

36

Build Machine Learning Apps Using Streamlit

Streamlit 搭建机器学习 Apps

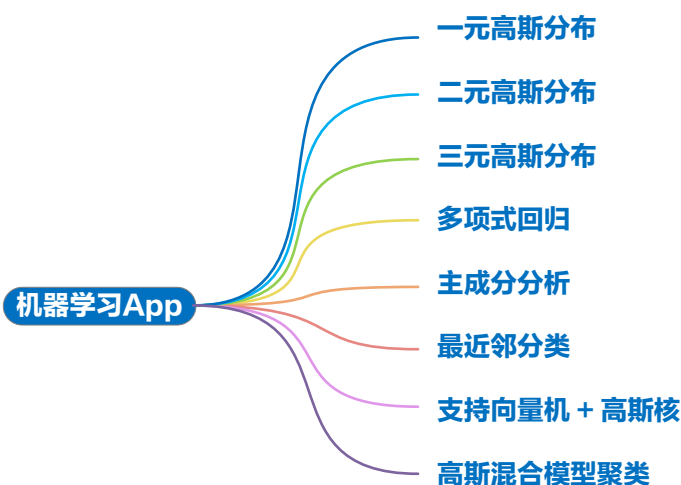
统计描述、数据可视化、概率模型、随机过程模拟



一片幽林，野径两条；而我
踏上了人迹罕至的那条。人生轨迹的
千差万别，由此而起。

*Two roads diverged in a wood, and I,
I took the one less traveled by,
And that has made all the difference.*

—— 罗伯特·佛洛斯特 (Robert Frost) | 美国诗人 | 1874 ~ 1963



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

36.1 搭建应用 App: 编程 + 数学 + 可视化 + 机器学习

本书最后一章用 Streamlit 搭建了 8 个机器学习 App，用来总结本书前文讲解的主要内容。本章正文不提供 Python 代码，请大家用 Spyder 自行打开配套代码查看并逐行注释。

此外，请大家按照上一章介绍的方法打开这几个 App，并想办法根据本书前文所学丰富这些 App 的功能。

36.2 一元高斯分布

大家可能好奇，高斯分布可以归类到概率统计板块，也不是机器学习算法。

是这样！

但是，机器学习很多机器学习算法都离不开高斯分布。

本书前文提过，**高斯朴素贝叶斯** (Gaussian Naive Bayes)、**高斯判别分析** (Gaussian discriminant analysis)、**高斯过程** (Gaussian process)、**高斯混合模型** (Gaussian mixture model)，甚至是协方差估计、随机数发生器、回归分析、主成分分析、马氏距离，也都和高斯分布有着千丝万缕的联系。

因此，把高斯分布搞的清清楚楚、明明白白格外重要。

本章用了三节分别设计了三个 Apps，分别展示一元、二元、三元高斯分布。



鸢尾花书《统计至简》将专门介绍高斯分布。

简单来说，**一元高斯分布** (univariate Gaussian distribution) 是一种对称的概率分布，常见于自然界和统计学中。它呈钟形曲线，数据集中在均值附近，随着距离均值的增加而减小。

图 1 所示为一元高斯分布概率密度函数曲线的 App。这个 App 很简单，我们通过调节期望、标准差来观察概率密度函数 PDF 曲线的变化。本书前文提过，期望影响图中曲线的位置，而标准差影响曲线“高矮胖瘦”。



请大家利用 `numpy.random.normal()` 生成服从 App 中输入参数的一元高斯分布的随机数，在 App 中绘制随机数分布的直方图。然后，自行了解什么是一元高斯分布的 CDF，然后用 `scipy.stats.norm()` 在图 1 中增加一幅图，展示 CDF 曲线随期望、标准差变化。

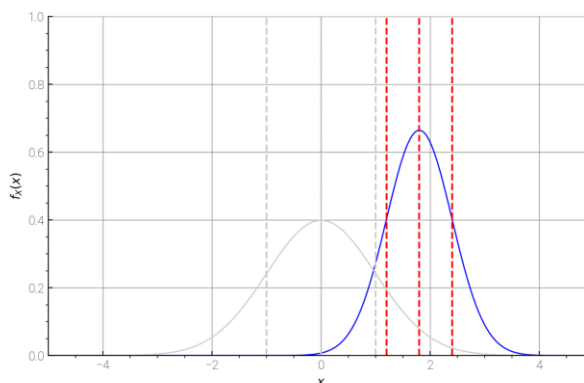
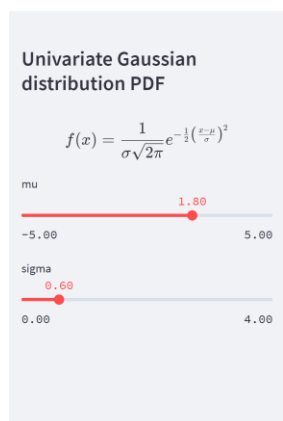


