目录

第1章 绪 论

机器学习与模式识别是相关性很强的两个研究领域。尤其在最近十余年，由于机器学习方法的快速发展，两者之间的交叉越发紧密。无论在理论学习还是技术应用中，硬性区分这两个概念并没有太大的意义。因此，本章将通过前两个小节的介绍，从任务和方法两个不同的角度去理解 “机器学习”与“模式识别”，通过介绍这两个领域中常见的概念和术语为读者提供描述和解构这两个主题的基本工具，通过介绍相关理论、方法与技术的发展历史和应用现状帮助读者建立起对机器学习与模式识别的直观感受与理性认识。

1.1 模式识别任务

1.1.1 分类

1935 年，美国植物学家 Edgar Anderson 在加拿大 Gaspe Peninsula 岛上采集了三种不同的鸢尾属花卉（具体种类分别为 setosa，versicolor 和 virginica）的几何形态数据。其目的是为了调查不同地理区域的植物异变现象。次年，英国统计学家 Ronald Fisher 在他的经典论文《The use of multiple measurements in taxonomic problems》中引入了 Anderson 采集的数据，将其整理为著名的 Fisher’s Iris 数据集[1]（也称 Anderson’s Iris 数据集，下文简称 Iris）用于统计分类方法的研究。Iris 数据集一般被认为是模式识别领域早期最具影响力的数据集之一，它包含了从 150 株鸢尾属花卉上测量到的五种不同的数据，包括花萼宽度、花萼长度、花瓣宽度、花瓣长度和花卉种类，如图 1-1 所示。其中左侧十六幅方形图表示使用 iris 数据库四种属性中的任意两种作为横纵坐标画出的散点图，其中三种颜色分别代表三种鸢尾花类别，实物如右侧三幅照片所示。为了便于后面的介绍，我们把数据库中的每一朵花称作是一个“样本”。如何通过对 Iris 数据集的分析与学习，根据鸢尾花的花萼长度、宽度与花瓣长度宽、度去判断一朵鸢尾花属于哪个类别就是模式识别领域最基本的任务之一——“分类”（Classification）。分类任务可以简单定义为：“根据观测数据判断一个对象的类别。”

在现实生活中的分类任务几乎无处不在。早上起床后，睡眼惺忪的你胡乱地看了一眼搭在床头的 T 恤就能分清正反面。你用鼻子闻了一下发现是昨天穿过的脏衣服。吃早餐时，你吸取了昨天的教训，尝了尝佐料罐里装的到底是糖还是盐，避免喝到令人倒胃口的咸咖啡。上学路上，在公交车拥挤的人群中，一段熟悉的音乐响起，你马上意识到是自己的手机响了。你把手伸进装满了各种杂物的书包，仅凭触觉就很快翻出了自己的手机。上述日常行为全部属于模式识别中的分类任务，只不过任务的执行主体是人类的大脑。在模式识别技术应用领域，分类任务也广泛存在。你的电子邮箱网站会自动帮助你区分当前邮件是不是垃圾邮件；银行的自助 ATM 机会利用摄像头与人脸识别技术判断你是否与借记卡持有者是同一个人；内科医生会借助计算机辅助诊断系统判断 CT 影像中的肿瘤是良性还是恶性；钢铁企业会借

[1] https://archive.ics.uci.edu/ml/datasets/Iris

助工业相机拍摄到的图片与视觉检测算法判断一块钢板是否符合质量标准；最新的无人驾驶汽车会在前方安全范围内出现行人时自动刹车。上述案例均属于模式识别领域的分类任务。

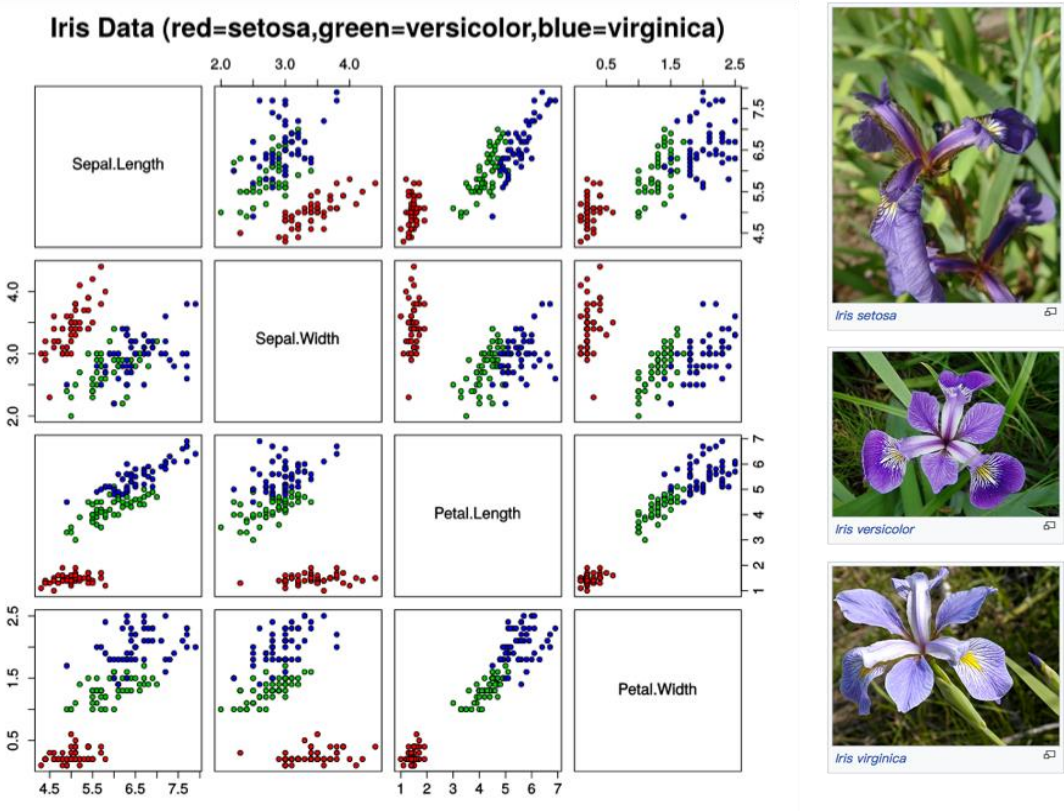


图 1-1　Iris 数据集分布图及实物图。

1.1.2　聚类

在我们谈论分类问题时，总是假设我们已经定义好类别，并掌握了关于类别的某些先验知识或者判别标准。比如在 Iris 数据库中，我们清楚地知道数据库中每朵花的几何形态与类别，我们可以从这些信息中建立起观测数据到类别之间的某种映射关系。但很多时候，我们只能观测到样本，却无法掌握其类别信息。仍以 Iris 数据库为例，假设这个数据库的原始采集者不是那位资深的植物学家 Anderson，而是当地一位正在准备科学课作业的小学生 Peterson。他仍然能测得每朵花的花萼与花瓣的长度和宽度，却无法标记其类别。因此在 Peterson 采集整理的数据集（简称 P-Iris）中，每朵花只记录了 4 项数据，而没有类别标记。假设仍然是那位大名鼎鼎的统计学家 Fisher 拿到了 P-Iris 数据集，他恐怕没有办法准确地给出这些样本的类别定义，但他可以根据这些样本的数据分布把它们人为地分为几类，让每一类内的样本具有更好的相似性，不同类之间具有较大的差异性。这一操作在模式识别领域就被称为聚类（Clustering）任务，其中的类一般称为“簇”（cluster），簇的定义一般没有统一的或预先制定的标准，是根据当前数据的分布情况产生的。

现实生活中有很多潜在的聚类任务。例如：早期的动植物学家们根据观测记录将生物分为不同的门类；一个公司业务员感觉自己的通讯录联系人太多不易查找，需要将所有联系人

分成几个子类分别存储；一个水果店售货员要把刚刚进货的桃子按照品相分成两类，一类卖高价，一类卖低价；视频网站会根据网上浏览信息将用户群体分为多类，并推送不同的广告；复杂工业过程会将系统故障分为几种不同的类型，并针对每一类给出处理预案与检修策略。

