

微课  
视频版

作者 20 多年软件开发、大数据实战、企业应用经验积累，10 万忠实粉丝读者群，继《Python 编程从零基础到项目实战》之后的 Python 进阶力作！

# Python 编程

## 从数据分析到机器学习实践

刘瑜 ⊕ 著

**视频讲解 + 示例案例 + 源代码 + 习题实验 + 在线服务**  
随“三酷猫”一起学 Python，去探索机器学习的乐趣

**视频讲解：**608 分钟微视频讲解，实战专家带你畅游 Python。

**示例案例：**“三酷猫”带你进入 Numpy、Scipy、Pandas、Matplotlib、Scikit-learn 库学习科学计算、数据分析与机器学习。

**源码解析：**本书提供实例解析的源代码，方便读者学习与操练。

**习题实验：**每章均配有习题及实验，便于巩固与提高。

**在线服务：**QQ 群、公众号在线服务，让 Python 学习无后顾之忧。



中国水利水电出版社  
www.waterpub.com.cn

# Python 编程从数据 分析到机器学习实践 ( 微课视频版 )

刘 瑜 著



中国水利水电出版社  
www.waterpub.com.cn

· 北京 ·

## 内 容 提 要

《Python 编程从数据分析到机器学习实践（微课视频版）》是一本基于 Python 语言进行数据分析和机器学习的入门与应用类图书，也是一本兼顾实战要求的视频教程。具体内容包括：Jupyter Notebook 应用，Numpy 科学计算、矩阵、线性代数和高级技术，Matplotlib 基础知识和高级应用，Scipy 基础知识和高级应用，Pandas 基础知识、数据处理和基于时间应用，Scikit-learn 基础知识与应用等。本书突出了代码编写的实战要求，为每一章提供了生动有趣的实践内容，包含了文字处理、图像识别、音频编辑、数据分析及预测等实际案例。本书的编写基于 Python 3.7 的最新版本。另外，本书配备了 608 分钟的微视频讲解、提供完整的源代码及 PPT 课件下载。具体下载方法见“前言”中的相关介绍。

《Python 编程从数据分析到机器学习实践（微课视频版）》适合具有 Python 编程基础的 IT 编程工程师、计算机相关专业的学生、专业科学研究人员、数据工程师、高校老师等使用。本书可作为高校、相关培训机构的教材使用。

### 图书在版编目（CIP）数据

Python 编程从数据分析到机器学习实践：微课视频  
版 / 刘瑜著. -- 北京：中国水利水电出版社，2020.2  
ISBN 978-7-5170-8152-4

I. ①P... II. ①刘... III. ①软件工具—程序设计  
IV. ①TP311.561

中国版本图书馆 CIP 数据核字（2019）第 248847 号

书 名	Python 编程从数据分析到机器学习实践（微课视频版） Python BIANCHENG CONG SHUJU FENXI DAO JIQI XUEXI SHIJIAN
作 者	刘瑜 著
出版发行	中国水利水电出版社 （北京市海淀区玉渊潭南路 1 号 D 座 100038） 网址：www.waterpub.com.cn E-mail: zhiboshangshu@163.com
经 售	电话：（010）62572966-2205/2266/2201（营销中心） 北京科水图书销售中心（零售） 电话：（010）88383994、63202643、68545874 全国各地新华书店和相关出版物销售网点
排 版	北京智博尚书文化传媒有限公司
印 刷	三河市龙大印装有限公司
规 格	190mm×235mm 16 开本 28.25 印张 697 千字
版 次	2020 年 2 月第 1 版 2020 年 2 月第 1 次印刷
印 数	0001—5000 册
定 价	99.80 元

凡购买我社图书，如有缺页、倒页、脱页的，本社营销中心负责调换

版权所有·侵权必究

# 前言

## Preface

进入 21 世纪后，随着大数据、人工智能、物联网、云平台等新技术的发展，利用各种各样的数据进行科学分析，成为可能。而这里最引人关注的是基于 Python 语言的各种分析工具，这也是本书将要介绍的主要工具，它们可以轻而易举地解决网页、表格、图片、音频、视频等数据的处理和应用，不再受关系数据和非关系数据的制约，并且，大多数情况下在普通计算机上就可以执行。这是一件非常令人兴奋的事情！我们可以借助这些工具做数据工程师、数据分析及研究科学家、人工智能专家所做的工作。这在以前是不可想象的，总觉得那是一件非常遥远或高不可攀的事情，如今一切皆有可能。

### 一、本书主要涉及工具

#### 1. Jupyter Notebook

该工具的前身 `Ipynb` 是科学家科学计算可视化操作的专业工具，如今被普遍应用到数据处理及分析当中。本书主要介绍其后续项目 `Jupyter Notebook` 的使用功能。

