Machine Learning in Scikit-Learn

28 Scikit-Learn 机器学习

利用 Scikit-Learn 库完成回归、降维、分类、聚类

合理即存在，存在即合理。

What is rational is actual and what is actual is rational.

—— 黑格尔 (Hegel) | 德国哲学家 | 1770 ~ 1831

28.1 什么是机器学习?

人工智能、机器学习、深度学习、自然语言处理

人工智能的外延十分宽泛，泛指指计算机系统通过模拟人的思维和行为，实现类似于人的智能行为。人工智能领域包含了很多技术和方法，如机器学习、深度学习、自然语言处理、计算机视觉等。

机器学习（Machine Learning，ML）是人工智能（Artificial Intelligence，AI）的一个子领域，是通过计算机算法自动地从数据中学习规律，并用所学到的规律对新数据进行预测或者分类的过程。本书这个板块将会着重介绍 Python 中 Scikit-Learn 这个机器学习工具。

深度学习是一种机器学习的子领域，它是通过建立多层神经网络模型，自动地从原始数据中学习到更高级别的特征和表示，从而实现对复杂模式的建模和预测。Python 中常用的深度学习工具有 TensorFlow、PyTorch、Keras 等，这些工具不在本书讨论范围内。

自然语言处理（Natural Language Processing，NLP）是计算机科学与人工智能领域的一个重要分支，旨在通过计算机技术对人类语言进行分析、理解和生成。自然语言处理主要应用于自然语言文本的处理和分析，如文本分类、情感分析、信息抽取、机器翻译、问答系统等。

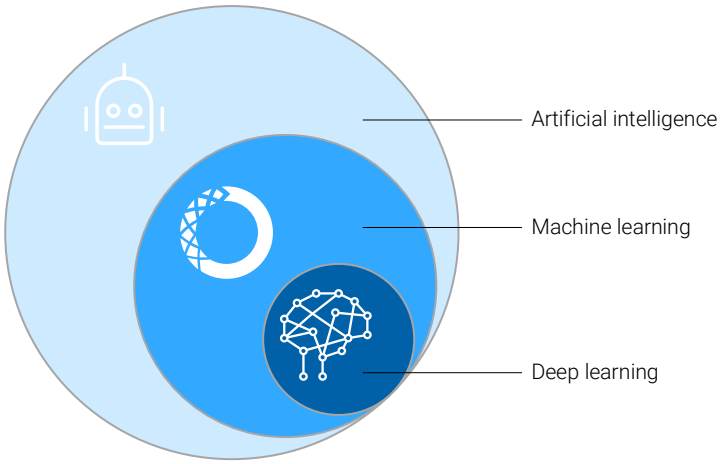


图 1．人工智能、机器学习、深度学习

机器学习适合处理的问题有如下特征：（a）大数据；（b）黑箱或复杂系统，难以找到**控制方程**（governing equations）。机器学习需要通过数据的训练。

机器学习分类

如图 2 所示，简单来说，机器学习可以分为以下两大类：

- **有监督学习**（supervised learning），也叫监督学习，训练有标签值样本数据并得到模型，通过模型对新样本进行推断。有监督学习可以进一步分为两大类：回归（regression），分类（classification）。本书第 30 章介绍常用回归算法，第 32 章介绍常用分类算法。

◀ **无监督学习**（unsupervised learning）训练没有标签值的数据，并发现样本数据的结构和分布。无监督学习可以分类两大类：降维（dimensionality reduction）、聚类（clustering）。本书第 31 章介绍常用降维算法，第 32 章介绍常用聚类算法。

