

1

All Is Number

万物皆数

从数据分析、机器学习视角再看数字



但凡满足以下两个条件的理论，便可以称之为优质理论：基于几个有限的变量，准确描述大量观测值；能对未来观测值做出确定的预测。

A theory is a good theory if it satisfies two requirements: it must accurately describe a large class of observations on the basis of a model that contains only a few arbitrary elements, and it must make definite predictions about the results of future observations.

—— 史蒂芬·霍金 (Stephen Hawking) | 英国理论物理学家、宇宙学家 | 1942 ~ 2018



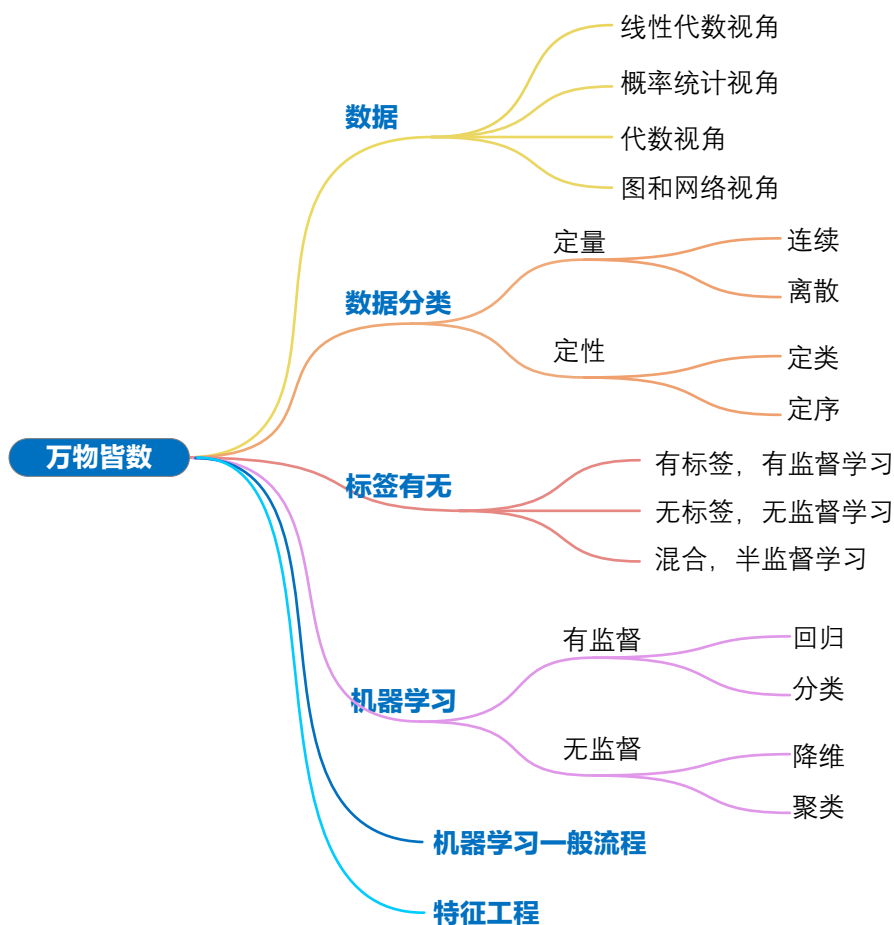
本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

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



1.1 万物皆数：从矩阵说起

这是一个有关数字的故事，故事的开端便是形如图 1 所示的表格数据。生活中大部分数据都可以用表格形式呈现、保存、运算、分析。

任何表都可以看成是由行 (row) 和列 (column) 构成。

从线性代数角度来看，我们管图 1 这个表格叫做矩阵。鸢尾花书中，矩阵这个数学概念无处不在。

《矩阵力量》介绍过矩阵的每一行可以看成是一个行向量 (row vector)，每一列为列向量 (column vector)。

将图 1 这个矩阵记做 X ， X 可以写成一组列向量 $X = [x_1, x_2, \dots, x_D]$ 。

X 当然也可以写成一组行向量 $X = [x^{(1)}, x^{(2)}, \dots, x^{(n)}]^T$ 。

⚠ 注意，在《机器学习》一册中，为了方便 $x^{(1)}, x^{(2)}, \dots, x^{(n)}$ 偶尔也会被视作为列向量，到时候会具体说明。

从统计角度来看，表格的每一列还可以视作一个随机变量的样本数据。图 1 则代表 D 个随机变量 (X_1, X_2, \dots, X_D) 的样本数据。

随机变量 X_1, X_2, \dots, X_D 可以构成 D 元随机变量列向量 $\chi = [X_1, X_2, \dots, X_D]^T$ 。

从代数角度来看，图 1 表格的每一列相当于变量 (x_1, x_2, \dots, x_D) 的取值。比如，我们会在回归分析的解析式中看到这种记法 $y = b_0 + b_1x_1 + b_2x_2 + \dots + b_Dx_D$ 。

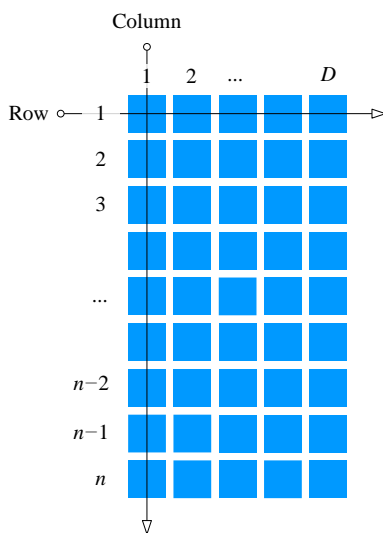


图 1. 多行、多列矩阵

图 2 所示为鸢尾花样本数据，这是鸢尾花书最常用的数据集，鸢尾花书也因此命名。表格中四列特征 (花萼长度、花萼宽度、花瓣长度、花瓣宽度) 就可以看成一个矩阵。而表格最后一列为鸢尾花分类标签。

