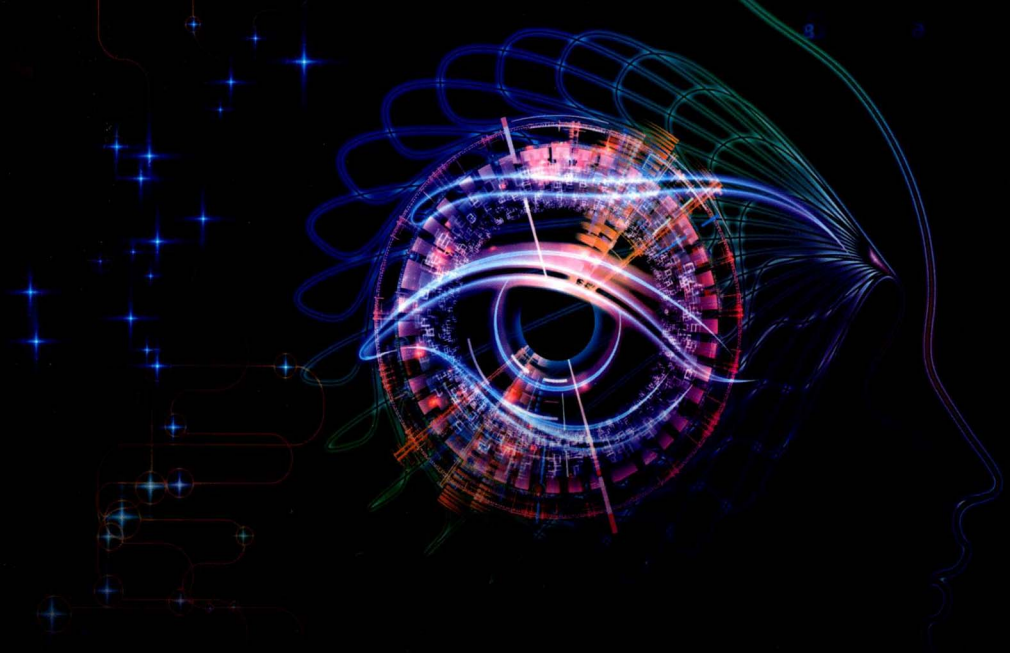

黄永昌◎编著

scikit-learn 机器学习

常用算法原理及编程实战

Machine Learning by scikit-learn: Algorithms and Practices

拨开笼罩在机器学习上方复杂的数学“乌云”，让读者以较低的门槛入门机器学习

涵盖机器学习的应用场景、编程步骤、开发包、算法模型性能评估、8个常用算法、7个实战案例

孙言东	刘凡	陈源	戴剑	.....	共同推荐
阿里云栖社区技术专家	百度高级研发工程师	蒙牛乳业数据分析总监	神州数码云计算公司技术总监		

机械工业出版社
China Machine Press

scikit-learn 机器学习

常用算法原理及编程实战

Machine Learning by scikit-learn: Algorithms and Practices

黄永昌◎编著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

scikit-learn机器学习：常用算法原理及编程实战/黄永昌编著. —北京：机械工业出版社，2018.1

ISBN 978-7-111-59024-8

Ⅰ. s… Ⅱ. 黄… Ⅲ. 机器学习 Ⅳ. TP181

中国版本图书馆CIP数据核字（2018）第016718号

内 容 简 介

本书通过通俗易懂的语言、丰富的图示和生动的实例，拨开了笼罩在机器学习上方复杂的数学“乌云”，让读者以较低的代价和门槛轻松入门机器学习。

本书共分为11章，介绍了在Python环境下学习scikit-learn机器学习框架的相关知识，涵盖的主要内容有机器学习概述、Python机器学习软件包、机器学习理论基础、k-近邻算法、线性回归算法、逻辑回归算法、决策树、支持向量机、朴素贝叶斯算法、PCA 算法和k-均值算法等。