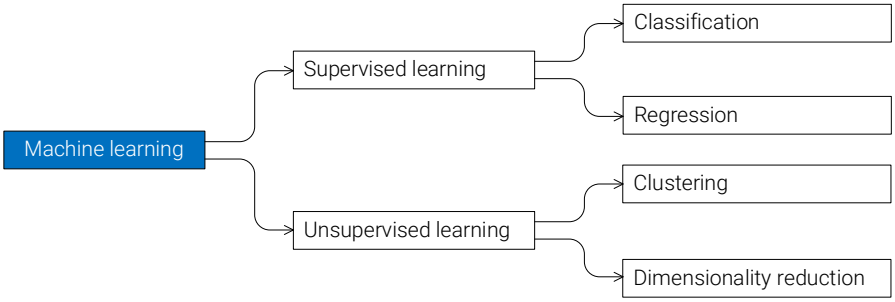


图 2．机器学习分类

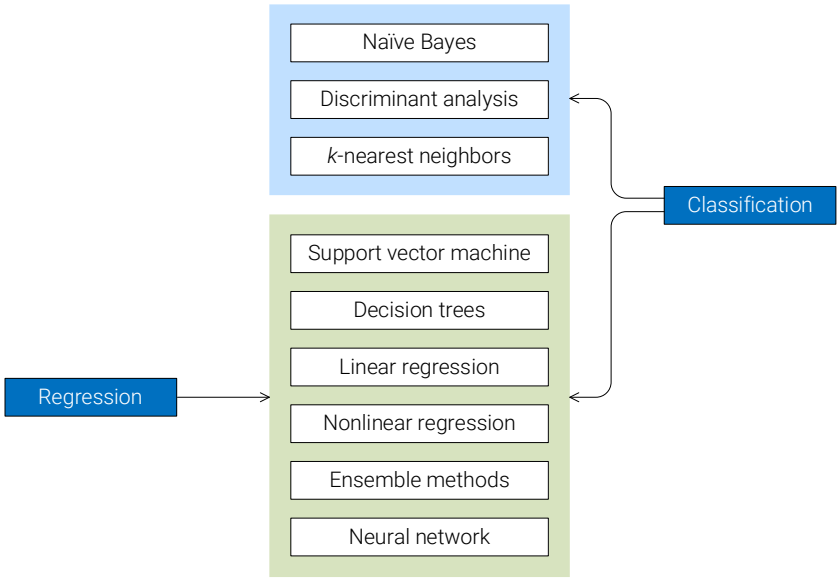


图 3．监督学习常见方法

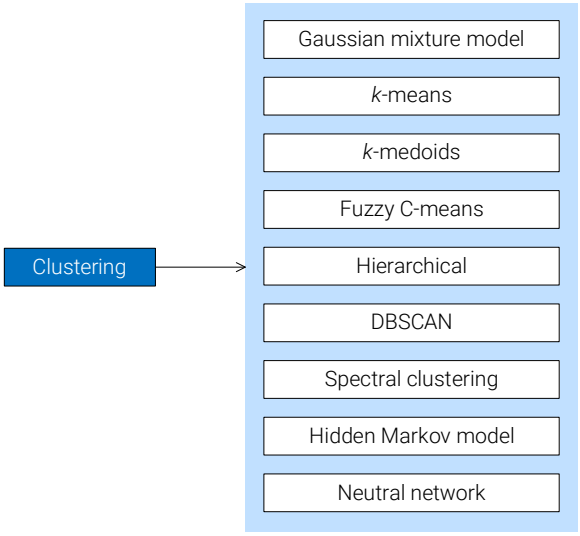


图 4．常用聚类方法

机器学习流程

图 5 所示为机器学习的一般流程。

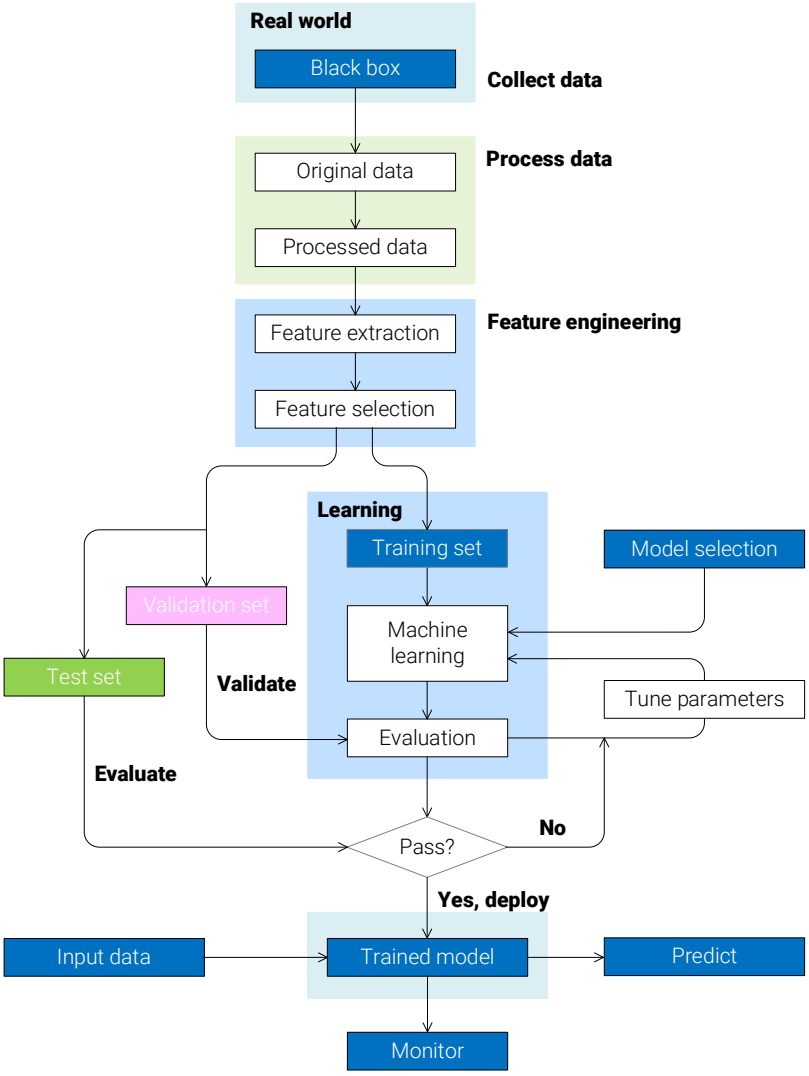


图 5．机器学习一般流程

具体分步流程通常包括以下步骤：

- **收集数据**：从数据源获取数据集，这可能包括数据清理、去除无效数据和处理缺失值等。
- **特征工程**：对数据进行预处理，包括数据转换、特征选择、特征提取和特征缩放等。
- **数据划分**：将数据集划分为训练集、验证集和测试集等。训练集用于训练模型，验证集用于选择模型并进行调参，测试集用于评估模型的性能。
- **选择模型**：选择合适的模型，例如线性回归、决策树、神经网络等。

- **训练模型**：使用训练集对模型进行训练，并对模型进行评估，可以使用交叉验证等方法进行模型选择和调优。
- **测试模型**：使用测试集评估模型的性能，并进行模型的调整和改进。