#### 2. Numpy

`Numpy` 是 Python 技术体系下科学计算的基础工具，具有基石的作用，其他很多工具都是基于该工具进行功能扩展的。

#### 3. Scipy

`Scipy` 是在 `Numpy` 基础上开发而成的，是一种高级科学计算工具，除了继承 `Numpy` 大量的功能外，还拥有一系列高级计算功能。

#### 4. Pandas

`Pandas` 是进行数据处理和分析的主流技术工具，尤其擅长二维表格的数据处理，业内人士认为 `Pandas` 将二维数据处理技术用到了极致。

#### 5. Matplotlib

对数据进行图像可视化处理的首选工具，其提供了丰富多彩的二维、三维处理功能，可以看作 `Numpy`、`Scipy`、`Matplotlib` 的组合，可以替代 `MATLAB` 的相关功能。

#### 6. Scikit-learn

`Scikit-learn` 是人工智能入门的首选工具，可以借助它深入理解人工智能的一些基础知识和原理，而且该工具相对简单易学，所提供的功能强大、案例丰富。

当然，上述工具都是免费的、开源的，用户可以直接进行商业应用。

## 二、本书构建思路

### 1. 让读者相对容易地入门

本书可以使读者方便地入门，并能持续地深入学习。基于此，本书对基础知识进行了详细讲解，并通过图、表、提示等技巧使读者更加容易接受；另外，本书遵循由浅入深、层层推进的原则对知识点进行了深入浅出的分析；本书还突出了代码示例、案例的作用。

### 2. 让读者有相对清晰的知识结构

由于科学计算涉及工具众多，如何解释清楚它们之间的关系，并说明工具本身的知识范围，是一件非常具有挑战意义的事情。本书在这方面做了很多工作，可以让读者在一步步掌握知识点的同时，也能很清晰地掌握相关知识结构。

### 3. 让读者能具备一定的实战能力

学完本书的读者可以具备基本扎实的数据处理能力，这也是作为数据工程师所必需具备的。读者在掌握了基本数据处理能力并有一定实操经验后可以比较顺利地进入互联网企业、大数据中心等就业，当然，也可以进一步深入学习和研究。

### 4. 让读者觉得本书有趣味性

为了增加趣味性，本书引入“三酷猫”的故事，“三酷猫”是电影《九条命》主题曲 *Three Cool Cats* 的中文翻译。因为作者和作者的孩子都喜欢那只可爱的猫咪。

## 三、本书适合读者对象

(1) 高校学生。本书主体定位于科学计算入门教材，主要面向高校中已具有 Python 编程基础，并立志于从事数据分析、人工智能研究的在校学生。

(2) IT 行业编程人员。本书还适合已经掌握 Python 语言，并且想在数据工程、人工智能应用方面有所了解或发展的技术人员。

(3) 相关培训机构的教师和学员。本书可作为相关培训机构的培训教材使用。

(4) 专业科研机构的研究人员、高校教师。

(5) 编程爱好者。

## 四、本书相关资源及获取方式

(1) 为方便读者学习，本书提供视频源文件，方便读者下载后在电脑上学习观看。

(2) 为方便读者巩固知识，本书免费提供配套电子版《习题及答案》。

(3) 为方便读者实战学习，本书附赠所有章节的案例源代码。

(4) 为方便教师教学，本书制作了教学课件 PPT。

### 资源下载方式

(1) 读者通过扫描下面的二维码，关注微信公众号后，输入“Py524”即可获得本书资源下载链接。



(2) 读者可以加入本书 QQ 交流群: 797965584 (若群满, 会创建新群, 请注意加群时的提示), 按群公告中的提示获取本书资源下载链接。

## 五、作者介绍

刘瑜, 具有 20 多年 C、ASP、Basic、Foxbase、Delphi、Java、C#、Python 等编程经验, 著有《战神——软件项目管理深度实战》《NoSQL 数据库入门与实践》《Python 编程从零基础到项目实战》, 作者是高级信息系统项目管理师、软件工程硕士、CIO、硕士企业导师、协会理事。

## 六、编写内容约定

(1) 虽然本书主体代码是在 Windows 操作系统下通过的测试, 但是在 Linux、Mac 操作系统下也可以实现 (书中主要通过“说明”给予提示)。

(2) 书中代码行首出现的“\$”若不作特殊说明, 指 Linux 下的命令提示符。

(3) 本书主要内容是在 Python 3.7.2 基础上结合最新第三方库来实现的。

(4) 编排格式及阅读方式提示。本书所有代码在 Notebook 里执行并通过, 采用如下编排风格:

```
s1=[100,100,50]           #列表对象 s1
s1                         #执行 s1
[100, 100, 50]           #列表执行结果
sum(s1)                   #求列表元素和
250                       #求和结果
```

