第一章 机器学习概述

近年来，人工智能、机器学习的发展十分迅速并对社会生活产生越来越多的影响，专门从事这方面理论研究或应用开发人员也越来越多。人工智能、机器学习这些名词或概念究竟表达怎样的含义？实现机器学习究竟具有哪些基本方式或方法？机器学习在解决实际问题时究竟能够发挥怎样的作用？任何一个从事机器学习理论研究或系统开发的专业人员都应对这些问题具备明确而正确的认识。本章将对这些问题展开讨论，为读者提供机器学习最基本的概念和知识框架。首先介绍机器学习的基本概念，包括人工智能与机器学习的关系、机器学习的基本术语及误差分析；然后以机器学习的发展历程为主线简要介绍连接学习、符号学习和统计学习的基本思想，它们分别代表机器学习三种不同的基本类型；最后分析讨论机器学习的三个基本问题，即特征提取、规则构建和模型评估。

1.1 机器学习基本概念

从外部环境中学习所需知识或技能是人类的一项重要能力，机器学习要解决的问题就是如何使得机器能够像人类一样具有这种学习能力。目前，机器学习作为实现人工智能的一项核心技术，已在数据挖掘、计算机视觉、搜索引擎、语音识别、游戏博弈、经济预测与投资分析等众多领域得到广泛应用。本节主要介绍机器学习的基本概念，包括人工智能与机器学习的关系、机器学习的定义、有关机器学习的若干基本术语以及机器学习的误差分析。

1.1.1 人工智能与机器学习

发明创造某种工具来延伸人类器官功能是实现人类科技进步的一种重要手段。例如，汽车、轮船和飞机等工具的发明延伸了人腿功能，极大提升了人类的交通能力；摄像机、望远镜和显微镜等工具的发明延伸了人眼功能，极大提升了人类的视觉能力。为摆脱复杂繁重的科学与工程计算任务，人们发明了计算机代替人脑进行计算。实践证明计算机不仅能够胜任科学与工程计算工作，而且算得比人脑更快、更准确。那么计算机是否可以进一步承担人脑的推理或思维等智能任务呢？受此启发，以麦卡赛、明斯基、罗切斯特和申农等一批具有远见卓识的科学家共同探究使用机器模拟人类思维或人类智能的一系列问题，并在 1956 年夏季首次提出人工智能的概念，标志着人工智能学科的诞生。

人工智能的主要目标是通过计算机这台机器模拟人的某些思维能力或智能行为，如推理、证明、识别、感知、认知、理解、学习等思维能力或活动，让计算机能够像人类一样进行思考。六十多年来，人工智能取得了长足的发展，目前在机器翻译、智能控制、图像理解、语音识别、游戏博弈等领域有着广泛应用。纵观人工智能的发展历程，可依据所用核心技术的不同将其大致分为逻辑推理、知识工程和机器学习这三个基本阶段。

20 世纪 50 年代至 70 年代是人工智能发展的早期阶段，那时人们普遍认为实现人工智能的关键技术在于自动逻辑推理，只要机器被赋予逻辑推理能力就可以实现人工智能。因此，早期人工智能主要通过谓词逻辑演算模拟人类智能。这个阶段的人工智能的主流核心技术是

符号逻辑计算，在数学定理自动证明等领域获得了一定成功。

然而，人们逐步意识到如果没有一定数量专业领域知识支撑，则很难实现对复杂实际问题的逻辑推理。因此，以知识工程为核心技术的专家系统在 20 世纪 70 年代至 90 年代逐步成为人工智能的主流。专家系统使用基于专家知识库的知识推理取代纯粹的符号逻辑计算，在故障诊断、游戏博弈等领域取得了巨大成功。

专家系统需要针对具体问题的专业领域特点建立相应的专家知识库，利用这些知识来完成推理和决策。例如，如果让专家系统做疾病诊断，就必须把医生的诊断知识建成一个知识库，然后使用该库中知识对病情进行推断。然而，把专家知识总结出来并以适当的方式告诉计算机程序有时非常困难，通常需要针对每个具体任务手工建立相应的知识库。例如在图像识别领域，为识别图像中目标是否为猫而建立的知识库不能用于对目标是否为狗的识别，若要实现对图像中狗的识别，就必须专门建立用于识别狗的知识库。因此，专家知识的人工获取和表示方式严重制约了人工智能的进一步发展。

俗话说，授人以鱼不如授人以渔。既然把专家知识总结出来再灌输给计算机的知识工程方式非常困难甚至在很多场合不可行，那么可以考虑让人工智能系统自己从数据中学习领域知识。从外部环境中学习所需知识或技能是人类的一项重要能力，机器学习要解决的问题就是如何使得机器能够像人类一样具有这种学习能力。事实上，机器学习的思想可以追溯到 20 世纪 50 年代的感知机数学模型，该模型可以通过样本数据调整连接权重的方式保持模型对外部环境变化的自适应性。专家系统的知识工程困境使得机器学习思想和技术逐步得到重视，并在 20 世纪 80 年代初步形成一套相对完备的机器学习理论体系。

20 世纪 90 年代中期以来，机器学习得到迅速发展并逐步取代传统专家系统成为人工智能的主流核心技术，使得人工智能逐步进入机器学习时代。特别是近十几年来，数据量爆发式增长、计算机运算能力的巨大提升和机器学习新算法（深度学习）的出现，使得人工智能获得飞跃式迅猛发展。目前，以机器学习为主流核心技术的人工智能在多个领域取得的巨大成功已使其成为社会各界关注的焦点和引领社会未来的战略性技术。

图 1-1 表示一种典型的人工智能系统计算框架，其中机器学习模块通过适当算法解析数据，从数据中获取知识和模型参数，输出可用于决策或预测的数学模型，为人工智能系统提供核心算法支撑。计算机视觉、语音工程等专业应用模块使用机器学习算法提供的数学模型完成对相关对象的识别、合成、分析、理解、决策等信息处理任务。

