

Preface

前言

感谢

首先感谢大家的信任。

作者仅仅是在学习应用数学科学和机器学习算法时，多读了几本数学书，多做了些思考和知识整理而已。知者不言，言者不知。知者不博，博者不知。水平有限，把自己有限所学所思斗胆和大家分享，作者权当无知者无畏。希望大家在 B 站视频下方和 Github 多提意见，让这套书成为作者和读者共同参与创作的优质作品。

特别感谢清华大学出版社的栾大成老师。从选题策划、内容创作、装帧设计，栾老师事无巨细、一路陪伴。每次和栾老师交流，我都能感受到他对优质作品的追求、对知识分享的热情。

出来混总是要还的

曾经，考试是我们学习数学的唯一动力。考试是头悬梁的绳，是锥刺股的锥。我们中的绝大多数人从小到大为各种考试埋头题海，数学味同嚼蜡，甚至让人恨之入骨。

数学给我们带来了无尽的折磨。我们憎恨数学，恐惧数学，恨不得一走出校门就把数学抛之脑后、老死不相往来。

可悲可笑的是，我们其中很多人可能会在毕业的五年或十年以后，因为工作需要，不得不重新学习微积分、线性代数、概率统计，悔恨当初没有学好数学、走了很多弯路、没能学以致用，从而迁怒于教材和老师。

这一切不能都怪数学，值得反思的是我们学习数学的方法、目的。

再给自己一个学数学的理由

为考试而学数学，是被逼无奈的举动。而为数学而数学，则又太过高尚而遥不可及。

相信对于绝大部分的我们来说，数学是工具、是谋生手段，而不是目的。我们主动学数学，是想用数学工具解决具体问题。

现在，这套书给大家一个“学数学、用数学”的全新动力——数据科学、机器学习。

数据科学和机器学习已经深度融合到我们生活的方方面面，而数学正是开启未来大门的钥匙。不是所有人生来都握有一副好牌，但是掌握“数学 + 编程 + 机器学习”绝对是王牌。这次，学习数学不再是为了考试、分数、升学，而是投资时间、自我实现、面向未来。

未来已来，你来不来？

本套丛书如何帮到你

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

为了让大家学数学、用数学，甚至爱上数学，作者可谓颇费心机。在创作这套书时，作者尽量克服传统数学教材的各种弊端，让大家学习时有兴趣、看得懂、有思考、更自信、用得着。

为此，丛书在内容创作上突出以下几个特点：

- ◀ **数学 + 艺术**——全彩图解，极致可视化，让数学思想跃然纸上、生动有趣、一看就懂，同时提高大家的数据思维、几何想象力、艺术感；
- ◀ **零基础**——从零开始学习 Python 编程，从写第一行代码到搭建数据科学和机器学习应用；
- ◀ **知识网络**——打破数学板块之间的壁垒，让大家看到数学代数、几何、线性代数、微积分、概率统计等板块之间的联系，编织一张绵密的数学知识网络；
- ◀ **动手**——授人以鱼不如授人以渔，和大家一起写代码、用 Streamlit 创作数学动画、交互 App；
- ◀ **学习生态**——构造自主探究式学习生态环境“微课视频 + 纸质图书 + 电子图书 + 代码文件 + 可视化工具 + 思维导图”，提供各种优质学习资源；
- ◀ **理论 + 实践**——从加减乘除到机器学习，丛书内容安排由浅入深、螺旋上升，兼顾理论和实践；在编程中学习数学，学习数学时解决实际问题。

虽然本书标榜“从加减乘除到机器学习”，但是建议读者朋友们至少具备高中数学知识。如果读者正在学习或曾经学过大学数学（微积分、线性代数、概率统计），这套书就更容易读了。

聊聊数学

数学是工具。锤子是工具，剪刀是工具，数学也是工具。

数学是思想。数学是人类思想的高度抽象的结晶体。在其冷酷的外表之下，数学的内核实际上就是人类朴素的思想。学习数学时，知其然，更要知其所以然。不要死记硬背公式定理，理解背后的数学思想才是关键。如果你能画一幅图、用大白话描述清楚一个公式、一则定理，这就说明你真正理解了它。