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

如图 3 所示，一幅照片本质上也是数据矩阵。

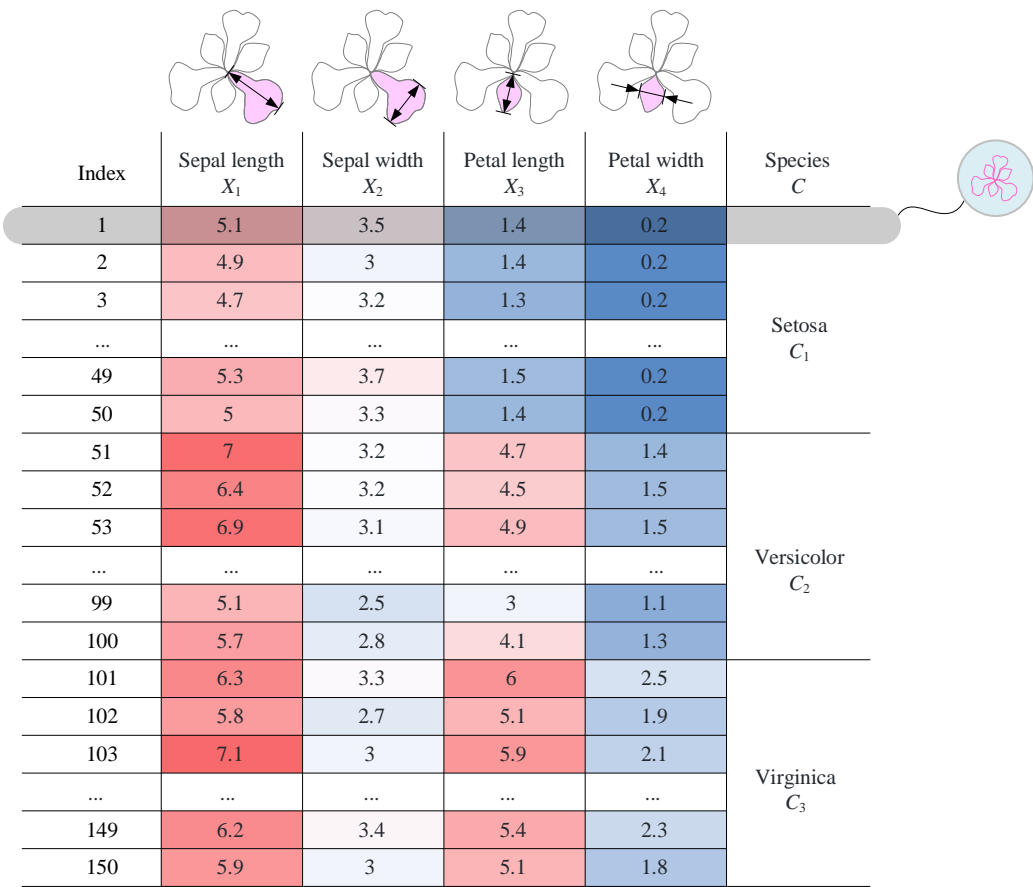


图 2. 鸢尾花数据表格，特征数据单位为厘米 (cm)