- **应用模型**：将模型应用到新数据中进行预测或分类等任务。
- **模型监控**：监控模型在实际应用中的性能，并进行调整和改进。

以上是机器学习的一般分步流程，不同的任务和应用场景可能会有一些变化和调整。在实际应用中，还需要考虑数据的质量、模型的可解释性、模型的复杂度和可扩展性等问题。

28.2 有标签数据、无标签数据

根据输出值有无标签，如图 6 所示，数据可以分为**有标签数据** (labelled data) 和**无标签数据** (unlabelled data)。鸢尾花数据显然是有标签数据。删去鸢尾花最后一列标签，我们便得到无标签数据。有标签数据和无标签数据是机器学习中常见的两种数据类型，它们在不同的应用场景中有不同的用途。

简单来说，有标签数据对应有监督学习，无标签数据对应无监督学习。

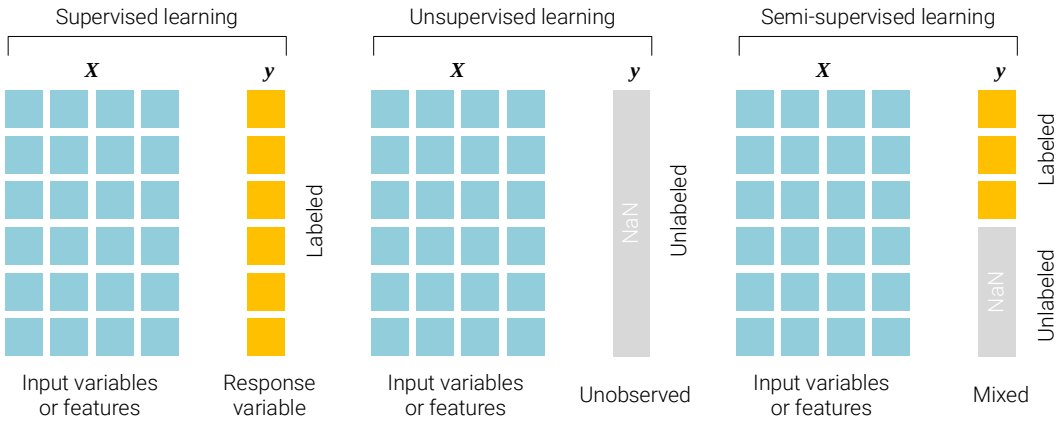


图 6. 根据有无标签分类数据

有监督学习中，如果标签为连续数据，对应的问题为**回归** (regression)，如图 7 (a)。如果标签为分类数据，对应的的问题则是**分类** (classification)，如图 7 (c)。