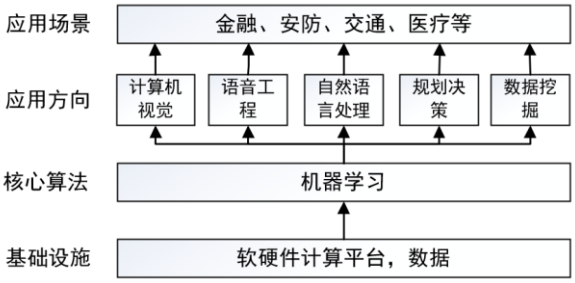


图 1-1 人工智能系统典型框架

由以上分析可知，机器学习为人工智能系统提供基础性的模型和算法支撑，是实现人工智能系统必备的核心技术。下面具体讨论机器学习的基本含义。

我们知道，在使用计算机解决实际问题的时，通常需要对实际问题建立数学模型，将实

际问题的求解转化为对数学模型的求解。此时不可避免地会出现一些模型参数，这些参数的取值情况往往会对模型及其求解结果产生很大的影响，一般需要调整参数以便取得更好的结果。然而，模型参数较多或者取值状态比较复杂时，手工调整参数就会变得非常困难和费时。为解决这个问题，可以考虑从实际问题中采集适当的样本数据，通过对这些样本数据进行解析自动计算出所需模型参数，并随着样本数据变化自动调整参数取值，使得数学模型和求解算法具有良好的普适性和自适应性。上述做法类似于人类向周围环境学习知识或规则的行为，样本数据相当于周围环境，模型参数相当于学习获得的知识或规则，由此产生机器学习理论和算法的基本思想。

从外部环境中学习所需知识或技能是人类的一项重要能力，获取知识或技能的根本目的在于提高自身的判断、推理、决策或识别等思维水平。因此，从本质上说，机器学习就是通过样本数据等适当的经验信息来改善模型的性能。例如，在使用模型M_0识别猫或狗的图片时，可采用适当方式将一些关于猫或狗的带标注图片输入模型M_0中，通过改进M_0参数或结构产生一个新的模型M_p，使得模型M_p 的识别正确率高于M_0。这就是一个机器学习的过程。此时，经验信息表现为猫或狗的带标注图片，模型的性能即为识别的正确率。

机器学习对初始模型M_0的改善不仅体现在模型参数方面，有时还会对模型结构进行改进。因此，通常用改进模型M_p泛指机器学习的输出结果。由此得到如下机器学习定义：

机器学习是一种通过先验信息来提升模型能力的方式。具体地说，对于给定的任务和性能度量标准，使用先验信息E，通过某种计算方式T改进初始模型M_0，获得一个性能更好的改进模型M_p，即有$M_p = T(M_0, E)$。

机器学习定义中任务所界定的范畴非常广泛，在不同应用领域有着不同的具体含义。例如，如果编写一个机器学习程序让机器人能够行走，那么机器人行走就是一个任务。但是机器学习本身不是任务，因为机器学习是获取或提升完成某项任务所需能力的一种途径。

从上述机器学习概念可知，机器学习的目标就是通过计算手段从经验数据等先验信息E中产生一个性能改善的新模型M_p。因此，机器学习的研究内容就是使用计算机从经验数据等先验信息中产生模型的算法，即学习算法。如果说计算机科学是一门关于算法的学问，那么机器学习就是一门关于学习算法的学问。

【例题 1.1】已知样本数据集为：

$$D = \{(x, y)|(1.1,1.9), (2.7,2.3), (3.2,3.4), (3.6,2.9), (4.7,3.4), (5.1,4.3)\}$$

D中数据点在坐标系中的分布如图 1-2 所示。令初始模型M_0：$y = ax + b$，试根据数据集D优化M_0并计算$x = 6$时的模型输出$\hat{y}(6)$。

【解】对于初始模型M_0，令M_0对每个样本x_i的预测输出$\hat{y}_i$与其真实值y_i之间的误差平方为e_i，即$e_i = (\hat{y}_i - y_i)^2$，则模型$M_0$对所有样本的累计误差为：

$$Q(a, b) = \sum_{i=1}^{6} e_i = \sum_{i=1}^{6} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{6} (\hat{y}_i - ax_i - b)^2$$

由于对模型M_0进行优化的依据是D中所有样本的真实取值，故当模型对所有样本预测值与真实值之间的累计误差最小时，模型对样本的预测输出最准确，此时的模型就是所求的优化模型，即机器学习定义中具有更好性能的新模型。基于以上分析，可将模型M_0的参数求解转化为计算累计误差最小值的优化问题。

将模型M_0的参数a,b看作累计误差函数Q的变量，由于在多元函数极值点处，函数对其所有变量的偏导数均为 0，故分别对参数a,b求偏导，并令偏导数为 0，得到：

$$\frac{\partial Q}{\partial a}=2\sum_{i=1}^{6}(y_i-ax_i-b)(-x_i)=0$$

$$\frac{\partial Q}{\partial b}=2\sum_{i=1}^{6}(y_i-ax_i-b)(-1)=0$$

联立上述等式并代入D中样本数据，解得:$a\approx 0.66$，$b\approx 0.789$。

由此得到优化模型M_p：$y=0.66x+0.789$，图 1-3 中的实线表示M_p的函数图像，将$x=6$代入M_p，可求得优化模型的预测值：$\hat{y}(6)=4.749$。□