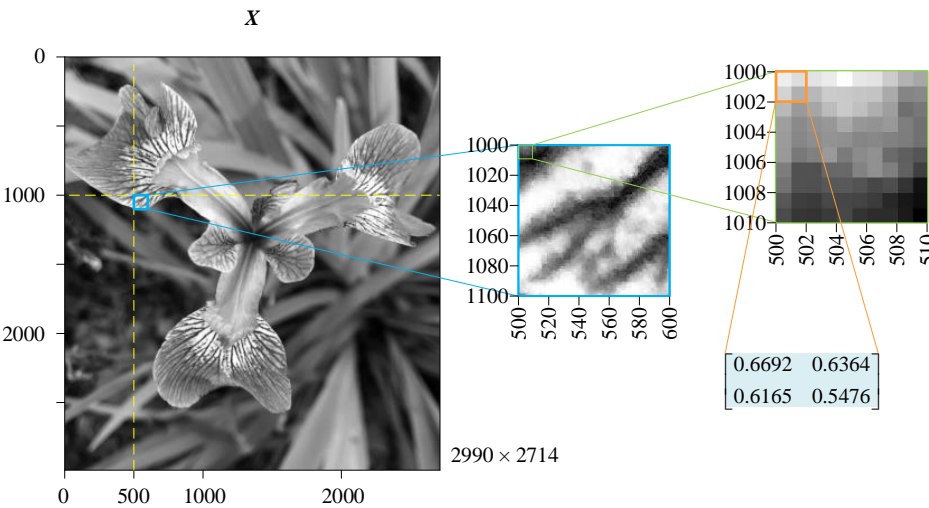


图 3. 照片也是数据矩阵，图片来《矩阵力量》

一个矩阵还可以衍生得到其他形式，具体如图 4 所示。

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

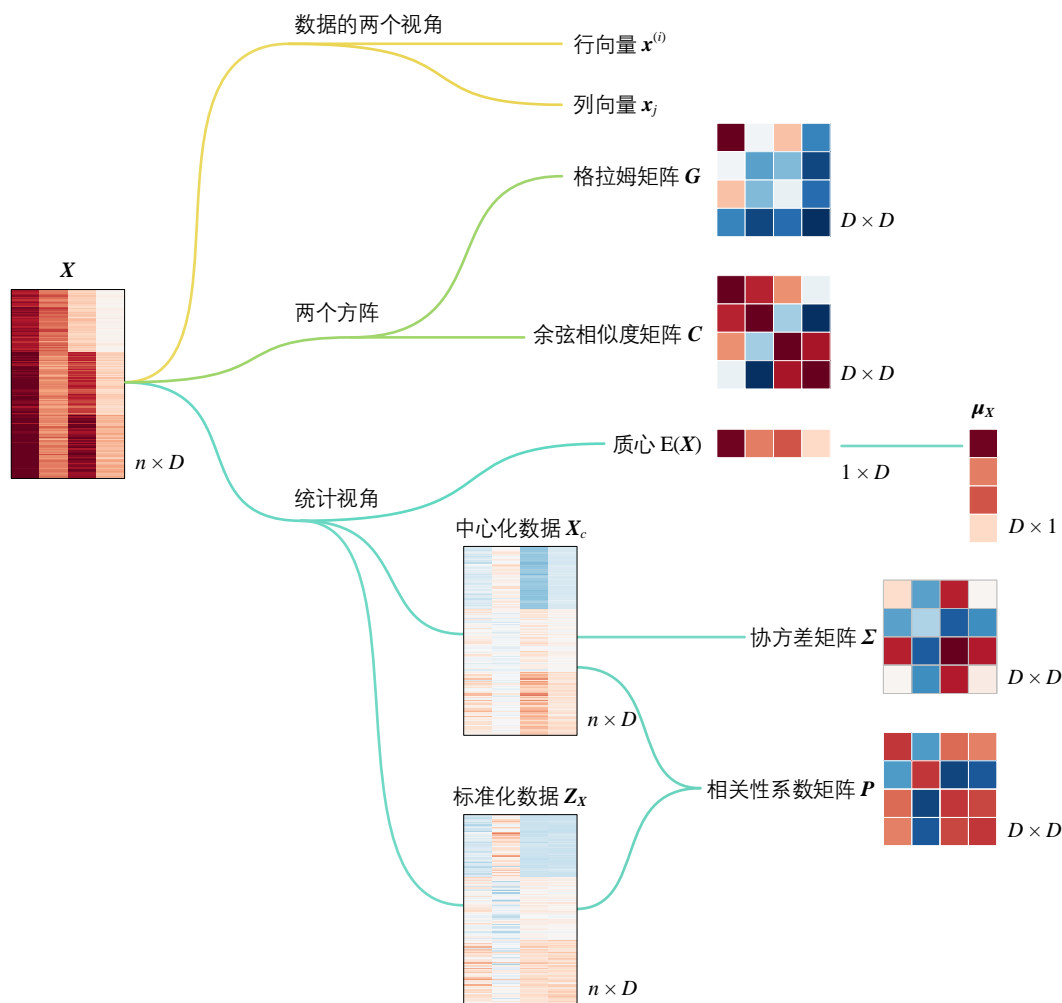


图 4. 鸢尾花数据衍生得到的几个矩阵，图片来自《矩阵力量》

图 5 则总结了和数据矩阵 X 有关的向量、矩阵、矩阵分解、空间等概念。

这幅图的数据分为两个部分：第一部分以 X 为核心，向量以 0 为起点；第二部分是统计视角，以去均值数据 X_c 为核心，向量以质心为起点。

一般情况， X 为细高型矩阵，形状为 $n \times D$ ，样本数 n 一般远大于特征数 D 。对 X 进行 SVD 分解可以得到四个空间。

格拉姆矩阵 G 含有 X 列向量模、向量夹角两类重要信息。余弦相似度矩阵 C 仅仅含有向量夹角信息。对格拉姆矩阵 G 进行特征值分解只能获得两个空间。对格拉姆矩阵 G 进行 Cholesky 分解得到上三角矩阵 R 可以“代表” X 列向量坐标。

在统计视角下， X 有两个重要信息——质心、协方差矩阵。质心确定数据中心位置，协方差矩阵描述数据分布。协方差矩阵 Σ 同样含有“标准差向量”的模（标准差大小）、向量夹角（余弦值为相关性系数）两类重要信息。相关性系数矩阵 P 仅仅含有向量夹角（相关性系数）信息。

X_c 是中心化数据矩阵，即每一列数据都去均值。 Z_x 是标准化数据矩阵，即 X 的 Z 分数。在几何视角下， X 到 X_c 相当于质心“平移”， X 到标准化数据 Z_x 相当于“平移 + 缩放”。

协方差矩阵 Σ 相当于 X_c 的格拉姆矩阵。相关性系数矩阵 P 相当于 Z_X 的格拉姆矩阵。此外，注意样本数据缩放系数 $(n-1)$ 。 X_c 进行 SVD 分解也得到四个空间。这四个空间因 X_c 而生，一般情况不同于 X 的四个空间。