数学是语言。就好比世界各地不同种族有自己的语言，数学则是人类共同的语言和逻辑。数学这门语言极其精准、高度抽象，放之四海而皆准。虽然我们中绝大多数人没有被数学女神选中，不能为人类的对数学认知开疆扩土；但是，这丝毫不妨碍我们使用数学这门语言。就好比，我们不会成为语言学家，我们完全可以使用母语和外语交流。

数学是体系。代数、几何、线性代数、微积分、概率统计、优化方法等等，看似一个个孤岛，实际上都是数学网络的一条条织线。建议大家学习时，特别关注不同数学板块之间的联系，见树，更要见林。

数学是基石。拿破仑曾说“数学的日臻完善和这个国强民富息息相关。”数学是科学进步的根基，是经济繁荣的支柱，是保家卫国的武器，是探索星辰大海的航船。

数学是艺术。数学和音乐、绘画、建筑一样，都是人类艺术体验。通过可视化工具，我们会在看似枯燥的公式、定理、数据背后，发现数学之美。

数学是历史，是人类共同记忆体。“历史是过去，又属于现在，同时在指引未来。”数学是人类的集体学习思考，她把人的思维符号化、形式化，进而记录、积累、传播、创新、发展。从甲

骨、泥板、石板、竹简、木牍、纸草、羊皮卷、活字印刷、纸质书，到数字媒介，这一过程持续了数千年，至今绵延不息。

数学是无穷无尽的**想象力**，是人类的**好奇心**，是自我挑战的**毅力**，是一个接着一个的**问题**，是看似荒诞不经的**猜想**，是一次次胆大包天的**批判性思考**，是敢于站在前人的臂膀之上的**勇气**，是孜孜不倦地延展人类认知边界的**不懈努力**。

家园、诗、远方

诺瓦利斯曾说：“哲学就是怀着一种乡愁的冲动到处去寻找家园。”

在纷繁复杂的尘世，数学纯粹的就像精神的世外桃源。数学是，一束光，一条巷，一团不灭的希望，一股磅礴的力量，一个值得寄托的避风港。

打破陈腐的锁链，把功利心暂放一边，我们一道怀揣一分乡愁、心存些许诗意、踩着艺术维度，投入数学张开的臂膀，驶入她色彩斑斓、变幻无穷的深港，感受久违的归属，一睹更美、更好的远方。

Acknowledgement

致谢

To my parents.

谨以此书献给我的母亲父亲

How to Use the Book

使用本书

丛书资源

本系列丛书提供的配套资源有以下几个：

- ❖ 纸质图书；
- ❖ PDF 文件，方便移动终端学习；请大家注意，纸质图书经过出版社五审五校修改，内容细节上会和 PDF 文件有出入。
- ❖ 每章提供思维导图，纸质书提供全书思维导图海报；
- ❖ Python 代码文件，直接下载运行，或者复制、粘贴到 Jupyter 运行；
- ❖ Python 代码中有专门用 Streamlit 开发数学动画和交互 App 的文件；
- ❖ 微课视频，强调重点、讲解难点、聊聊天。

在纸质书中为了方便大家查找不同配套资源，作者特别设计了如下几个标识。



数学家、科学家、
艺术家等语录



代码中核心Python
库函数和讲解



思维导图总结本章
脉络和核心内容



配套Python代码完
成核心计算和制图



用Streamlit开发制
作App应用



介绍数学工具、机
器学习之间联系



引出本书或本系列
其他图书相关内容



提醒读者格外注意
的知识点



每章配套微课视频
二维码



相关数学家生平贡
献介绍



每章结束总结或升
华本章内容



本书核心参考和推
荐阅读文献

微课视频

本书配套微课视频均发布在 B 站——生姜 DrGinger：

❖ <https://space.bilibili.com/513194466>

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

微课视频是以“聊天”的方式，和大家探讨某个数学话题的重点内容，讲讲代码中可能遇到的难点，甚至侃侃历史、说说时事、聊聊生活。

本书配套的微课视频目的是引导大家自主编程实践、探究式学习，并不是“照本宣科”。

纸质图书上已经写得很清楚的内容，视频课程只会强调重点。需要说明的是，图书内容不是视频的“逐字稿”。

代码文件

本系列丛书的 Python 代码文件下载地址为：

◀ <https://github.com/Visualize-ML>

Python 代码文件会不定期修改，请大家注意更新。图书配套的 PDF 文件和勘误也会上传到这个 GitHub 账户。因此，建议大家注册 GitHub 账户，给书稿文件夹标星 (star) 或分支克隆 (fork)。