背景颜色, 浅灰色为输入并执行的代码, 深灰色为执行结果输出。本书对代码进行了注释, 除了方便读者阅读和理解外, 也是为了尽量利用纸张的空间, 在有限的空间提供尽量多的知识内容。所以提醒读者, 代码中的注释内容与正文内容同等重要, 必须严格阅读和掌握。在涉及数据使用时, 尽量在一章节的开始, 先实现数据存储及展现的过程, 后续同一节的内容, 避免反复展现相同的数据内容, 这要求读者在同一节能前后对照着阅读代码执行内容。如 s1 数组变量, 可以用于同一节的后续不同代码案例, 而不用反复定义 s1。

## 七、习题及实验使用说明

《习题及答案》主要是为高校学生提供知识巩固测验之用。作者将为购买本书作为教材的学校提供习题标准答案 (可以通过 QQ 群或微信公众号联系获取)。

实验是针对所有读者的，无论在校学生，还是编程从业人员，均应该认真完成每章所提供的实验任务，以切实掌握每章的核心编程内容（对于实验结果，学校可以从作者处获得标准答案；编程从业人员可以参与 QQ 群讨论和咨询）。

## 八、致谢

这里先要致谢视频支持老师——屈晓渊，榆林学院副教授，为本书录制了配套视频讲座。




同时感谢其他参与本书编写的人员——裴英尚、董树南、阚伟、刘勇。




本书受作者知识水平局限，虽然尽心尽力确保书中内容的质量，但难免存在疏漏之处。欢迎广大读者提出宝贵意见，谢谢！

编 者

# 目 录

## Contents

第 1 章 入门准备.....	1	2.2.3 常用菜单和快捷键 功能 .....	29
 视频讲解: 31 分钟		2.2.4 导出文件 .....	32
1.1 基本知识概述.....	2	2.3 Jupyter Magic (魔法) 命令 .....	32
1.1.1 背景知识.....	2	2.3.1 接触魔法命令 .....	32
1.1.2 智能概述.....	3	2.3.2 行魔法命令 .....	33
1.1.3 主要库功能.....	4	2.3.3 单元魔法命令 .....	35
1.2 工具安装.....	5	2.4 习题及实验 .....	37
1.2.1 安装准备工作.....	5	第 3 章 Numpy 科学计算基础 .....	39
1.2.2 Windows、Linux、 Mac 下安装过程.....	7	 视频讲解: 252 分钟	
1.2.3 Anaconda 功能使用.....	8	3.1 接触 Numpy .....	40
1.2.4 其他安装.....	11	3.1.1 什么是 Numpy .....	40
1.3 数据.....	13	3.1.2 开始使用 .....	40
1.3.1 数据分类.....	13	3.2 建立数组 .....	41
1.3.2 数据处理流程.....	15	3.2.1 用 array 建立数组 .....	41
1.4 对读者的建议.....	16	3.2.2 其他常见数组建立 方法 .....	44
1.4.1 学习要求.....	16	3.2.3 数组属性的使用 .....	47
1.4.2 发展方向.....	16	3.2.4 数组方法的使用 .....	48
1.5 公共约定.....	17	3.2.5 数组对接、分割 .....	49
1.6 习题及实验.....	18	3.2.6 案例 1 [建立学生成绩 档案].....	51
第 2 章 Jupyter Notebook 应用基础 .....	20	3.3 索引与切片 .....	52
 视频讲解: 56 分钟		3.3.1 基本索引 .....	52
2.1 接触 Jupyter Notebook .....	21	3.3.2 切片 .....	54
2.1.1 什么是 Jupyter Notebook .....	21	3.3.3 花式索引 .....	55
2.1.2 配置 Jupyter .....	21	3.3.4 迭代 .....	57
2.2 图形界面使用.....	22	3.3.5 案例 2 [完善学生成绩 档案].....	57
2.2.1 主界面功能.....	22		
2.2.2 代码编辑界面.....	25		

