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 本册在全套丛书的定位

“鸢尾花书”有三大板块——编程、数学、实践。数据科学、机器学习各种算法都离不开数学，因此“鸢尾花书”在数学板块着墨颇多。

本册《统计至简》是“数学三剑客”的第三本，也是最后一本。“数学”板块的第一本《数学要素》是各种数学工具的“大杂烩”，可谓数学基础。第二本《矩阵力量》专门讲解机器学习中常用的线性代数工具。本册《统计至简》则介绍机器学习和数据分析中中常用的概率统计工具。

《统计至简》的核心是“多元统计”，离不开《矩阵力量》中介绍的线性代数工具。在开始本册内容学习之前，请大家务必掌握《矩阵力量》的主要内容。

在完成本册《统计至简》学习之后，我们便正式进入“实践”板块，开始《数据有道》、《机器学习》两册的探索之旅。

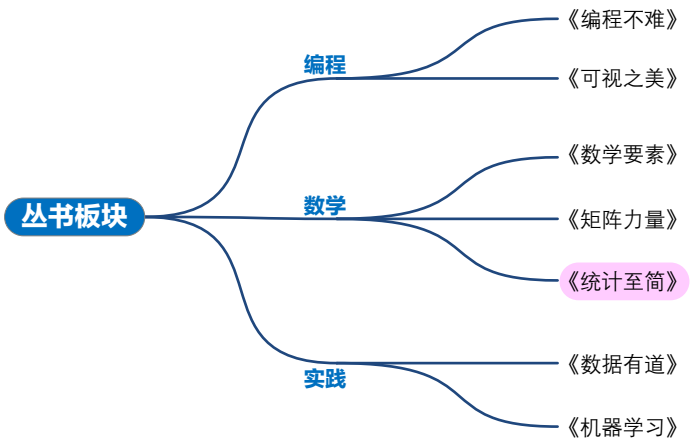

图 1. 本系列丛书板块布局

0.2 结构：7 大板块

本书可以归纳为 7 大板块——统计、概率、高斯、随机、频率派、贝叶斯派、椭圆。

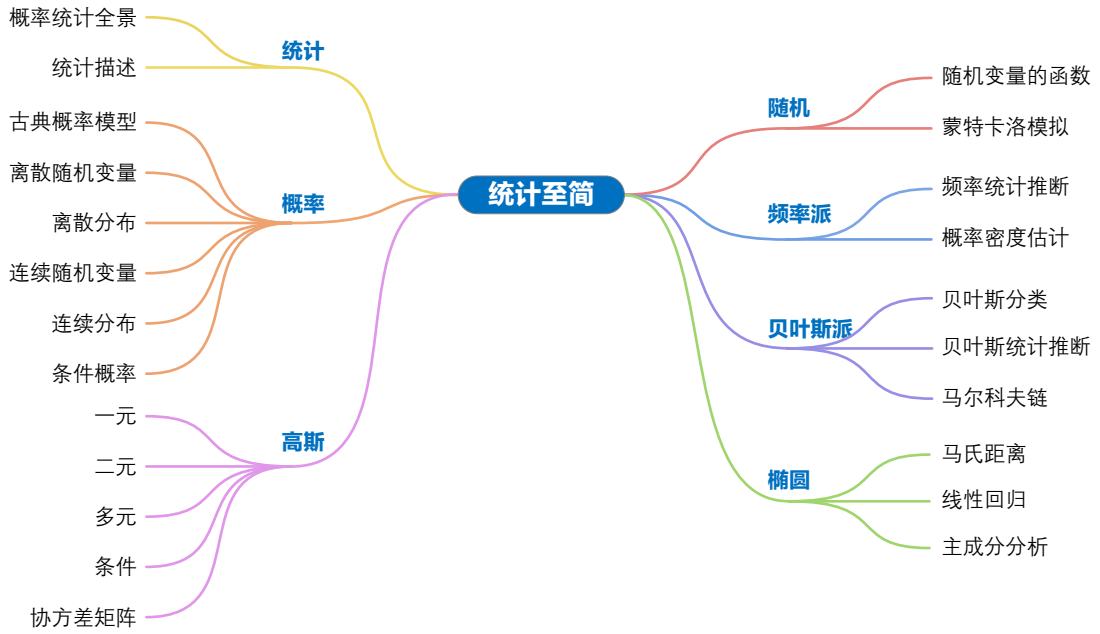

图 2.《统计至简》板块布局

统计

本书第 1 章可能是整个“鸢尾花书”系列中“最无聊”的一章。这章首先给大家出了一个线性代数的小测验，如果顺利通过测验的话就可以开始本册内容学习。如果不顺利，建议大家回顾《矩阵力量》一册相关内容。然后，这章总结了《统计至简》重要的公式，大家可以把这些内容当成“公式手册”来看待，学习本册时或学完本册后回看参考。

第 2 章介绍统计描述。这一章用图像、量化汇总等方式描述样本数据重要特征。学习这章时，建议大家回顾《矩阵力量》第 22 章。

概率

概率是统计推断的基础数学工具。“概率”这个板块将主要介绍离散、连续两大类随机变量及常见的概率分布。

第 3 章介绍古典概型，重中之重是贝叶斯定理。本书“厚”贝叶斯派，“薄”频率派，因此本书很多内容在展示贝叶斯定理的应用。希望大家从第 3 章开始就格外重视贝叶斯定理。

第 4、5 两章介绍离散随机变量、离散分布。第 6、7 两章介绍连续随机变量、连续分布。第 4、6 章特别用鸢尾花数据为例讲解随机变量，建议大家对比阅读。第 8 章特别介绍离散、连续随机变量的条件期望、条件方差。学习各种分布时，请大家格外注意它们的 PDF、CDF 形状。二项分布、多项分布、高斯分布、Dirichlet 分布这几种分布将会在本书后续发挥重要作用，希望大家留意。