图 1. 一元高斯分布 App | Bk1_Ch36_01.py

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>

本书配套微课程视频均发布在 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

36.3 二元高斯分布

图 2 所示**二元高斯分布** (bivariate Gaussian distribution) 的 PDF 曲面和椭圆紧密相连。质心 (期望值向量) 影响图中椭圆位置, 而协方差 $\begin{bmatrix} \sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 \\ \rho_{1,2}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$ 则影响椭圆的形状。

鸢尾花书会帮助大家“吃透”二元高斯分布, 因为这个分布可以帮助我们理解圆锥曲线、几何操作 (平移、旋转、缩放、剪切)、协方差矩阵、特征值分解、马氏距离、离群值、卡方分布、主成分分析、回归分析等等。

特别是, 特征值分解得到的特征向量告诉我们椭圆的长轴、短轴方向, 特征值和长半轴、短半轴长度直接相关。

 请大家在 App 中显示协方差矩阵的具体值。用 `numpy.random.multivariate_normal()` 生成服从 App 中输入参数的二元高斯分布的随机数, 在 App 中用 `seaborn.jointplot()` 绘制随机数分布的散点图和边缘分布。

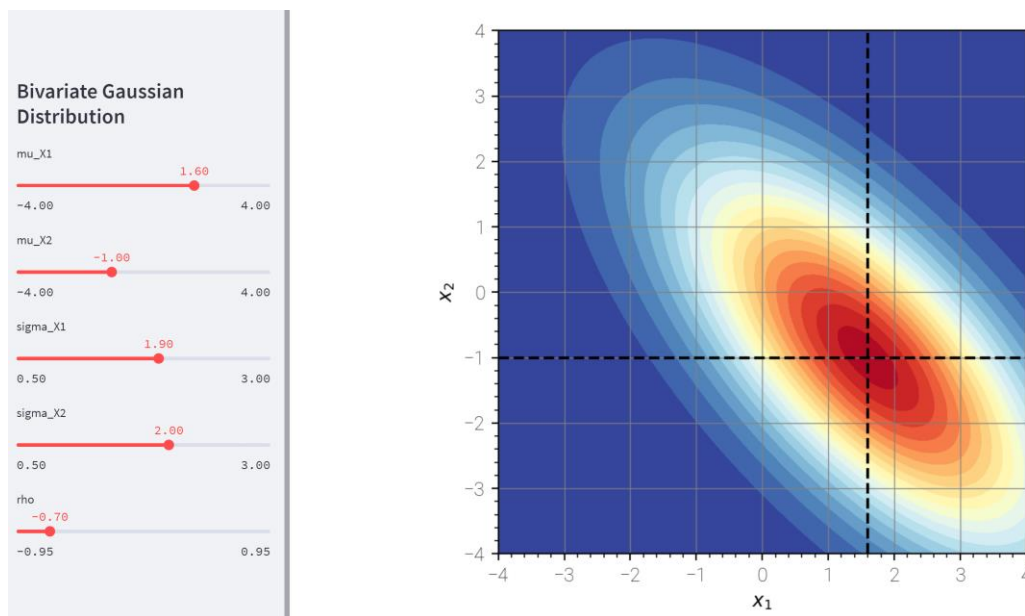




图 2. 二元高斯分布 App |  Bk1_Ch36_02.py

36.4 三元高斯分布

图 3 所示为用 `plotly.graph_objects.Volume()` 呈现**三元高斯分布** (trivariate Gaussian distribution), 请大家自行参考技术文档了解这个函数用法。几何角度来看, 三元高斯分布 PDF 相当于一层层椭球。