考虑再三，作者还是决定不把代码全文印在纸质书中，以便减少篇幅，节约用纸。

本书编程实践例子中主要使用“鸢尾花数据集”，数据来源是 Scikit-learn 库、Seaborn 库。此外，系列丛书封面设计致敬梵高《鸢尾花》，要是给本系列丛书起个昵称的话，作者乐见“鸢尾花书”。

App 开发

本书几乎每一章都至少有一个用 Streamlit 开发的 App，用来展示数学动画、数据分析、机器学习算法。

Streamlit 是个开源的 Python 库，能够方便快捷搭建、部署交互型网页 App。Streamlit 非常简单易用、很受欢迎。Streamlit 兼容目前主流的 Python 数据分析库，比如 NumPy、Pandas、Scikit-learn、PyTorch、TensorFlow 等等。Streamlit 还支持 Plotly、Bokeh、Altair 等交互可视化库。

本书中很多 App 设计都采用 Streamlit + Plotly 方案。此外，本书专门配套教学视频手把手和大家一起做 App。

大家可以参考如下页面，更多了解 Streamlit：

◀ <https://streamlit.io/gallery>

◀ <https://docs.streamlit.io/library/api-reference>

实践平台

本书作者编写代码时采用的 IDE (integrated development environment) 是 Spyder，目的是给大家提供简洁的 Python 代码文件。

但是，建议大家采用 JupyterLab 或 Jupyter notebook 作为本系列丛书配套学习工具。

简单来说，Jupyter 集合“浏览器 + 编程 + 文档 + 绘图 + 多媒体 + 发布”众多功能与一身，非常适合探究式学习。

运行 Jupyter 无需 IDE，只需要浏览器。Jupyter 容易分块执行代码。Jupyter 支持 inline 打印结果，直接将结果图片打印在分块代码下方。Jupyter 还支持很多其他语言，比如 R 和 Julia。

使用 markdown 文档编辑功能，可以编程同时写笔记，不需要额外创建文档。Jupyter 中插入图片和视频链接都很方便。此外，还可以插入 Latex 公式。对于长文档，可以用边栏目录查找特定内容。

Jupyter 发布功能很友好，方便打印成 HTML、PDF 等格式文件。

Jupyter 也并不完美，目前尚待解决的问题有几个。Jupyter 中代码调试不方便，需要安装专门插件 (比如 debugger)。Jupyter 没有 variable explorer，要么 inline 打印数据，要么将数据写到 csv 或 Excel 文件中再打开。图像结果不具有交互性，比如不能查看某个点的值，或者旋转 3D 图形，可以考虑安装 (jupyter-matplotlib)。注意，利用 Altair 或 Plotly 绘制的图像支持交互功能。对于自定义函数，目前没有快捷键直接跳转到其定义。但是，很多开发者针对这些问题都开发了插件，请大家留意。

大家可以下载安装 Anaconda，JupyterLab、Spyder、PyCharm 等常用工具都集成在 Anaconda 中。下载 Anaconda 的地址为：

◀ <https://www.anaconda.com/>

学习步骤

大家可以根据自己的偏好制定学习步骤，本书推荐如下步骤。



学完每章后，大家可以在平台上发布自己的 Jupyter 笔记，进一步听取朋友们的意见，共同进步。这样做还可以提高自己学习的动力。

意见建议

欢迎大家对本系列丛书提意见和建议，丛书专属邮箱地址为：

◀ jiang.visualize.ml@gmail.com

也欢迎大家在 B 站视频下方留言互动。

Contents

目录

0

Introduction

绪论

动手编程；知其然，不需要知其所以然

0.1 本册在全套丛书的定位

鸢尾花书共有 7 册，分为三大板块——编程、数学、实践。

《编程不难》是鸢尾花书的第一本，也是“编程”板块的第一本，着重介绍如何零基础入门学 Python 编程。“编程”板块的第二本则探讨如何用 Python 完成数学、数据可视化。

