

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：

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

- ◀ **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富民富息息相关。”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。”

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

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		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>

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 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

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。

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

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

绪论

图解 + 编程 + 实践 + 数学板块融合

0.1 本册在全套丛书的定位

欢迎大家来到“鸢尾花书”7本书最后一本——《机器学习》！

特别对于从“鸢尾花书”第1本《编程不难》一直读到本册的读者，请大家为自己坚持到底的精神鼓掌！

回过头来看，“鸢尾花书”的前两本强调编程基础，《编程不难》是零基础入门 Python，《可视之美》和大家探讨“数学 + 美学”。“数学”板块三本书为“实践”打下了坚实的数学基础。因此，数学基础不强的读者，不建议跳过“数学”板块直接学习本册。而《数据有道》则强调数据直觉，和大家探讨数据探索背后的数学和可视化。

希望鸢尾花书的前六本帮大家解决了很多编程、可视化、数学方面的痛点问题。

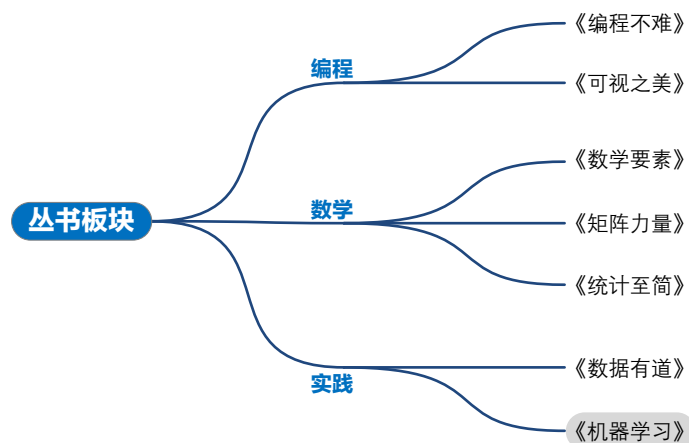


图 1. 本系列丛书板块布局

0.2 结构：4 大板块

《机器学习》设置了 24 个话题，对应四大类机器学习经典算法（回归、分类、降维、聚类）。每类算法不多不少正好 6 个话题。此外，“鸢尾花书”之前的 6 本书铺垫的内容确保大家能够完全理解、充分掌握这 24 个机器学习算法。

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

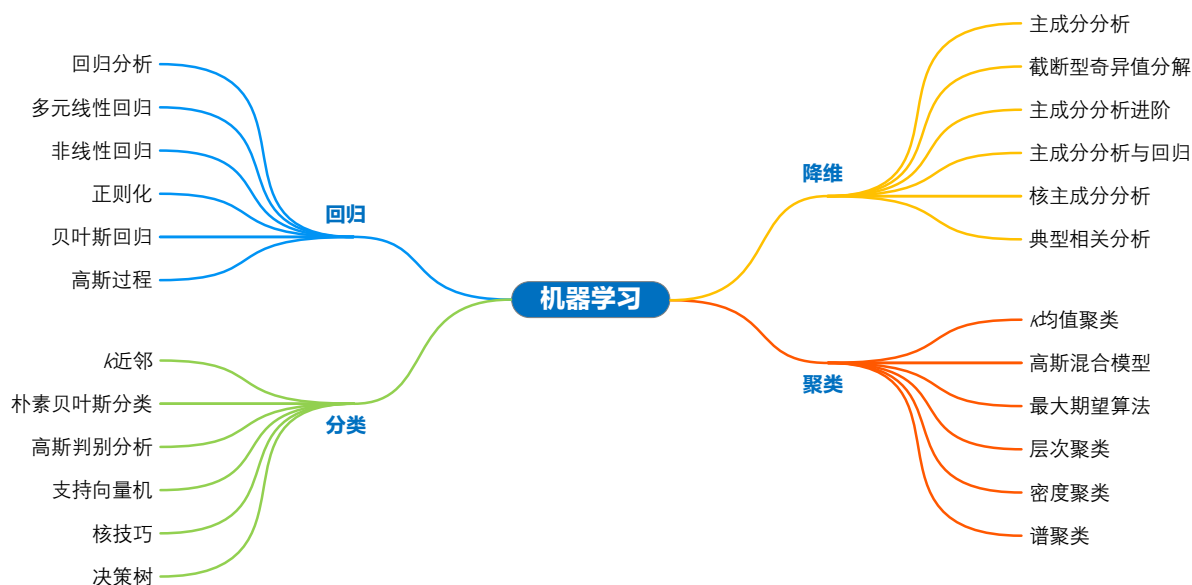


图 2. 《机器学习》板块布局

本书第 1 章不属于上述任何一个板块，这章相当于是“综述”，和大家聊聊机器学习、人工智能、深度学习之间的关系，然后盘点《机器学习》要介绍的核心内容。

回归

第 2 章首先利用一元 OLS 线性回归讲解回归分析，本章中大家会学到方差分析、拟合优度、 F 检验、 t 检验、置信区间、预测区间、对数似然函数、信息准则等概念。这一章相对较为无聊，建议大家学习时没有必要全部掌握。实践时再回来有针对性地学习。