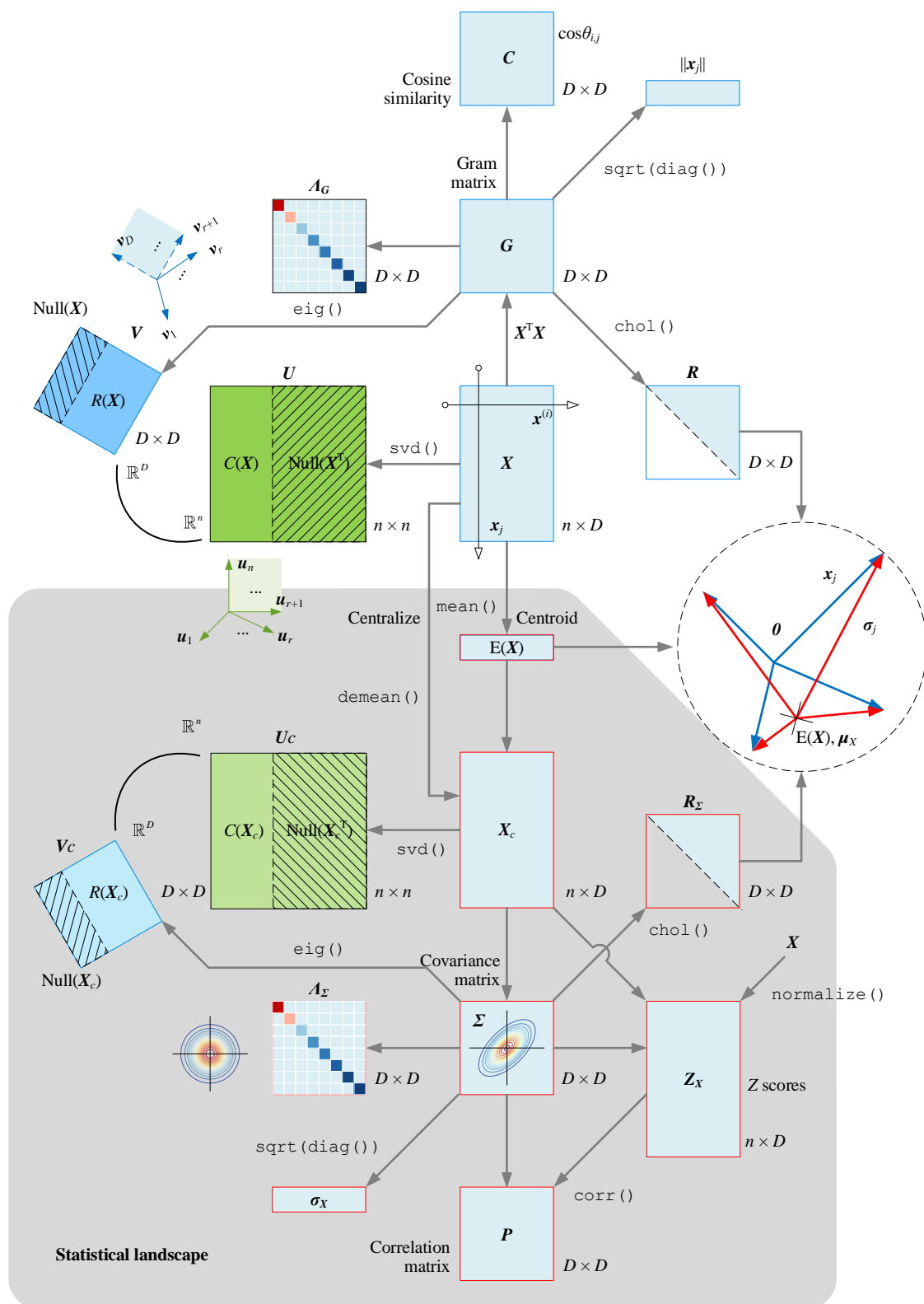


图 5. 矩阵、矩阵分解、空间, 图片来自《矩阵力量》

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。

代碼及 PDF 文件下載：<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 6 则“鸟瞰”鸢尾花书介绍各种有关矩阵的运算、分析、可视化工具。本书则要给鸢尾花数据这个矩阵增加一个全新视角——图！

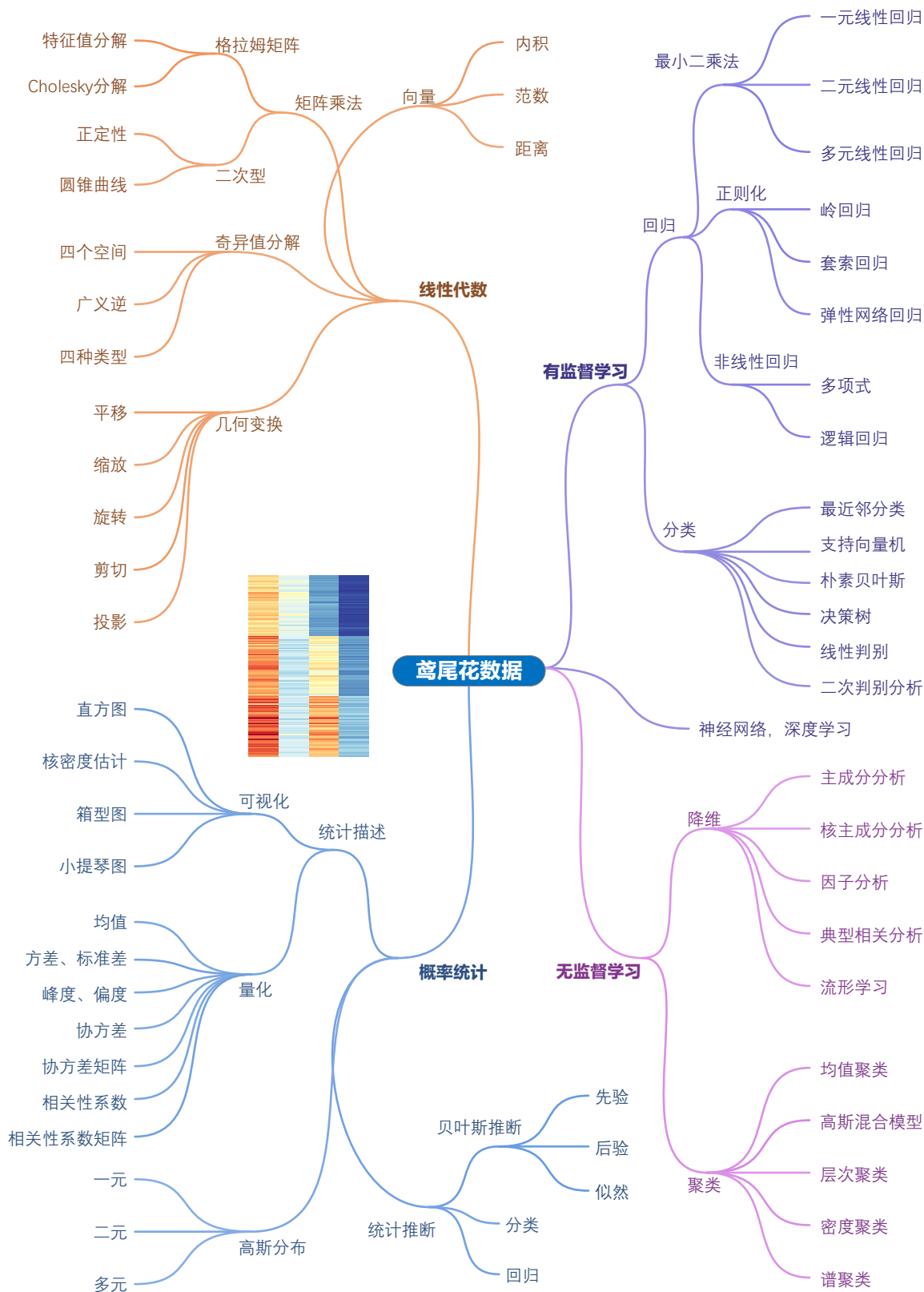


图 6. 有关鸢尾花数据的可视化“头脑风暴”，图片来自《可视之美》

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

图 7 所示为根据鸢尾花数据前两个特征欧氏距离矩阵创建的无向图。图论 (graph theory) 中，无向图 (undirected graph) 是一类图 (graph)，其中的节点 (nodes) 通过边 (edge) 相连，但这些边无方向 (undirected)。图和网络是本册的重要话题，所占篇幅超过一半。