与分类任务相比，由于缺少类别信息的指导，聚类任务似乎更加困难，聚类结果的好坏似乎也缺少明确的评价标准。例如当 Peterson 搜集到 P-Iris 数据库时，他似乎没有特别的理由必须按照植物学体系将其分为 setosa，versicolor 和 virginica 三类，他更有可能按照老师的要求或是个人喜好将其分为大花和小花，朝南的花和朝北的花，山顶的花和山下的花等等。上述挑战性使得聚类任务在实现方法上与分类任务有明显的差别（这一点我们在 1.2 节的机器学习方法中会进一步深入讨论），因此在模式识别领域通常认为聚类与分类是两个截然不同的任务类型。

1.1.3 异常检测

尽管分类和聚类两种任务在现实生活中覆盖了大多数与类别判断相关的任务，但仍然存在一些问题无法简单地定义为分类或聚类任务？让我们仍以 Iris 数据库为例，考虑一个新问题。假设 Anderson 有一位叫做 Iverson 的学生，他对老师整理的 Iris 数据库了如指掌，现在他因为一个新的科研项目来到了老师曾经调查过的 Gaspe Peninsula 岛进行花卉数据采集，当他看到一朵花时，他能够根据他的经验判断这朵花是当年老师整理的三类鸢尾花中的具体哪一类吗？

这个问题看起来跟分类问题的标准形态很相似，但问题在于这朵花可能根本就不是鸢尾花，又或者属于鸢尾花中另一个未知的子类。从类别的定义上看，Iverson 可以把 setosa，versicolor 和 virginica 这三类鸢尾花统一为已知类别或“正常”类，而把不属于这三类的所有花朵统称为未知类别或“异常类”。Iverson 需要首先解决的问题是如何判断一朵花属于“正常”类还是“异常”类。这类任务被称为“异常检测”（Anomaly detection）。现实中的异常检测任务也十分常见，例如系统故障诊断、互联网中的用户异常登录检测、面向无人驾驶的路面异常检测、智能视频监控系统中的行人异常行为检测等等。

异常检测与分类和聚类任务的差别在于后两者需要掌握所有参与类别的部分信息（分类任务掌握了所有类别的观测数据和类别标记信息，聚类通常只掌握观测数据），而异常检测通常只掌握了“正常”这一个类别的信息，对“异常”类别却一无所知。因此异常检测也常被称为单类分类问题。当然这一假设在具体应用中也不是绝对的，从已知数据的条件看，异常检测可以分为三种情况：1）正常类别数据已知，异常类别数据未知；2）观测数据同时包括了正常和异常样本，但每一个具体样本的类别未知；3）观测数据同时包含正常和异常样本，且每个样本的类别已知。上述三类问题中第一类属于标准的异常检测任务；第二类本质上属于聚类任务，只不过聚类规则中需要适当考虑异常性的定义；第三类本质上就是分类问题，只不过这类数据集中正常样本的数量通常要远多于异常样本，因此会引发机器学习领域的“样本不均衡”问题。

1.1.4 回归

如果把模式识别看作是根据对事物的观测得到某种结论的过程，那么分类、聚类和异常检测任务的执行结果通常是获得一种关于类别信息的离散化表达，比如人的性别、职业或者

工业产品的种类等。然而在实际生产生活中，还有大量的模式识别任务要求得到更加连续和精细的结论。例如，气象台需要根据观测到的气象数据预测明天某市的 PM2.5 的数值；二手车网站需要根据车的各项参数来估计成交价格；铁路运营商要对春运期间各大城市铁路运输客流量进行精准估计以便提前做好车辆调度。这类问题在模式识别领域一般被称为“回归”（Regression）任务。回归一词来源于统计学中的一个专有名词——回归分析，即确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法。从模式识别角度看，回归任务主要关注如何依据某些变量的观测数据估计或预测出其他变量的精确数值。

经典的波士顿房价预测问题可以帮助我们更加形象地理解模式识别领域的回归任务。美国波士顿房地产市场竞争激烈，如果你想成为该地区最好的房地产经纪人，为了更好地与同行竞争，你需要运用模式识别算法帮助客户为自己的房产定下最佳售价。幸运的是，你找到了一个包含 506 套房产信息的波士顿郊区房产交易信息的数据集[2]。除房价外，该数据集还提供了每套房产的住宅房间数、是否靠河、到市中心的距离、地区犯罪率、地区房产税率、地区教育水平等 13 项数据。我们要完成的回归任务，就是寻找这 13 项数据与房价之间存在的某种未知的函数关系，从而尽可能准确地估计出波士顿地区一套房产的价格。需要特殊提到的是，虽然相当多的回归任务需要估计或预测某一种连续变量（例如价格、流量、温度等），但回归算法从理论上并不要求输出变量必须为连续值。如果把分类问题中的类别看作是一种可以被回归的离散变量，则分类问题也可以看作是回归问题的一种特例。

1.1.5 概率估计

在上面提到的分类和回归两种模式识别任务中，我们总是希望模式识别系统给出一个确定的答案，但在现实生活中有些问题可能无法或者很难给出确定的回答。一个典型的例子就是天气预报。在作者还是学生的时候，中央电视台的天气预报通常只会告诉你明天的天气是晴、多云、阴天、下雨、下雪等典型气象中的一种，但现在的天气预报通常还会增加一项叫做“降水概率”的数据。其实大多数气象分类算法的结果都是基于降水概率估计值所作出的二次判断。之所以采用降水概率直接作为输出变量，可能出于两方面的考虑：一是因为气象分类是一个非常困难的任务，很多时候降水概率的估计值在 50%左右，这意味着如果一定要给出是否下雨的判断，则错判的情况将会经常出现；二是因为降水概率估计值定量地描述了是否下雨这一判断中所包含的不确定性，信息更加丰富，更有利于人们作出符合自己需求的判断。与降水概率估计相似的问题还有很多，例如保险、银行、博彩业普遍存在的商业风险估算，医疗健康领域的很多基因检测业务会给出由某一种基因缺陷引发某一种疾病的患病率，工业领域有大量的贵重设备需要估算其在不同的工作环境、年限和任务量下的设备故障率与报废率。这类任务在模式识别领域统称为概率估计（Probability estimation）问题。概率估计任务是模式识别领域十分常见的基础任务，它在形式上属于回归任务，但在功效上通常用于支撑分类任务。

1.1.6 其他任务与模式识别的广义理解

上述模式识别任务从结果上看倾向于为人们提供一种结论性的表述，例如类别、预测值、

[2] https://github.com/rupakc/UCI-Data-Analysis/tree/master/Boston%20Housing%20Dataset/Boston%20Housing

估计值等等，也存在一些常见的模式识别任务关注于为其他的系统或者算法提供一些阶段性的表述：以人脸识别为例，一幅人脸彩色图像通常包含几万个像素，这意味着一个人脸样本的观测数据维度将高达数万维，一般的模式识别算法很难直接处理这样的高维数据，因此需要另一个算法将高维的人脸数据转化为数百或数十维的数据，这个任务称为**数据降维**，是一种经典的模式识别任务。与降维任务类似，还有一些任务的目的是从复杂冗余的原始数据中提取和转化出少数更加有效和精炼的数据，我们通常称其为**特征选择与特征学习**。

除了上述任务外，不同的观测数据类型和差异化的用户需求还会催生出一些非典型性的、广义上的模式识别任务。例如视频网站或者新闻网站经常需要根据多媒体文件（文本、图片、音乐和视频等）的内容提取或构造出部分关键信息作为多媒体文件检索或浏览的标签（关键词、类型标签、封面图片等）；智能手机中的各类短视频和拍照 APP 中所包含的图像美化与增强现实功能；互联网大数据的关键词分析与可视化；工业大数据中的工艺参数选择与相关性分析；医疗大数据中的基因缺陷与各类疾病的关联性分析；无人驾驶汽车中的路面行人、车辆与交通标志检测、场景图像分割与行为理解；自然图像、医学影像、人脸图像的超分辨率重建与修复；各类互联网搜索引擎中的搜图功能；基于麦克风阵列的语音盲源分离与说话者定位；智能音箱中文本与语音的相互转化；复杂环境下的人形智能机器人运动控制、人机交互与人机协同等。

