18 降维

Dimensionality Reduction

鸢尾花书有关降维算法模型的综述

人类的历史，本质上是思想的历史。

Human history is, in essence, a history of ideas.

—— 赫伯特·乔治·威尔斯 (Herbert George Wells) | 英国小说家和历史学家 | 1866 ~ 1946

- `sklearn.decomposition.PCA()` 主成分分析函数
- `sklearn.decomposition.TruncatedSVD()` 截断奇异值分解
- `sklearn.decomposition.FastICA()` 独立成分分析
- `sklearn.decomposition.IncrementalPCA()` 增量主成分分析

18.1 一张“降维”版图

机器学习中的降维 (dimensionality reduction) 是指将高维数据转换为低维数据的过程，即将包含大量特征的数据集通过某种方式转化为特征较少但仍能保留原有信息的数据集。降维的目的是减少特征数量，降低计算复杂度，并提高模型的准确性和泛化能力。降维算法在尽可能地保留数据的重要信息的同时，将高维数据映射到低维空间中。图 1 总结几种常见降维的算法。

相信大家对下面这几种算法已经熟悉：主成分分析 (Principal Component Analysis, PCA)、典型相关分析 (Canonical Correlation Analysis, CCA)。这幅图也是本章的思维导图。

本书前文介绍的线性判别分析 (Linear Discriminant Analysis, LDA) 也可以视作一种降维方法。本章还要简单介绍核主成分分析 (Kernel Principal Component Analysis, KPCA)、独立成分分析 (Independent Component Analysis)、流形学习 (Manifold Learning) 这几种方法。

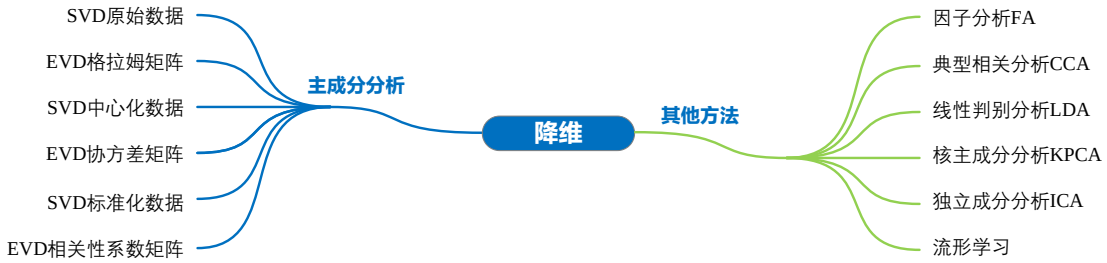

图 1. 降维方法分类

主成分分析

鸢尾花书对主成分分析着墨颇多。主成分分析是一种常用的数据降维方法，通过将高维数据投影到低维空间中，尽可能保留原始数据的重要信息。PCA 将原始数据的特征转换为新的特征，这些新特征按照重要性递减排列。通过选取前面的几个主成分，可以实现对数据的压缩和可视化。主成分分析常用于数据预处理、数据可视化和特征提取等领域。它能够剔除冗余的特征信息，简化数据模型，提高模型的效率和准确性，是机器学习中非常重要的技术之一。