虽然《编程不难》主要讲解 Python 编程，但是也离不开数学。本书尽量避免讲解数学概念公式，而是用图形、近乎口语化的语言描述程序设计、数据分析、机器学习背后常用数学思想。我们把理解这些数学工具的任务放在了鸢尾花书“数学”板块，也叫“数学三剑客”——《数学要素》《矩阵力量》《统计至简》。

《编程不难》正文提供代码示例和讲解，而且会提供练习题目。每章还会配套 Jupyter Notebook 代码文件。

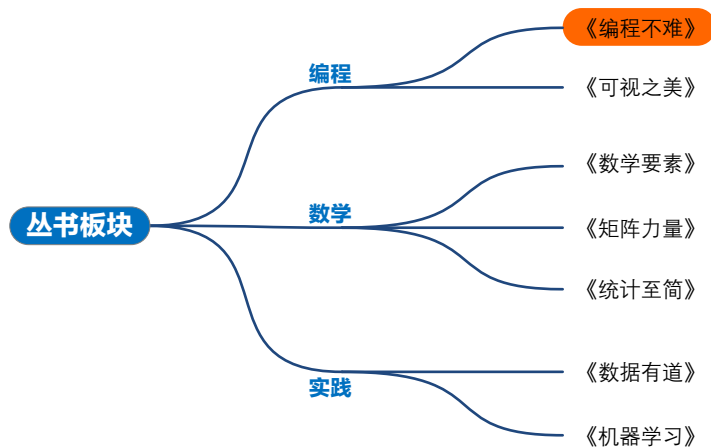


图 1. “鸢尾花书”板块布局

0.2 结构：8大板块

本书一共有 36 个话题，可以归纳为 8 大板块——预备、语法、绘图、数组、数据、数学、机器学习、应用。

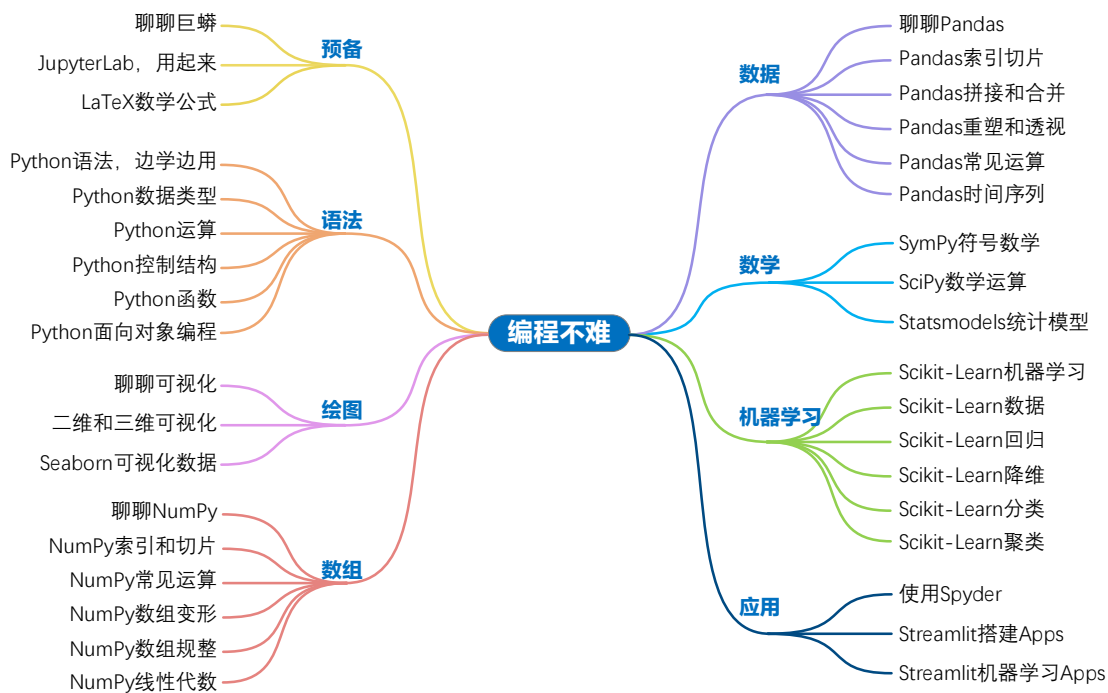


图 2. 《编程不难》板块布局

预备

这部分有 3 章，占全书 1/12。全书第 1 章聊了聊 Python 编程，Python 和可视化、数学、机器学习有什么关系。这一章最关键的任务是成功安装并测试 Anaconda。