很多同学目前还不能准确地理解上述五花八门的各类任务为什么可以看作是广义上的模式识别任务。我们可以先试探着以一个通用的“输入-输出”系统来理解模式识别任务。这个系统的输入通常情况下是对某一个对象的观测数据，而输出信号的类型可以有很多选择。在分类、聚类和异常检测任务中，输出信号是类别标签；在回归任务中，输出通常是某一个变量的估计值或预测值；在概率估计任务中，输出是概率质量或者概率密度估计值。而对于上述广义上的模式识别任务，输出信号的形式可以是十分复杂和多样化的，例如一张图片、一段视频、一个表格、一个时空范围、一段语音和文本甚至是机器人的动作。因此从本质上看，模式识别任务的广义概念就是根据输入信号求取用户需求的输出信号的过程。输出信号的数据类型与格式在很大程度上决定了模式识别任务的类型，但模式识别作为一个整体的研究方向并不要求输出信号在形式上必须符合某种既定的规则，这大大扩展了模式识别的研究范围。

1.2 机器学习引论

1.2.1 从模式识别到机器学习

在上一节中我们将广义模式识别描述为一种通用的“输入-输出”系统，利用数学语言可以描述为公式(1.1)。这里X表示系统的输入，通常是对任务对象的某种观测；Y表示用户需求或任务定义的系统输出；F表示输入输出之间的映射关系。模式识别的核心工作就是如何找到、建立和描述映射关系F。

$$Y = F(X) \tag{1.1}$$

获得映射关系 F 的数学描述的方法有三种不同的思路，其中比较传统的思路是**基于知识推理的方法**，即将映射关系F看作是一系列可以用科学原理准确描述的简单关系的某种组合。

例如在一场乒乓球比赛中，如果一个模式识别系统能够观测到乒乓球离开球拍时精确的线速度和角速度，就可以根据运动学方程精确地预测出乒乓球的落点。这里乒乓球的初始运动状态就是观测数据X，乒乓球的落点坐标就是系统输出 Y，而运动学方程解的表达式就是模式识别系统要寻找的映射关系F。这个问题可以被看作是一个预测型的回归任务，其中映射关系F的求取可以被看作是模式识别系统的数理模型建立过程。尽管基于知识推理的系统建模方法解决模式识别问题在理论上是可行的，但在实际应用中却存在诸多困难。首先，系统所需的输入——也就是对象的精确观测数据——通常很难获取或获取成本很高，例如乒乓球的角速度的观测就非常困难；其次，对于不确定性太强的对象——例如人脑、互联网、股票等——很难建立起精准的数理模型；再次，系统的输入输出关系有时会因为太过复杂而无法精确地求解，例如气象预测问题以及天文学上的“三体”问题等；最后，很多实际问题属于“黑盒子”问题，即只能观测系统的输入输出，但无法直接观测输入输出之间的因果关系的问题，例如加密芯片、高温电炉、深海生态系统等等。鉴于上述原因，基于知识推理的模式识别方法通常只适用于问题模型相对确定、输入输出关系比较简单且观测数据比较完备的物理系统。

第二种建模思路是基于经验的方法，将熟悉系统和任务特性的专家或用户的经验总结为一系列的数学规则，进而形成相应的数学模型。例如谚语“早霞不出门，晚霞行千里”就可以看作一种基于经验的模式识别系统，模糊推理系统、专家系统、规则学习都属于这类方法。在这类方法中，映射关系F通常是由领域专家根据经验总结出的某一个函数或者一系列规则的组合，形式上简单直接，在很多缺乏数理模型支撑但却积累了大量专家经验的问题上经常能取得令人惊喜的效果。这类方法的本质是利用人脑从经验中总结规律并数学化，其研究的核心工作在于如何筛选和总结人的经验并将其转化为某种数学描述。然而在实际应用中，人的经验经常是不充分、不准确也不可靠的。因此对于一些复杂问题，如何建立起完备的经验库，如何判断经验是否有效且鲁棒，如何将人的经验准确地转化为数学描述都是非常棘手的问题。因此，此类方法通常只适用于问题相对简单，系统内在数理模型却不甚明确或观测数据的数学描述难以建立的问题。

最后一种思路是基于学习的方法。这种方法不关注输入输出关系的内在数理逻辑，其核心思想是通过某种学习算法找到一个数学映射F，使其能够更好地吻合已经观测到的输入输出数据。此类方法通常要先建立一个通用的参数模型，然后确定一个与观测数据密切相关的学习目标，最后使用某种优化方法找到一组最优的模型参数对映射关系F进行数学描述。因此，基于学习的模式识别方法的本质是优化问题。但是这类优化问题与经典的优化问题有比较大的差别，它通常会模拟人类的自主学习过程，它在问题的覆盖面上更广且更具一般性，它的目标函数更关注于数据而不是物理模型，它的学习过程更具自主性和适应性，我们称这类方法为“机器学习”。

在上述三类方法中，机器学习是模式识别领域现阶段最主要也最先进的研究方法，比较常见的机器学习类型包括有监督学习(Supervised Learning)、无监督学习(Unsupervised Learning)、弱监督学习(Weakly Supervised Learning)和强化学习(Reinforcement Learning)等，以下我们将逐一进行简单介绍。

1.2.2 有监督学习

有监督学习(Supervised Learning)，又称监督学习，是机器学习中技术最成熟、应用最广

泛同时也是平均性能最好的一种学习方法。前面提到过，机器学习主要采用从数据中学习的思路，因此“监督”这个词主要是指数据中是否直接具有“监督信号”。

以鸢尾花分类任务为例，输入数据是花的四种几何形态数据（花瓣的长和宽，花萼的长和宽），输出数据是花的类别。标准的 Fisher Iris 数据库中既包含了输入数据，又包含了正确的输出数据。这个正确的输出数据就是我们需要的“监督信号”，也就是公式(1.1)中的Y的标准答案。充分利用 Iris 数据库中的观测数据与监督信号设计一种学习算法进而得到分类模型的过程就是典型的有监督学习。从另一个角度看，有监督学习可以理解为由老师直接教导的学习过程，老师的教导是以“告知正确答案”的形式实现的，因此“监督信号”有时也称为“教师信号”。

有监督学习由于能够从标准答案中直接学习，因此学习效率非常高。但所谓“成也萧何、败也萧何”，由于需要标准答案，有监督学习的数据获取成本也非常高，其中工作量最大的是人工标记过程。仍以 Iris 数据库为例，150 朵花的类别需要植物学专家 Anderson 逐一对样本所述的鸢尾花类别进行标注，而 Anderson 的标注过程是否仅仅依靠鸢尾花的 4 种几何形态参数就能实现呢？恐怕不行，他可能还需要依靠其他更加具有可区分性的植物学指标进行判断，因此也就顺带增加了观测成本。如果仅仅是几百个样本，这种人工标注的工作量还可以接受。但随着大数据与深度学习技术的不断发展，现在的机器学习算法动辄需要几百万乃至上亿个样本，人工标注工作量巨大到难以承受，例如计算机视觉领域最具代表性的 ImageNet 图像数据库包含几百万张图片，分为几千个类别，需要数百名专业人员花费几个月甚至几年的时间进行人工标注。不断加剧的人工成本在很大程度上限制了有监督学习方法的发展，因此近几年来其他几类标记成本较低的机器学习方法也开始得到了学术界和工业界的普遍关注。