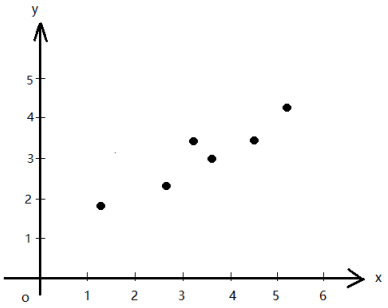


图 1-2 样本分布图

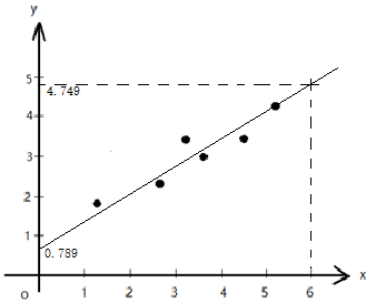


图 1-3 优化模型M_p

例题 1.1 所示的机器学习实例主要通过对初始模型参数的优化估计求出具有更好性能的新模型。事实上，机器学习有时还可以根据实际需要改变初始模型的结构。例如，对于样本数据集$D=\{(x_i,y_i)|i=1,2,\ldots,n\}$,假设初始模型为如下$k$次多项式：

$$N_0：y=\theta_0+\theta_1x+\theta_2x^2+\cdots+\theta_kx^k（k>2）$$

则可以通过适当方式将N_0简化为如下二次多项式模型N_p：$y=\hat{\theta}_0+\hat{\theta}_1x+\hat{\theta}_2x^2$。

事实上，要将初始模型N_0变为简化模型N_p的形式，只需通过适当方式调整N_0的参数，使得参数$\theta_3,\theta_4,\ldots,\theta_k$的取值趋向于 0 即可。可用均方误差最小化的思想实现这个效果，为此构造如下以$\theta_0,\theta_1,\ldots,\theta_k$为自变量的函数：

$$Q(\theta_0,\theta_1,\ldots,\theta_k)=\sum_{i=1}^{n}(\hat{y}_i-y_i)^2+C\sum_{j=3}^{k}\theta_k^2 \qquad (1-1)$$

其中$\hat{y}_i=\theta_0+\theta_1x_i+\theta_2x_i^2+\cdots+\theta_kx_i^k$。

函数$Q(\theta_0,\theta_1,\ldots,\theta_k)$中第一项为模型$M_0$对所有样本取值的累计误差，第二项是对参数$\theta_3,\theta_4,\ldots,\theta_k$添加的限制条件。在最小化函数$Q(\theta_0,\theta_1,\ldots,\theta_k)$的过程中，可将$C$定义为一个非常大的取值，使得参数$\theta_3,\theta_4,\ldots,\theta_k$的取值趋向于 0，以尽量消除$C$值对函数$Q(\theta_0,\theta_1,\ldots,\theta_k)$最小化取值的影响。此时用函数$Q(\theta_0,\theta_1,\ldots,\theta_k)$代替累计误差函数求最小值，相当于在对参数$\theta_3,\theta_4,\ldots,\theta_k$做趋向于 0 的限制条件下解出优化模型的全部参数$\hat{\theta}_0,\hat{\theta}_1,\hat{\theta}_2$。具体地说，就是由于函数$Q$中第二项权重过大，为使得函数$Q$整体取值最小，该项所涉及参数$\theta_3,\theta_4,\ldots,\theta_k$均会趋向于 0，由此即可获得结构调整后的优化模型$y=\hat{\theta}_0+\hat{\theta}_1x+\hat{\theta}_2x^2$。

1.1.2 机器学习的基本术语

如前所述，机器学习主要通过样本提供的信息提升模型性能以完成给定的学习任务，即从样本中学习。对于任意一个给定的样本对象ξ，一般需要对其提取若干属性形成对该样本的数据描述或表征，并将这些属性值作为机器学习模型的输入。令：

$$x_1 = \psi_1(\xi), x_2 = \psi_2(\xi), \dots, x_m = \psi_m(\xi)$$

为样本ξ的m个属性提取函数，则可通过这些函数将样本ξ映射成一个m元**表征向量**X，即：

$$X = X(\xi) = (x_1, x_2, \dots, x_m)^T$$

其中x_i为样本ξ的第i个属性值，$i = 1,2,\dots,m$。

显然，表征向量X是对样本对象ξ的一个数据抽象，从数学的角度看，两者没有本质上的差异。因此，为方便表达，在不产生混淆的情况下，通常将表征向量为X的样本ξ简称为样本X，即不加区分地使用表征向量X和表征向量为X的样本ξ这两个没有本质差异的概念。

机器学习的任务是指所要解决的问题，主要包括回归、分类和聚类等。回归任务是通过若干带有标注的样本数据构造出一个预测模型$R(X)$，使得$R(X)$的预测输出尽可能符合真实值，并称$R(X)$为**回归模型**。通常将用于构造模型的样本称为**训练样本**，用于测试模型效果的样本称为**测试样本**。一般使用两组不同的样本集合分别作为训练样本集和测试样本集。

设$\xi_1, \xi_2, \dots, \xi_n$是任意给定的$n$个训练样本，$X_k = (x_{1k}, x_{2k}, \dots, x_{mk})^T$和$y_k$分别表示$\xi_k$的表征向量和标注值$(k = 1,2,\dots,n)$，则由这$n$个样本构成的训练样本集$D$可以表示为：

$$D = \{\langle X_k, y_k \rangle | k = 1,2,\dots,n\}$$

回归模型$R(X)$的初始模型是一个带有参数的计算模型，机器学习的模型训练算法使用训练样本集D中的数据信息计算出$R(X)$的全部参数，得到具体的回归模型。有了回归模型的具体参数，就可以使用该模型完成回归任务。例如，例题 1.1 解决的就是一个机器学习回归任务，通过训练样本数据计算所得模型$y = 0.66x + 0.789$就是一个具体的回归模型。

日常生活和工作中会经常遇到一些分类问题，例如有时需要将产品按质量分为优等品、合格品和次品，将公司客户分为贵宾客户和普通客户等等。可以使用机器学习方式实现这种分类任务，即根据带标注训练样本构建相应的分类模型，然后根据分类模型实现对目标的自动分类。显然，如果回归模型的预测输出是离散值，则机器学习的回归任务就转化为分类任务。也就是说，分类其实是预测输入样本所在类别的一类特殊回归任务，特殊性在于要求预测结果为离散的类别值而不是连续值。