 鸢尾花书《统计至简》还会介绍如何将三元高斯分布椭球投影到不同平面, 以及用特征值分解帮我们找到椭球的主轴方向和半轴长度。

? 请大家也用 `numpy.random.multivariate_normal()` 生成服从 App 中输入参数的三元高斯分布的随机数，在 App 中用 `plotly.express.scatter_3d()` 绘制随机数分布的三维散点图。

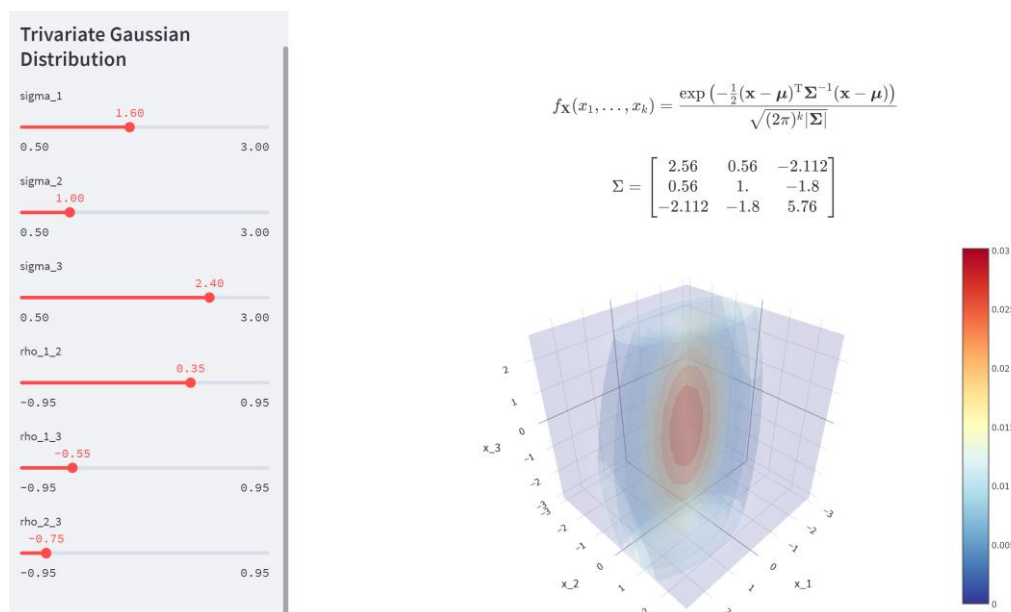


图 3. 三元高斯分布 App | `Bk1_Ch36_03.py`

36.5 多项式回归

图 4 所示为展示**多项式回归** (polynomial regression) 的 App，我们可以调节次数来观察拟合曲线变化。

简单来说，多项式回归利用多项式函数来拟合数据关系。

与线性回归不同，它可以捕捉到数据中的非线性模式，通过增加项的次数灵活地适应复杂模型。但是，随着次数增加，多项式回归模型容易出现过拟合。

➡ 鸢尾花书《数据有道》将详细介绍各种线性和非线性回归方法。

本书前文提过，所谓的**过拟合** (overfitting) 是一种机器学习模型过度学习训练数据的现象，导致在新数据上表现不佳。模型过于复杂，拟合了训练数据中的噪声和细节，严重影响**泛化能力** (generalization capability, generalization)。

而本书第 30 章介绍的**正则化** (regularization)，比如**岭回归** (ridge regression)，可以帮助我们降低过拟合的影响。