本书适合有一定编程基础的读者阅读，尤其适合想从事机器学习、人工智能、深度学习及机器人相关技术的程序员和爱好者阅读。另外，相关院校和培训机构也可以将本书作为教材使用。

scikit-learn 机器学习

常用算法原理及编程实战

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：欧振旭 李华君 责任校对：姚志娟

印 刷：中国电影出版社印刷厂 版 次：2018 年 3 月第 1 版第 1 次印刷

开 本：186mm × 240mm 1/16 印 张：13.75

书 号：ISBN 978-7-111-59024-8 定 价：59.00 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 88379649 68995259 读者信箱：hzit@hzbook.com

前言

机器学习是近年来非常热门的方向，然而普通的程序员想要转行机器学习却困难重重。回想起来，笔者在刚开始学习机器学习时，一上来就被一大堆数学公式和推导过程所折磨，这样的日子至今还历历在目。当时笔者也觉得机器学习是个门槛非常高的学科。但实际上，在机器学习的从业人员里，究竟有多少人需要从头去实现一个算法？又有多少人有机会去发明一个新算法？从一开始就被细节和难点缠住，这严重打击了想进入机器学习领域新人的热情和信心。

本书就是要解决这个问题。笔者希望尽量通过通俗的语言去描述算法的工作原理，并使用 scikit-learn 工具包演示算法的使用，以及算法所能解决的问题，给那些非科班出身而想半路“杀进”人工智能领域的程序员，以及对机器学习感兴趣的人提供一本入门的书籍。

当然，这里不是否认数学和算法实现的重要性，毕竟它们是人工智能领域的基础学科方向。万事开头难，只有打开了一扇门，才能发现一个新的五彩缤纷的世界。在这个世界里，我们可以吃到新口味的面包，也能认识那些做面包给别人吃的人。希望这本书能帮助读者打开机器学习的这扇门。

本书特色

1. 用通俗易懂的语言介绍机器学习算法的原理，符合初学者的认知规律

本书讲解时首先会用通俗易懂的语言介绍常用的机器学习算法，帮助读者直观地理解每个算法的基本原理，然后用大量的图示及实例介绍如何使用 scikit-learn 工具包解决现实生活中的机器学习问题。这种由浅入深、循序渐进的讲授方式，完全遵循了初学者对机器学习算法的认知规律。

2. 丰富的示例图片，可以帮助读者更加直观地理解算法背后的原理

机器学习以其背后复杂的数学原理及异常复杂的算法推导和证明过程而吓退了一大批读者。一图胜千言，本书给出了大量的图示，用图片的方式形象地介绍了算法的基本原理，让读者对算法有更加直观的理解。这样就把复杂的数学公式和冗长的文字描述浓缩到一张张图片中，有效地降低了学习的门槛。

3. 实例丰富，可以帮助读者使用机器学习算法解决工程应用问题

手写识别程序怎么做？怎么实现人脸识别系统？怎么过滤垃圾邮件？电子商务网站上猜你喜欢的商品是什么原理？怎么实现的？电影网站怎样去推荐符合用户喜好的电影？怎么利用机器学习对消费者的特性进行细分，从而更好地服务好各细分市场的消费者？银行怎样去检测用户的信用卡可能被盗了？通过阅读本书，读者将了解到这些复杂问题背后的原理，甚至你都可以自己解决这些问题。

本书内容介绍

第 1 章机器学习介绍，涵盖了机器学习的定义、应用场景及机器学习的分类，并通过一个简单的示例，让读者了解机器学习的典型步骤和机器学习领域的一些专业术语。

第 2 章 Python 机器学习软件包，介绍了 scikit-learn 开发环境的搭建步骤，以及 IPython、Numpy、Pandas 和 Matplotlib 等软件包的基础知识，并通过一个 scikit-learn 机器学习实例介绍了 scikit-learn 的一般性原理和通用规则。