此外概率"这个板块强调，概率质量函数、概率密度函数无非就是对 1 (样本空间对应的概率)的不同“切片、切块”、“切丝、切条”方式，请大家注意理解。

高斯

“高斯”是数据科学、机器学习算法中如雷贯耳的名字，大家会在回归分析、主成分分析、高斯朴素贝叶斯、高斯过程、高斯混合模型等等算法中遇到高斯分布。因此本书中高斯分布“戏份”格外吃重。

“高斯”这一板块分别介绍一元 (第 9 章)、二元 (第 10 章)、多元 (第 11 章)、条件高斯分布 (第 12 章)。几何视角是理解高斯分布的利器，大家学习这几章时，请特别注意高斯分布、椭圆、椭球之间的联系。第 13 章则介绍高斯分布中的重要成分——协方差矩阵。

这个板块，特别是在讲解多元高斯分布、协方差时，大家会看到无所不在的线性代数。

随机

第 14 章介绍随机变量的函数，请大家特别注意从几何视角理解线性变换、主成分分析。第 15 章讲解几个蒙特卡罗模拟试验，请大家掌握产生满足特定相关性的随机数的两种方法。这两种方法分别对应《矩阵力量》中介绍的 Cholesky 分解、特征值分解，建议大家在学习时回看《矩阵力量》相关内容。

频率派

本书中有关频率派的内容着墨较少，这是因为机器学习中贝叶斯统计应用场合更为广泛。第 16 章介绍常见经典统计推断方法，请大家务必掌握最大似然估计 MLE。第 17 章讲解概率密度估计，请大家特别注意高斯核概率密度估计。

贝叶斯派

这一部分先从贝叶斯分类开始。第 18、19 章介绍如何利用贝叶斯定理完成鸢尾花分类，请大家掌握后验概率、证据因子、先验概率、似然概率这些概念。在贝叶斯分类算法中，优化问题可以最大化后验概率，也可以最大化联合概率，即“似然概率 × 先验概率”。注意，《机器学习》会深入介绍“朴素贝叶斯分类”算法。

第 20、21 章讲解贝叶斯派推断。贝叶斯推断所体现出来的“学习过程”和人类认知过程极为相似，请大家注意类比。贝叶斯推断把总体的模型参数看作随机变量。贝叶斯推断中，后验 $\propto$ 似然 × 先验，无疑是最重要的关系。请大家务必掌握最大后验概率 MAP。

第 22 章简单介绍 Metropolis-Hastings 采样，并讲解如何使用 Pymc3 获得服从特定后验分布的随机数。

椭圆

本书最后一个板块可以叫“椭圆三部曲”，因为最后三章都和椭圆有关。这三章也开启了下一册《数据有道》三个重要话题——数据处理、回归、降维。

第 23 章讲解马氏距离，请大家特别注意马氏距离、欧氏距离、标准化欧氏距离的区别，以及马氏距离和卡方分布的联系。

第 24 章中，我们将从最小二乘法 OLS、优化、投影、线性方程组、条件概率、最大似然估计 MLE 这几个视角讲解线性回归。这一章相当于是《数学要素》第 24 章的扩展。

预告一下，《数据有道》将铺开介绍更多回归算法，比如多元回归分析、正则化、岭回归、套索回归、弹性网络回归、贝叶斯回归、多项式回归、逻辑回归，以及基于主成分分析的正交回归、主元回归等算法。

第 25 章以概率统计、几何、矩阵分解、优化为视角介绍主成分分析。《数据有道》将会深入讲解主成分分析，以及典型性分析、因子分析。

0.3 特点：多元统计

《统计至简》一册最大特点就是，多元统计。

当前多数概率统计教材都侧重于“一元”，而数据科学、机器学习中处理的问题几乎都是多特征，即“多元”。从一元到多元，有一道鸿沟。能帮助我们跨越这道鸿沟的正是线性代数工具。这就是为什么一再强调大家要学好《矩阵力量》之后再开始本书学习。

概率统计是个庞杂的知识系统，本书只能选取机器学习中最常用的数学工具。“大而全”的数学公式手册范式不是本书的追求，这也就是本书书名“至简”二字的来由。本书“轻巧”知识体系骨架足够撑起丛书后续数据科学、机器学习内容，也方便大家进一步扩展填充。

本书“繁复”的一点是丰富的实例和可视化方案，它们可以帮助大家理解常用概率统计工具，力求让大家学透每一个公式。学习《统计至简》时，请大家注意使用几何视角，提升自己空间想象力。

阅读本册时，大家注意两个“斯”——高斯、贝叶斯。高斯分布可能是最重要的连续随机变量分布。本书把高斯分布从一元扩展到多元，关键在掌握多元高斯分布。此外，全书每个板块几乎都有“贝叶斯定理”投下的影子。

“图解 + 编程 + 机器学习应用”是丛书的核心特点，本册也不例外。这套书用“编程 + 可视化”取代“习题集”。为了达到更好的学习效果，希望大家一边阅读，一边编程实践。

大多数概率统计的图书给大家的印象是公式连篇。《统计至简》为了打破这种刻板印象，尝试直接给核心公式“配图”，以强化理解。这也是本册的一个小实验，效果好的话再版的时候将推广应用到“鸢尾花书”其他分册。

此外，鸡、兔、猪这三个“小伙伴”也会来到《统计至简》客串出演，帮助大家理解复杂的概率统计概念。

“有数据的地方，必有统计！”

在《统计至简》这本书中，大家会看到微积分、线性代数、概率统计等数学工具“济济一堂”，但是没有丝毫的违和感！

下面，我们就开始“数学三剑客”的收官之旅！