用于分类任务的机器学习模型称为**分类模型**或**分类器**，分类任务的目标是通过训练样本构建合适的分类器$C(X)$，完成对目标的分类。分类类别只有两类的分类任务称为**二值分类或二分类**，这两个类别分别称为**正类**和**负类**，通常用+1 和−1 分别指代。分类类别多于两类的分类任务通常称之为多**值分类**。

对于一个具体的回归或分类任务，所有可能的模型输入数据组成的集合称为**输入空间**，所有可能的模型输出数据构成的集合称为**输出空间**。显然，回归或分类机器学习任务的本质就是寻找一个从输入空间到输出空间的映射，并将该映射作为预测模型。从输入空间到输出空间的所有可能映射组成的集合称为**假设空间**。

回归或分类模型的训练计算可以看成是一个在假设空间中搜索所需模型的过程，模型训

练算法在假设空间中搜索合适的映射，使得该映射的预测效果与训练样本所含先验信息相一致。事实上，满足条件的映射通常不止一个，此时需要对多个满足条件的映射做出选择。在没有足够依据进行唯一性选择的情况下，有时需要做出具有主观倾向性的选择，即更愿意选择某个映射作为预测模型。这种选择的主观倾向性称为机器学习算法的**模型偏好**。例如在多个映射与训练样本所包含的先验信息一致时，可选最简单的映射作为预测模型，此时模型偏好为最简单映射。这种在同等条件下选择简单事物的倾向性原则称为**奥卡姆剃刀原则**。

自然界和社会生活中经常会出现物以类聚、人以群分的现象，例如羊、狼等动物总是以群居的方式聚集在一起，志趣相同的人们通常会组成特定的兴趣群体。机器学习的**聚类任务**就是对样本数据实现物以类聚的效果。显然，聚类的类别由不同样本之间的某种相似性确定，因而聚类类别所表达的含义通常是不确定的，聚类样本也不带特定的标注表示样本所属的类别。这是聚类与回归或分类任务之间的本质区别。通常将不带标注的样本称为**示例**，而带标注的样本称为**样例**。

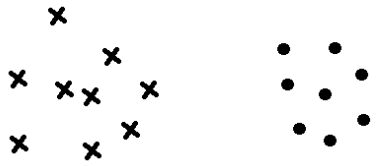


图 1-4 包含两个簇的示例集

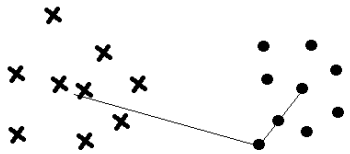


图 1-5 聚类示意图

在聚类任务中，所有输入示例的集合称为**示例集**，被划分为同一类别的示例所构成的集合称为一个**簇**。图 1-4 表示一个具有两个簇的示例集。

对于任意给定的一个n元示例集$S=\{x_1,x_2,\ldots,x_n\}$，假设Δ是一个对S的划分，则有：

$$\Delta=\{S_1,S_2,\ldots,S_t\} \qquad (1-2)$$

其中$S_1,S_2,\ldots,S_t$均是由S中示例构成的簇且满足：

$$S=S_1\cup S_2\cup\ldots\cup S_t$$

聚类的目标是寻找一个对S的适当划分Δ，使得划分的各簇内部的示例之间相似性尽可能地小。通常采用欧式距离或余弦距离等作为样本之间相似性的度量标准。图 1-5 表示一个示例依据相似性度量标准被划分到与之相似的簇中。

聚类任务使用的先验信息与回归或分类任务有着很大差别。聚类任务的先验信息为示例，即不带标注的样本，而回归和分类任务的先验信息均为带标注的样本。事实上，除了带标注样本和不带标注样本之外，先验信息有时还以某种反馈信息的形式存在。可根据先验信息的不同形式，将机器学习分为监督学习、无监督学习和强化学习这三种基本方式。

监督学习是指利用一组带标注样本调整模型参数，提升模型性能的学习方式。监督学习的基本思想是通过标注值告诉模型在给定输入的情况下应该输出什么值，由此获得尽可能接近真实映射方式的优化模型。监督学习不像传统计算机问题求解那样需要根据实际问题具体情况设计一个固定流程进行计算，而是由计算机根据带标记的样本集自动获得一个问题的求解模型并由此实现对问题求解。图 1-6 表示监督学习的基本流程。

无监督学习通过比较样本之间的某种联系实现对样本的数据分析。相比于监督学习，无监督学习最大特点是学习算法的输入是无标记样本。例如现有一些图片，其中每张图片内容是两类不知名花卉之一，通过观察花卉特点将同类的花卉图片放到一起，这便是无监督学习。

在实际问题中遇到样本缺失标记的情况或者人工标注成本过高的情况下，可以使用无监督学习方式实现对这些数据自动分析，将所得到分析结果作为参考信息。

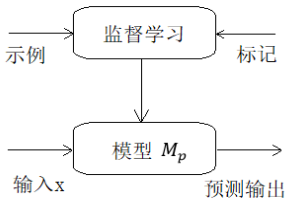


图 1-6 监督学习流程图

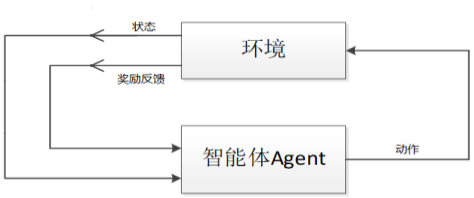


图 1-7 强化学习流程图

强化学习是根据反馈信息来调整机器行为以实现自动决策的一种机器学习方式。一个强化学习系统主要由智能体和环境两个部分组成。智能体是行为的实施者，由基于环境信息的评价函数对智能体的行为做出评价，若智能体的行为正确，则由相应的回报函数给予智能体正向反馈信息以示奖励，反之则给予智能体负向反馈信息以示惩罚。强化学习的基本流程如图 1-7 所示，智能体根据环境的当前状态选择下一个动作，环境对这个动作做出评价并反馈给智能体，同时更新环境状态，不断重复这一过程直至达到某种设定，选取累计奖励值最大的一组动作作为所求的最终策略。