? 请大家在 App 左侧控制栏增加岭正则化的惩罚因子，用来抑制过拟合。

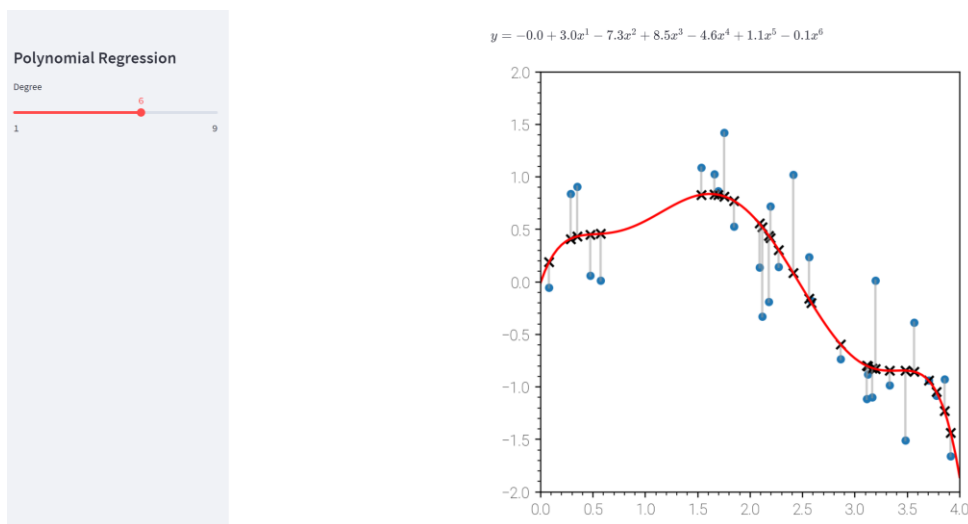


图 4. 多项式回归 App | Bk1_Ch36_04.py

36.6 主成分分析

图 5 所示为展示**主成分分析** (Principal Component Analysis, PCA) 的 App，我们可以通过调节主成分数量观察数据“复刻”情况。

主成分分析是一种重要的降维技术，通过找到数据中的主要特征，将信息压缩到较少的维度。它用于简化复杂数据集，保留关键信息。

想要理解主成分分析，就必须掌握**协方差矩阵估计** (Estimation of Covariance Matrix)、**特征值分解** (Eigen Value Decomposition, EVD)、**奇异值分解** (Singular Value Decomposition, SVD)，这是鸢尾花书后续要帮大家攻克的难关。



鸢尾花书《数据有道》将专门介绍主成分分析的不同技术路线以及其他降维方法。



请大家在 App 中增加散点图和热图两种可视化方案来比较原始数据和还原数据。

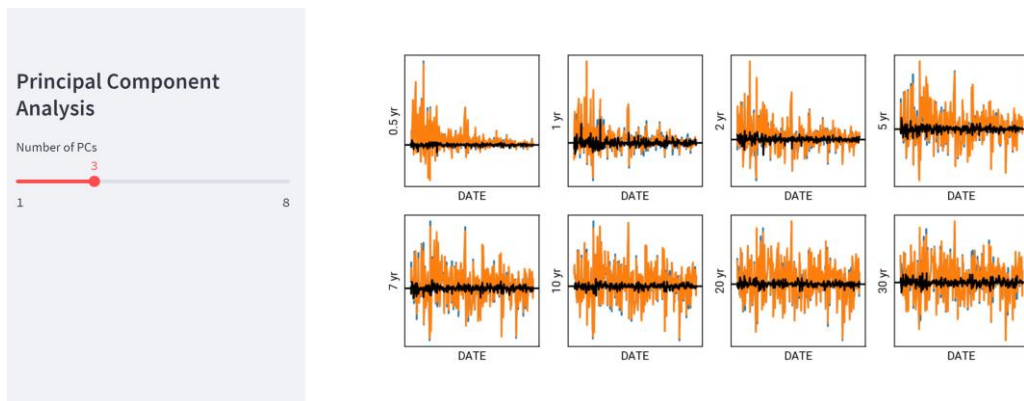


图 5. 主成分分析 App | Bk1_Ch36_05.py

36.7 最近邻分类

图 6 展示的是最 **k 最近邻分类** (k -Nearest Neighbors, kNN) 算法, 我们可以调节近邻数量来观察决策边界。本书前文提过, 最近邻分类算法实际上体现的就是“近朱者赤近墨者黑”这个朴素的思想。

