

1

All Is Number

万物皆数

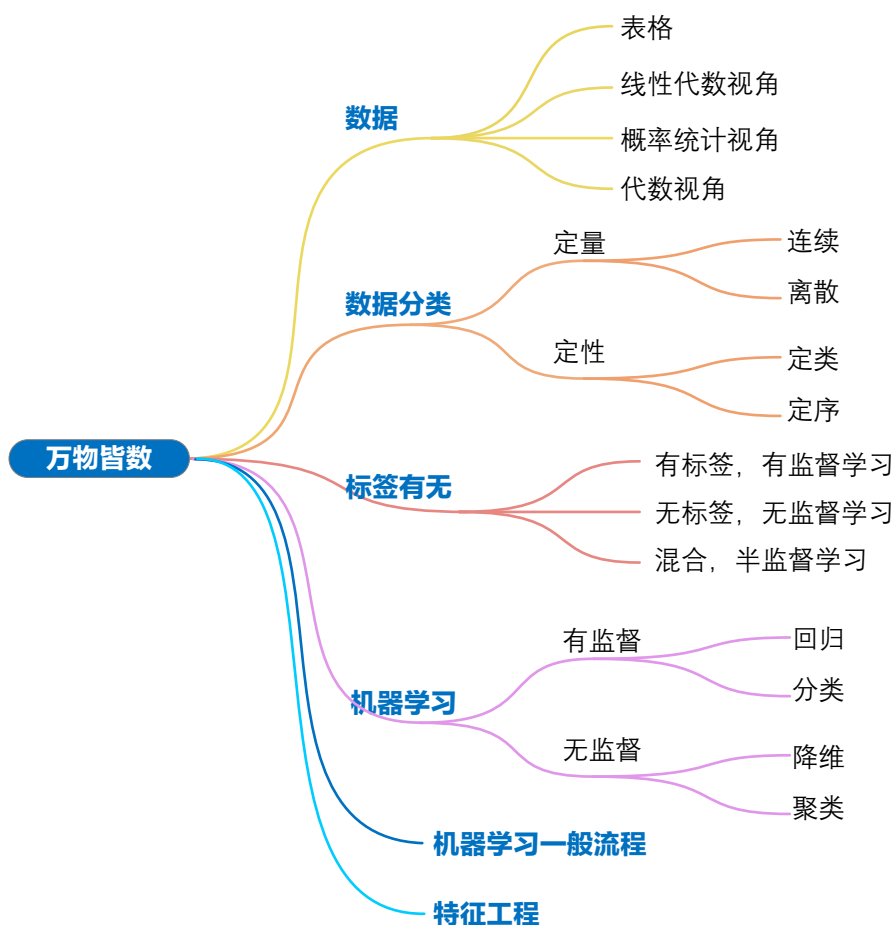
从数据科学、机器学习视角再看数字



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 从表格说起

四个视角

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。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，图 1 这个表格本质上是一个矩阵。《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)。

比如，将图 1 这个矩阵记做 \mathbf{X} ， \mathbf{X} 可以写成一组列向量 $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_D]$ 。 \mathbf{X} 当然也可以写成一组行向量 $\mathbf{X} = [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(n)}]^T$ 。

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

从统计角度来看，表格的每一列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。

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

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 $y = b_0 + b_1x_1 + b_2x_2 + \dots + b_Dx_D$ 。