1.1.3 机器学习误差分析

机器学习模型是对实际求解问题的一种数学抽象，模型的输出结果与其对应的真实值之间往往会存在一定的差异，这种差异称之为该模型的**输出误差**，简称为**误差**。机器学习的一个重要手段就是以模型输出误差为基本依据不断优化或校正模型，使得模型输出误差尽可能变小。因此，对机器学习模型进行误差分析，从误差分析角度分析寻找影响机器学习模型性能的关键因素，是机器学习的重要研究内容。

为便于误差分析，通常需要构造某种函数用于度量模型对于单个样本的输出误差，这样的函数称之为**损失函数**。具体地说，对于给定的机器学习模型f，假设该模型对应于输入样本X的输出为$\hat{y}=f(X)$，与X对应的实际真实值为y，则可用以y和$f(X)$为自变量的某个函数$L(y,f(X))$作为损失函数来度量模型f在输入样本X下的输出误差。

损失函数的具体形式有很多种，可根据实际问题需要构造或选用适当的损失函数进行误差分析。例如$L(y,f(X))=(y-f(X))^2$和$L(y,f(X))=|y-f(X)|$是两种经常用于度量回归模型输出误差的损失函数，分别称之为**平方损失函数**和**绝对值损失函数**。

在机器学习中，面向单个样本的损失函数度量的只是模型在某个特定样本下的输出误差，不能很好地反映模型在某个样本集上对所有样本的整体计算准确度。因此，需要进一步定义面向某个特定样本集的综合误差，通常称之为该样本集上的**整体误差**。

对于任意给定的一个n元样本集$S=\{(X_1,y_1),(X_2,y_2),\dots,(X_n,y_n)\}$，模型$f$在$S$上的整体误差$R_S(f)$定义为：

$$R_S(f)=E\left[L(y,f(x))\right]=\frac{1}{n}\sum_{i=1}^{n}L(y_i,f(X_i)) \qquad (1-3)$$

即将$R_S(f)$定义为S中所有单个样本所分别对应损失函数值的平均值。

对于某个给定的机器学习任务，假设与该任务相关的所有样本构成的样本集合为D，则机器学习模型在样本集合D上的整体误差称为该模型关于该学习任务的**泛化误差**。具体地说，令样本集合D中所有样本的概率分布为$P(D)$，模型f对输入样本X的输出为$\hat{y}=f(X)$，X所对应的实际真实值为y，则可将模型f的泛化误差定义为：

$$R_{exp}(f)=E_{P(D)}\left[L\left(y,f(X)\right)\right] \tag{1-4}$$

泛化误差表示机器学习模型在整个样本集合D上的平均误差，是刻画机器学习模型普适性的重要指标，作为模型求解和模型评估的基本依据在机器学习过程中发挥极为重要的作用。然而，精确计算模型的泛化误差需要知道整个样本集合D中所有样本的真实取值和概率分布，这通常是不可行的。因此，一般无法计算泛化误差的精确值，需要采用某些便于计算的度量指标作为泛化误差的近似代替值。

机器学习模型训练的目标是尽可能获得普适性或泛化性最好的模型，理论上要求模型的泛化误差达到最小。然而，通常无法直接计算模型的泛化误差，更难以直接对泛化误差进行优化分析。由于训练样本通常采样自整个样本集合D，训练样本集通常与D有着比较相似的样本概率分布，故一般采用训练误差近似替代泛化误差对模型进行训练。

所谓**训练误差**，是指模型在训练样本集上的整体误差，也称之为**经验风险**。具体地说，对于任意给定的n元训练样本集合$G=\{(X_1,y_1),(X_2,y_2),\dots,(X_n,y_n)\}$，假设模型$f$对输入样本$X$的预测输出为$\hat{y}=f(X)$，则该模型关于训练样本集$G$的训练误差定义为：

$$R_{\text{emp}}(f)=\frac{1}{n}\sum_{k=1}^{n}L(y_k,f(X_k)) \tag{1-5}$$

其中X_k表示训练集中的第k个样本，$f(X_k)$表示模型对输入样本X_k的输出$\hat{y}_k$，y_k为机器学习任务中输入X_k对应的实际真实取值。

因此，机器学习中的模型训练或优化通常使用最小化训练误差的方法完成。该方法称之为**经验风险最小化**方法，由此得到的优化模型为：

$$\hat{f}=\arg_{f\in F}\min R_{\text{emp}}(f) \tag{1-6}$$

其中F为假设空间。

对于已训练出的模型，通常使用测试误差近似替代泛化误差的对该模型进行测试。所谓**测试误差**，是指模型在测试样本集上的整体误差。具体地说，对于任意给定的v元测试样本集合$T=\{(X_1^t,y_1^t),(X_2^t,y_2^t),\dots,(X_v^t,y_v^t)\}$，该模型关于$T$的测试误差定义为：

$$R_{test}=\frac{1}{v}\sum_{k=1}^{v}L(y_k^t,f(X_k^t)) \tag{1-7}$$

其中X_k^t表示测试集中第k个样本，$f(X_k^t)$表示模型对输入X_k^t的输出$\hat{y}_k^t$，y_k^t为机器学习任务中输入X_k^t对应的实际真实值。

对于训练样本集合中的每个样本，该样本都会存在一些普适于整个样本集D的共性特征和一些仅仅适合于特定训练样本集的个性特征。机器学习中模型训练的最理想效果就是充分提取训练样本的共性特征而尽量避免提取其个性特征，使得训练出来的模型具有尽可能广泛的普适性，即具有尽可能好的泛化性能。

然而，模型的训练通常以最小化训练误差为标准，此时对于固定数量的训练样本，随着

训练的不断进行，训练误差会不断降低，甚至趋向于零。如果模型训练误差过小，就会使得训练出来的模型基本上完全适应于训练样本的特点。此时，训练模型不仅拟合了训练样本的共性特征而且也拟合了训练样本的个性特征，反而降低了训练模型的泛化性能，使得泛化误差不断增大。这种同时拟合训练样本的共性特征和个性特征的现象，在机器学习领域通常称之为模型训练的**过拟合**现象。