无监督学习中，样本数据没有标签。如果目标是寻找规律、简化数据，这类问题叫做**降维** (dimensionality reduction)，比如主成分分析目的之一就是找到数据中占据主导地位的成分，如图 7 (b)。如果模型的目标是根据数据特征将样本数据分成不同的组别，这种问题叫做**聚类** (clustering)，如图 7 (b)。

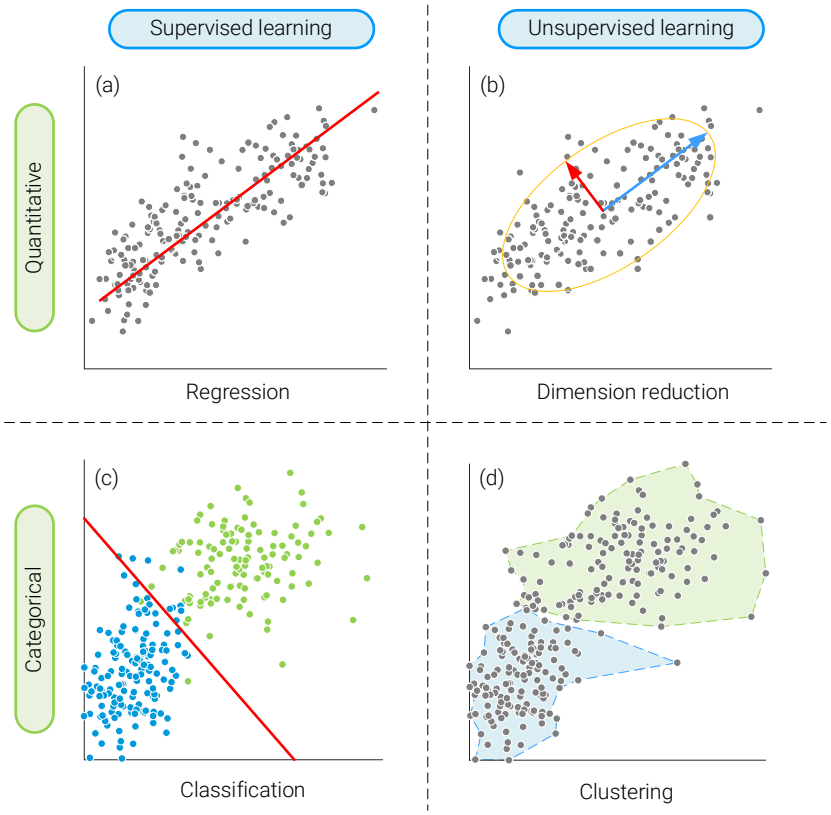


图 7. 根据数据是否有标签、标签类型细分机器学习算法

28.3 回归：找到自变量与因变量关系

回归是机器学习中一种常见的任务，用于预测一个连续变量的值。常见的回归算法包括线性回归、非线性回归、正则化、贝叶斯回归和基于分类算法的回归。

线性回归（linear regression）通过构建一个线性模型来预测目标变量。最简单的线性回归算法是一元线性回归，多元线性回归则是利用多个特征来预测目标变量。

非线性回归（nonlinear regression）目标变量与特征之间的关系不是线性的。多项式回归（polynomial regression）是非线性回归的一种形式，通过将特征的幂次作为新的特征来构建一个多项式模型。逻辑回归（logistic regression）既是一种二分类算法，可以用于非线性回归。

正则化（regularization）正则化通过向目标函数中添加惩罚项来避免模型的过拟合。常用的正则化方法有岭回归、Lasso 回归、弹性网络回归。岭回归通过向目标函数中添加 L2 惩罚项来控制模型复杂度。Lasso 回归通过向目标函数中添加 L1 惩罚项，它不仅能够控制模型复杂度，还可以进行特征选择。弹性网络是岭回归和 Lasso 回归的结合体，它同时使用 L1 和 L2 惩罚项。