第 3 章机器学习理论基础，介绍了算法模型性能评估的指标和评估方法等理论基础。本章内容是本书最关键的理论基础知识，对理解本书其他章节的内容非常重要。

第 4 章 k-近邻算法，介绍了一个有监督的机器学习算法，即 k-近邻算法。该算法可以解决分类问题，也可以解决回归问题。

第 5 章线性回归算法，介绍了单变量线性回归算法和多变量线性回归算法的原理，以及通过梯度下降算法迭代求解线性回归模型，并给出一个房价预测的实例。另外，本章对成本函数和使用线性回归算法对数据进行拟合也做了讲解。

第 6 章逻辑回归算法，介绍了逻辑回归算法的原理及成本函数。在本章中主要解决的问题有：逻辑回归算法的原理是什么？怎样使用梯度下降算法解决迭代求解逻辑回归算法的模型参数？什么是正则化？正则化能解决什么问题？L1 范数和 L2 范数作为模型正则项有什么区别？如何使用逻辑回归算法解决乳腺癌检测问题？

第 7 章决策树，主要介绍了决策树的算法原理和算法参数，并给出了一个预测实例，最后对集合算法做了必要讲解。

第 8 章支持向量机，主要介绍了支持向量机的基本算法原理及常用核函数，并给出了用支持向量机来解决乳腺癌检测问题的实例。

第 9 章朴素贝叶斯算法，首先从贝叶斯定理谈起，引入了朴素贝叶斯分类法；然后通过一个简单的例子说明了算法的基本原理；接着介绍了概率分布的概念及几种典型的概率分布；最后通过一个文档分类实例来说明朴素贝叶斯算法的应用。

第 10 章 PCA 算法，首先介绍了 PCA 的算法原理；然后通过一个简单的模拟运算过程帮助读者理解该算法的原理和实现步骤；最后介绍了 PCA 算法背后的物理含义。本章

在讲解的过程中顺便给读者推荐了一些优秀的线性代数资源，供读者参考。

第 11 章 k-均值算法，首先介绍了该算法的基本原理及关键迭代步骤；然后通过一个简单的例子，介绍了如何使用 scikit-learn 中的 k-均值算法解决聚类问题；最后使用一个文本聚类分析的例子介绍了 k-均值算法的应用，并介绍了典型的无监督机器学习算法的性能评估指标。

如何更好地使用本书

如果你只是好奇机器学习背后的原理，大可只阅读书中的文字部分，而跳过代码实现环节；如果你是想用本书敲开机器学习这扇大门，并且未来想从事这一行业，那么建议你系统地阅读本书，而且要亲自动手完成书中的所有实例。本书提供了书中所有实例的完整源代码，建议你认真阅读这些源代码，并亲自动手运行这些代码，还可以调整参数，看看结果有什么变化，最后再独立把这些实例实现一遍。

阅读本书需要的知识储备

阅读本书，建议你最好学习过 Python 语言，即便是两年前学的，学过后就算没怎么用也没有关系。如果你不熟悉 Python 语言，那么最好有其他编程语言基础，如 C++或 Java 语言等。

本书读者对象

1．有一定编程经验，而不满足于永远在“搬砖”的软件工程师

你是不是厌倦了每天重复“搬砖”的过程？你是不是想提高职业的溢价？本书或许可以帮助你打开一扇大门。人工智能在可以预见的未来有巨大的发展前景。特别是近几年，层出不穷的开源机器学习框架不断涌现出来，云计算和分布式计算能力的进一步提升，为人工智能应用于更广泛的领域提供了必要的基础。在可以预见的未来，人工智能领域对机器学习工程师的需求将急剧上升。如果本书能帮助你打开机器学习领域的这扇大门，让你能利用机器学习的知识解决实际问题，这将是笔者最大的荣耀。

2．对这个世界充满好奇的人

笔者之前在某电商网站上搜索了某款手机，之后上网时有大量的网站广告都在向笔者展示手机及其相关产品。这些网站是怎么知道笔者近期想买手机的？笔者常去的电影网站每次都能给笔者推荐一些符合笔者“口味”的电影。这是如何做到的？本书便可以让你以很低的门槛了解这些问题背后的原理，甚至你也可以自己动手做一个，玩一玩。