第 2 章介绍如何使用 JupyterLab。对于鸢尾花书系列图书，JupyterLab 特别适合大家进行探究式学习。第 2 章中，学习并熟练使用快捷键可以极大提高生产力。

第 3 章介绍如何用 LaTeX 语言在 JupyterLab markdown 中编写常用数学表达。注意，本章不会介绍如何利用 LaTeX 撰写论文等文档，感兴趣的读者可以自行学习。

语法

这部分有 6 章，占全书 1/6，主要介绍 Python 基本语法。

第 4 章主要介绍注释、缩进、变量、包、Python 风格等基础概念。

第 5 章讲解 Python 常用数据类型，比如数字、字符串、列表、字典。这一章还介绍了线性代数中的矩阵和向量这两个概念。

第 6 章讲解 Python 常见运算，比如算术、比较、逻辑、赋值、成员、身份等运算符。学习时，请大家注意这些运算符的优先级。

第 7 章介绍 Python 控制结构，比如条件语句、循环语句、迭代器等概念。这一章还介绍了线性代数中的向量内积和矩阵乘法这两种重要数学概念，请大家务必掌握其运算规则。

第 8 章介绍 Python 函数，比如自定义函数、匿名函数，以及如何构造模块、库。

第 9 章简介 Python 面向对象编程，其中包括属性、方法、装饰器、父类、子类等概念。这一章仅仅介绍了 Python 面向对象编程的冰山一角。

绘图

可视化是鸢尾花书核心的特色之一，所以特别创作了《可视之美》一册专门讲解数学、数据可视化。《编程不难》中的绘图仅仅蜻蜓点水介绍了本册常用的可视化工具，因此这部分仅仅安排了 3 章，占全书 1/12，主要介绍 Matplotlib、Plotly、Seaborn 这三个库中最常用的几种可视化函数。

第 10 章首先介绍了一幅图的重要组成元素，并讲解如何用 Matplotlib 和 Plotly 绘制线图。

第 11 章介绍几种最常用的二维和三维可视化方案，比如散点图、等高线图、热图、网格面等等。大家如果对可视化特别感兴趣的话，也可以平行学习《可视之美》。

第 12 章主要介绍如何用 Seaborn 完成样本数据统计描述，这章讲解的可视化方案包括直方图、小提琴图、箱型图散点图、概率密度分布等等。

数据

这个板块主要介绍 NumPy，一共有 6 章，占全书 1/6。NumPy 是一个用于科学计算和数据分析的 Python 库。它提供了高效的多维数组对象，以及用于对这些数组执行各种数学、逻辑、统计操作的函数。在机器学习中，NumPy 具有重要的作用，因为它为数据处理、数值计算和数组操作提供了强大的工具，为机器学习算法的实现和优化提供了基础支持。

第 13 章介绍数组、数列、网格数据、随机数、导入、导出等 NumPy 库基本概念。

第 14 章介绍如何对 NumPy 数组进行索引和切片。请大家务必注意视图、副本这两个概念。

第 15 章介绍 NumPy 常见运算，比如基本算术、代数、统计运算。请大家务必掌握广播原则。

第 16 章介绍 NumPy 中常用的各种数组变形方法。第 17 章进一步介绍了数组堆叠、重复、分块等规整操作。

第 18 章走马观花介绍地介绍 NumPy 的 linalg 模块中常用的线性代数工具，比如向量的模、向量内积、矩阵乘法、Cholesky 分解、特征值分解、奇异值分解等。想要进一步深入学习线性代数工具，请大家参考鸢尾花书《矩阵力量》。

数组

这个板块主要介绍 Pandas，一共有 6 章，占全书 1/6。Pandas 是一个用于数据分析和数据处理的 Python 库，它提供了高效的数据结构和数据操作工具，特别适用于处理和分析结构化数据。在机器学习中，Pandas 具有重要的作用，因为它能高效地加载、处理、清洗、转换、探索、分析数据，为机器学习建模和分析提供了强大的支持。