➡ 鸢尾花书《机器学习》将专门介绍不同分类算法。

? 请大家在 App 中增加选项, 分别指定横轴、纵轴特征, 这两个特征数据将会被用来完成最近邻分类。

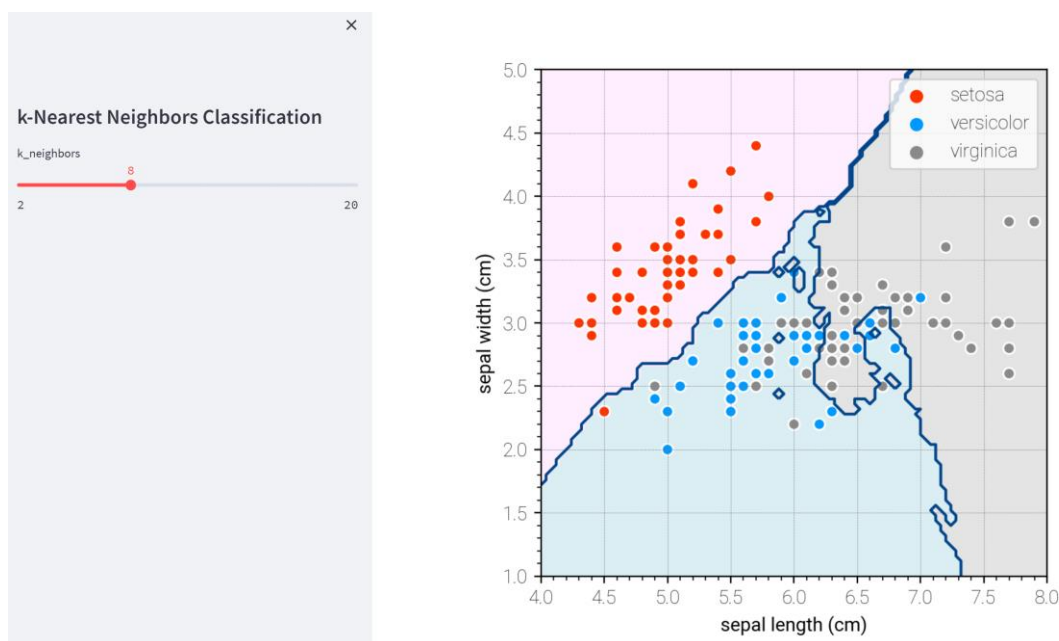


图 6. 最近邻分类 App | Bk1_Ch36_06.py

36.8 支持向量机 + 高斯核

图 7 所示为“支持向量机 + 高斯核”分类 App, 我们可以修改 γ 修改。

支持向量机 (Support Vector Machine, SVM) 可以用来完成分类和回归任务。

支持向量机的**高斯核** (Gaussian kernel), 也叫**径向基核** (radial basis function kernel), 是一种**核函数** (kernel function)。通过引入**核技巧** (kernel trick), 我们可以将数据映射到高维空间, 从而有效处理非线性关系。

➡ 鸢尾花书《机器学习》将专门介绍不同分类算法。

? Scikit-learn 中 SVD 算法函数 `sklearn.svm.SVC()` 中 `kernel` 参数主要有 'linear'、'poly'、'rbf'、'sigmoid' 这几个选择。请大家在 App 左侧增加一个选项卡用来选择不同的核。注

意，'poly' 是多项式核，默认的次数为 3，请大家增加一个选项用来调节多项式核次数。此外，请大家注意，参数 gamma 适用于 'poly'、'rbf'、'sigmoid' 这三个核函数。