本书虽然有大量的程序示例代码，但是笔者通过通俗易懂的讲述，并配以大量的图示，让这本书的阅读门槛很低，甚至可以作为本科普读物去阅读。可以说，这本书几乎适合所有对这个世界充满好奇的人阅读，尤其是那些对人工智能充满好奇的人，以及对机器学习算法感兴趣的人。

本书源代码获取方式

本书涉及的源代码文件需要读者自行下载。请读者登录机械工业出版社华章公司的网站 www.hzbook.com，然后搜索到本书页面，找到下载模块下载即可。

本书作者

本书由黄永昌组织编写，其他参与编写的人员还有张昆、张友、赵桂芹、张金霞、张增强、刘桂珍、陈冠军、魏春、张燕、孟春燕、项宇峰、李杨坡、张增胜、张宇微、张淑凤、伍云辉、孟庆宇、马娟娟、李卫红、韩布伟、宋娟、郑捷、方加青、曾桃园、曾利萍、谈康太、李秀、董建霞、方亚平、李文强、张梁、邓玉前、刘丽、舒玲莉、孙敖、王善芬、杨淑芬、刘玉平、孙家宝。

因作者水平和成书时间所限，本书难免存有疏漏和不当之处，敬请各位读者指正。读者在阅读本书时若有疑问，可以发电子邮件到 hzbook2017@163.com，以获得帮助。

目录

第 1 章　机器学习介绍

本章简要介绍了机器学习的定义、应用场景及机器学习的分类，并通过一个简单的示例介绍了机器学习的典型步骤，以及机器学习领域的一些专业术语。本章涵盖的内容如下：

- 机器学习的概念；
- 机器学习要解决的问题分类；
- 使用机器学习解决问题的一般性步骤。

1.1　什么是机器学习

机器学习是近年来的一大热门话题，然而其历史要倒推到半个多世纪之前。1959 年 Arthur Samuel 给机器学习的定义是：

Field of study that gives computers the ability to learn without being explicitly programmed

即让计算机在没有被显式编程的情况下，具备自我学习的能力。

Tom M. Mitchell 在操作层面给出了更直观的定义：

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.

翻译过来用大白话来说就是：针对某件事情，计算机会从经验中学习，并且越做越好。从机器学习领域的先驱和“大牛”们的定义来看，我们可以自己总结出对机器学习的理解：**机器学习是一个计算机程序，针对某个特定的任务，从经验中学习，并且越做越好。**

从这个理解上，我们可以得出以下针对机器学习最重要的内容。

数据：经验最终要转换为计算机能理解的数据，这样计算机才能从经验中学习。谁掌握的数据量大、质量高，谁就占据了机器学习和人工智能领域最有利的资本。用人类来类比，数据就像我们的教育环境，一个人要变得聪明，一个很重要的方面是能享受到优质的教育。所以，从这个意义来讲，就能理解类似 Google 这种互联网公司开发出来的机器学习程序性能为什么那么好了，因为他们能获取到海量的数据。

模型：即算法，是本书要介绍的主要内容。有了数据之后，可以设计一个模型，让数据作为输入来训练这个模型。经过训练的模型，最终就成了机器学习的核心，使得模型成为了能产生决策的中枢。一个经过良好训练的模型，当输入一个新事件时，会做出适当的

反应，产生优质的输出。

1.2 机器学习有什么用

受益于摩尔定律，随着计算机性能的提高，以及计算资源变得越来越便宜，机器学习在诞生半个世纪后的今天，得到了越来越广泛的应用。你可能感受不到，但是你的日常生活已经与人工智能密不可分了。

早晨起床，用 iPhone 打开 Siri，问：“今天天气怎么样？”。Siri 会自动定位到当前你所在的城市，并且把天气信息展现出来。这个功能用起来很简单，但其背后的系统是异常复杂的。