3.4	基本数学计算.....	58	4.3.2	求最小二乘解 .....	102
3.4.1	加、减、乘、除.....	58	4.3.3	求张量方程 .....	102
3.4.2	求余、求幂、取整、 复数运算.....	60	4.4	向量、特征向量、特征值 ....	103
3.4.3	数组比较运算.....	63	4.4.1	向量 .....	103
3.4.4	数组位运算.....	64	4.4.2	特征值、特征向量 ....	104
3.4.5	案例 3 [三酷猫 种树] .....	65	4.4.3	特征值分解 .....	106
3.5	通用函数.....	66	4.5	案例 5 [三酷猫求三维空间 面积].....	107
3.5.1	初等函数.....	66	4.6	习题及实验 .....	108
3.5.2	随机函数.....	70	第 5 章	Numpy 高级技术 .....	110
3.5.3	数组集合运算.....	73		视频讲解: 59 分钟	
3.5.4	基础统计函数.....	74	5.1	处理数据文件 .....	111
3.5.5	高级统计函数.....	78	5.1.1	文本文件 .....	111
3.5.6	排序.....	80	5.1.2	二进制文件 .....	115
3.5.7	将数值替换到数组 指定位置.....	81	5.1.3	其他方式处理文件 ....	117
3.5.8	增加和删除行(列) ....	82	5.2	数组原理 .....	119
3.5.9	数值修约等杂项 函数.....	84	5.2.1	数组结构 .....	119
3.5.10	案例 4 [班级成绩 分析] .....	86	5.2.2	副本与视图 .....	120
3.6	习题及实验.....	87	5.2.3	广播原理 .....	121
第 4 章	Numpy 矩阵和线性代数 .....	89	5.3	字符串处理 .....	123
	视频讲解: 81 分钟		5.3.1	字符串操作方法 .....	123
4.1	行列式建立及计算.....	90	5.3.2	字符串信息查找及 判断 .....	124
4.1.1	基本行列式.....	90	5.4	案例 6 [三酷猫制订减肥 计划].....	125
4.1.2	特殊值行列式建立及 对角线获取.....	90	5.5	习题及实验 .....	127
4.2	矩阵计算.....	93	第 6 章	Matplotlib 基础 .....	129
4.2.1	构建矩阵.....	93		视频讲解: 129 分钟	
4.2.2	矩阵转置及维数 调整.....	96	6.1	开始绘图 .....	130
4.2.3	求逆矩阵.....	98	6.1.1	绘制第一张图 .....	130
4.2.4	矩阵积.....	99	6.1.2	画家眼中的绘图 .....	131
4.3	求线性方程组.....	101	6.1.3	图上的那支笔—— plot().....	133
4.3.1	求线性方程组解 .....	101	6.1.4	颜色、图标和线型 ....	134
			6.1.5	注释 .....	135
			6.1.6	在绘图中显示中文 ....	138
			6.1.7	移动刻度线 .....	139

6.1.8	无坐标绘图.....	140	7.2.2	用 animation 工具 (二维) .....	178
6.1.9	多画板多绘图区域....	140	7.2.3	draw()方法 (二维) .....	179
6.2	绘制图形.....	143	7.2.4	随机散点漫步 (三维) .....	182
6.2.1	绘制不同形状的 图形.....	143	7.2.5	旋转三维空间 .....	183
6.2.2	绘制条形图.....	145	7.3	工程化.....	184
6.2.3	绘制直方图.....	147	7.3.1	Web 项目 .....	185
6.2.4	绘制饼状图.....	149	7.3.2	GUI 项目 .....	188
6.2.5	绘制散点图.....	150	7.4	参数配置.....	190
6.2.6	绘制极坐标图.....	151	7.4.1	参数配置文件的 配置 .....	190
6.2.7	绘制极等高图.....	152	7.4.2	常用参数配置示例 .....	192
6.2.8	图形填充.....	153	7.4.3	配置文件其他相关 操作 .....	194
6.3	处理图像.....	155	7.5	案例 8 [三酷猫设计机械零 配件].....	194
6.3.1	读写图像文件.....	155	7.6	习题及实验.....	197
6.3.2	图像伪彩色、灰度 处理.....	157	第 8 章	Scipy 基础 .....	199
6.3.3	给伪彩色加背景色....	158	8.1	接触 Scipy.....	200
6.3.4	根据特征取值.....	159	8.1.1	Scipy 库组成 .....	200
6.3.5	利用矩阵技术处理 图像.....	159	8.1.2	常量使用 .....	201
6.3.6	剪切图像.....	160	8.2	特殊数学函数 (special) .....	202
6.4	案例 7 [三酷猫戴皇冠].....	161	8.2.1	special 分类.....	202
6.5	习题及实验.....	162	8.2.2	逻辑回归模型 .....	203
第 7 章	Matplotlib 高级应用 .....	164	8.2.3	求立方根 .....	204
7.1	绘制三维图形.....	165	8.3	读写数据文件 (io) .....	205
7.1.1	建立三维坐标.....	165	8.3.1	可读写文件函数 .....	205
7.1.2	绘制点、线、面.....	166	8.3.2	WAV 文件处理.....	205
7.1.3	给面打光源.....	169	8.3.3	矩阵文件处理 .....	207
7.1.4	设置标签、标题、 图例.....	171	8.4	线性代数 (linalg) .....	208
7.1.5	旋转三维坐标系.....	172	8.4.1	LU 分解.....	208
7.1.6	绘制三维网线、 条形.....	172	8.4.2	西尔维斯特方程 .....	209
7.1.7	三维像素体.....	175	8.4.3	建立块对角矩阵 .....	210
7.2	动画.....	177	8.5	统计 (stats) .....	210
7.2.1	原始动画绘制 (二维) .....	177	8.5.1	随机变量 .....	210