贝叶斯回归（Bayesian regression）是一种基于贝叶斯定理的回归算法，它可以用来估计连续变量的概率分布。

基于分类算法的回归，比如 *k*NN 算法是一种基于距离度量的分类算法，但也可以用于回归任务。支持向量回归（Support Vector Regression，SVR）则是一种基于支持向量机（Support Vector Machine，SVM）的回归算法，它通过寻找一个最优的边界，来预测目标变量。

图 8 比较线性回归、多项式回归、逻辑回归三种回归算法。

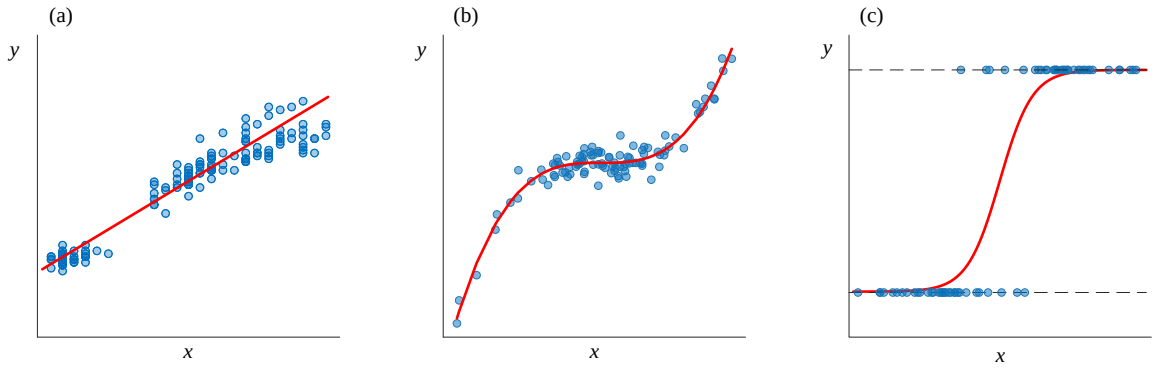


图 8. 比较回归算法，线性回归、多项式回归、逻辑回归

28.4 降维：降低数据维度，提取主要特征

降维是指将高维数据转换为低维数据的过程，这个过程可以提取出数据的主要特征，并去除噪声和冗余信息。降维可以有效地减少计算成本，加速模型训练和预测，并提高模型的准确性和可解释性。

以下是机器学习中常用的降维算法：

主成分分析（Principal Component Analysis，PCA）通过线性变换将高维数据映射到低维空间。利用特征值分解、奇异值分解都可以完成主成分分析。

核主成分分析（Kernel Principal Component Analysis，KPCA）是一种非线性降维算法，它使用核函数将数据映射到高维空间，然后使用 PCA 在新的空间中进行降维。

典型相关分析（Canonical Correlation Analysis，CCA）是一种统计学习算法，它通过最大化两个变量之间的相关性来降低维度。

流形学习（Manifold Learning）是一种非线性降维算法，它通过保持局部结构的连续性来将高维数据映射到低维空间。流形学习可以发现数据中的非线性关系和流形结构。