和 OLS 线性回归类似，主成分分析也可以从几何 (图 2)、投影、数据、线性组合、特征值分解、SVD 分解、优化、概率统计等视角来理解。

《数据有道》第 18、19 两章还介绍利用主成分分析进行回归的两种方法：正交回归、主元回归。

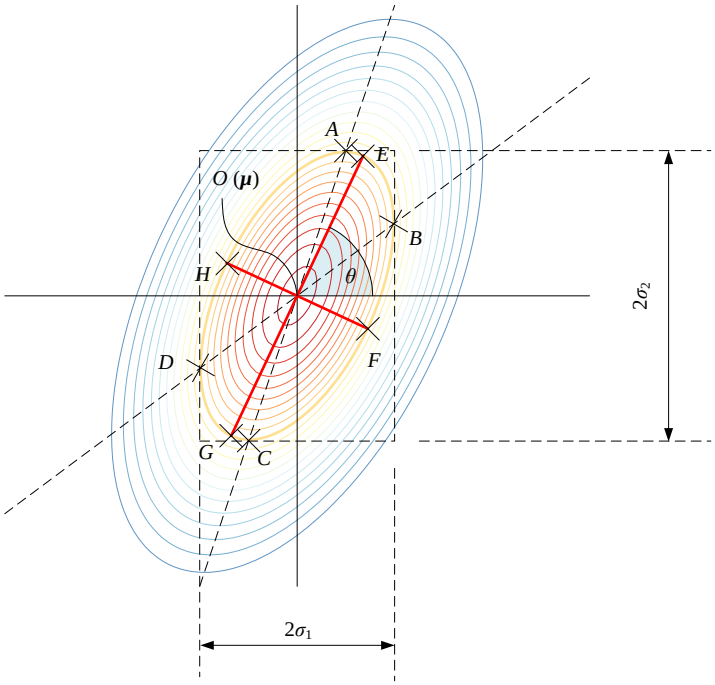

图 2. 主成分分析和椭圆的关系，图片来自《统计至简》第 25 章

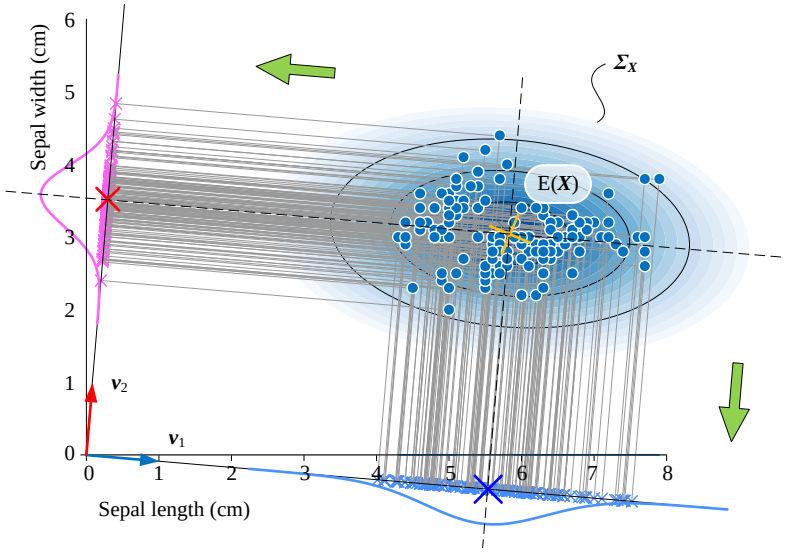

图 3. 投影视角看 PCA，图片来自《统计至简》第 14 章

此外，《数据有道》第 17 章还专门比较过主成分分析的六条技术路线，如表 1 所示。

表 1. 六条 PCA 技术路线，来自《矩阵分解》第 25 章

对象	方法	结果

原始数据矩阵 $\boldsymbol{X}$	奇异值分解	$\boldsymbol{X} = \boldsymbol{U_X S_X V_X}^{\mathrm{T}}$
格拉姆矩阵 $\boldsymbol{G} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 本章中用“修正”的格拉姆矩阵 $\boldsymbol{G} = \frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}}{n-1}$	特征值分解	$\boldsymbol{G} = \boldsymbol{V_X \Lambda_X V_X}^{\mathrm{T}}$
中心化数据矩阵 $\boldsymbol{X}_c = \boldsymbol{X} - \mathrm{E}(\boldsymbol{X})$	奇异值分解	$\boldsymbol{X}_c = \boldsymbol{U_c S_c V_c}^{\mathrm{T}}$
协方差矩阵 $\boldsymbol{\Sigma} = \frac{(\boldsymbol{X} - \mathrm{E}(\boldsymbol{X}))^{\mathrm{T}}(\boldsymbol{X} - \mathrm{E}(\boldsymbol{X}))}{n-1}$	特征值分解	$\boldsymbol{\Sigma} = \boldsymbol{V_c \Lambda_c V_c}^{\mathrm{T}}$
标准化数据 (z 分数) $\boldsymbol{Z}_{\boldsymbol{X}} = (\boldsymbol{X} - \mathrm{E}(\boldsymbol{X}))\boldsymbol{D}^{-1}$ $\boldsymbol{D} = \mathrm{diag}(\mathrm{diag}(\boldsymbol{\Sigma}))^{\frac{1}{2}}$	奇异值分解	$\boldsymbol{Z_X} = \boldsymbol{U_Z S_Z V_Z}^{\mathrm{T}}$
相关性系数矩阵 $\boldsymbol{P} = \boldsymbol{D}^{-1}\boldsymbol{\Sigma}\boldsymbol{D}^{-1}$ $\boldsymbol{D} = \mathrm{diag}(\mathrm{diag}(\boldsymbol{\Sigma}))^{\frac{1}{2}}$	特征值分解	$\boldsymbol{P} = \boldsymbol{V_Z \Lambda_Z V_Z}^{\mathrm{T}}$