8.5.2	描述性统计 .....	213	9.4.4	边缘检测 .....	256
8.5.3	核密度估计 .....	214	9.4.5	图像缩放 .....	257
8.6	积分 (integrate) .....	217	9.5	聚类 (cluster) .....	258
8.6.1	integrate 模块 .....	217	9.5.1	K-Means 算法 .....	259
8.6.2	用积分求面积 .....	218	9.5.2	分层聚类算法 .....	260
8.6.3	积分求体积 .....	219	<b>9.6 案例 10 [三酷猫图像文字</b>		
8.6.4	复合梯形积分 .....	220	<b>切割].....</b>	<b>262</b>	
8.6.5	常微分方程求解 .....	221	9.7	习题及实验 .....	265
8.7	空间算法和数据结构		<b>第 10 章 Pandas 基础.....</b>	<b>267</b>	
	(spatial) .....	223	10.1	接触 Pandas.....	268
8.7.1	快速查找最近邻点 .....	224	10.1.1	Pandas 概述 .....	268
8.7.2	凸壳计算 .....	225	10.1.2	数据结构 .....	268
8.8	稀疏矩阵 (sparse) .....	226	10.2	Series 基本操作 .....	269
8.8.1	创建面向列的稀疏		10.2.1	创建 Series .....	269
	矩阵 .....	226	10.2.2	索引 Series 数据 .....	271
8.8.2	创建基于坐标格式的		10.2.3	修改、删除	
	稀疏矩阵 .....	229		Series .....	272
<b>8.9 案例 9 [三酷猫统计岛屿</b>			10.3	DataFrame 基本操作 .....	273
<b>面积].....</b>	<b>230</b>		10.3.1	创建 DataFrame .....	273
8.10	习题及实验 .....	231	10.3.2	读取 DataFrame	
<b>第 9 章 Scipy 高级应用.....</b>	<b>233</b>			指定位置数据 .....	275
9.1	信号处理 (signal) .....	234	10.3.3	修改 DataFrame	
9.1.1	过滤 .....	234		数据 .....	277
9.1.2	快速傅里叶变换 .....	236	10.3.4	删除、增加	
9.1.3	信号窗函数 .....	238		DataFrame 数据 .....	280
9.1.4	卷积 .....	239	10.3.5	排序和排名 .....	283
9.2	插值 (interpolate) .....	240	10.3.6	其他基本功能 .....	287
9.2.1	单变量插值 .....	240	10.4	DataFrame 数据索引	
9.2.2	多变量插值 .....	241		深入 .....	289
9.2.3	样条插值 .....	243	10.4.1	调整行列索	
9.3	优化与拟合 (optimize) .....	244		引值 .....	289
9.3.1	最小二乘拟合 .....	245	10.4.2	多层级索引 .....	291
9.3.2	B-样条拟合 .....	247	10.5	数据计算 .....	292
9.4	多维图像处理 (ndimage) .....	250	10.5.1	常用基础数值	
9.4.1	读写图像 .....	250		运算 .....	292
9.4.2	截取、翻转、旋转 .....	251	10.5.2	比较运算和布尔	
9.4.3	图像滤波 .....	252		值判断 .....	293

10.6	读写数据.....	295	11.4.2	专业样本统计 .....	323
10.6.1	CSV 格式导入		11.5	数据分组和聚合运算 .....	327
	导出.....	295	11.5.1	groupby.....	327
10.6.2	JSON 格式导入		11.5.2	聚合 .....	328
	导出.....	296	11.5.3	分组转换 .....	329
10.6.3	HTML 格式导入		11.5.4	分组过滤 .....	330
	导出.....	297	11.6	数据可视化 .....	331
10.6.4	Excel 格式导入		11.6.1	plot 绘图 .....	332
	导出.....	298	11.6.2	绘制统计图形 .....	333
10.6.5	Clipboard 格式导		11.6.3	用 Matplotlib	
	入导出.....	299		绘图 .....	337
10.6.6	Pickling 格式导入		11.7	字符串数据处理 .....	337
	导出.....	300	11.7.1	字符串对象方法	
10.6.7	HDF5 格式导入			处理 .....	337
	导出.....	300	11.7.2	正则表达式	
10.6.8	SQL 格式导入			处理 .....	339
	导出.....	301	11.8	案例 12 [三酷猫分析	
10.6.9	NoSQL 格式导入			简历].....	340
	导出.....	302	11.9	习题及实验 .....	344
10.7	案例 11 [三酷猫发布交易		第 12 章	Pandas 基于时间应用 .....	346
	公告].....	303	12.1	时间处理基础 .....	347
10.8	习题及实验.....	306	12.1.1	时间基础 .....	347
第 11 章	Pandas 数据处理.....	308	12.1.2	时间表示 .....	347
11.1	缺失数据处理.....	309	12.1.3	时间序列 .....	348
11.1.1	缺失数据产生 .....	309	12.1.4	时间转换 .....	351
11.1.2	缺失数据判断和		12.1.5	时间检索 .....	352
	统计 .....	310	12.2	时间增量处理 .....	354
11.1.3	缺失数据清理 .....	310	12.2.1	时间增量基本	
11.2	多源数据操作.....	312		操作 .....	354
11.2.1	合并.....	312	12.2.2	增量数学运算 .....	356
11.2.2	连接.....	314	12.2.3	时间增量属性、	
11.2.3	指定方向合并 .....	315		增量索引 .....	357
11.3	数据转置和透视表.....	317	12.3	时间周期处理 .....	359
11.3.1	数据转置.....	317	12.3.1	时间周期建立 .....	359
11.3.2	数据透视表 .....	319	12.3.2	时间周期序列 .....	360
11.4	数据统计.....	320	12.4	日期偏移处理 .....	361
11.4.1	基础数学统计 .....	320	12.4.1	时间偏移量建立 .....	361

