，学习是在外部刺激下记住大部分以往的经验，从而能够

实现改变的能力。因此，机器学习是一种工程方法，对于增加或提高自适应变化的各项技术都十分重要。例如，机械手表是一种非凡的工件，但其结构符合静止定律，当外部发生变化时会变得没有任何用处。学习能力是动物——特别是人特有的，根据达尔文的理论，它也是所有物种生存和进化的关键要素。机器虽然不能自主进化，但似乎也遵循同样的规律。

因此，机器学习的主要目标是学习、策划和改进数学模型，该数学模型可以使用由环境提供的相关数据通过一次或连续多次的训练得到，利用该数学模型推断未来并做出决定而不需要所有影响因素(外部因素)的全部知识。换句话说，智能体(从环境中接收信息的软件实体，选择达到特定目标的最佳行动并观察其结果)采用统计学习方法，通过确定正确的概率分布，来预测最有可能成功(具有最小错误)的动作(值或决策)。

我更喜欢使用术语“推断”而不是“预测”，只是为了避免把机器学习看成是一种现代魔法(这种看法并不罕见)。此外，可以引入一个基本的声明：一个算法只有在影响实际数据时，才能推断出一般的规律，并以相对较高的精度来学习算法的结构。虽然术语“预测”可以自由使用，但其具有与物理学或系统理论相同的含义。在复杂场景下，例如使用卷积神经网络的图像分类问题，即使信息(几何、颜色、特征、对比度等)已经存在于数据中，模型也必须足够灵活以便提炼和永久学习。

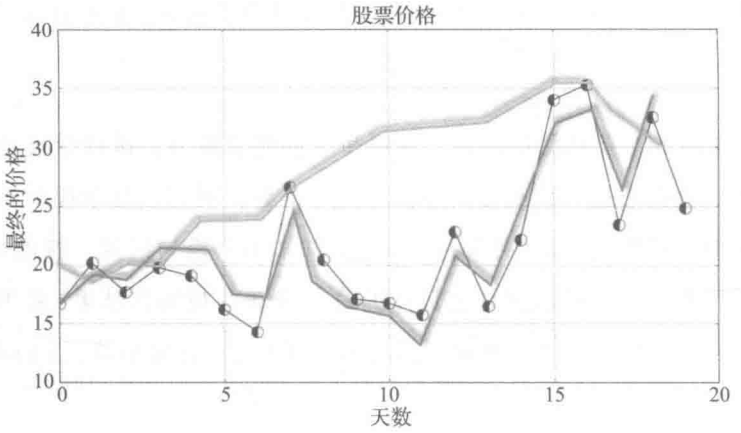
在接下来的章节中，将会简要介绍常见的机器学习方法，而数学模型、算法和实例将在后面的章节中讨论。

1.2.1 监督学习

监督学习中有教师或监督者的概念，其主要功能是提供误差的精确度量(直接与输出值相比)。在实际算法中，该功能由多组对应值(输入和期望输出)组成的训练集提供。基于训练集，可以修正模型参数以减少全局损失函数。在每次迭代之后，如果算法足够灵活并且数据是一致的，则模型总体精度增加，并且预测值和期望值之间的差距变得接近于零。当然，监督学习的目标是训练一个系统，使得该系统能够预测以前从未见过的样本。因此，有必要让模型具有泛化能力，以避免一个常见的称为过拟合的问题。过拟合将导致由于拟合能力过剩而导致过度学习。这一问题将在第 2 章中详细讨论，可以说，过拟合的主要影响是虽然能够正确预测用于训练的样本，但其他

样本的误差却很大。

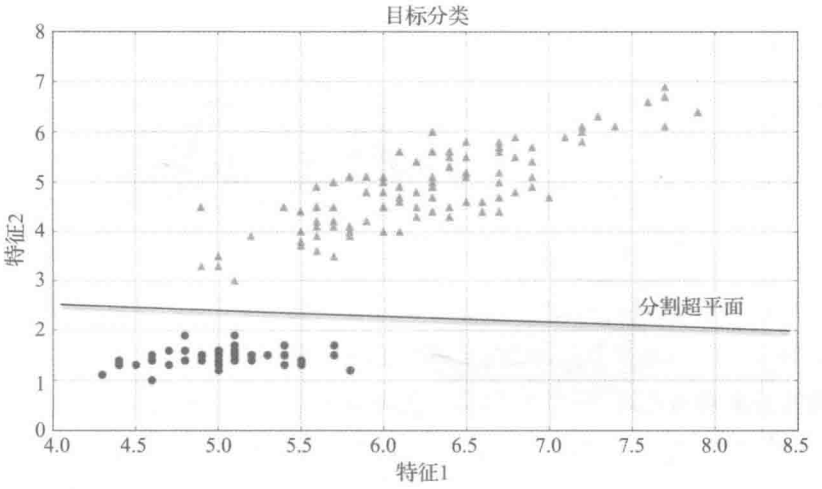
下图中，圆圈表示训练样本点，细蓝线表示完美的泛化(图中连接成简单的线段)：



使用相同的数据集训练两个不同的模型，分别对应于红线和绿线。红线所表示的模型不能被接受，因为它的泛化能力差且预测精度低，而绿线代表的模型在趋势和预测分析中，对残差和泛化性能有非常好的折中。

一般来讲，基于连续的输出值时前面所示的例子称为回归。相反，如果只有一个离散量表示的结果(称为类别)，则该过程被称为分类。有时，除了预测实际的类别，最好是确定其概率分布。例如，可以训练一种算法来识别手写字母，其输出(英文中的26个字母)是分类。另一方面，即使对于人类来说，当手写不够清楚无法确定是哪一个字母时，会产生多于一个可能的结果。这意味着通过离散概率分布可以更好地描述实际输出(例如，使26个字母表示的连续值归一化，使得它们的总和为1)。

下图是一个具有两个特征的数据集的分类实例，该实例是一个线性问题。大多数算法尝试通过施加不同的条件来找到最佳的分割超平面。在分类过程中，目标是相同的，即减少错误分类的数量并增加对噪声的鲁棒性。例如，对于接近分割超平面(此处为直线)的三角形的点(其坐标约为[5.1-3.0])，当第二个特征受到噪声影响的坐标值远小于3.0时，略高一点的超平面就可能会错误地将该点分类。我们将会在后面的章节中讨论一些强大的技术来解决这类问题。



常见的监督学习的应用包括：

- 基于回归的预测分析或分类
- 垃圾邮件检测
- 模式检测
- 自然语言处理
- 情绪分析
- 自动图像分类
- 自动序列处理(例如音乐或语音)

1.2.2 无监督学习

无监督学习方法没有任何监督，只能基于对绝对误差的衡量。当需要对一组数据根据其相似度(或距离)进行分组(聚类)时，需要采用无监督学习方法。例如，前面的分类图中，人们不需要考虑颜色或形状就可以立即识别出两个类。事实上，圆点(以及三角形)确定了一个集合，不管集合内的点之间的分离程度如何，圆点所代表的集合很容易与三角形代表的集合分离开来。这就像当理想的样本是海洋时，仅仅考虑岛屿之间的相互位置和内部联系就可以将海洋分成几个区域。

下图中，每个椭圆表示一个聚类，聚类中的点用相同的记号标记，类之间的边界点(例如，与圆形区域重叠的三角形)通过特定标准(通常是权衡距离度量)来确定所属的类别。对于模糊的分类(如 P 和残缺的 R)，好的聚类方法应该考虑异常值的存在并对它们进行处理，以增加内部一致性(意味着选择使局部密度最大化的分类)和聚类之间的距离。