避免过拟合现象的一个有效措施是尽可能扩大训练样本的数量，尽可能降低样本在训练样本集与整个样本集上概率分布的差异，以充分增强训练样本的共性特征，弱化训练样本的个性特征。近年来计算机运算能力的巨大提升及在各行各业中不断涌现的大数据使得通过扩大训练样本数量以避免过拟合的措施变得可行，这正是机器学习在如今互联网和大数据时代得到迅猛发展的重要原因。

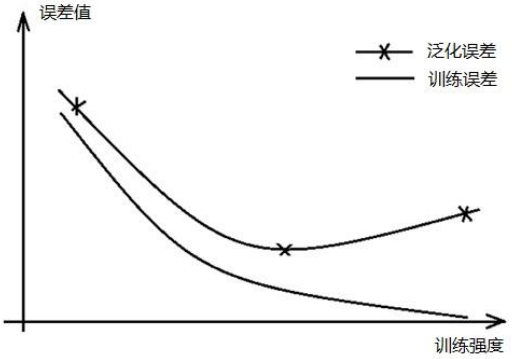


图 1-8 训练误差和泛化误差关系图

由以上分析可知，在机器学习的模型训练中，随着训练过程的进行，训练误差会一直不断降低，但泛化误差则会先减小然后因产生过度拟合现象而导致不断增大，具体如图 1-8 所示。在训练的初始阶段，由于模型尚未充分拟合训练样本的共性特征，故此时模型的泛化误差较大。这种由于未能充分拟合训练样本共性特征造成模型泛化误差较大而导致模型泛化能力较弱的现象称为模型训练的**欠拟合**现象。随着训练过程的不断进行，训练误差和泛化误差均不断减少，欠拟合现象通常会逐渐消失。

对于给定的训练样本集合，如果对模型训练强度不做适当控制，就会在模型训练的后期将训练样本的个性特征引入模型当中而引起泛化误差的增大，产生过拟合现象。因此，泛化误差由下降变为上升的转折点处对应的训练模型具有最好的泛化性能。也就是说，对于给定的训练样本集合，可以在适当训练强度下获得具有最好泛化性能的训练模型。

现在进一步分析讨论不同训练样本集合的差异对模型训练结果的影响，考察训练模型对训练样本集合变化的稳定性。

对于任意给定的一个初始模型f，假设$D_1, D_2, \dots, D_s$是s个不同的训练样本集合，其中每个训练样本均采样自整个样本集合D，通过训练样本集合 D_i训练初始模型f所得到的优化模型记为$f_i, i \in (1,2,\dots,s)$，　$\hat{y}_i = f_i(X)$表示第i个模型对于输入样本X的模型输出，X所对应的实际真实值为y，则这s个优化模型对于输入样本X的期望输出为：

$$E[F(X)] = \frac{1}{s}\sum_{i=1}^{s} f_i(X) \qquad (1-8)$$

其中$F(X) = (f_1(X), f_2(X), \dots, f_s(X))^T$，可将其看成一个关于$f(X)$的离散随机变量。

此时，模型$f(X)$对于测试样本集合变化的稳定性可用相应的方差指标进行度量。模型$f(X)$在训练样本集$D_1, D_2, \dots, D_s$下所得优化模型$f_1(X), f_2(X), \dots, f_s(X)$输出的方差为：

$$\mathrm{var}[F(X)] = E\{[F(X) - E[F(X)]]^2\} = \frac{1}{s}\sum_{i=1}^{s}[f_i(X) - E[F(X)]]^2$$

对于任意一个给定的初始模型f，如果该模型变化的自由度较大，例如模型参数的数目较多或者参数的取值范围较大，则能够更好地适应训练样本数据的变化，能对多种不同的训练样本集合获得较好的拟合效果；反之，如果该模型参数的变化自由度较小，则模型适应训练数据变化的能力就比较小，可以有效拟合的训练数据范围也就比较有限。机器学习模型这种适应训练数据变化的能力，称之为模型的**学习能力**或**模型的容量**。

显然，模型的容量主要反映该模型对数据的拟合能力。容量越大的模型对数据的拟合能力就越强，能够更好地适应训练样本数据的变化。可以使用模型输出在不同训练样本集合下的综合偏差对其进行度量，这种综合偏差称为**模型输出的偏差**，简称**为偏差**。

对于模型$f(X)$在训练样本集$D_1, D_2, \dots, D_s$下的优化模型$F(X) = (f_1(X), f_2(X), \dots, f_s(X))^T$，$F(X)$作为一个离散随机变量与$X$所对应实际真实值$y$之间的偏差$\mathrm{bias}[F(X)]$为：

$$\mathrm{bias}[F(X)] = E[F(X)] - y \qquad (1-9)$$

对基于平方损失函数的泛化误差$R_{exp}(f) = E\left[L\left(y, F(X)\right)\right] = E[(F(X) - y)^2]$，对其进行偏差-方差分解，可得：

$$\begin{aligned} E[(F(X) - y)^2] &= E\{[F(X) - E[F(X)] + E[F(X)] - y]^2\} \\ &= E\{[F(X) - E[F(X)]]^2\} + E\{[E[F(X)] - y]^2\} \\ &\quad +2E\{F(X) - E[F(X)]\}\{E[F(X)] - y\} \end{aligned}$$

由于$E\{F(X) - E[F(X)]\} = E[F(X)] - E[F(X)] = 0$，故有：

$$\begin{aligned} E[(F(X) - y)^2] &= E\left\{\left[F(X) - E[F(X)]\right]^2\right\} + E\{[E[F(X)] - y]^2\} \\ &= E\left\{\left[F(X) - E[F(X)]\right]^2\right\} + [E[F(X)] - y]^2 \\ &= \mathrm{var}[F(X)] + \left[\mathrm{bias}[F(X)]\right]^2 \end{aligned}$$