第 19 章介绍如何创建数据帧 DataFrame，以及常见数据帧操作。

第 20 章讲解如何对 Pandas DataFrame 进行索引和切片，比如提取特定行列、特定行列、条件索引、多层索引等等。

第 21 章讲解如何利用 concat()、join()、merge() 方法对 DataFrame 进行拼接和合并。

第 22 章介绍如何用 pivot()、stack()、unstack() 方法对 DataFrame 进行重塑和透视。

第 23 章介绍 Pandas 中各种运算，比如四则运算、统计运算，以及用 groupby()、apply() 方法完成聚合和自定义操作。

第 24 章讲解 Pandas 时间序列数据，包括缺失值、移动平均、统计分析等操作。

数学

这个板块主要介绍 SymPy、SciPy、Statsmodels 三个库，一共有 3 章，占全书 1/12。

第 25 章介绍 SymPy，SymPy 是一个 Python 的符号数学计算库。大家可以用这一章回顾或了解常用的代数、微积分、线性代数概念。

第 26 章讲解如何用 SciPy 完成插值、积分、线性代数、优化、统计等运算。

第 27 章介绍 Statsmodels 模块，并介绍如何利用 Statsmodels 完成线性回归、主成分分析、概率密度估计。

机器学习

这个板块主要介绍 Scikit-learn，一共有 6 章，占全书 1/6。Scikit-learn 是一个用于机器学习和数据挖掘的 Python 库，它建立在 NumPy、SciPy 和 Matplotlib 等库的基础之上，提供了丰富的机器学习算法、工具和函数，用于实现各种机器学习任务，如分类、回归、聚类、降维、模型选择等。

第 28 章简述了有标签数据、无标签数据、回归、降维、分类、聚类等机器学习概念。

第 29 章介绍了 Scikit-learn 中数据集、产生样本数据、处理缺失值、处理离群值、特征缩放等方法。

第 30 ~ 33 章分别介绍了回归、降维、分类、聚类四个机器学习问题。

应用

这一板块有 3 章，占全书 1/12。

第 34 章介绍如何使用 Spyder 完成 Python 编程开发。这一章介绍的 Spyder 是为下一章开发 Streamlit 提供 IDE 工具。

第 35 章介绍如何用 Streamlit 搭建应用 App。Streamlit 是一个用于创建交互式数据应用程序的 Python 库。它的主要目标是让数据科学家、工程师和开发人员能够快速、轻松地将数据融入到应用程序中，而无需深入了解前端开发。使用 Streamlit，可以将数据可视化、机器学习模型、分析结果等内容转化为具有用户界面的应用，从而方便地与用户进行交互。

第 36 章中，我们将用 Streamlit 开发几个数学学习、机器学习应用 Apps。

0.3 特点：知其然，不需要知其所以然

《编程不难》极力避免“Python 语法工具书”这种大而全范式。

《编程不难》想要以轻松、图解方式，为零基础入门读者提供可读性高、学以致用内容。学习时，期待大家立刻有“获得感”，并有持续动力、兴趣继续深入学习。因此在创作这本书时，作者定下的目标是——力争让大家在阅读本书时兴致勃勃、眼界大开，读完本书后感觉收获满满、意犹未尽。

作为鸢尾花书系列的第一本，《编程不难》格外强调“零基础入门”学习 Python，因此本书力争给大家提供“保姆式”手把手式的教学体验；鉴于此，如果 Python 有经验的读者觉得本书在行文上显得“婆婆妈妈”，请体谅。

值得反复强调的是，学习 Python 编程时，希望大家一定要吸取英语学习失败的教训，不能死磕语法。千万不要死记硬背，一定要边学边用、活学活用、以用为主。

由于《编程不难》强调“零基础”学习 Python 编程，不需要大家掌握 Python 库中常用函数背后的数学工具、数学思想；也就是说，“知其然，不需要知其所以然”。即便《编程不难》提到了某些数学工具，我们也只用文字和图像方式介绍，因此大家在本书中不会看到各种编号公式。

下面，让我们正式开启鸢尾花书第一本《编程不难》的学习之旅。