12.4.2	时间偏移量别名表.....	362	13.5	聚类.....	404
12.5	日期重采样.....	363	13.5.1	聚类基础.....	405
12.5.1	重采样方法.....	364	13.5.2	鸢尾花无监督学习.....	407
12.5.2	降采样.....	364	13.6	降维.....	409
12.5.3	升采样.....	366	13.6.1	降维基础.....	409
12.6	基于时间的绘图处理.....	367	13.6.2	手写数字图像降维.....	411
12.6.1	模拟股票.....	367	13.7	模型选择.....	413
12.6.2	GDP 统计.....	368	13.7.1	模型选择基础.....	413
12.7	案例 13 [三酷猫分析历年分数线].....	370	13.7.2	交叉验证及模型选择.....	414
12.8	习题及实验.....	372	13.7.3	模型固定.....	416
第 13 章	Scikit-learn 基础.....	373	13.8	数据预处理.....	417
13.1	机器学习入门.....	374	13.8.1	数据预处理基础.....	417
13.1.1	从垃圾邮件说起.....	374	13.8.2	手写数字的预处理.....	419
13.1.2	相关概念.....	375	13.9	Scikit-learn 与 TensorFlow 的比较.....	422
13.1.3	Scikit-learn 库.....	377	13.10	案例 14 [三酷猫预测手写数字].....	422
13.2	数据准备.....	378	13.11	习题及实验.....	423
13.2.1	国内外专业在线数据源.....	379	附录一	数据类型.....	425
13.2.2	Scikit-learn 数据源.....	380	附录二	数组常量.....	427
13.2.3	业务数据库数据.....	384	附录三	Matplotlib 的线型、线色、图标.....	429
13.2.4	随机自生成数据.....	386	附录四	机器学习数据集详细说明.....	431
13.2.5	指定文件读取数据.....	392	附录五	本书附赠代码清单.....	434
13.3	分类.....	394	参考文献.....	438	
13.3.1	分类基础.....	394	后记.....	439	
13.3.2	手写字识别.....	398			
13.4	回归.....	400			
13.4.1	回归基础.....	400			
13.4.2	鸢尾花相似度预测.....	402			

# 第 1 章



## 入门准备

Python 语言作为“胶水”语言，“粘连”了大量的第三方库，这些第三方库在科学计算、数据分析、人工智能等方面都发挥了强大的应用功能，其中，包括了 Jupyter Notebook、Numpy、Scipy、Pandas、Matplotlib、Scikit-learn 等。对于已经有 Python 语言基础，想从事数据分析、科学计算、人工智能等工作的读者，本章提供了基础入门知识。

### 学习内容

- 自然智能、人工智能、机器学习、深度学习等概念
- 工具安装
- 数据
- 对读者的建议
- 公共约定
- 习题及实验



扫一扫，看视频

## 1.1 基本知识概述

本节介绍目前 IT 行业内比较流行的且跟本书紧密相关的一些背景知识，同时把 Jupyter Notebook、Numpy、Scipy、Pandas、Matplotlib、Scikit-learn 的基本情况 & 特点进行整体介绍，而本书的主体内容也是围绕上述六大库进行系统讲解和应用的。

### 1.1.1 背景知识

进入 21 世纪，Python 语言得到业内越来越多人的青睐，在全球范围内程序软件使用排名一路上升。截至 2019 年 2 月，其在 TIOBE 网<sup>①</sup>上排名居于前 3，且仍具有强烈的上升势头。Python 的火爆主要基于其两个优点，一是简单易学，赢得了大量编程初学者的认可；二是其粘连了大量的第三方库，如 Jupyter Notebook、Numpy、Scipy、Pandas、Matplotlib、Scikit-learn 等<sup>②</sup>。借助这些第三方库，Python 可以做（大）数据分析、科学计算、人工智能等相关的工作，而且这些第三方库都是免费的、开源的，这给了很多技术人员、研究人员更多的机会，他们可以更加轻松地学习，并通过上述工具，来解决看起来非常“高、大、上”的事情。