即有：

$$E[(F(X) - y)^2] = \mathrm{var}[F(X)] + \left[\mathrm{bias}[F(X)]\right]^2 \qquad (1-10)$$

由(1-10)式可知，模型的泛化误差等于模型输出偏差平方与方差之和。如前所述，模型输出的偏差反映模型容量的大小或者说模型学习能力的强弱，模型输出的方差则反映模型对训练样本变化的敏感程度。一般而言，对于容量较大的模型，由于其拟合能力较强，因而会使得模型输出的偏差相对较小。然而，大容量模型的变化自由度通常较大，会导致模型参数对样本数据的变化比较敏感，使得模型输出的方差较大。因此，同时减小模型输出的方差和偏差是不可行的，图 1-9 表示泛化误差与偏差和方差之间的关系。

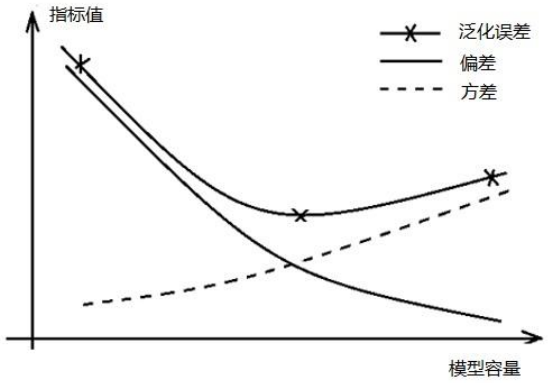


图 1-9 泛化误差与方差和偏差的关系

从图 1-9 中可以看出，随着模型容量的增加，模型输出的偏差随之减小，模型输出的方差却随之增大。因此，模型的泛化误差会出现先减后增的情况。当模型容量较低时，其拟合能力较弱，难以对训练样本的共性特征进行有效拟合，故欠拟合现象会较为严重。当模型容量过高时，模型对数据的变化太过敏感，具有过强的拟合能力，对训练样本的个性特征也进行了拟合，此时过拟合现象较为严重。由此可知，对于具体的机器学习任务而言，模型容量并非越高越好，一个容量适中的机器学习模型通常更能满足任务需求。

1.2 机器学习发展历程

机器学习作为人工智能的一个重要研究领域随着人工智能的产生而产生并随着人工智能理论发展而发展。目前，机器学习理论大致分为连接学习、符号学习和统计学习这三种基本类型。符号学习和连接学习分别源自人工智能的符号主义和连接主义，统计学习源自符号学习中的归纳学习。从历史上看，连接学习是机器学习最初采用的策略，感知机和神经网络是机器学习初创时期的代表性成果。20 世纪 80 年代，随着人工智能符号主义的发展，符号学习逐步成为机器学习的主流技术。20 世纪 90 年代以来，统计学习方法逐步走向成熟，并以其巨大理论创新和良好应用效果逐步取代符号学习成为机器学习的研究热点。近年来，得益于计算机运算能力的巨大提升和数据量的快速增长，以深度学习为代表的连接学习再次兴起，涌现出一大批优秀的理论和应用成果。统计学习和深度学习的巨大成功使得人工智能全面进入机器学习时代并成为引领社会未来的战略性技术。

1.2.1 感知机与连接学习

人工智能最早期的探索从模仿人类或动物大脑的生物结构开始。人类或动物大脑神经系统最基本的组成结构是神经元，相互连接的多个神经元通过相互传送某些化学物质改变电位的方式实现信息传递与交互。麦卡洛克和皮茨在 1943 年发表论文《A Logical Calculus of the Ideas Immanent in Nervous Activity》首次提出模拟生物神经元的数学模型，名为**MP模型**。图 1-10 表示该模型的基本结构，其中$\{x_1,\dots,x_m\}$为m个模型输入变量，$x_i \in (0,1)$，θ为阈值，模型输出y有两种可能的取值状态：当$\sum_{i=1}^{m} x_i > \theta$时，$y=1$，否则，$y=0$。

MP模型是对单个神经元的简单模拟，模型的输出值仅为 0 或 1，没有区分m个输入在重要性方面的差异，不能很好地体现神经元对外部环境变化的自适应性。为此，赫布在 1949

年提出赫布学习规则，指出神经系统对一个信号进行响应的同时，相关被激活神经元之间的联系也会随之得到增强，各单个神经元的输入之间应存在某种重要性差别。为此，赫布对MP模型进行改进，对其m个输入变量$\{x_1,\dots,x_m\}$添加了权重信息，使得MP模型能够根据实际情况自动修改这些权重达到模型优化或机器学习目的。这是机器学习最早期的思想萌芽。

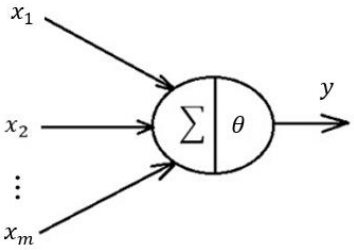


图 1-10 MP 模型示意图

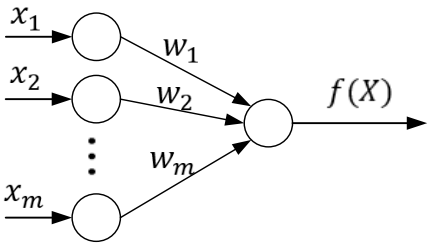


图 1-11 单层感知机示意图

人类或动物大脑的神经网络是一种由大量神经元通过层级连接构成的复杂网络。通过模仿生物大脑神经网络结构的方式实现机器智能是人工智能研究一个基本思想，称之为人工智能的**连接主义**思想，简称为连接主义。连接主义分别将每个神经元模型作为一个简单的独立计算单元，并按照一定的结构和规则对多个神经元模型进行组合和连接，形成一个复杂庞大的神经元网络模型用于模拟生物大脑的神经网络，实现人工智能。

