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

曾几何时，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

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。

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

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

Introduction

0 绪论

图解 + 编程 + 实践 + 数学板块融合

0.1 本册在鸢尾花书的定位

首先祝贺大家完成“数学”板块的学习，同时欢迎大家来到鸢尾花书第三板块——实践。

“实践”这个板块，我们将会把学到的编程、可视化，特别是数学工具应用到具体的数据科学、机器学习算法中，并在实践中加深对这些工具的理解。

“实践”这个板块有两本书：《数据有道》、《机器学习》。

《数据有道》则以数据为视角，《数据有道》的很多

《机器学习》着重介绍机器学习中最经典的四大类算法——回归、分类、降维、聚类。

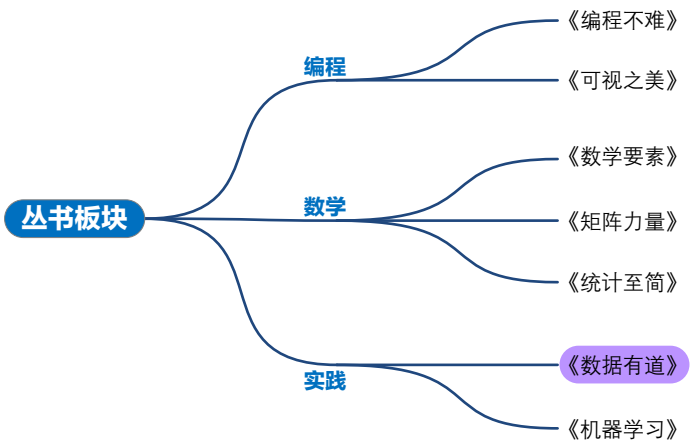

图 1. 鸢尾花书板块布局

0.2 结构：6 大板块

《数据有道》可以归纳为 6 大板块——数据说、数据处理、时间数据、图论基础、图与矩阵、图论实践。这 6 个板块都紧紧围绕一个主题——数据！

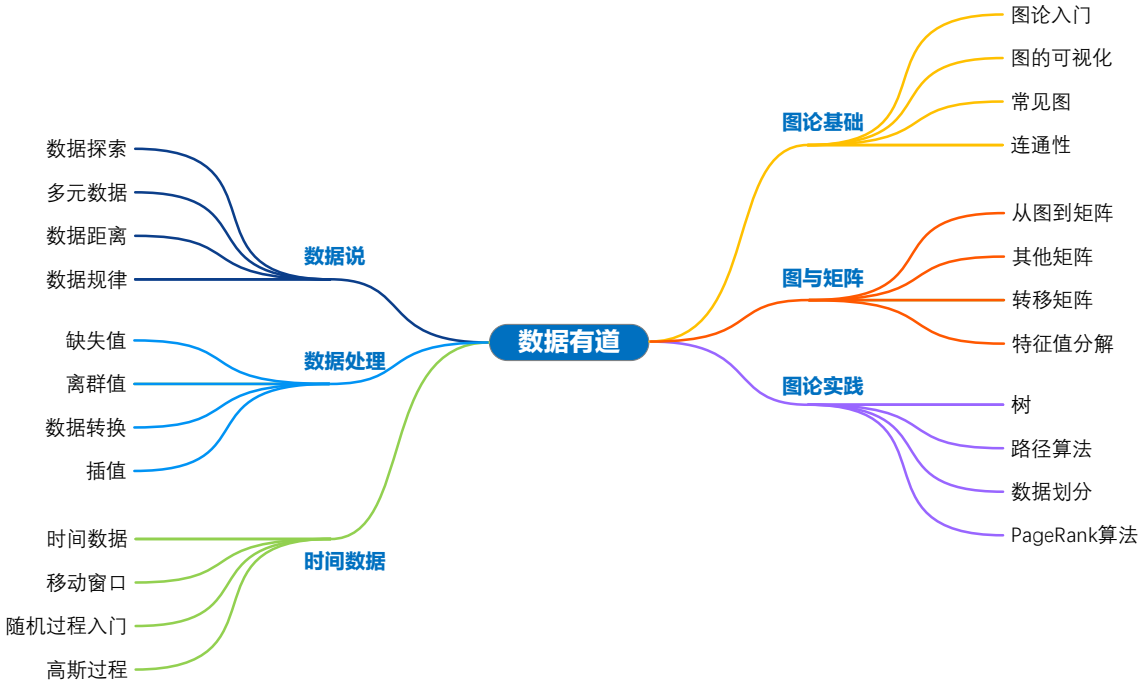

图 2. 《数据有道》板块布局

本书第 1 章不属于上述任何一个板块，这章相当于是本册“综述”，和大家鸟瞰数据。

数据说

这个板块安排了 4 章内容，这 4 章相当于从数据视角回顾本书前 5 册内容。

第 2 章是从数据探索角度回顾统计描述。第 3 章则介绍多元数据描述，特别是协方差矩阵。鸢尾花书一而再再而三不厌其烦地从各个角度讲解协方差矩阵，这是因为协方差矩阵在机器学习算法中扮演太重要的角色。第 4 章介绍数据距离，机器学习算法几乎都离不开距离度量。而距离度量丰富多彩，本章带大家回顾各种距离度量。第 5 章讲解数据规律，我们将用各种数据实例，展示机器学习四大算法的能量所在。