这些降维算法都有不同的优点和适用场景，根据数据的特点和需求选择适合的算法进行建模。

28.5 分类：针对有标签数据

在机器学习中，分类是指根据给定的数据集，通过对样本数据的学习，建立分类模型来对新的数据进行分类的过程。下面简述一些常用的分类算法。

最近邻算法（KNN）：基于样本的特征向量之间的距离进行分类预测，即找到与待分类数据距离最近的 K 个样本，根据它们的类别进行投票决策。

朴素贝叶斯算法（Naive Bayes）：利用贝叶斯定理计算样本属于某个类别的概率，并根据概率大小进行分类决策。

支持向量机（SVM）：利用间隔最大化的思想来进行分类决策，可以通过核函数将低维空间中线性不可分的样本映射到高维空间进行分类。

决策树算法（Decision Tree）：通过对样本数据的特征进行划分，构建一个树形结构，从而实现对新数据的分类预测。

我们可以通过比较决策边界的形状大致知道采用的是哪一种分类算法，图 9 给出四个例子。本书第 30 章将专门介绍几种分类算法。

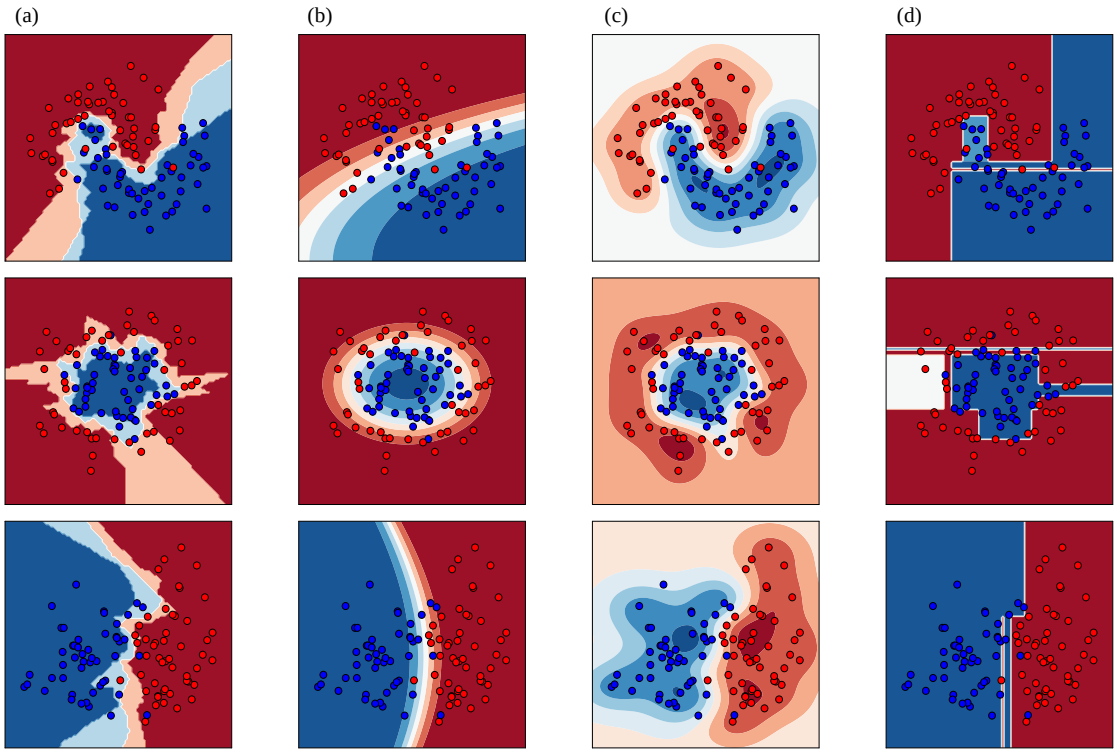


图 9. 比较分类算法决策边界，最近邻、朴素贝叶斯、支持向量机、决策树

28.6 聚类：针对无标签数据

在机器学习中，聚类是指将数据集中的样本按照某种相似性指标进行分组的过程。常用的聚类算法包括。

k 均值算法（kMeans）：将样本分为 k 个簇，每个簇的中心点是该簇中所有样本点的平均值。

高斯混合模型（Gaussian Mixture Model, GMM）：将样本分为多个高斯分布，每个高斯分布对应一个簇，采用 EM 算法进行迭代优化。

层次聚类算法（Hierarchical Clustering）将样本分为多个簇，可以使用自底向上的凝聚层次聚类或自顶向下的分裂层次聚类。

DBSCAN（Density-Based Spatial Clustering of Applications with Noise）是基于密度的聚类算法，可以自动发现任意形状的簇。