第 3 章讲解多元线性回归，回归分析的维度提高。这一章请大家多从几何、数据视角思考回归分析。

第 4 章讲解非线性回归，需要大家掌握多项式回归，并理解过拟合。此外，这一章还介绍了逻辑回归，逻辑回归既可以用来回归分析，也可以用来分类。

第 5 章利用正则化解决多元线性回归过拟合、多重共线性的问题。这一章一共介绍三种正则化：a) 岭回归；b) 套索回归；c) 弹性网络回归。

第 6 章介绍如何将贝叶斯推断用在回归分析中。学习这一章时，建议大家回顾《统计至简》第 20 ~ 22 章。这一章最后从贝叶斯推断视角理解正则化。

第 7 章介绍高斯过程，这种算法集合了高斯分布、条件概率、协方差矩阵、随机过程等数学工具，理解上不是很容易。高斯过程可以解决分类、回归两类问题。

分类

第 8 章首先介绍 k 近邻算法，这个算法基本思想是“小范围投票，少数服从多数”，它可以用来分类，也可以用来回归。这章还总结了常用距离度量。

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

第 9 章介绍朴素贝叶斯分类。有关朴素贝叶斯分类算法，希望大家记住“假设特征之间条件独立，最大化后验概率”。

第 10 章介绍高斯判别分析，算法特点是“假设后验概率为高斯分布，最小化分类错误”。线性判别、二次判别都包含在高斯判别之中。

第 11、12 章介绍支持向量机。支持向量机的特点是间隔最大化，支持向量确定决策边界。第 12 章着重介绍核技巧，将样本数据映射到高维特征空间中，使数据在高维空间中线性可分。本书后续还会用到核技巧，需要大家注意。

支持向量机既可以用来分类，也可以用来回归。想要理解支持向量机绝对离不开《矩阵力量》中各种线性代数工具，特别是《矩阵力量》第 19 章内容、以及有关格拉姆矩阵的知识。

第 13 章讲解决策树，大家注意理解信息熵、信息增益等概念。

降维

第 14、15、16 章讲解主成分分析。第 14 章侧重从利用特征值分解完成主成分分析；第 15 章则介绍用截断型奇异值分解完成主成分分析；第 16 章则区分六种不同的技术路线。鸢尾花书在不同的板块都或多或少地介绍过主成分分析，这样安排的目的是当大家从线性代数、概率统计、优化、数据等不同角度透彻理解主成分分析。对读者来说，这种抽丝剥茧、逐层深入的讲解方式，不至于信息过载。

第 17 章分别介绍以主成分分析为基础的两回归方法：正交回归、主元回归。虽然这章介绍的是回归方法，但是它们都离不开主成分分析。此外，第 17 章还简单介绍了偏最小二乘回归。

第 18 章讲解核主成分分析，这种方法也用到了核技巧。核主成分分析是一种在高维数据中提取关键信息的统计技术。它通过转换原始特征，寻找数据中的主要结构，减少数据维度。与传统主成分分析不同，核主成分分析能有效处理非线性关系，为数据降维提供更灵活的手段，广泛应用于模式识别、特征提取等领域。

第 19 章介绍典型相关分析。典型相关分析方法的目的是找到两组数据的整体相关性的最大线性组合。

无监督学习

第 20 章介绍 k 均值聚类，算法特点是簇内距离和最小、迭代求解。注意， k 均值聚类的 k 不同于 k NN 中的 k 。

第 21 章介绍高斯混合模型。高斯混合模型组合若干高斯分布，期望最大化。高斯混合模型求解离不开第 22 章讲解的最大期望算法。最大期望算法的特点是迭代优化两步走：E 步，M 步；最大化对数似然函数。

第 23 章介绍层次聚类。层次聚类基于数据之间距离，自下而上聚合，或自上而下分裂。

第 24 章介绍密度聚类 DBSCAN，算法特点是利用数据分布紧密程度聚类。Scikit-learn 中 OPTICS 算法类似 DBSCAN。

第 25 章讲解谱聚类。谱聚类通过构造无向图，降维聚类。本章略微介绍有关图论的内容，但是没有展开。

0.3 特点：经典 + 综述

机器学习、深度学习算法不断涌现，让人目不暇接。限于作者知识水平、本书篇幅，本册在选取算法模型的标准只有一个——经典。从“经典”算法角度切入，《机器学习》目标是覆盖 Scikit-learn 库的常用机器学习算法函数，让大家充分理解算法理论，又能理论联系实际。

因此，在学习本书时，特别希望大家不仅仅满足于会“调包”，也就是调用 Scikit-learn 各种函数，更要理解这些算法背后的数学理论。因此，本书给出适度的数学推导以及扩展阅读。

《机器学习》并不“贪大求全”，目的就是让大家完全理解这 24 个经典算法，确保现在“充分掌握”，方便日后“有机提升”。

不“贪大求全”，这也是《机器学习》的短板。本书不涉及集成学习、神经网络、强化学习、深度学习、自然语言处理等话题。其次，本书也不涉及机器学习理论。虽然《数据有道》一册介绍过很多特征工程的工具，但是本书没有专门讲解特征工程章节。还有，本书也没有讨论如何部署机器学习模型。这些话题留给大家“按需”探索学习。

最后，欢迎大家来到“鸢尾花书”的收官之旅——《机器学习》！