其一是语音识别，这是机器学习最早的应用研究领域，Siri 需要先把你说的话转换为文字。大家知道，语音从本质上是一系列幅度不同的波，要转换为文字，就需要设计一个模型，先通过大量的语音输入来训练这个模型，等模型训练好了，把语音作为输入，就可以输出文字了。语音识别在 20 世纪 50 年代就开始研究了，其模型是不断演变的。一个比较大的演变，就是由基于模式识别的算法演变为基于统计模型的算法，这一转变大大提高了语音识别的准确率。

其二是自然语言处理，这是机器学习和人工智能又一个非常重要的研究方向。Siri 把语音转成文字后，软件需要理解文字的意思才能给出准确的回答。要让计算机理解文字可不是简单的事情。首先要有大规模的语料库，其次要有相应的语言模型，然后通过语料库来训练语言模型，最终才能理解文字的部分语义。关于自然语言处理以及搜索引擎的相关技术，可以参阅吴军老师的《数学之美》，这是一本把高深的数学讲得通俗易懂、妙趣横生的科普读物。

我们接着讲前面起床的故事。在洗漱期间，你抽空浏览手机上的新闻，发现新闻下方有感兴趣的行车记录仪的广告，点进去后打开了某知名电商网站，你看了一下产品的价格和评价，顺手就买了。接着浏览新闻，发现这个新闻客户端越来越人性化，自动把你感兴趣的 IT 新闻及体育新闻排在了首页。好不容易收拾完毕可以出门了，你坐在地铁上，打开音乐播放器，浏览了一遍曲库，没有找到特别想听的歌，于是就让系统给你推荐一些歌。系统推荐的歌还挺“靠谱”的，虽然很多都没听过，但都很对你的“胃口”。

在这段体验描述里，背后的功臣就是推荐系统，这也是机器学习的一个重要应用方向。推荐系统的核心，是不断地学习用户的使用习惯，从而刻画出用户的画像，根据用户的画像去推荐用户感兴趣的商品和文章。

公司新上线了人脸识别系统，在这个“刷脸”的时代，已经没有“忘带工牌”这个签卡的借口了。你走到公司大门口，人脸识别系统自动把你识别出来，然后开门，并准确地通过语音播报的方式和你打招呼。

目前最先进的人脸识别系统基本上都是基于深度学习模型的算法实现的。这一领域也

由早期的传统方法慢慢地被深度学习模型所替代。

当然，机器学习不止这些应用场景。我们在介绍具体算法的时候，会再详细列举出每个算法的应用场景。

延伸阅读：强人工智能

未来学家 Ray Kurzweil 预言，人类将在 2045 年实现强人工智能，就是说到时人工智能将远远强于人类。那个时候人类与强人工智能的差距，要比蚂蚁与人类的差距大几个数量级。这是个让人“脑洞”大开的想象。网上有一篇很火的翻译过来的文章“为什么最近有很多名人，比如比尔盖茨，马斯克、霍金等，让人们警惕人工智能？”，推荐读者阅读一下，其比普通的科幻小说要好看得多。喜欢阅读英文原文的读者，可以在 waitbutwhy.com 上搜索“The AI Revolution”。

1.3　机器学习的分类

机器学习可以分成以下两类。

有监督学习（Supervised learning）通过大量已知的输入和输出相配对的数据，让计算机从中学习出规律，从而能针对一个新的输入做出合理的输出预测。比如，我们有大量不同特征（面积、地理位置、朝向、开发商等）的房子的价格数据，通过学习这些数据，预测一个已知特征的房子价格，这种称为**回归学习（Regression learning）**，即输出结果是一个具体的数值，它的预测模型是一个连续的函数。再比如我们有大量的邮件，每个邮件都已经标记是否是垃圾邮件。通过学习这些已标记的邮件数据，最后得出一个模型，这个模型对新的邮件，能准确地判断出该邮件是否是垃圾邮件，这种称为**分类学习（Classfication learning）**，即输出结果是离散的，即要么输出 1 表示是垃圾邮件，要么输出 0 表示不是垃圾邮件。