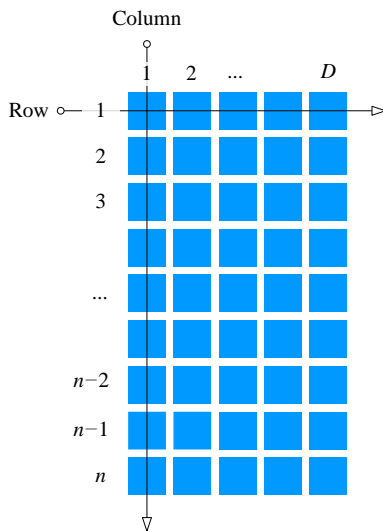


图 1. 表格数据

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 2 所示。

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。

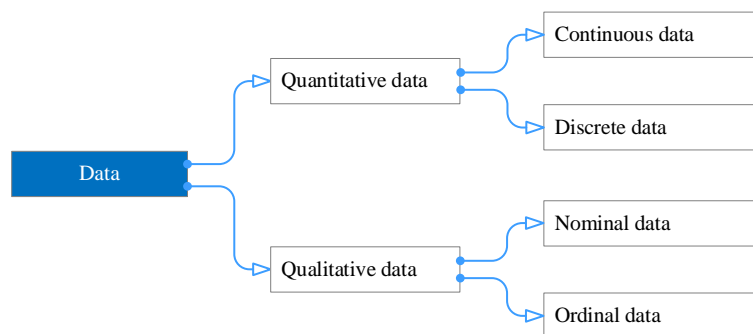


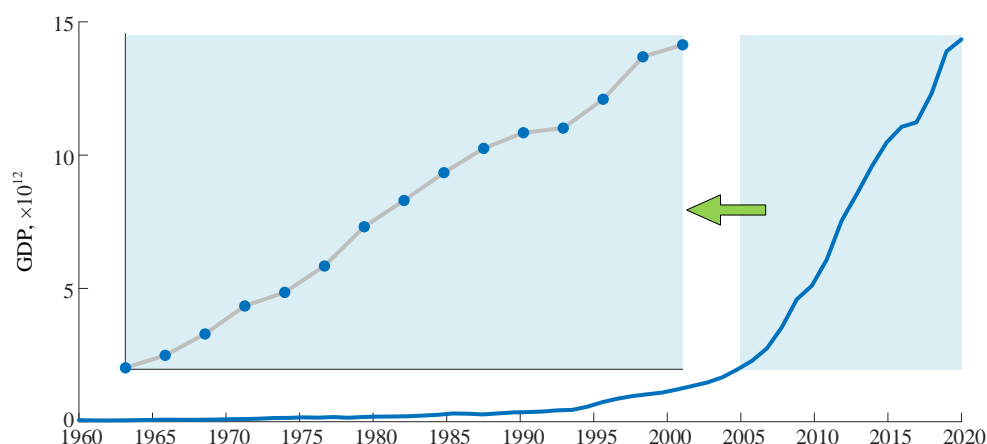
图 2. 数据分类

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 3 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”。



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 3. 采样数据

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) 也可以看做时一种定序数据，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。

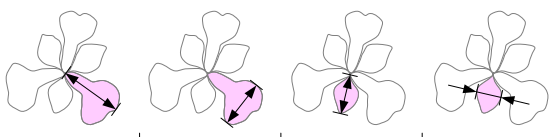
混合

很多时候，一个表格常常是各种数据的集合体。如图 4 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 5 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 4. 学生数据



Index	Sepal length X_1	Sepal width X_2	Petal length X_3	Petal width X_4	Species C
1	5.1	3.5	1.4	0.2	Setosa C_1
2	4.9	3	1.4	0.2	
3	4.7	3.2	1.3	0.2	
...	
49	5.3	3.7	1.5	0.2	
50	5	3.3	1.4	0.2	Versicolor C_2
51	7	3.2	4.7	1.4	
52	6.4	3.2	4.5	1.5	
53	6.9	3.1	4.9	1.5	
...	
99	5.1	2.5	3	1.1	Virginica C_3
100	5.7	2.8	4.1	1.3	
101	6.3	3.3	6	2.5	
102	5.8	2.7	5.1	1.9	
103	7.1	3	5.9	2.1	
...	
149	6.2	3.4	5.4	2.3	
150	5.9	3	5.1	1.8	

图 5. 鸢尾花数据表格，单位为厘米 (cm)

有标签、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**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。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。