数据处理

这个板块主要介绍机器学习中的数据处理。第 6 章讲解如何处理数据中的缺失值。第 7 章介绍处理离群值的常用工具，这一章和机器学习算法联系紧密。第 8 章讲解常用数据转换方法，本章也相当于对统计知识的回顾。第 9 章特别介绍插值，注意插值和回归的区别；另外，请大家注意《可视之美》中介绍的贝塞尔曲线和插值的联系。

时间数据

这个板块介绍一类特殊数据——具有时间戳的数据，也叫时间序列。第 10 章讲解如何处理时间数据、发现数据的趋势、时间序列分解等内容。时间数据的特征随时间动态变化，这是第 11 章特别强调的一点。第 7 章中，大家会看到均值、标准差 (波动率)、相关性系数、回归系数、协方差矩阵都可以随

移动窗口变化。第 12 章是随机过程入门，介绍布朗运动、几何布朗运动，以及用几何布朗运动完成股价走势的蒙特卡罗模拟。这一章是《统计至简》第 15 章的延伸。第 13 章介绍高斯过程。高斯过程可谓高斯分布、贝叶斯定理的集大成者。这一章相对来说难度较高，需要大家下一点功夫理解。《机器学习》还会介绍高斯过程，并且介绍如何用 Scikit-learn 中高斯过程工具完成回归和分类。

本书最后三个板块都和图论有关。可以说，图是一种特别有趣的数据结构。而且，阅读这部分内容，大家会发现，图就是矩阵，矩阵就是图。这 12 章内容都会采用 NetworkX 库创作各种和理论紧密结合的实操实例，希望大家边学边练。

图论基础

第 14 章主要介绍了图论当中的有向图、无向图及相关概念。第 15 章专门介绍如何用 NetworkX 绘制图，特别是对节点、边、标注等元素的修饰。第 16 章结合 NetworkX 介绍图论中常见的几种图及特性。第 17 章介绍图论中连通性这个概念。在图论中，连通性用来分析图中节点之间路径相互连接关系。它用于研究图的整体结构和网络中信息的传递，对于解决网络设计、路径规划等问题具有重要意义。

图与矩阵

这个板块很有意思——把图和矩阵紧密联系在了一起。图就是矩阵，矩阵就是图，这一点值得反复强调。

第 18 章主要介绍邻接矩阵和图之间的关系。第 19 章介绍更多和图有关的矩阵，比如关联矩阵、度矩阵、拉普拉斯矩阵等等。这一章还介绍各种成对矩阵和图的关系，比如成对欧氏距离矩阵、成对相似度距离矩阵。第 20 章则把邻接矩阵和转移矩阵联系在一起，并引出了马尔科夫链；这实际上是《数学要素》中“鸡兔互变”的扩展。。第 21 章首先回顾特征值分解，然后介绍图论中特征值的应用。

图论实践

本书最后一个板块则是图论实践，一共设置了 4 个话题——树 (第 22 章)、路径问题 (第 23 章)、数据划分 (第 24 章)、PageRank 算法 (第 25 章)。第 22 章讲了 *kd* 树，顺便会提到 *k* 临近算法，一种分类算法。这一章还会涉及到决策树 (分类算法)、层次聚类 (聚类算法)，这是《机器学习》要展开讲解的两种重要算法。第 24 章讲解数据划分时用到的算法是谱聚类，《机器学习》还会简单介绍这种算法。

0.3 特点：以好奇心为驱动力

《数据有道》几易其稿。稿件不断大修大改的过程中，笔者不断问自己，《数据有道》怎么写才把鸢尾花书之前五本书的内容融合在一起，又能用数据视角扩展知识网络，还能帮助大家铺平学习第 7 册《机器学习》的道路？

想来想去，想到一个办法——以数据为视角，强调实践应用中可能出现的数据问题为导向。

鸢尾花书前五本书分别介绍了编程、可视化、数学、线性代数、概率统计这五个板块。虽然每本书穿插了很多应用案例，但是“工具”还是“主”，“应用”则是“宾”。

打个比方，前五本书好像个“学长”，时时刻刻在告诉大家“认真读，好好学，这些编程工具、可视化工具、数学工具以后有大用途”。

这可能也是课堂“被动”教学的弊端，包括数学在内的课堂学习都是发生在我们的实际生活、工作、探索世界的“需求”之前。先学着，好好学，以后可能用得着。至于什么时候用，怎么用，这都不是现在要操心的事情。

以奇异值分解为例，在《编程不难》中，我们仅仅介绍了 NumPy 中完成奇异值分解的 Python 函数。《可视之美》则利用几何变换展示奇异值分解背后的数学之美。而《矩阵力量》则花了两章内容专门讲解这个奇异值分解这个数据工具，然后又利用奇异值分解介绍了四个空间。

而《数据有道》试图做的就是逆转这种被动的“主宾”关系！

以数据为名，以好奇心和疑问为驱动，主动探索使用“编程 + 可视化 + 数学”工具，回答各种各样问题。

在《数据有道》中我们将会回顾鸢尾花书前五本主要的“编程 + 可视化 + 数学”工具，让大家对很多概念从似懂非懂变成如数家珍；同时，我们还会掌握更多工具，扩展大家的知识网络。

至于《数据有道》是否成功实现了这一目标，就要看大家阅读后学习体验了，希望本册不会辜负大家期待。