1.2.3 无监督学习

无监督学习指没有监督信号的学习类型。观测数据中只包含输入数据，没有输出数据。以我们在 1.1.2 节的聚类任务中提到的小学生 Peterson 搜集的 P-Iris 数据库为例，该数据库中也包含了 150 朵花的属性特征，但没有记录花的类别。通过对这样的数据库的聚类学习得到分类模型的过程就属于典型的无监督学习范畴。比较常见的无监督学习任务包括聚类、降维、特征学习与选择、可视化等。由于缺少监督信号的指导，无监督学习通常效率偏低，且最终结果的准确性也较差。但无监督学习最大的优势在于无需人工标记，因此数据库获取成本很低，因此可以使用更多的数据参与模型训练。此外，无监督学习也适用于一些特殊的、难以直接获得教师信号的任务。

1.2.4 弱监督学习

有监督学习效率高，结果好，但数据采集和标记成本也高；无监督学习数据采集成本低，但学习效率和识别性能难以令人满意。有没有一种方法能够适当地调和上述两种学习模式的优缺点呢？答案是：弱监督学习。

弱监督学习是指在监督信号不够完善的情况下可以使用的机器学习方法。所谓“不够完善”，有时表现为监督信号数量不足，有时表现为监督信号存在错误或者较大的误差，有时则表现为监督信号不够确切。在上述情况下，如何更加充分利用标签数量不足的、不够准确或不够具体的监督信号进行模型的学习，就是弱监督学习方法要解决的问题。

相比于有监督学习，弱监督学习对于监督信号数量和质量的要求较低，可以节省大量的标记成本；而与无监督学习相比，弱监督学习由于引入了一定的监督信号所以学习结果的平均性能更好。目前深度学习方法对于数据的需求量越来越大，而全球范围内基于互联网与物联网的大数据获取途径虽然能够提供海量的观测样本，但无法提供在质量和数量上与之相匹配的监督信号。这一现状与弱监督学习的假设条件非常吻合。因此近年来弱监督学习逐渐成为机器学习与模式识别领域的研究热点之一。

1.2.5 强化学习

在有监督学习的介绍中，我们发现监督信号对于提升学习效率是非常重要的，但遗憾之处在于监督信号的获取通常伴随着大量的人力成本。一个顺理成章的想法是，有没有可能找到一种成本低廉的监督信号获取方式呢？最好是不需要人工参与，但又能保证监督信号的基本质量。这看起来似乎是一种“鱼与熊掌，可以得兼”的奢望，因为如果我们已经掌握了一种可以自动给出监督信号的方法，又何必去使用机器学习算法去预测这个标准答案呢？但强化学习为这个看起来似乎是悖论的奢望，探索出了一条非常巧妙的途径。

让我们通过一个案例——婴儿学走路——来简要介绍强化学习的思路。婴儿学习走路的方法显然不可能是有监督学习，因为如果将父母视为教师，他们其实也不知道具体该如何控制肌肉才能走得稳当。即便他们知道答案也无法把这种复杂的知识教给婴儿。那么婴儿是通过模仿父母走路的姿势学会走路的吗？可能有一点点这方面的因素，但以婴儿的大脑发育水平，也不大可能掌握模仿（或者学术范一点，称为示教学习）的窍门，更何况如何控制肌肉这件事从外表是看不出来的。其实真正教会婴儿走路的，是他与环境之间的交互。婴儿在环境中进行动作的探索，一旦他做出的动作错误，就会导致跌倒，环境会通过碰撞给他一个负面的反馈，我们称之为惩罚信号。而一旦婴儿偶然走对了步伐，直立行走的新奇体验和父母的赞扬与鼓励就会形成一个正面的反馈，我们称之为奖励信号。婴儿在尝试行走的过程中探索不同的动作，并根据环境的反馈，避开那些会导致惩罚信号的动作，重复和增强那些会得到奖励的动作，最终学会走路。这个过程就是强化学习的过程。“强化（reinforcement）”一词，是指环境的正面反馈对学习者正确行为的不断加强这一现象。

强化学习中通过行为与环境交互作用来提供监督信号的思路非常巧妙，但这种学习机制也依赖于某些特定要求。首先强化学习需要一个合适的环境，可以是真实的，也可以是虚拟的，但它必须跟学习模型的应用环境具有良好的一致性，这样从学习环境中得到的经验才能够适应于应用环境。其次，这个环境必须能够自动地给出比较可靠的惩罚或奖励信号来引导强化学习的进程。这两个要求看似简单，但其实并不容易获得。比如对于 iris 数据库的花卉分类问题，就很难构造出一个满足上述条件的环境，因为通过某种试探行为来构造出一朵鸢尾花非常困难，而要判断这朵花的分类结果是否正确也仍然需要大量的人工成本。因此，强化学习通常应用于一些相对封闭、可以模拟或真实环境实验成本较低的问题。例如机器人运动控制策略的学习，可以直接使用真实物理世界来获得站立和跌倒的反馈；围棋软件的学习，可以利用基于围棋规则构建的虚拟下棋软件给出输和赢的反馈。

近年来，深度学习与强化学习相结合产生的深度强化学习方法成为机器学习领域发展非常迅速的一个分支，尤其在机器博弈、无人驾驶、运动控制等领域都取得了令人瞩目的成绩。但深度强化学习仍然主要关注控制与决策问题，在模式识别领域的应用还有非常广阔的空间。

1.3 基本概念与术语

在利用机器学习方法解决模式识别问题的过程中，经常会出现一些对初学者来说比较陌生的概念、术语及其定义，虽然准确深刻地掌握和理解它们并不容易，但在学习之初对他们有一个感性的认识对于未来的学习是有帮助的。如 **Error! Reference source not found.**所示，本节涉及到的基本概念分为数据和模型两个部分。

1.3.1 数据相关概念

在模式识别领域，数据是对象观测结果的一种符号化描述，形式上可以使用由数值或符号组成的矢量、矩阵等结构化形式，也可以使用图像、声音、文本等非结构化形式。对于对象的一次观察所形成的数据，作为数据集的一个基本单元，被称为**样本**(Sample)或**实例**(Instance)。在机器学习中，一部分样本作为学习的依据对模型进行训练，被称为**训练样本**(Training samples)；另一部分样本不参与学习，但会用于评估学习结果的好坏，被称为**测试样本**(Testing samples)，所有可能的样本的集合称为**样本空间**(Sample Space)。例如，Iris 数据库中一共有 150 个鸢尾花样本，分为三类，每类 50 个样本。如果每类用 30 个样本进行训练，另外 20 个样本作为测试，则一共有 90 个训练样本和 60 个测试样本。

用于描述样本某一方面特性和表现的数据称为**特征**(Feature)或**属性**(Attribute)。一个样本的多个特征如果可以用一个向量来描述，则称为该样本的**特征向量**(Feature Vector)。这个特征向量的长度称为样本的**特征维度**(Feature Dimension)。对于一类样本而言，所有可能观测到的样本的特征向量所支撑的空间，称为**特征空间**(Feature Space)。例如，Iris 数据库的每个样本都包括花瓣长度、花瓣宽度、花萼长度和花萼宽度，共 4 个特征，所以特征向量是一个维度为 4 的向量，其特征空间是一个 4 维空间。而每朵花的类别在分类任务中称为类别的**标签**(Label)。

如果将一个样本的特征向量看作是对表征某一类对象的随机特征向量的一次采样，那么这个随机特征向量可以由定义在特征空间上某种**概率分布**(Probability distribution)来描述。在没有进行观测前，这个类别出现的概率称为**先验概率**(prior probability)；如果已经进行了观测，以观测结果为条件得到的关于这个类别出现的条件概率称为**后验概率**(posterior probability)。如果该类样本的特征是连续随机变量，根据概率论中的相关定义，可以用**概率密度函数**(Probability Density Function，PDF)来描述其在特征空间中的概率分布。另一种在数学形式上与概率或概率密度函数相同，但主要用于描述模型参数的可能性的概念称作**似然性**(Likelihood)。似然性是概率模型参数的函数，而概率密度函数是观测特征值的函数，两者形似而神非（关于似然性的详细定义将在第三章中给出）。