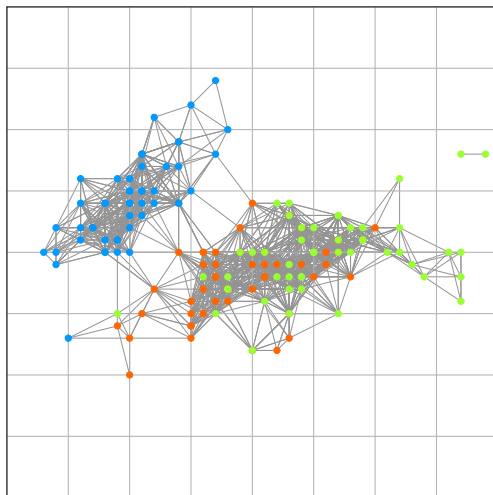


图 7. 根据鸢尾花数据前两个特征欧氏距离矩阵创建的无向图

简单来说，图是表示关系的节点和边的集合。节点是对应于对象的节点；边是连接对象之间的关系。图的边有时带有权重，表示节点之间连接的强度或其他属性。

图 8 所示的这幅无向图中，128 个节点代表美国主要城市，节点大小代表人口数；而节点的位置为城市的真实相对地理位置。这幅图中节点之间的边代表临近的两两城市距离。

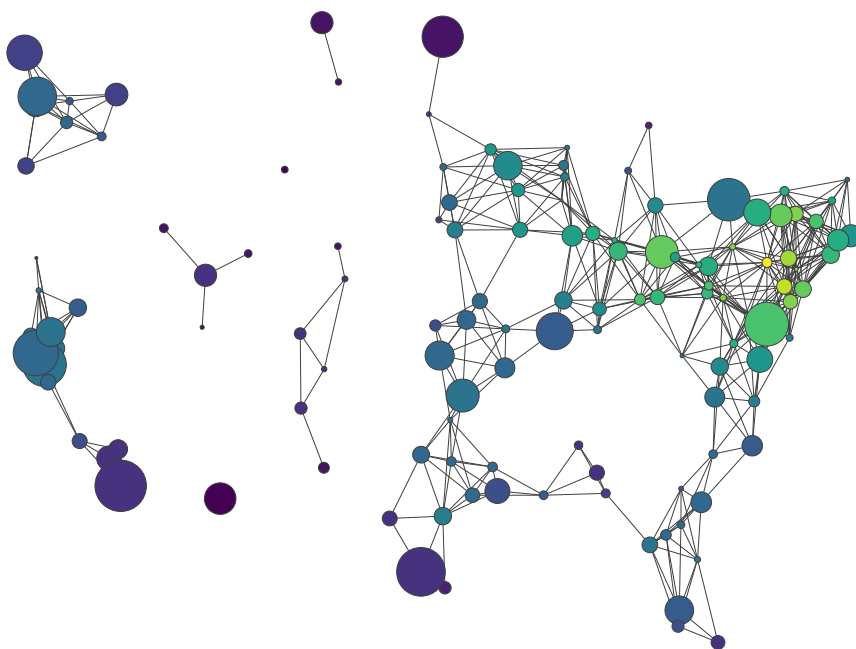


图 8. 128 个美国城市人口和距离组成的无向图

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

再举个例子，我们可以使用图来模拟社交网络中的友谊关系，如图 9 所示。图的节点是人，而边表示友谊关系。图与物理对象和情境的对应关系意味着我们可以使用图来模拟各种各样的系统。



图 9. 社交网络图

排版时，请替换为矢量图，见附件 SVG 文件

1.2 数据分类：定量、定性、连续、离散

定量数据、定性数据

数据一般可以分为**定量数据** (quantitative data) 和**定性数据** (qualitative data)，具体分类如图 10 所示。

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

定量数据指的是，可以采用数值表达的数据，比如股票价格、人体高度、气温等等。

定性数据，也叫**类别数据** (categorical data)，指的是描述事物的特征、属性等文字或符号，比如姓名、颜色、国家、性别、五星评价等等。

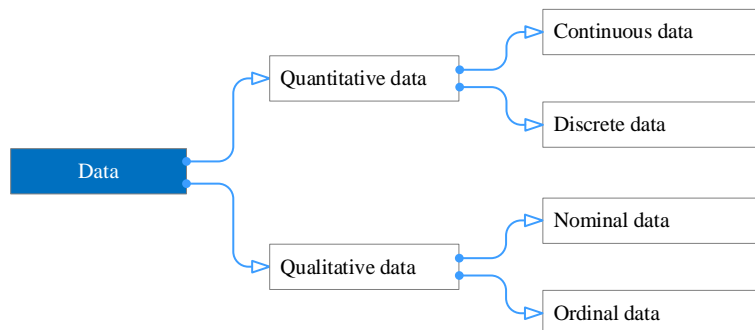


图 10. 数据分类

连续数据、离散数据

定量数据，还可以进一步分为**连续数据** (continuous data) 和**离散数据** (discrete data)。

连续数据是指在一定区间内可以任意取值的数据，比如气温、GDP 数据等等。离散数据只能采取特定值，比如说个数（整数）、一到五星好评、骰子点数等等。

一天 24 小时之内的温度数据不可能被持续记录，按一定时间频率需要采样。举个例子，比如，每小时记录一个温度数值。图 11 所示为某国家 GDP 数据，虽然为年度数据，当数据量足够大时，GDP 增长曲线看上去是连续曲线；但是，当展开局部数据时，可以发现这条所谓的连续数据实际上是相邻点相连构成的“折线”。

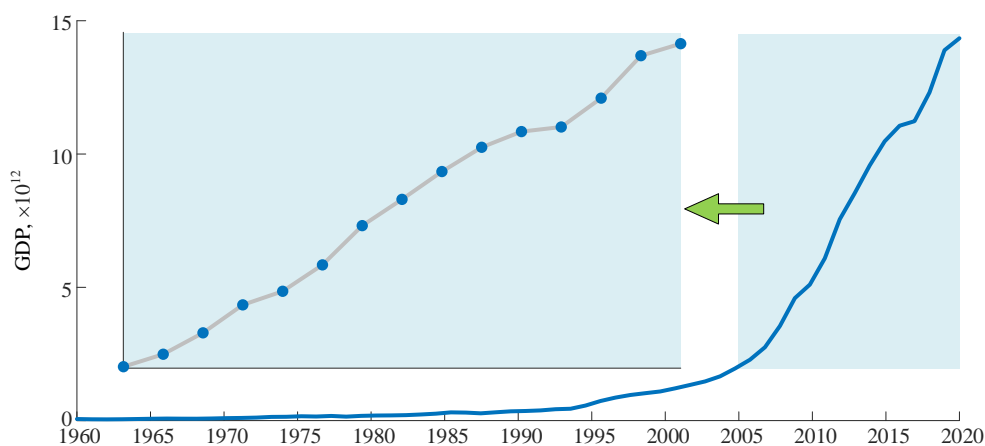


图 11. 采样数据

时间序列 (timeseries) 是指按照时间顺序排列的一系列数据点或观测值，通常是等时间间隔下的测量值，如每天、每小时、每分钟等。时间序列数据通常用于研究时间相关的现象和趋势，例如股票价格 (如图 12)、气象数据、经济指标等。本书专门有一个板块介绍和时间序列相关内容。