奇异值分解

表 1 中前两种 PCA 方法，又叫截断奇异值 (truncated SVD)。

sklearn.decomposition.TruncatedSVD() 这个函数支持这两种技术路线。

《矩阵力量》第 16 章介绍了四种奇异值分解，图 4 ~ 图 7 展示了它们之间的关系。此外，请大家回顾《矩阵力量》第 6 章有关分块矩阵乘法相关内容。

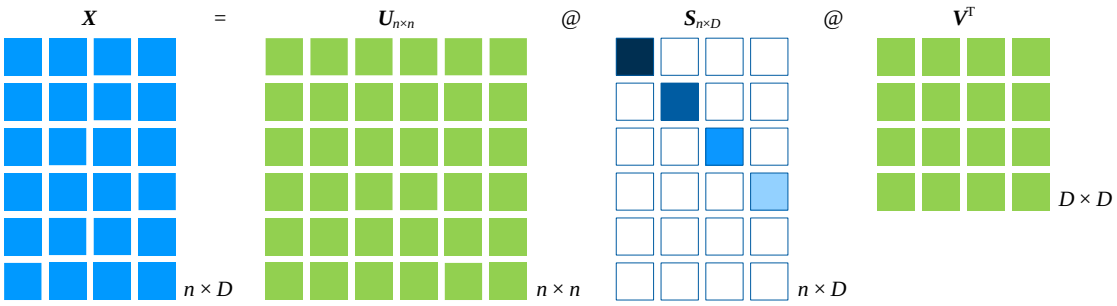

图 4. 完全型 SVD 分解

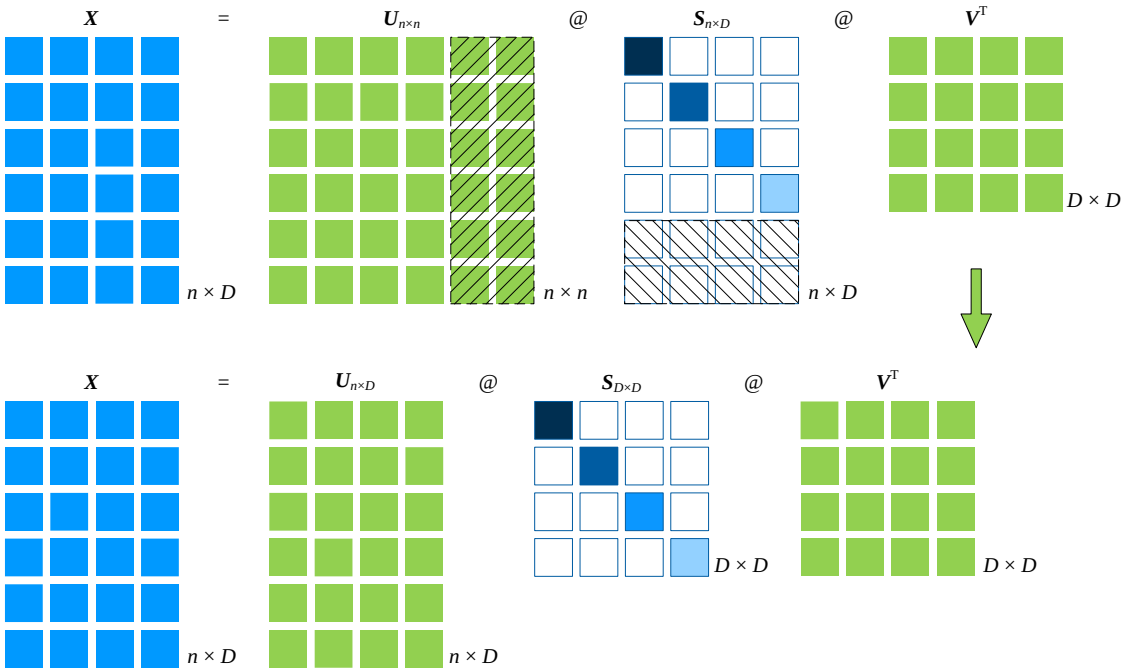

图 5. 从完全型到经济型

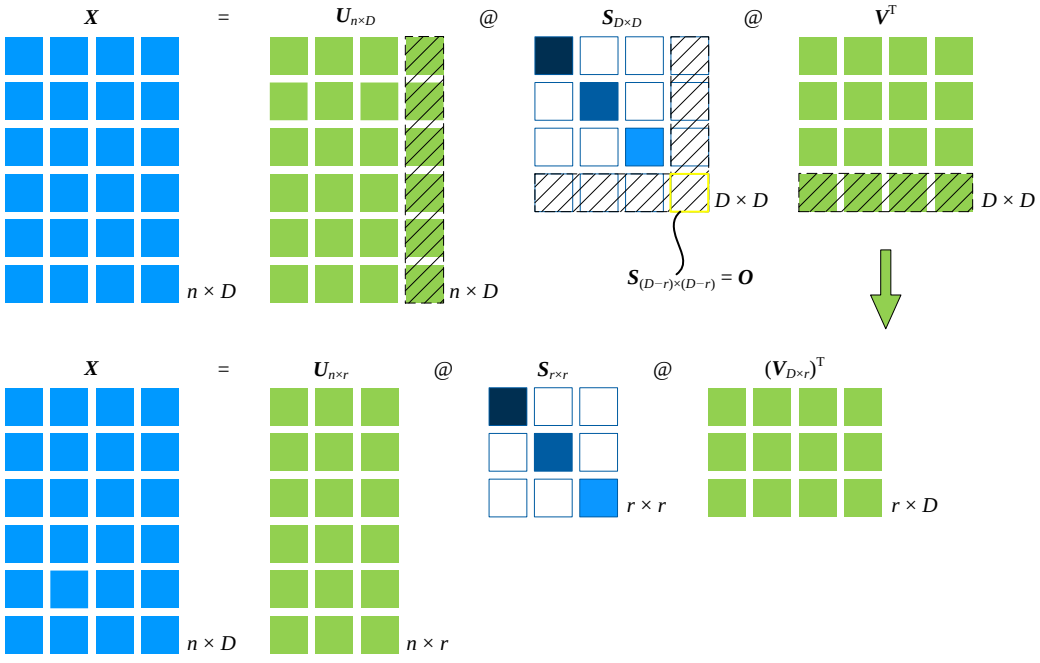

图 6. 从经济型到紧缩型

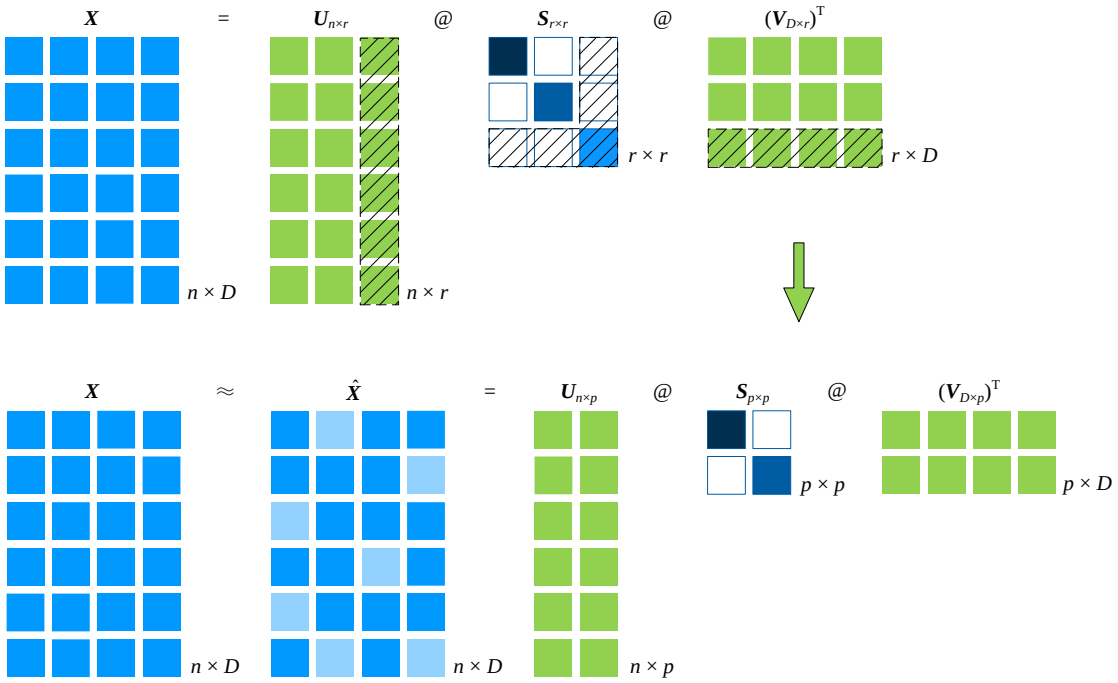

图 7. 从紧缩型到截断型