在模式识别任务中，**模式**(Pattern)这个概念看似玄妙，其实就是指识别对象。更直观地说，就是特征向量所描述的对象。而**模式类**(Pattern Class)则是模式组成的类别，它通常与分类问题中的类别对应。仍以 Iris 数据库为例，150 朵花就对应于 150 个模式，分别从属于 3 个模式类。判断一个“模式”属于哪一个“模式类”就是模式识别的核心任务——分类。

1.3.2 模型相关概念

要解决模式识别任务，需要建立一个模式识别系统，描述这个模式识别系统的数学形式称为**模型**(Model)，如果这个模型需要通过机器学习的方法得到，也可以称为学习模型。

模式识别系统需要根据对象的观测数据，得到相应的类别或某种估计。所以模式识别系统的模型一般都具有“输入”和“输出”。输入通常是样本的特征向量，而输出则是对当前样本的分类标签或某个相关变量的评估值。例如在 Iris 数据库分类任务中，输入为花瓣宽度等四个特征组成的 4 维特征向量，输出为鸢尾花类别的标签。而在另一个任务中，我们也可以使用花瓣长度、花瓣宽度、花萼长度以及类别作为输入，以当前样本花萼宽度的估计值作为输出。因此，输入和输出由数据集和模式识别任务共同决定。

当输入与输出确定后，通过学习得到的具体模型反映了输入与输出之间的某种关系。这个关系被称为**假设**(Hypothesis)。与其对应的，在物理世界中输入输出之间实际存在的关系或规律被称为**真实**(Ground Truth)。假设在数学形式上一般采用带有**参数**(Parameter)的函数来表示，例如线性函数、多项式函数、对数几率函数等。在机器学习过程中，假设的函数形式一般不会发生变化，但参数会随着学习而改变。所以不同的参数取值对应着不同的假设。对于一个模型而言，所有可能的参数取值的集合称为**参数空间**（Parameter Space），而所有符合定义的假设的集合称为**假设空间**(Hypothesis Space)。

模型学习过程就是通过对参数的学习，让假设逼近真实的过程。显而易见，只有当假设空间包含“真实”时，才有可能通过机器学习方法使假设逼近“真实”。在“真实”不可知的情况下，一个模型的假设空间越大，它拟合“真实”的能力就越强。因此，我们可以进一步定义模型的**容量**(Capacity)，即该模型拟合各种“真实”的能力，容量的大小通常由模型的假设空间决定。以上涉及的概念比较抽象，我们可以通过一个示例加以解释。

- **模型相关概念示例**

假设一个物理系统，其真实的输入输出关系是$y = 5x^2$，其中x为输入，y为输出。观测者对该系统的 100 组输入输出信号进行了测量与记录，在考虑了测量误差与噪声的情况下，记录数据如图 1-2 中散点所示，记为训练集$\{x_i, y_i | i = 1,2,\dots,100\}$。另一个负责建模的工程师尝试基于上述数据学习一个模式识别系统$h(x)$，使其能够根据输入量x去估计该物理系统的输出量y。建模工程师根据经验以及对数据分布的观测，将该任务判定为一个回归任务，并设计了一个模式识别系统，其输入输出关系设定为三次多项式函数$h(x) = \alpha_1 x^3 + \alpha_2 x^2 + \alpha_3 x + \beta$。工程师利用某种机器学习算法，基于观测数据对该模式识别系统进行训练，得到拟合效果最好的 4 个参数取值为：$\alpha_1 = 0.0001, \alpha_2 = 4.98, \alpha_3 = 0.002, \beta = 0.03$，其对应的曲线如图 1-2 中蓝色实线所示，可以看到其余红色虚线表示的真实$y = 5x^2$存在一定的差异，这通常与观测误差、假设空间和学习算的优化性能有关。在该示例中，模型相关概念的解释如表 1-1 所示。

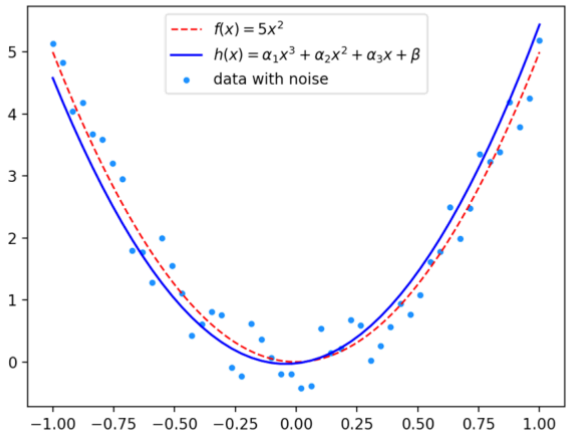


图 1-2　模型相关概念示意图

表 1-1　模型相关概念解释

概念	**解释**
真实	$y = f(x) = 5x^2$
模型	$h(x) = \alpha_1 x^3 + \alpha_2 x^2 + \alpha_3 x + \beta$
参数	$[\alpha_1, \alpha_2, \alpha_3, \beta]$的一组具体取值
参数空间	$[\alpha_1, \alpha_2, \alpha_3, \beta]$的所有可能取值的集合
假设	$[\alpha_1, \alpha_2, \alpha_3, \beta]$的一组具体取值所对应的三次多项式函数$h(x; \alpha_1, \alpha_2, \alpha_3, \beta)$
假设空间	所有可能的三次多项式函数

在上面的示例中，如果观测者将另外记录的 50 组带有噪声的实测数据$\{x_j', y_j' | j = 1,2, \ldots, 50\}$作为测试集，将其中的输入量$x_j'$送入工程师训练出的模式识别系统，系统给出的输出值$\hat{y}_j'$与真实测量的输出值$y_j'$会存在一定的差异，这是模型通过对数据的学习以获取逼近真实的假设过程中存在一个巨大的挑战，这是因为有限的训练样本集大多数时候不可能完全充分和准确地反映“真实”。

以 Iris 数据库为例，在模型学习过程中我们只有 150 个样本可以参考，即便我们学习到的“假设”对于当前的 150 个样本全部成立，也不意味着该假设对于样本空间中其他没有被观测到的鸢尾花样本都成立。因此，机器学习领域提出了**泛化**(Generalization)能力的概念用于描述学得模型适应新样本的能力。具有良好泛化能力的模型，能够适用于样本空间的大部分区域，因此对于训练过程中没有学习过的新样本也具有良好的识别性能。然而想获取好的泛化能力并不容易。当模型的容量远远大于实际任务所需的假设空间，模型学习的结果通常会对训练样本过于敏感，换言之模型利用其庞大的容量记住了训练样本的所有具体的、微小的、甚至可能是由观测噪声引发的特征。只有当输入样本与训练样本非常一致的时候，它才能被模型正确识别。这意味着模型对新样本的泛化能力会明显下降。这种情况被称为**“过拟合”**（Overfitting）。反之，当模型容量不够大时，即便对训练样本也达不到良好的识别效果，对新样本的泛化能力就更无从谈起，这种现象称为**“欠拟合”**（Underfitting）。一般来说，过

拟合现象在缺少验证样本的时候，不太容易被发现，而欠拟合现象在训练样本集上就能验证，比较容易发现。因此如何避免出现过拟合现象一直是模式识别与机器学习领域一个巨大的挑战。关于过拟合的问题，我们将会会在后续章节中进行更加深入的探讨。

无论是泛化能力还是过拟合现象，只是对模型是否能够胜任识别任务的一种定性描述。从专业研究的需求出发，我们必须建立一整套精细的、定量的评价体系来描述模型的性能。我们称这一任务为**模型评估**。模型评估涉及到相当多的**性能指标**，例如均方误差、PR 曲线、ROC 曲线、AUC、识别率、真阳率、假阳率、虚警率、漏检率、F1-measure 等等，这些指标的具体定义我们将会在后面的“模型评估”一章给出具体的介绍。

1.4 发展历史

在对模式识别与机器学习的基本概念有了简单的认识后，本书将对相关领域研究历史做一个简单的回顾与梳理，以方便读者从发展的角度对模式识别与机器学习建立起更加系统性的认知，同时也有利于读者理解本书章节安排的内在逻辑。