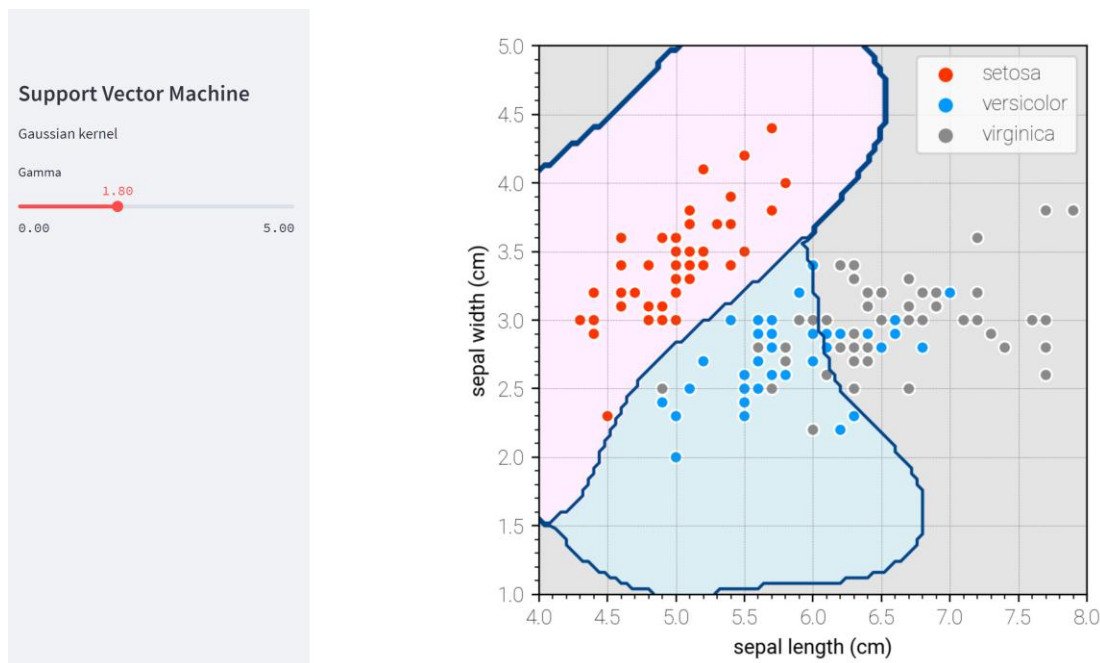


图 7. “支持向量机 + 高斯核”分类 App | Bk1_Ch36_07.py

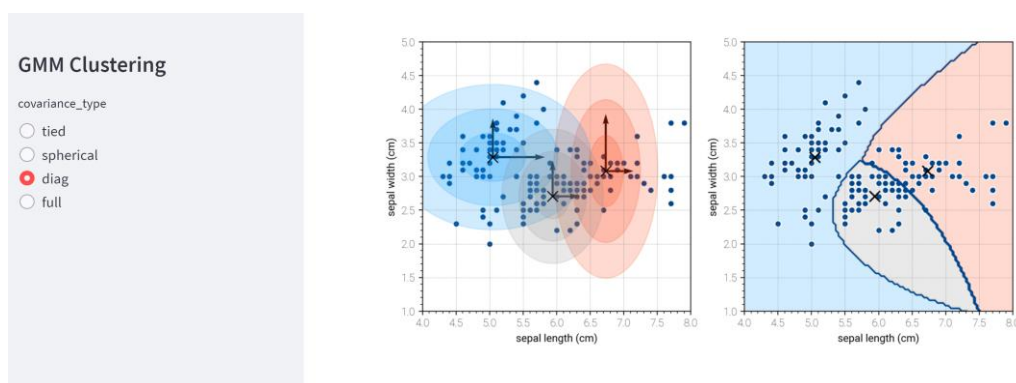
36.9 高斯混合模型聚类

图 8 所示为高斯混合模型 (Gaussian Mixture Model, GMM) 聚类 App，我们可以选择不同的协方差矩阵类型。

简单来说，高斯混合模型是一种概率模型，假设数据是由多个高斯分布组合而成。它常用于聚类和密度估计，灵活地适应不同形状和大小的数据簇。

➡ 鸢尾花书《机器学习》将专门介绍不同聚类算法。

? 类似前文，请大家在 App 中增加选项，分别指定横轴、纵轴特征，这两个特征数据将会被用来完成高斯混合模型聚类。



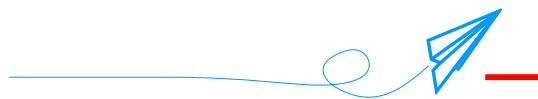
本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

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 8. 高斯混合模型聚类 App |  Bk1_Ch36_08.py

首先祝贺大家完成了《编程不难》的“修炼”!

作为鸢尾花书的第一本,《编程不难》相当于从“Python 编程”角度全景展示鸢尾花书整套内容;因此,《编程不难》内容跨度极大、涉猎话题广泛。

本书从零基础入门 Python 语法,到可视化,然后又介绍了各种数据处理方法以及完成复杂数学运算的工具,深入到常用机器学习算法,最后又聊了聊如何搭建 App 应用。

大家能够坚持到最后,实属不易!

希望大家读到这里,会有一种自信——Python 也不过如此嘛!

《编程不难》在开篇强调,本书只要求大家知其然,不需要大家知其所以然;即便如此,本书还是见缝插针地不用任何公式讲解了很多数学工具和算法。相信大家读完本册,数学修炼也有质的提高。请大格外家注意线性代数工具,尤其是矩阵乘法。

特别希望大家读完这本书后,开始试着利用几何图形来解释数学工具。这便引出鸢尾花书的下一分册——《可视之美》。

《可视之美》是鸢尾花书中一本真正意义的“图册”,她的目的只有一个——尽显数学之美!

《可视之美》会从美学角度展示科技制图、计算机图形学、创意编程、趣味数学实验、数学科学、机器学习等等内容。

请大家相信“反复 + 精进”的力量!让我们在《可视之美》一册,不见不散!