增量 PCA

当 PCA 需要处理的数据矩阵过大，以至于内存无法支持，可以使用增量主成分分析 (Incremental PCA, IPCA) 替代主成分分析。IPCA 分批处理输入数据，以便节省内存使用。Scikit-learn 中专门做增量 PCA 的函数为 sklearn.decomposition.IncrementalPCA()。

有关增量 PCA，大家可以参考下例：

https://scikit-learn.org/stable/auto_examples/decomposition/plot_incremental_pca.html

典型相关分析 CCA

典型相关分析也可以视作一种降维算法。典型相关分析是一种用于探究两组变量之间相关关系的统计方法，通常用于多个变量之间的关系分析。典型相关分析可以找出两组变量中最相关的线性组合，从而找到它们之间的相关性。典型相关分析的目的是提取出两组变量之间的共性信息，用于预测和解释数据。CCA 也可以从几何、数据、优化、线性组合、统计几个不同视角来理解。

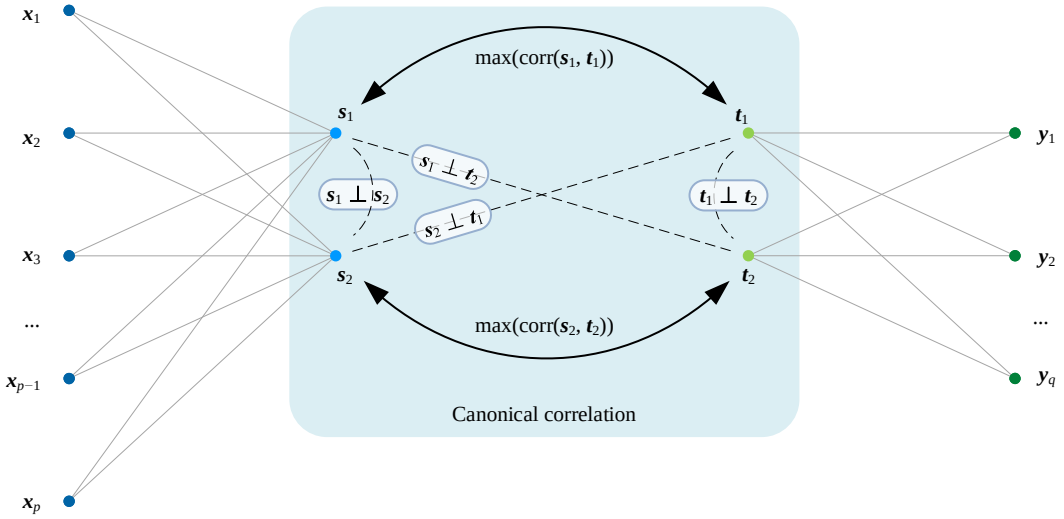

图 8. 线性组合角度看 CCA，图片来自《数据有道》第 20 章

下面这个例子比较偏最小二乘法 PLS、CCA，请大家参考：

https://scikit-learn.org/stable/auto_examples/cross_decomposition/plot_compare_cross_decomposition.html

18.2 核主成分分析

核主成分分析 (Kernel PCA) 是一种非线性的主成分分析方法，它通过使用核技巧将高维数据映射到低维空间中，从而提取出数据中的主要特征。与传统的 PCA 相比，Kernel PCA 可以更好地处理非线性数据，更准确地保留数据中的非线性结构。