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、降维和异常检测等任务。

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）。

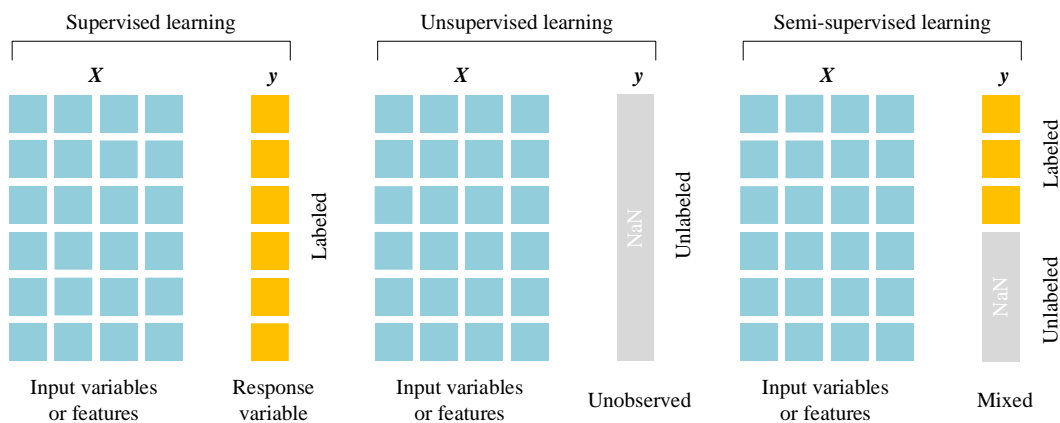


图 6. 根据有无标签分类数据

1.2 机器学习方法分类

人工智能 (Artificial Intelligence, AI) 是一套算法系统，它通过模拟人类智慧，感知环境，经过分析计算，进而可以执行设定的行为动作。

机器学习

机器学习是实现人工智能的一大类方法和技术。机器学习算法的特点是，从样本数据中分析并获得某种规律，再利用这个规律对未知数据进行预测。它是涉及概率、统计、矩阵论、代数学、优化方法、数值方法、算法学等多领域的交叉学科。

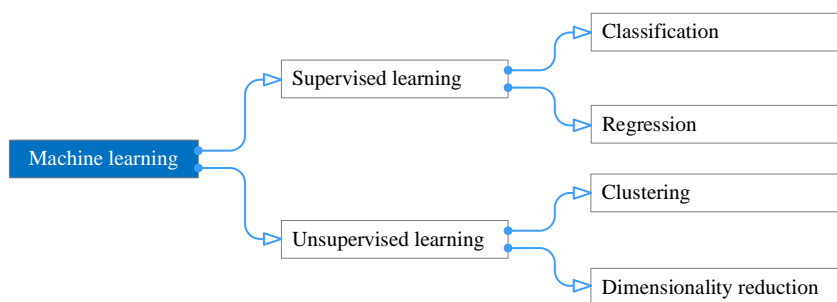


图 7. 机器学习分类

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。

如图 7 所示，简单来说，机器学习可以分为以下两大类：

- ◀ **有监督学习**，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。
- ◀ **无监督学习**训练没有标签值的数据，并发现样本数据的结构和分布。