- **起源与诞生**

“模式识别”（Pattern Recognition）这个术语的起源可以追溯到 19 世纪末。最早提出这个概念的人之一是英国统计学家 Sir Francis Galton。Francis Galton 于 1883 年在他的一篇论文中首次提到了“合成肖像”（Composite Portraits）的概念。他通过将多张人脸照片叠加在一起来创建一种平均脸型，以探索人类面部特征的平均表现。该项工作反映了研究者们对“模式”这一抽象概念在统计学和心理学方面的早期认识。然而受限于当时的计算工具，人们还无法建立起科学的、定量的模式识别方法。

1940-50 年代，Claude Shannon 等人构建了信息论的理论基础，其中**“熵”**（Entropy）的概念的提出，对于以统计学为根基的早期模式识别工作具有重要的奠基作用。1956 年，John McCarthy，Marvin Minsky 等人共同主办的**达特茅斯会议**被认为是人工智能学科奠基的标志性事件，尽管会议的主要焦点是人工智能，但会议在学习与自适应系统、自然语言处理、机器感知等议题上的讨论对模式识别领域产生了深远的影响。

早期模式识别研究的一个重要支柱来源于统计学理论，主要集中于**最大似然估计**（Maximum Likelihood Estimation，MLE）与**贝叶斯分类器**（Bayesian Classifier）在模式识别领域的应用，并逐步形成了**贝叶斯决策论**的理论架构。MLE 理论最早可以追溯到数学家 Carl Friedrich Gauss 在 19 世纪早期的工作。1940-60 年代，Jerzy Neyman、Egon Pearson 和 Ronald A. Fisher 等统计学家对 MLE 理论在模式识别领域的应用起到了重要推动作用。贝叶斯分类器的基本原理——**贝叶斯定理**——是由 Thomas Bayes 在 18 世纪晚期提出的，Thomas Cover 等人则在 1960 年代开始尝试将其应用于模式识别的具体应用。

模式识别早期研究中另一项具有里程碑意义的工作是美国心理学家 Frank Rosenblatt 于 1957 年提出**感知器**（Perceptron）模型，它是基于生物神经元工作原理构建的一个单层神经网络模型。这项工作的重要意义在于首次赋予神经元模型学习的能力，并在应用层面实现了模式二分类任务。因此无论在人工智能的大领域，还是在机器学习与模式识别的小领域，感知器模型都具有开创性的历史意义，本书将在第四章线性模型部分对这一算法进行详细的介绍。

贝叶斯决策论与感知器既是模式识别领域的早期代表性工作，也分别引领了机器学习与模式识别理论研究的两个流派——统计学习与联结主义。前者以概率论为基础，通过构架模式类的概率分布模型的实现模式识别；后者以神经网络为主要工具，通过模仿人脑神经系统的结构与功能实现模式识别。

● 发展与挑战

1970 年代到 20 世纪末可以看做是模式识别与机器学习领域逐渐发展的阶段，诞生了一大批具有一定实用价值的经典算法，其理论方法与应用技术的体系也逐渐完善，并随着一系列典型应用案例和经典教材的诞生，逐渐开始进入工业界与教育界。与此同时，联结主义学派却一度遭遇了非常大的挑战，但终究又柳暗花明，豁然开朗。

基于统计学习的模式识别方法研究发展比较平稳，陆续出现了一批具有重要理论意义和巨大应用价值的研究成果。在贝叶斯决策论的框架下，为了解决复杂分布下的分类问题，Arthur. Dempster 和 N. M. Laird 等人提出了**高斯混合模型**（Gaussian Mixture Model，GMM）和**期望最大化**（Expectation-Maximization，EM）算法，大幅度提高了贝叶斯分类器的实用性；沿着这一路线，**朴素贝叶斯分类器**（Naïve Bayesian Classifier, NBC）也被应用于具体的模式识别问题以应对高维数据的挑战；Judea Pearl 和 David Heckerman 等人则将基于概率图模型的**贝叶斯网络**（Bayesian Network）应用于概率回归与不确定性推理任务，并沿此方向逐渐衍生出**隐马尔科夫模型**（Hidden Markov Model, HMM）、**马尔科夫随机场**（Markov Random Fields，MRF）和**条件随机场**（Conditional Random Fields，CRF）等方法。与此同时，基于统计回归的后验概率估计方法也得到了长足发展，代表性的方法包括**线性回归**（Linear Regression）与**逻辑回归**（Logistic Regression）。1973 年， Richard O. Duda、Peter E. Hart 和 David G. Stork 合著的经典书籍**《模式分类与场景分析》**（Pattern Classification and Scene Analysis）正式出版。该书从特征选择、与分布无关的分类算法、统计分类方法、非监督学习和顺序学习等方面对当时相对成熟的各类模式识别算法进行了汇总、整理与系统化介绍，对该领域的教育和研究产生了深远的影响。该书籍的出版也被认为是模式识别领域进入成熟阶段的标志性事件之一。1984 年，计算机科学家 Leslie Valiant 首次提出了 **PAC**（Probably Approximately Correct）学习理论，为机器学习提供了定量化统计分析的基本框架。至此基于统计学习的模式识别理论体系基本建立起来。

到了 20 世纪后期，统计机器学习方法百花齐放，在线性分类器和**结构风险最小化**（Structural Risk Minimization，SRM）原则的基础上，Vladimir Vapnik 提出了大名鼎鼎的**支持向量机**（Support Vector Machine，SVM），并拓展了统计学习理论，从而将线性分类技术真正意义上扩展到非线性分类领域，并在实际应用中取得了令人瞩目的成绩。在 Shannon 的信息论基础上发展起来的**决策树**（Decision Tree）方法，则在离散数据的分类任务中大放异彩，Ross Quinlan 提出的 **ID3，C4.5** 决策树算法，Jerome Friedman 等人提出的**分类回归树 (Classification and Regression Tree, CART)算法**大幅提升了决策树算法的应用范围，而 Leo Breiman 将决策树与**集成学习策略**相结合，于 2001 年提出了经典的**随机森林**（Random Forest, RF）算法，显著提高了统计学习分类算法的泛化能力，有力地推动了模式识别算法在实际工程中的应用。

以神经网络为主要研究对象的联结主义学派的发展则是一波三折。受到人工智能整体发

展趋势的影响，联结主义学派的思想受到了巨大的质疑，人工智能领域的开山鼻祖之一 Marvin Minsky 在 1969 年出版的**《Perceptron》**一书对感知器算法的局限性进行了尖锐的批判，英国数学家莱特希尔则在他的人工智能领域调研报告中对现有模型结构的局限性，大数据缺失和计算能力限制等问题给出了非常悲观的预测。一系列打击使得人工智能领域尤其是神经网络相关研究在 1970-80 年代进入了第一次寒冬。然而 Werbos 在 1974 年的博士论文中对多层感知器的学习算法进行了全新的尝试；1980 年代初，美国物理学家 John Hopfield 提出的 **Hopfield 网络**和加拿大学者 Geoffrey Hinton 提出的**玻尔兹曼机**（Boltzmann Machine）则为神经网络研究在模式识别和优化领域的研究探索了新的道路，揭开了联结主义学派复兴运动的序幕；1986 年，David Rumelhart、Geoffrey Hinton 和 Ronald Williams 正式提出了**反向传播算法**（Backpropagation，BP），该算法解决了多层感知器神经网络训练的关键问题，使联结主义学派的思想重新受到关注，神经网络在一些相对简单的工程问题上得到了实际应用。

- **突破与繁荣**

进入 21 世纪后，随着传感器技术、互联网和高速无线通信技术的飞速发展，模式识别技术的应用对象从早期的几维到几十维的结构化数据逐渐转变为文本、语音、图像、视频等数千到数千万维的超高维数据，这给统计机器学习和联结主义学派都带来的巨大的挑战。