无监督学习（Unsupervised learning）通过学习大量的无标记的数据，去分析出数据本身的内在特点和结构。比如，我们有大量的用户购物的历史记录信息，从数据中去分析用户的不同类别。针对这个问题，我们最终能划分几个类别？每个类别有哪些特点？我们事先是不知道的。这个称为**聚类（Clustering）**。这里需要特别注意和有监督学习里的分类的区别，分类问题是我们已经知道了有哪几种类别；而聚类问题，是我们在分析数据之前其实是不知道有哪些类别的。即分类问题是在已知答案里选择一个，而聚类问题的答案是未知的，需要利用算法从数据里挖掘出数据的特点和结构。

网络上流传一个阴谋论：如果你是一个很好说话的人，网购时收到有瑕疵的商品的概率会比较高。为什么呢？理由是电商库存里会有一部分有小瑕疵但不影响使用的商品，为了保证这些商品顺利地卖出去并且不影响用户体验，不被用户投诉，他们会把有瑕疵的商品卖给那些很好说话的人。可问题是，哪些人是好说话的人呢？一个最简单的方法是直接

把有小瑕疵的商品寄给一个用户，如果这个用户没有投诉或退货，并且还给出了好评，就说明他是个好说话的人。还可以通过机器学习来优化这一过程。电商网站有你的大量交易记录和行为记录，如果你从来没有投诉过，买之前也不会和卖家沟通太久，买之后也没有上网评价，或者全部给好评，那么机器学习算法从你的行为特征中会判定你为“好对付”的人。这样你就成了电商们的瑕疵商品的倾销对象了。在这个案例中，电商通过用户的行为和交易数据，分析出不同的用户特点，如哪些人是“老实”人、哪些人是有车一族、哪些人是“土豪”、哪些人家里有小孩等。这就属于无监督学习的聚类问题。

这两种机器学习类别的最大区别是，有监督学习的训练数据里有已知的结果来“监督”；而无监督学习的训练数据里没有结果“监督”，不知道到底能分析出什么样的结果。

1.4 机器学习应用开发的典型步骤

本节通过一个例子来介绍一下机器学习应用开发的典型步骤，以及机器学习领域的一些常用概念。假设，我们要开发一个房价评估系统，系统的目标是对一个已知特征的房子价格进行评估预测。建立这样一个系统需要包含以下几个步骤。

1.4.1 数据采集和标记

我们需要大量不同特征的房子和所对应的价格信息，可以直接从房产评估中心获取房子的相关信息，如房子的面积、地理位置、朝向、价格等。另外还有一些信息房产评估中心不一定有，比如房子所在地的学校情况，这一特征往往会影响房子的价格，这个时候就需要通过其他途径收集这些数据，这些数据叫做**训练样本**，或**数据集**。房子的面积、地理位置等称为**特征**。在数据采集阶段，需要收集尽量多的特征。特征越全，数据越多，训练出来的模型才会越准确。

通过这个过程也可以感受到数据采集的成本可能是很高的。人们常说石油是黑色的“黄金”，在人工智能时代，数据成了透明的“石油”，这也说明为什么蚂蚁金服估值这么高了。蚂蚁金服有海量的用户交易数据，据此他们可以计算出用户的信用指标，称为芝麻信用，根据芝麻信用给你一定的预支额，这就是一家新的信用卡公司了。而这还只是单单一个点的价值，真正的价值在于互联网金融。

在房价评估系统这个例子里，我们的房子价格信息是从房产评估中心获得的，这一数据可能不准确。有时为了避税，房子的评估价格会比房子的真实交易价格低很多。这时，就需要采集房子的实际成交价格，这一过程称为**数据标记**。标记可以是**人工标记**，比如逐个从房产中介那打听房子的实际成交价格；也可以是自动标记，比如通过分析数据，找出房产评估中心给的房子评估价格和真实成交价格的匹配关系，然后直接算出来。数据标记对有监督的学习方法是必须的。比如，针对垃圾邮件过滤系统，我们的训练样例必须包含