谱聚类算法（Spectral Clustering）是基于样本之间的相似度来构造拉普拉斯矩阵，然后对其进行特征值分解来实现聚类。

图 10 比较四种 k 均值、高斯混合模型、DBSCAN、谱聚类算法结果。

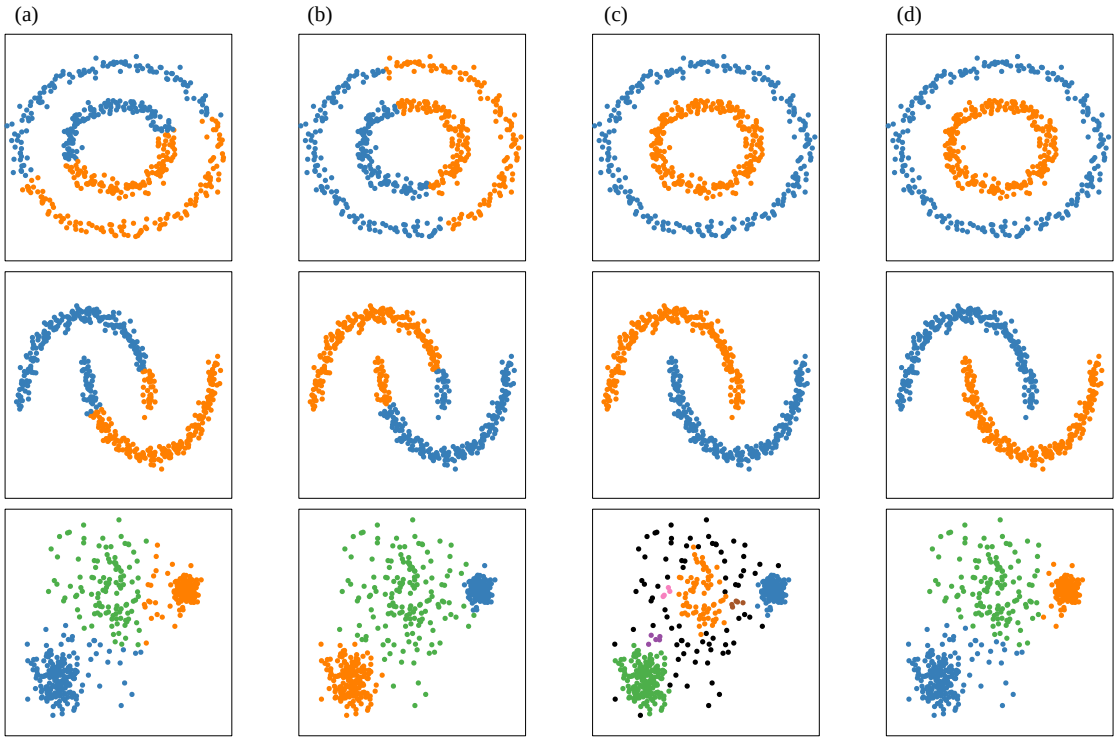


图 10. 比较聚类算法，k 均值、高斯混合模型、DBSCAN、谱聚类

28.7 什么是 Scikit-Learn?

Scikit-learn 是一个流行的 Python 机器学习库，提供完成机器学习任务各种工具。Scikit-learn 和前文介绍的 NumPy、SciPy、Pandas、Matplotlib 等重要工具联系紧密。

以下是 Scikit-learn 中的主要工具：

数据集：Scikit-learn 中包含多个标准数据集，还提供生成样本数据的函数。这些数据集可以用于测试和评估机器学习模型的性能。

数据预处理（data preprocessing）。数据预处理是机器学习的重要一步，它包括数据清洗、数据重构和数据变换。Scikit-learn 提供了各种数据预处理工具，包括特征缩放、归一化、标准化、处理缺失值、数据编码等。Scikit-Learn 数据本书下一章（第 29 章）要探讨的话题。

监督学习模型：Scikit-learn 支持多种监督学习模型，包括线性回归、逻辑回归、支持向量机、决策树、随机森林、神经网络等。

无监督学习模型：Scikit-learn 支持多种无监督学习模型，包括聚类、降维、密度估计等。这些模型可以用于在没有标签的情况下对数据进行分析和理解。

模型选择和评估：Scikit-learn 提供了各种工具，用于选择最佳模型和评估模型的性能。这些工具包括交叉验证、网格搜索、评估指标等。

管道：Scikit-learn 中的管道工具可用于将数据预处理和模型训练流程组合在一起，使得处理和训练过程更加高效和简单。

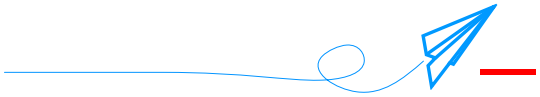
总的来说，scikit-learn 提供了一个全面的机器学习工具包，使得机器学习的建模和评估过程更加高效和方便。

请大家完成如下题目。

Q1．本章没有编程练习题，只要求把 Scikit-learn 的官方示例库浏览一遍，曲面了解 Scikit-learn 库能够完成的机器学习算法，具体页面如下。

https://scikit-learn.org/stable/auto_examples/index.html

* 这道题目不需要答案。



本章全景介绍有关机器学习的基本知识。需要大家理解的概念包括，有标签数据、无标签数据，以及机器学习四大任务（回归、降维、分类、聚类）。下面四章将按照这个顺序用示例展开如何利用 Scikit-learn 工具完成机器学习任务。