可以这样理解，PCA 是 Kernel PCA 的特列。PCA 中用到的格拉姆矩阵、协方差矩阵、相关性系数矩阵都可以看成是不同线性核。

图 9 (a) 所示数据线性不可分，我们先用非线性映射把数据映射到高维空间，使其线性可分。利用 KPCA 之后的结果如图 9 (b)。这一点和支持向量机中的核技巧颇为类似。

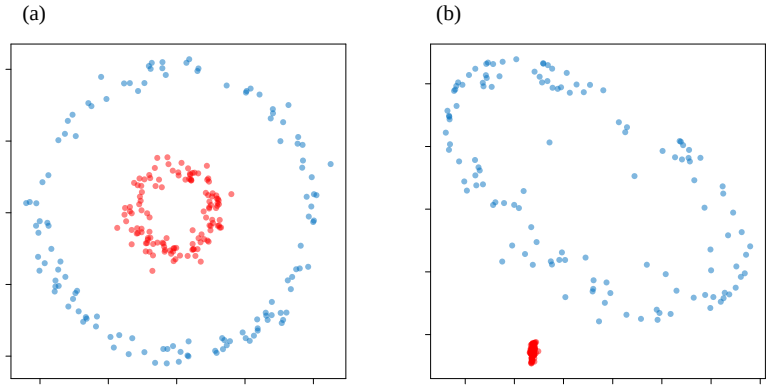

图 9. 核主成分分析

参考自如下示例，请大家自行学习：

https://scikit-learn.org/stable/auto_examples/decomposition/plot_kernel_pca.html

核主成分分析算法介绍，请参考：

https://people.eecs.berkeley.edu/~wainwrig/stat241b/scholkopf_kernel.pdf

18.3 独立成分分析

独立成分分析是一种用于从混合信号中恢复原始信号的数学方法。ICA 通过将混合信号映射到独立的成分空间中，从而恢复原始信号。独立成分分析将一个多元信号分解成独立性最强的可加子成分。因此，独立成分分析常用来分离叠加信号。

图 10 比较 PCA 和 CCA 对同一组数据的分解结果。与 PCA 不同的是，ICA 假设原始信号是独立的，而 PCA 假设它们是正交关系。

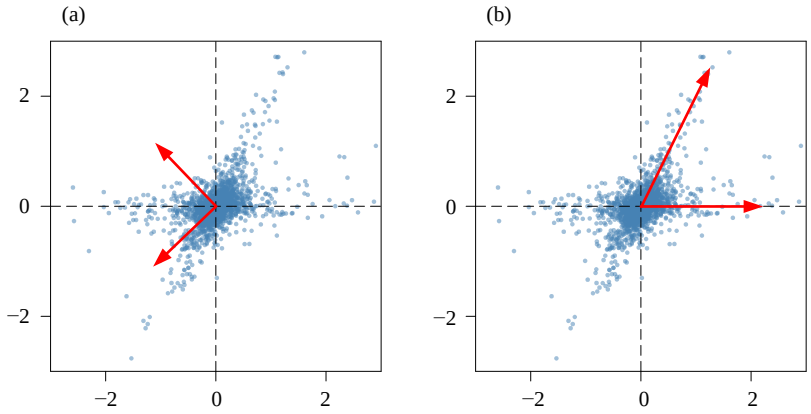

图 10. 比较 PCA 和 ICA

参考自如下示例，请大家自行学习：

https://scikit-learn.org/stable/auto_examples/decomposition/plot_ica_vs_pca.html

有关独立成分分析算法原理，请大家参考：

https://www.emerald.com/insight/content/doi/10.1016/j.aci.2018.08.006/full/html

18.4 流形学习

空间的数据可能是按照某种规则“卷曲”，度量点与点之间的“距离”要遵循这种卷曲的趋势。换一种思路，我们可以像展开“卷轴”一样，将数据展开并投影到一个平面上，得到的数据如图 12 所示。在图 12 所示平面上，A 和 B 两点的“欧氏距离”更好地描述了两点的距离度量，因为这个距离考虑了数据的“卷曲”。

流形学习 (manifold learning) 核心思想类似图 11 和图 12 所示展开“卷轴”的思想。流形学习用于发现高维数据中的低维结构，也是非线性降维的一种方法。于 PCA 不同的是，流形学习可以更好地处理非线性数据和局部结构，具有更好的可视化效果和数据解释性。