统计学习方法解决该问题的思路是将模式识别任务与数据降维算法相结合。对于图像、语音和文本等特殊的非结构化数据，采用手工特征提取方法降低数据维度，相关研究推动了图像处理、计算机视觉、语音识别、自然语言处理等领域的快速发展；对于非特定的高维数据，可以采用**主元分析**（Principal Component Analysis, PCA）、**独立元分析**（Independent Component Analysis，ICA）、**稀疏编码**（Sparse Coding）、**流形学习**（Manifold Learning）、**词袋模型**（Bag of Features, BoF）等统计特征提取算法实现数据降维。经过降维处理后，超高维数据通常被转化为几十到几百维的特征向量，再采用线性回归、逻辑回归、SVM、决策树和随机森林等算法进行模式识别。上述方法可以总结为特征提取与模式识别相互分离的两阶段策略，在一些规模较小且相对简单的问题上取得了不错的应用效果，但对于复杂开放环境下的视频、图像、声音和文本等超高维非结构化数据的模式识别任务的处理能力仍然非常有限，很难达到商用要求。

联结主义学派处理超高维复杂数据的核心思路是通过加深神经网络的层数来获取更强的数据表征能力。然而受到**“梯度消失”问题**的困扰，传统的 BP 算法很难实现三层以上的神经网络的迭代优化。Hinton 等人首先尝试在玻尔兹曼机的基础上通过启发式权重初始化、逐层无监督预训练、逐层贪婪训练与多层特征堆叠等技术手段构建了**深度置信网络**(Deep Belief Networks, DBN)；而 **ReLU**（Rectified Linear Unit）激活函数及其变种的出现彻底解决了深度网络的梯度消失问题。在 2012 年的 ImageNet 大规模视觉识别竞赛中，以 **AlexNet** 为代表的深度神经网络首次在 ImageNet 竞赛中获胜，将分类错误率降低到 16%以下，远低于传统方法。这一成就引起了广泛的关注，并被认为是人工智能进入深度学习时代的标志性事件。

- **深度时代**

随着深度神经网络的深入研究与广泛应用，模式识别与机器学习作为人工智能的核心领域，一起进入了深度学习时代，而基于统计学习的模式识别方法研究更多地体现为对深度模

型的理论引导与策略设计。在深度时代，模式识别与机器学习领域的发展特点是理论研究与技术应用的广泛深入融合。

在理论研究方面，由于深度神经网络为复杂的网络结构提供了广阔的设计空间，一系列各具特色的网络结构模型应运而生，包括具有不同计算单元结构的**卷积神经网络**（Convolutional Neural Network，CNN）、**循环神经网络**（Recurrent Neural Network，RNN）、**长短时记忆网络**（Long Short-Term Memory，LSTM）、**门控循环单元**（Gated Recurrent Unit，GRU）、**图卷积网络**（Graph Convolutional Network）、**残差网络**（Residual Network，ResNet），以及具有不同网络架构的**深度置信网络（DBN）**、**自编码器**（Autoencoder，AE）、**注意力机制**（Attention Mechanism）、**Transformer 网络**和**扩散模型**（Diffusion Model）等，其总体发展趋势表现为层数加深、结构复杂化、网络参数增加、计算需求大幅度提高、识别精度与泛化能力显著增强。此外，由于深度神经网络的训练需要海量数据的支撑，在有监督学习模式下会导致人工标注成本激增，因此近年来大量的研究工作开始从有监督方法转向弱监督、无监督和强化学习等无需或较少需要人工标注的学习方法，代表性工作包括**度量学习**（Metric Learning）、**自步学习**（Self-paced Learning）、**变分自编码器**（Variational AutoEncoder，VAE）、**生成对抗网络**（Generative Adversarial Network，GAN）、**孪生网络**（Siamese Networks）、**对比学习**（Contrastive Learning）、**自监督学习**（Self-Supervised Learning）以及各类数据增强与标签伪造技术等，其总体研究思路通过设计无须标签的损失函数或标签伪造策略解决标签缺失带来的不利影响。上述方法降低了模型训练对样本标签的巨大需求，因此可以更加充分地利用海量数据，从而达到提升深度模型性能的目的。

在技术应用层面，深度神经网络与大数据结合带来的强大识别能力使得深度模型能够妥善解决很多复杂的模式识别任务，并在表现上接近或超越了人类的平均水平，从而引发了人工智能技术应用革命的新浪潮。

在图像识别领域，人类在 **ImageNet 图像识别挑战赛**上的 Top-5 错误率约在 5%-10%之间，**GoogleNet** 于 2014 年在该比赛中取得了 6.7%的 Top-5 错误率，首次超越了人类平均水平，此后不断推出的 **ResNet、Inception-V4、SENet、EfficientNet** 等模型不断刷新这一指标，目前最新的 Top-5 错误率已经接近 1%，远超人类平均水平，这意味着以图像分类为代表的计算机视觉任务得到了实用性解决方案。此外，Facebook 在 2014 年推出的 **DeepFace** 模型使得人脸识别技术的大规模商用成为可能，Google 公司的 **FaceNet**、牛津大学的 **VGGFace** 和香港中文大学的 **DeepID** 进一步加速了这一过程。目前人脸识别作为深度学习最成熟的商业应用技术之一已经在门禁系统、智能手机、支付安全、互联网社交平台、安全监控和身份认证等领域取得了广泛的应用。

在语音识别领域，2017 年提出的 **WaveNet** 是一种基于本文生成高质量语音和音乐的模型，被广泛应用于文本到语音的转化应用；百度公司提出的 Deep Speech 模型在多语言和多方言识别任务上取得了出色的性能；Google、微软和亚马逊等公司都推出了自己的智能音箱和智能语音助手产品，在全球范围内深刻地改变了人类与机器之间信息交互模式。

在自然语言处理方面，2017 年出现的 **Transformer 网络**显著提升了自然语言处理算法的性能，**BERT** 和 **GPT** 模型的出现使得文本翻译、机器问答和文本摘要和扩写等技术的商用成为可能。2020 年 OpenAI 公司推出的 **GPT-3** 取得了惊人的自然语言处理成果，2022 年陆续推出的 **ChatGPT** 和 **GPT-4** 使语言大模型的性能取得历史性突破，并开始扩展到图像和语音等

多模态数据领域，语言大模型的突破性进展改变了人们在信息搜集、文本写作和工作决策等几乎所有与语言相关的任务上的行为模式，大幅度提高了人类的工作效率，是人工智能领域发展的一个里程碑，并将深远地影响人类社会发展的进程。

在数据生成领域，Ian Goodfellow 等人于 2014 年提出了**生成对抗网络**（GAN），可以生成逼真的图像、音频和视频数据，从而开辟了人工智能在数据生成方面的全新研究方向。目前以 Midjourney 公司为代表的图像生成平台在全球已经有超过数千万注册用户，并仍在高速增长。目前在全世界范围内基于 AI 生成技术进行广告设计、插画生成和短视频特效制作已经成为设计行业从业者的日常选择，而可以逼真地伪造人脸视频和语音的 **Deepfake** 技术则给人工智能伦理带来了新的挑战。