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

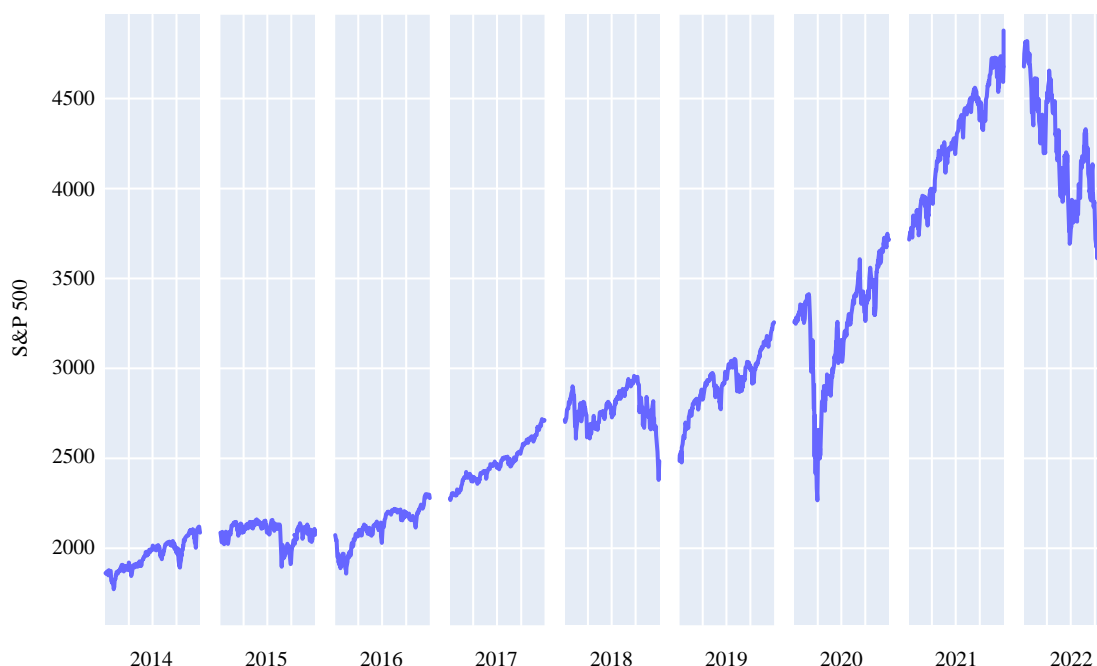


图 12. 标普 500 数据，按年观察趋势

定类数据、定序数据

定性数据也可以分为**定类数据** (nominal data) 和**定序数据** (ordinal data)。简单来说，定类数据没有任何内在顺序或排序，定序数据指具有内在顺序或排序的数据。

定类数据，也叫名义数据，用来表征事物类别，比如血型 A、B、AB 和 O。

定序数据，也叫有序数据，不仅能够代表事物的类别，还可以据此特征排序，比如学生成绩 A、B、C、D 和 F。此外，区间数据 (interval data) 也可以看做是一种定序数据，比如身高区间数据，160 cm 以下 (包括 160 cm)、160 cm 到 170 cm (包括 170 cm)、170 cm 到 180 cm (包括 180 cm) 和 180 cm 以上。

混合

很多时候，一个表格常常是各种数据的集合体。如图 13 所示，表格每一行代表一个学生的某些基本数据。表格第 1 列为学生姓名，表格第 2 列为性别 (定类数据)，表格第 3 列为身高 (连续定量数据)，第 4 列为成绩 (定序数据)，第 5 列为血型 (定类数据)。

大家已经很熟悉的鸢尾花数据也是混合数据表格。如图 2 所示，表格的第一列为序号，之后四列为花萼长度、花萼宽度、花瓣长度、花瓣宽度四个特征的连续数据。最后一列为鸢尾花分类标签。

Name	Gender	Height	Grade	Blood
James	Male	185	A	AB
Shawn	Male	178	A+	B
Mary	Female	165	A-	O
Alice	Female	175	A+	B
Bill	Male	171	B	A
Julia	Female	168	B+	A

图 13. 学生数据

1.3 机器学习：四大类算法

鸢尾花书不管是编程、可视化，还是数学工具、数据分析，都是为机器学习时间服务的。从机器学习算法角度来看，我们首先关心数据是否有标签。

有标签、无标签数据

根据输出值有无标签，如图 15 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**是指已经被人工或其他方式标注了类别或标签的数据。在有标签数据中，每个样本都有对应的标签或分类信息。如图 14 所示，每种动物可以以各种标签划分，比如冷血、温血。

有标签数据通常用于**监督学习** (supervised learning)，即机器学习模型可以利用已知的标签信息进行训练，并在后续的预测过程中使用这些信息进行分类或回归。