此外，**半监督学习**结合无监督学习和监督学习。

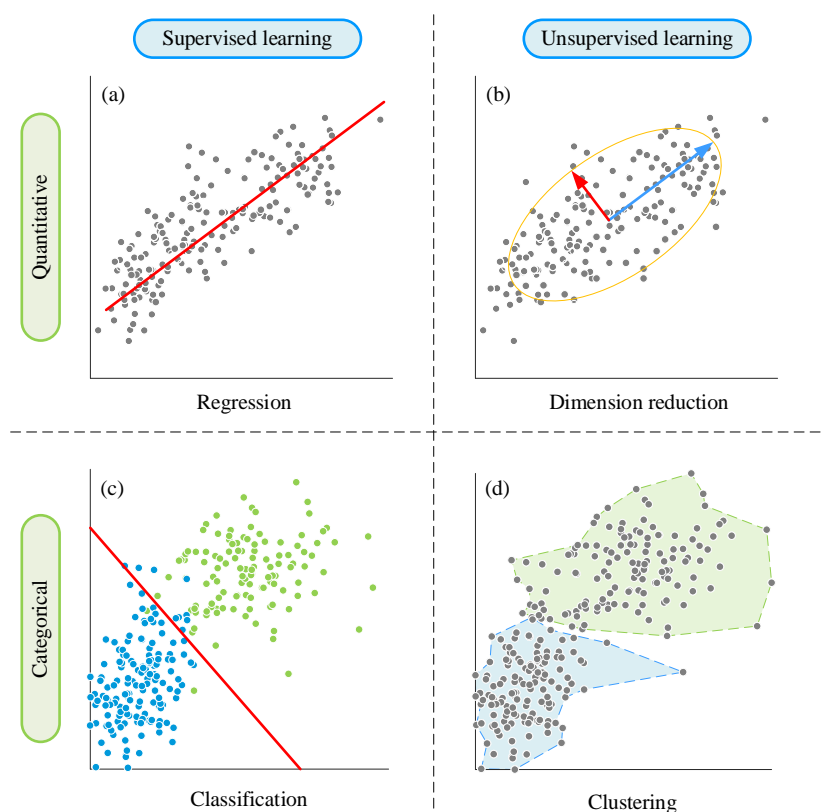


图 8. 根据数据是否有标签、标签类型细分机器学习算法，图片来自《矩阵力量》第 25 章

有监督学习

如图 8 所示，有监督学习可以进一步分为**分类** (classification)、**回归** (regression)。

分类问题是指将数据集划分为不同的类别或标签。给定一个输入，分类模型的目标是预测它所属的类别，如垃圾邮件分类、图像识别和情感分析等。分类问题的输出是一个离散值或类别标签。

回归问题是指根据已知的输入和输出数据，建立一个数学模型来预测输出值。给定一个输入，回归模型的目标是预测它的输出值，如房价预测、股票价格预测和天气预测等。回归问题的输出是一个连续的值或数值。

总的来说，分类问题与离散的输出相关，目标是将数据划分为不同的类别或标签，而回归问题与连续的输出相关，目标是预测数值型数据的结果。

本书将介绍如下几种回归算法：

- ◀ **线性回归** (linear regression)，本书第 10、11 章；
- ◀ **贝叶斯回归** (Bayesian regression)，本书第 12 章；
- ◀ **岭回归** (ridge regression)，本书第 13 章；
- ◀ **套索回归** (LASSO regression)，本书第 13 章；
- ◀ **弹性网络回归** (elastic net regression)，本书第 13 章；
- ◀ **多项式回归** (Polynomial regression)，本书第 14 章；
- ◀ **逻辑回归** (logistic regression)，本书第 15 章；
- ◀ **正交回归** (orthogonal regression)，本书第 18 章；
- ◀ **主元回归** (principal component regression)，本书第 19 章；
- ◀ **偏最小二乘回归** (partial least squares regression)，本书第 19 章。

《机器学习》一册将专门介绍分类算法。

▲ 注意，很多分类算法也可以用来完成回归分析，这也是《机器学习》一册要介绍的内容。

无监督学习

如图 8 所示，无监督学习主要分为**聚类** (clustering)、**降维** (dimensionality reduction)。

降维是指将高维数据映射到低维空间的过程，以便更好地理解和分析数据。通常情况下，高维数据在进行可视化、建模和处理时都会面临计算资源、时间复杂度和维数灾难等问题。通过降维可以减少数据维度，压缩数据，去除冗余信息，提高模型效率和准确度。

聚类是指将数据集中相似的数据分为一类的过程，以便更好地分析和理解数据。聚类分析是一种无监督学习方法，它不需要标记的训练数据，而是根据数据点之间的相似性或距离关系将它们分为不同的簇或群组。聚类可以用于数据挖掘、图像处理、文本分类、市场细分和生物信息学等领域。常见的聚类算法包括 K 均值聚类、层次聚类和 DBSCAN 等。

总的来说，降维是指将高维数据映射到低维空间的过程，目的是减少数据维度、压缩数据、去除冗余信息，而聚类是指将相似的数据分为一类的过程，目的是更好地分析和理解数据。

本书将主要介绍如下降维算法：

- ◀ **主成分分析** (principal component analysis), 本书第 15、16 章;
- ◀ **因子分析** (Factor Analysis), 本书第 19 章;
- ◀ **典型相关分析** (canonical correlation analysis), 本书第 20 章。