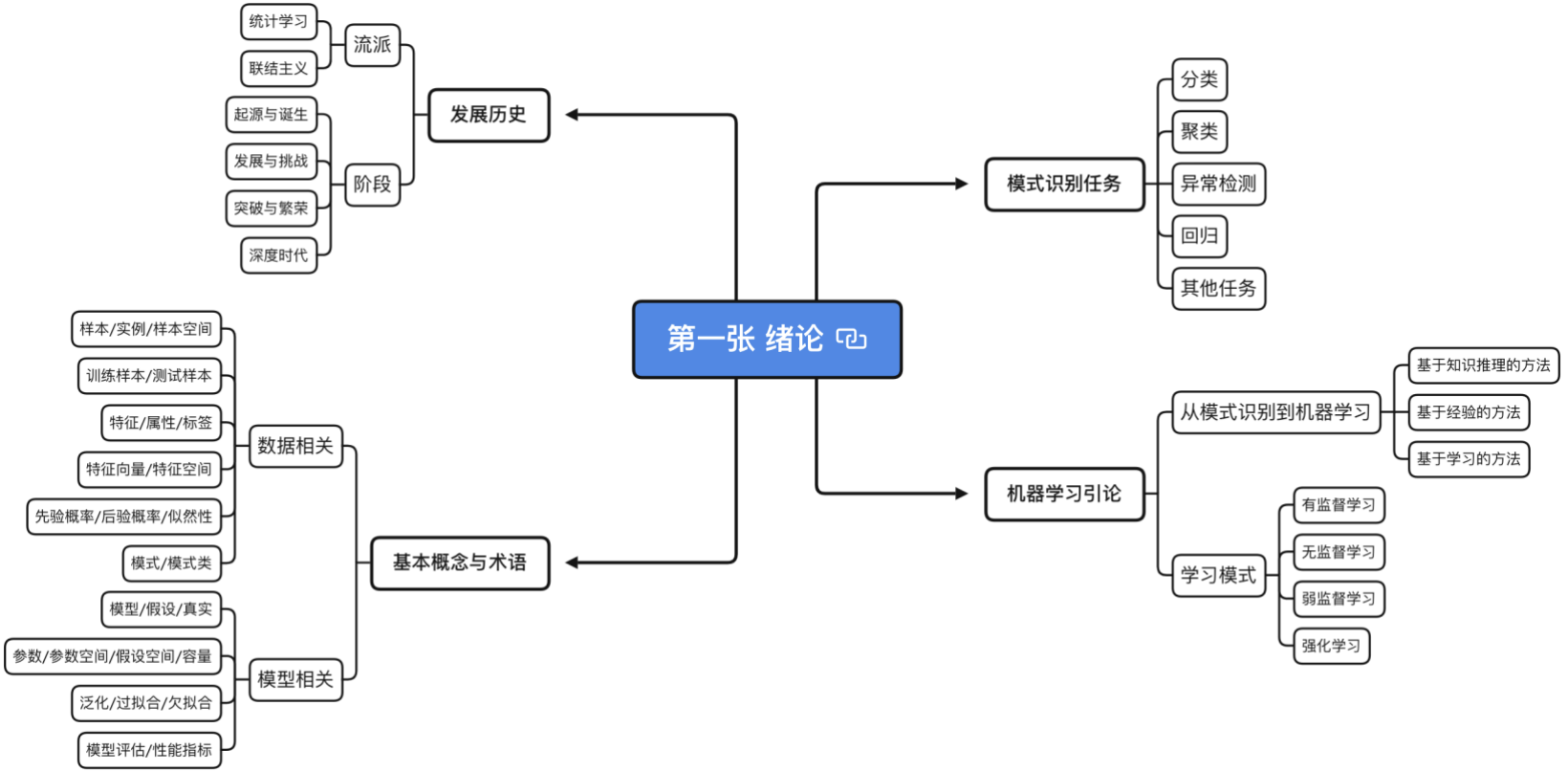
在机器博弈领域，DeepMind 团队开发的 **AlphaGo** 模型于 2016 年击败了世界围棋冠军李世石，引起了全世界广泛的关注，后续推出的 **Alpha Master** 和 **Alpha Zero** 更以绝对优势击败了柯洁为代表的全世界各国顶尖围棋高手，展示了深度学习在复杂策略游戏中的巨大优势，同时也让全世界对新一代智能方法产生了大范围的关注和兴趣。

在智能推荐领域，Google 的研究人员于 2016 年提出了 **Wide & Deep** 推荐模型，并在 Google 广告和 Play 商店等产品中得到了广泛应用，为深度学习在推荐系统中的采用提供了示范，此后 YouTube、Netflix、亚马逊、TikTok 和国内的抖音、今日头条、京东与淘宝等各类网络社交与电商平台开始大力投入深度推荐系统的研发与商用，深度推荐系统通过互联网平台深刻地影响了全世界人们的社交、消费、阅读与信息交互模式，已经成为影响世界文明发展进程的一股隐形但不可忽视的新力量。

在无人驾驶领域，Waymo、Tesla 和 Uber 等多家公司在 2015-2016 年间先后推出了基于深度学习技术的无人驾驶汽车产品，取得了一系列重要的突破。目前，大多数无人驾驶汽车的自动化水平已经达到了美国自动化交通协会（SAE）定义的 5 级自动化水平标准中的第 4 级，既“车辆可以在大多数情况下自主驾驶，包括大多数道路条件和交通情况。驾驶员只在特定条件下需要介入，但系统可以在大多数情况下独立操作。”无人驾驶与新能源已经成为汽车行业发展公认的两大主流趋势，并形成了广阔且富有活力的消费市场。

除上述领域外，深度学习驱动的人工智能技术在医疗健康、金融、零售与电子商务、教育、制造业、农业等诸多领域都已经取得了广泛的应用，应该说人类社会已经全面进入了人工智能的深度时代。但回顾历史不难发现，深度学习技术的爆炸式发展与模式识别与机器学习的经典理论与方法研究是密不可分的。本书的写作目的是通过对模式识别与机器学习领域的核心技术进行介绍与分析，帮助从事相关领域学习与研究的读者奠定扎实的理论基础，从而推动人工智能技术的进一步发展。

本章思维导图


第一张 绪论
发展历史
流派
统计学习
联结主义
阶段
起源与诞生
发展与挑战
突破与繁荣
深度时代
基本概念与术语
数据相关
样本/实例/样本空间
训练样本/测试样本
特征/属性/标签
特征向量/特征空间
先验概率/后验概率/似然性
模式/模式类
模型相关
模型/假设/真实
参数/参数空间/假设空间/容量
泛化/过拟合/欠拟合
模型评估/性能指标
模式识别任务
分类
聚类
异常检测
回归
其他任务
机器学习引论
从模式识别到机器学习
基于知识推理的方法
基于经验的方法
基于学习的方法
学习模式
有监督学习
无监督学习
弱监督学习
强化学习

本章习题

一、填空题

1 某班级同学计划在毕业晚会上自发分成 3 个小组表演节目，该任务属于模式识别中的（　　）任务；常用于完成该类任务的主流机器学习方法是（　　）学习。

2 将模式识别任务看做 1 个输入输出系统$y = f(\boldsymbol{x})$，则对于分类任务，y表示（　　）；对于回归任务，y 标识（　　）。

3 用一个模式识别算法对 Iris 数据库进行分类，其样本空间是（　　）的集合，样本维度是（　）；。

二、判断题

4 明日降水概率预测在形式上属于回归问题。（　）

5 通过照片判断自己是否认识该人属于异常检测问题。（　）

6 驯兽师用鞭打和喂食的方式教导老虎跳火圈属于有监督学习。（　）

7 一个数据集的特征维度总是小于该数据集的样本数量。（　）

8 模型容量大小与训练集样本数量无关。（　）

三、选择题

9 一个股票软件推荐是否买入某只股票属于一下哪种模式识别任务（　）

A. 分类　　B. 聚类　　C. 回归　　D. 以上均不是

10 以身高和性别为输入，构建一个带有常数项的线性模型用于预测目标的体重，则该模型有几个需要学习的模型参数？（　）

A. 3 个　　B. 2 个　　C. 1 个　　D. 不确定

11 儿童学习骑行自行车主要采用了以下哪种学习方式？（　）

A. 有监督学习　　B. 无监督学习　　C. 弱监督学习　　D. 强化学习

四、简答题

12 请分析聚类任务与分类任务的区别，尝试举出一个既是分类又是聚类的任务。

13 请分析分类任务是否可以被认为是回归任务的一种特例，并阐述理由。

14 利用 Iris 数据集有训练一个分类器时，你是否会将所有样本都用于训练，请阐述理由。

15 请分析有监督、无监督、弱监督和强化学习哪种模式更适合解决无人驾驶问题。

16 一个模型在训练集上表现良好，但在测试集上表现不佳，请分析原因。