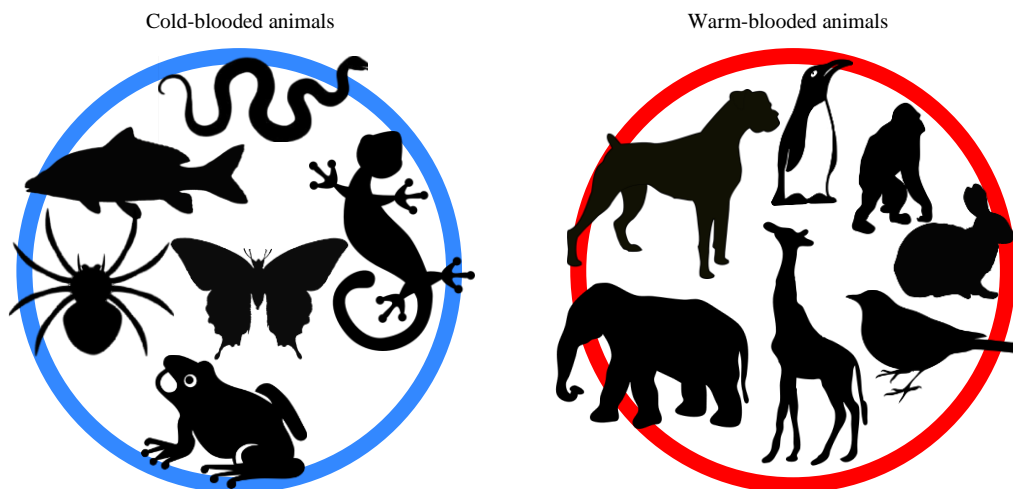


图 14. 动物的一种标签：冷血、温血

无标签数据是指没有标签或分类信息的数据。在无标签数据中，样本只有特征信息，而没有对应的标签信息。

无标签数据通常用于**无监督学习** (unsupervised learning)，即机器学习模型需要通过自己的学习过程，从数据中发现并学习出有意义的模式和结构。无监督学习通常包括聚类、降维、异常检测等任务。

在实际应用中，有标签数据和无标签数据往往同时存在。例如，在文本分类任务中，可以有大量已经标注好类别的文本数据 (有标签数据)，但同时还存在大量未分类的文本数据 (无标签数据)，可以利用这些无标签数据进行**半监督学习** (semi-supervised learning)。

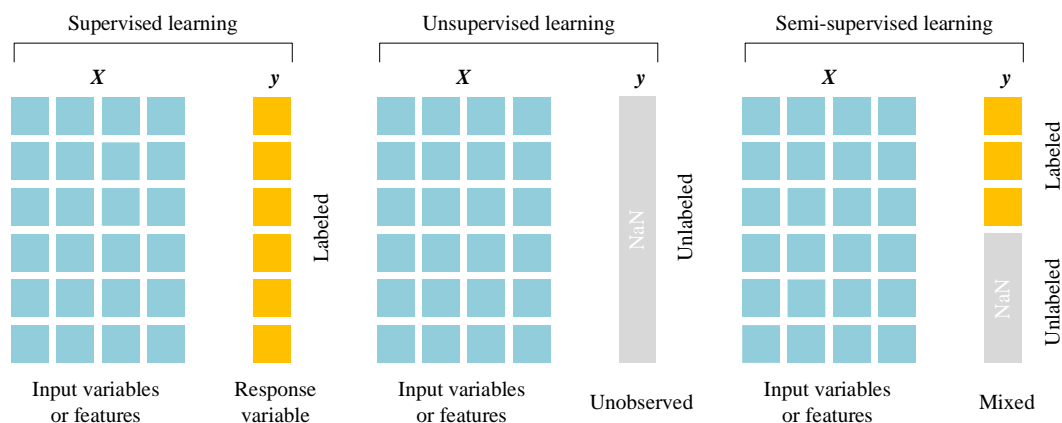


图 15. 根据有无标签分类算法类型

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 16 (a)。如果标签为分类数据，对应的问题则是**分类** (classification)，如图 16 (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 16 (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**簇** (cluster)，这种问题叫做**聚类** (clustering)，如图 16 (b)。