2003 年，两个美国科学家为了解决从网上爬取的以 10 亿页计的海量数据，开始研究大数据技术，并在 2008 年获得了成功<sup>③</sup>。而最近几年，我国各省市也陆续建立了省级或国家级的数据中心，累计的数据达到 TB、PB 甚至 EB 级别。另外，不同企业，甚至个人计算机上也都积累了数字、文字、网页、表格、图片、音频、视频等形式的大量数据。这些堆积的数据需要有专业的工具进行加工分析，从而产生对用户有价值的信息，而这正是 Python 所支持的第三方库所擅长的。

2016 年和 2017 年，阿尔法狗（AlphaGo）连续击败了世界围棋冠军李世石和排名第一的世界围棋冠军柯洁，这也宣告了目前人工智能在该领域彻底击败人类<sup>④</sup>。而阿尔法狗的工作原理是人工智能技术（主要是深度学习技术），要想在人工智能技术方面有所作为，利用基于 Python 语言的第三方库学习，是一个很好的选择。

另外，Python 语言在第三方库的支持下，还实现了强大的科学计算功能，如通过 Python 与 Numpy、Scipy、Matplotlib 的结合，可以替代 MATLAB 的部分功能，这对工程技术人员、科研人员来说，是一个极具诱惑的选择。

数学家说这个世界是数字的，今天我们距离可操控的数字世界是如此之近。

自从冯·诺依曼设计并制造出第一台通用电子计算机以来，“0”和“1”展现了这个世界越来越多的信息。从财务表格、数学公式、图片、音频、视频，到机器人智能博弈、音频图像识别、机器翻译、医疗诊断、自动驾驶、指纹和人脸识别等，计算机越来越多地替代人类的各种智能活动。让我们来看看董树南老师亲手研究并制作的“基于数字驱动三维碰撞模型”，如图 1.1 所

<sup>①</sup> TIOBE 网地址 <https://www.tiobe.com/tiobe-index/>。

<sup>②</sup> PyPi 网提供了超过 17 万个基于 Python 的项目（库），其地址为 <https://pypi.org/>。

<sup>③</sup> 刘瑜、刘胜松《NoSQL 数据库入门与实践》（基于 MongoDB、Redis），2018 年 2 月，第 4 页。

<sup>④</sup> 刘瑜《Python 编程从零基础到项目实战》第 375 页。

示，整个模型用数字搭建，再现了数字在三维世界里的仿真功能。

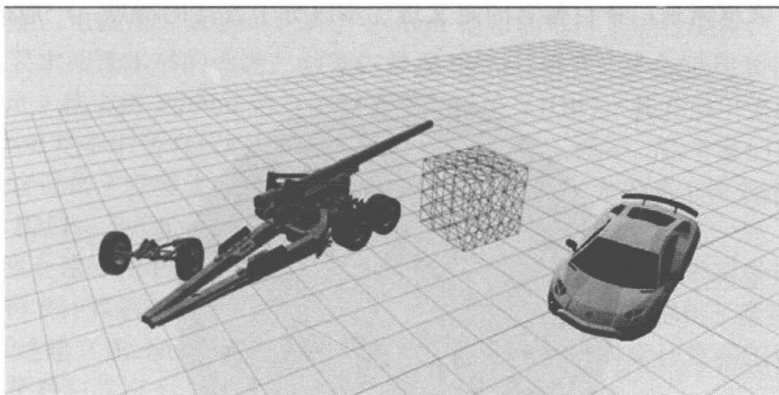


图 1.1 数字驱动三维动态仿真

动态的、随着时间变化的、碰撞的仿真，意味着什么呢？意味着整个世界都可以通过类似数字模型来展示。当读者耐心学完这本书时，大脑里会充满各种数字坐标和公式，也就是说这个世界可以用数字来表达，并可以通过二维、三维视觉来展现。

## 1.1.2 智能概述

在介绍相关工具前，需要先理清一些概念及关系，方便读者准确理解本书的知识。

自然智能（Natural Intelligence, NI），指人通过大脑的运算和决策产生有价值的行为。这些行为包括了人的大脑思考及决策、耳朵听力及判断、眼睛视觉及判断、鼻子嗅觉及判断、皮肤触觉及判断等，体现在人行为的方方面面。