在 scikit-learn 中，流形学习的函数是 sklearn.manifold 模块中的 Isomap、LocallyLinearEmbedding、SpectralEmbedding 和 TSNE 等。其中，Isomap 使用测地线距离来保留流形上的全局结构，LocallyLinearEmbedding 使用局部线性嵌入来保留局部结构，SpectralEmbedding 使用谱分解来发现流形的嵌入表示，TSNE 使用高斯分布来优化样本的嵌入表示，用于可视化高维数据。这些函数提供了一种方便、高效、易于使用的流形学习工具，可帮助大家更好地理解数据结构和特征。本书不展开讲解流形学习。

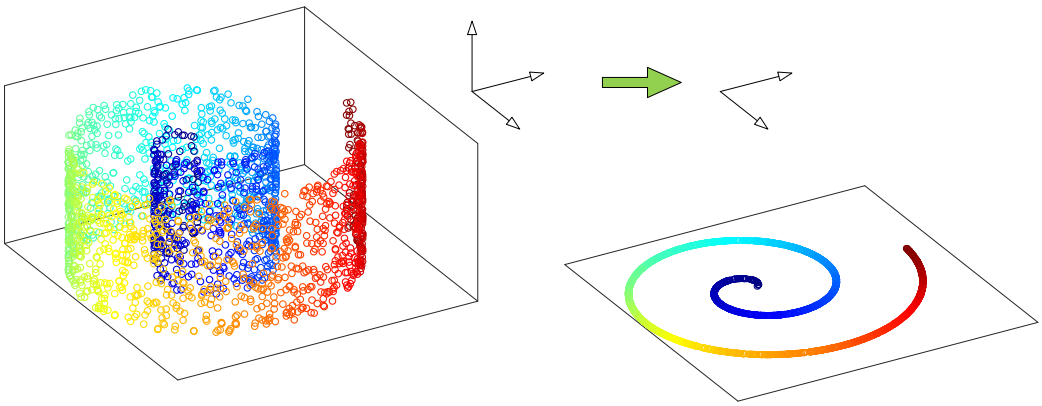

图 11. “卷曲”的数据

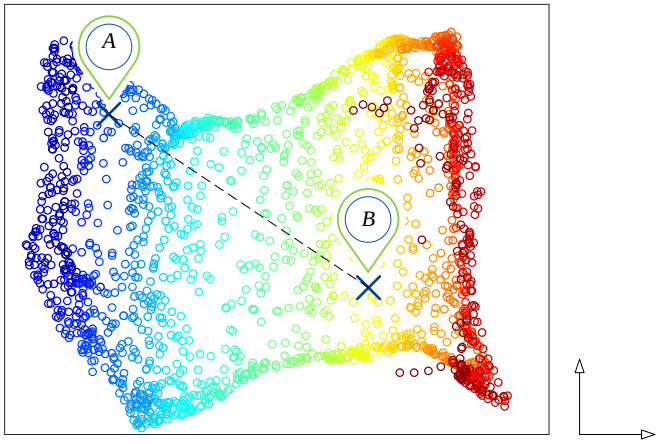

图 12. 展开“卷曲”的数据

机器学习中的降维是一种数据预处理技术，旨在通过减少特征数量来提高模型的训练效率和泛化能力。其中，主成分分析 PCA 是最常见的线性降维技术，前者通过将数据投影到最大方差方向上，后者则试图解释数据背后的因素。典型相关分析 CCA，是一种用于研究两个数据集之间的相关性的线性降维技术。它将两个数据集中的每个变量对应地进行线性组合，以使得这两个新的变量集之间的相关性最大。内核主成分分析 KPCA 是一种非线性降维方法，能够将数据映射到高维特征空间中，从而在新空间中找到数据的低维表示。独立成分分析 ICA 则试图从混合信号中恢复原始信号，假设原始信号是独立的。这些降维技术在机器学习、计算机视觉、信号处理、神经科学等领域都有广泛的应用，是提高数据处理和模型性能的重要工具。

Scikit-learn 中更多有关降维工具，请大家参考：

https://scikit-learn.org/stable/modules/decomposition.html

想要深入了解 Scikit-learn 中的流形学习工具，请大家参考：

https://scikit-learn.org/stable/modules/manifold.html

如下这篇文献介绍了流形学习的数学基础，请大家参考：

https://arxiv.org/pdf/2011.01307.pdf