《机器学习》一册将专门介绍聚类算法。

1.3 机器学习流程

图 9 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。

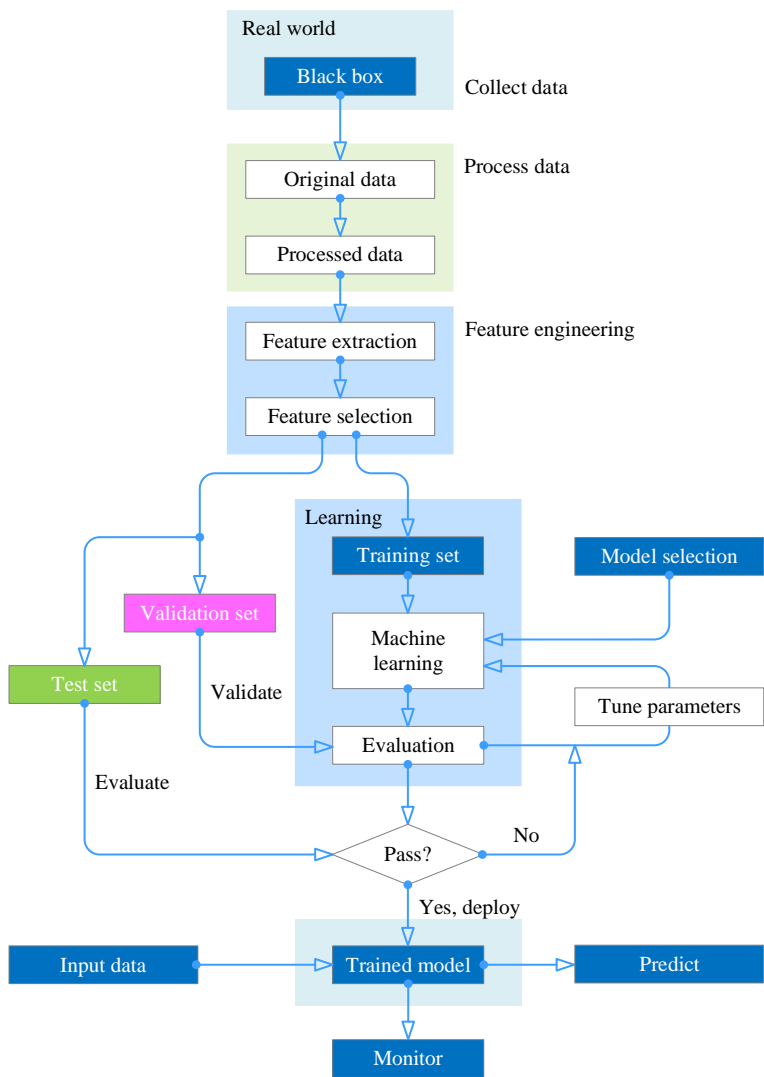


图 9. 机器学习一般流程

1.4 特征工程

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。
- ◀ **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。例如, 在房价预测问题中, 可以根据房屋面积和房龄创建新的特征。
- ◀ **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。例如, 在神经网络中, 可以使用标准化或归一化等方法对数据进行缩放。

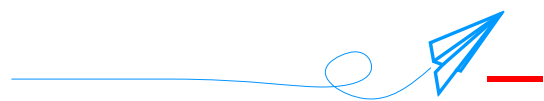
特征工程在机器学习中扮演着至关重要的角色, 它可以提高模型的精度、泛化能力和效率。在实际应用中, 需要根据具体问题选择合适的特征工程方法, 并不断尝试和改进以达到最佳效果。

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。

本书第一个板块“数据处理”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。

本书中的数据转换、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》一册的**独立成分分析** (independent component analysis, ICA)、**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。



本章首先简要介绍了观察数据的不同视角 (表格、线性代数、概率统计、代数)。然后, 讲解了数据分类。

大家特别需要注意根据数据有无标签可以把机器学习分成两个大类——有监督学习、无监督学习。而有监督学习又可以细分为回归、分类。无监督学习则进一步分为降维、聚类。《数据有道》主要讲解回归、降维, 《机器学习》则介绍分类、聚类。

本章最后又聊了聊机器学习的一般流程, 以及特征工程。本书几乎所有内容都可以服务特征工程。



有关特征工程，大家可以参考这本开源专著：

<http://www.feat.engineering/>

Scikit-learn 也有大量特征工程工具，请大家参考：

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