人工智能（Artificial Intelligence, AI），通过机器替代人，实现人具有的智能行为。这里的机器主要指计算机、数据、相关软件，也可以包括相关的智能终端设备。目前人工智能应用比较成熟的技术方向包括机器博弈（智能机器人）、声音识别、图像图片识别（文字、指纹、人脸等）、传感器等提供数据的分析与预测。人工智能研究的主要学科涵盖计算机科学、信息论、控制论、自动化、仿生学、生物学、心理学、数理逻辑、语言学、医学和哲学等。

机器学习（Machine Learning, ML）是算法和统计模型的科学研究，计算机系统使用它来有效地执行特定任务，无须使用明确的指令，而是依赖于模式和推理。它被视为人工智能的一个子集，也是人工智能的核心。<sup>①</sup>机器学习必须借助数据进行“学习”。机器学习的形式可以分为监督学习、半监督学习、无监督学习、强化学习。

深度学习（Deep Learning, DL）（也称为深度结构化学习或分层学习）是基于学习数据表示的机器学习方法系列的一部分，而不是特定于任务的算法。深度学习受生物神经系统中信息处理和通信模式的启发，但与生物大脑的结构和功能存在差异。目前，深度学习架构，如深度神经网络、深度置信网络和递归神经网络，已应用于计算机视觉、语音识别、自然语言处理、音频识别、社交网络过滤、机器翻译、生物信息学、药物设计和医学图像分析等领域。

<sup>①</sup> Machine Learning, [https://en.wikipedia.org/wiki/Machine\\_learning#cite\\_note-1](https://en.wikipedia.org/wiki/Machine_learning#cite_note-1)。

图 1.2 直观地体现了自然智能、人工智能、机器学习、深度学习之间的关系，它们是包含关系，同时也是部分继承关系，各自概念的定义进一步区分了各自的特点，一般读者对此有印象、会简单区分即可。

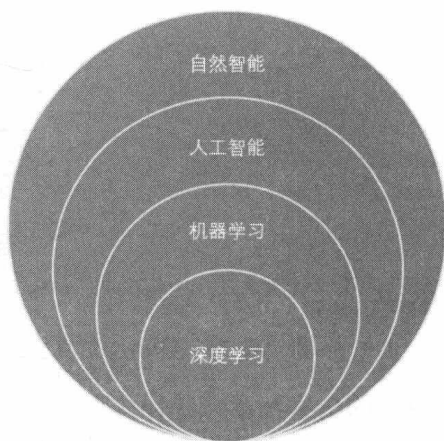


图 1.2 NI、AI、ML、DL 的关系

### 1.1.3 主要库功能

Python、Jupyter、Numpy、Scipy、Pandas、Scikit-learn、Matplotlib 等主要分别实现什么功能呢？它们之间是什么关系呢？如图 1.3 所示。

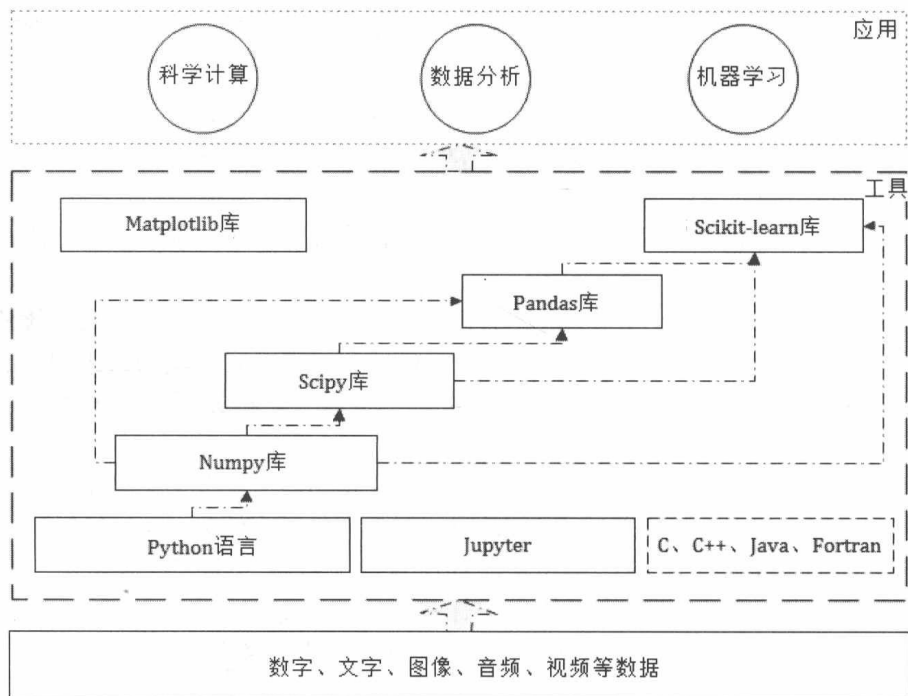


图 1.3 主要功能及相关关系