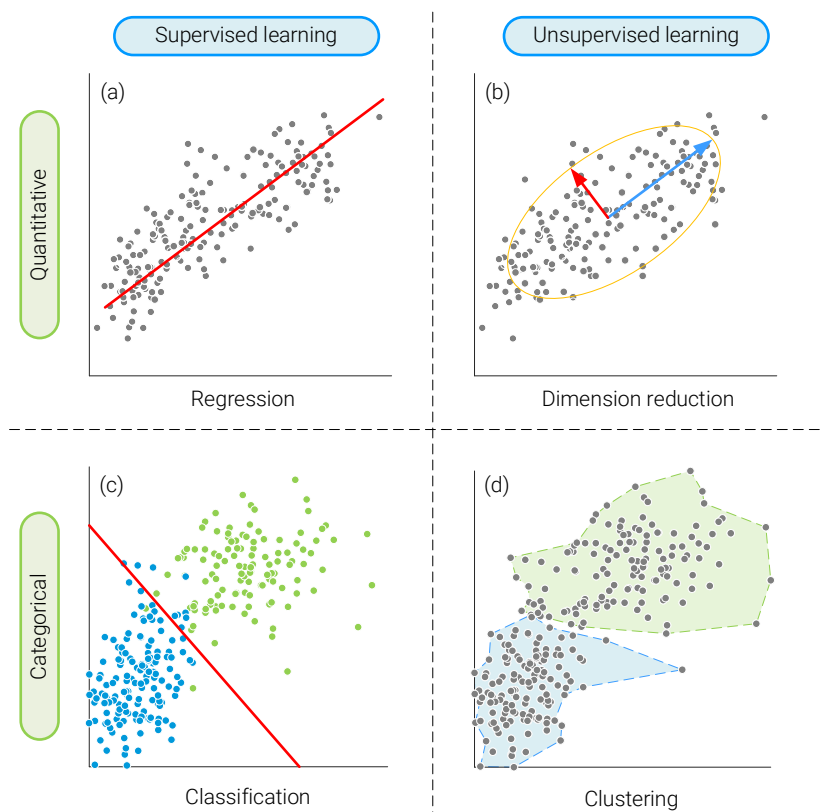


图 16. 根据数据是否有标签、标签类型细分机器学习算法

图 17 所示为机器学习的一般流程。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。

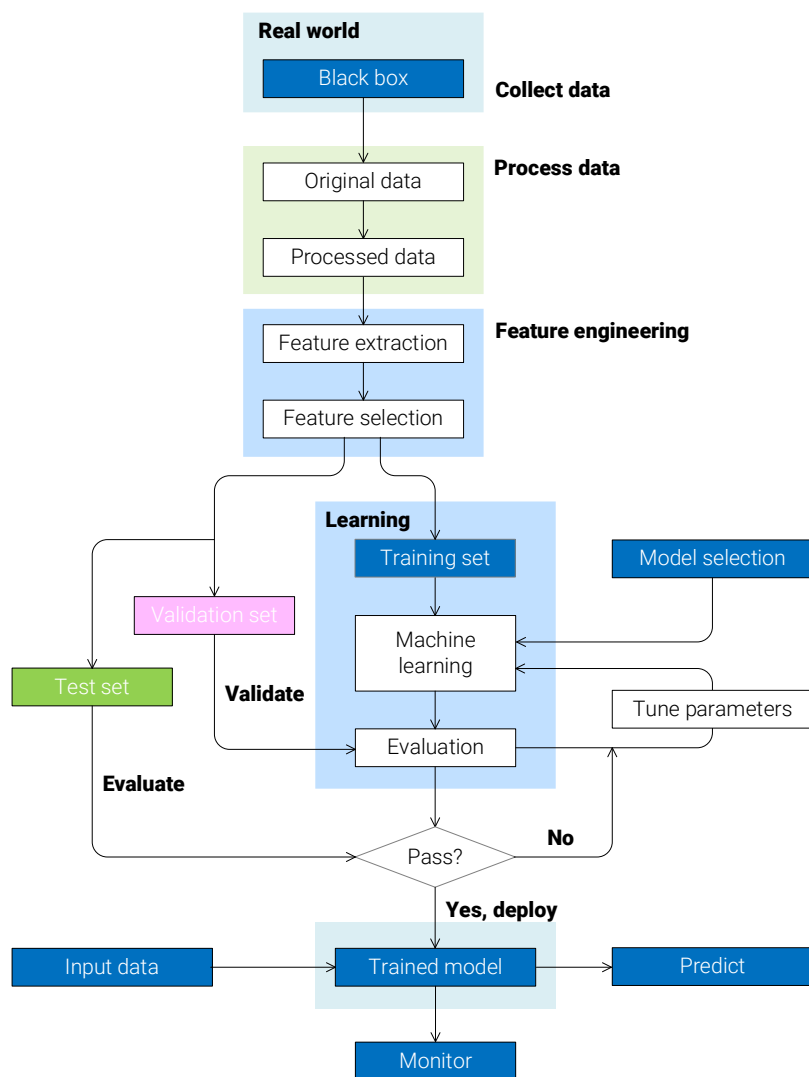


图 17. 机器学习一般流程

1.4 特征工程：提取、转换、构建数据

图 17 提到了特征工程，下面展开聊聊这个话题。

从原始数据中最大化提取可用信息的过程就叫做**特征工程** (feature engineering)。特征很好理解，比如鸢尾花花萼长度宽度、花瓣长度宽度，人的性别、身体、体重等，都是特征。

特征工程是机器学习中非常重要的一个环节，指的是对原始数据进行特征提取、特征转换、特征选择和特征创造等一系列操作，以便更好地利用数据进行建模和预测。

具体来说，特征工程包括以下方法：

- ◀ **特征提取** (Feature Extraction)：将原始数据转换为可用于机器学习算法的特征向量。注意，这个特征向量不是特征值分解中的特征向量。
- ◀ **特征转换** (Feature Transformation)：对原始特征进行数值变换，使其更符合算法的假设。例如，在回归问题中，可以对数据进行对数转换或指数转换等。

本书代码及 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

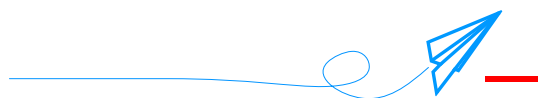
- ◀ **特征选择** (Feature Selection): 选择最具有代表性和影响力的特征。例如, 可以使用相关性分析、PCA 等方法选择最相关或最重要的特征。
- ◀ **特征创造** (Feature Creation): 根据原始特征创造新的特征, 比如特征相加减、相乘除等等运算。
- ◀ **特征缩放** (Feature Scaling): 将特征缩放到相同的尺度或范围内, 避免某些特征对模型训练的影响过大。

特征工程在机器学习中扮演着至关重要的角色, 它可以提高模型的精度、泛化能力和效率。在实际应用中, 需要根据具体问题选择合适的特征工程方法, 并不断尝试和改进以达到最佳效果。

相信大家都听过“**垃圾进, 垃圾出** (garbage in, garbage out, GIGO)”。这句话的含义很简单, 将错误的、无意义的输入数据输入计算机系统, 计算机自然也一定会输出错误、无意义的结果。在数据科学、机器学习领域, 很多时候数据扮演核心角色。以至于在数据分析建模时, 大部分的精力都花在了处理数据上。

特征工程很好的混合了专业知识、数学能力。虽然丛书不会专门讲解特征工程, 但是本书的很多内容都可以用于特征工程。

本书下一板块中介绍的缺失值、离散值处理可以视作特征预处理。而缺失值、离散值也经常使用各种机器学习算法。而数据转换、插值、正则化、主成分分析、因子分析、典型性分析也都是特征工程的利器。此外, 《统计至简》一册中的统计描述、统计推断, 《机器学习》一册介绍的**线性判别分析** (linear discriminant analysis, LDA)、**聚类算法**等也都可以用于特征工程。



本章从矩阵说起, 和大家聊了聊数据常见类型、机器学习四大类算法、特征工程等话题。

数据类型可以大致分为定量数据和定性数据。定量数据指的是可以通过计数或测量得到的数值数据, 进一步分为连续数据和离散数据。连续数据, 如温度和时间, 可以在任何范围内取无限多的值; 而离散数据, 如人数, 只能取有限或可数的值。定性数据描述的是事物的属性或类别。

有标签数据含有明确的输出标签, 适用于监督学习任务如分类和回归, 其中模型通过输入和输出的对应关系学习。无标签数据不含输出标签, 用于无监督学习如聚类和降维, 模型需自行发现数据的结构和模式。这两种数据类型直接影响算法选择, 以适应具体的学习任务和目标。

特征工程是在机器学习中优化模型性能的关键步骤, 涉及从原始数据中选择、修改和创建新的特征。通过特征提取、选择、转换、创造和缩放等技术, 它帮助改善模型的准确度和效率, 使模型能更好地理解数据的复杂结构, 从而做出更准确的预测。

下一章, 我们将进入数据处理板块, 聊聊缺失值、离群值、数据转换、距离度量等话题。