基于连接主义思想和赫布学习规则，罗森布拉特在 1957 年提出一种名为**感知机**的神经网络模型。如图 1-11 所示，该模型根据连接主义思想将多个神经元模型进行分层互联，基于赫布学习规则通过使用样例信息调整连接权重的方式实现模型优化。这种使用样例信息调节神经元之间连接权重的学习方式被称为**连接学习**。连接学习是机器学习初创时期的基本学习理论，感知机和神经网络模型是机器学习早期的代表性成果。

图 1-11 所示的感知机模型包括输入层与输出层，但只有输出层参与数值计算，故亦称之为**单层感知机**。单层感知机可用于二分类任务，若存在**超平面**可以将两类样本分隔开来，则单层感知机有能力确定一个这样的超平面，图 1-12 表示单层感知机的分类效果，但若两类样本不能通过简单超平面进行划分，则单层感知机亦无能为力。

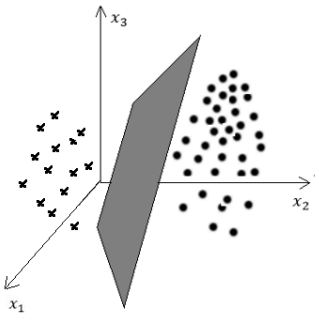


图 1-12 单层感知机分类效果

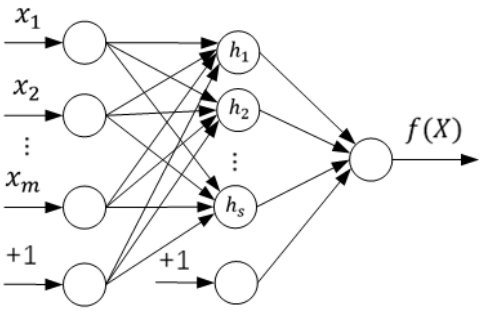


图 1-13 多层感知机模型

单层感知机能够完成一些简单的视觉处理任务，在当时引起了社会各界的广泛关注。然而，单层感知机的实际能力却名过其实。明斯基在其出版的著作《感知机：计算几何学导论》中证明单层感知机难以解决简单的异或分类问题。这使得单层感知机的实际能力被当时的社会所质疑，连接学习和连接主义的研究随之陷入低谷。

沃波斯在 1974 年提出**多层感知机**模型，有效解决了单层感知机无法解决的异或分类问题。如图 1-13 所示，多层感知机在单层感知机基础之上添加一个隐藏层，通过基于反向传

播的连接学习算法优化模型参数,图中输入数据 1 为线性组合中关于常数项的输入。然而,沃波斯的多层感知机模型未能给连接学习发展的低谷带来转机,其中影响连接学习甚至整个机器学习发展的反向传播学习算法也未获得应有的重视,因为当时整个神经网络和连接主义的研究正处低谷,基于符号逻辑推理的人工智能符号主义和专家系统的研究如日中天。

霍普菲尔德在 1982 年提出一种新的神经网络,即霍普菲尔德网络。它是一种基于神经动力学的神经网络模型,其连接学习过程可简单理解为模型稳定状态的优化搜索过程。霍普菲尔德网络不仅可以有效解决一大类的模式识别问题,还可以求得一类组合优化问题的近似解,这一成果振奋了连接学习领域。1989 年,塞班克在理论上证明了神经网络模型在本质上是一个通用的逼近函数,其拟合能力十分强大,包含一个隐藏层的神经网络模型可以逼近任意连续函数,而包含两个隐藏层的神经网络模型可以逼近任意函数。这使得连接学习在理论上向前迈进了一大步。这些成果使得连接主义和连接学习再次成为研究热点。

然而,连接学习的研究在二十世纪九十年代初遭遇发展瓶颈。由于在当时的计算条件下,随着网络层数的加深,网络模型变得难以收敛且超出计算能力,故具有强大拟合能力的浅层网络是当时最主要的研究对象。然而,关于浅层网络的理论研究进展缓慢,浅层网络模型在实际应用中难以取得满意的效果,连接主义和连接学习的研究再次进入低谷。

1.2.2 符号学习与统计学习

连接主义和连接学习从模拟生物大脑结构出发实现人工智能,认为智能的基本单元为神经元,智能活动过程是神经元之间的连接交互过程。这种基于模拟生物大脑结构的人工智能理论并非得到所有人的认可。很多研究者从人类思维内涵和过程出发研究人工智能,建立一套符号主义理论。符号主义认为思维的基本单元是符号信息,智能活动过程就是符号推理或符号计算的过程,生物大脑的本质就是一个能够高效处理符号信息的物理系统。基于符号主义理论,机器学习发展出另一套学习理论——**符号学习**。

人工智能的符号主义理论认为只要机器具备自动的逻辑推理能力便可拥有智能。因此,以谓词逻辑推理理论为基础的自动定理证明成为人工智能的重要研究领域。纽厄尔和西蒙等人在 1956 年编写完成的名为逻辑理论家的程序是符号主义的代表性成果。该程序自动证明了罗素的著作《数学原理》中全部 52 条定理,初步验证了用计算机来实现人类思维的可行性。纽厄尔和西蒙因这项成果获得 1975 年度图灵奖。

然而,人们逐渐发现仅赋予机器自动逻辑推理能力难以使其具有智能,拥有专业领域知识是实现人工智能不可或缺的条件。为此,费根鲍姆在 1965 年提出了**专家系统**的概念,并据此实现一个名为DENDRAL的历史上第一个专家系统。在此之后,出现了一大批成功应用于故障诊断、辅助设计等不同专业领域的专家系统,专家系统迅速成为人工智能的研究焦点和主流技术,专家系统的构建也提升到了工程的高度,名为**知识工程**。费根鲍姆由于在这方面的重大贡献被人们称为知识工程之父,并获得 1994 年度图灵奖。

所谓专家系统，是指一个拥有某领域专家级知识、能够模拟专家思维方式、能够像人类专家一样解决实际问题的计算机系统，其基本结构如图 1-14 所示。知识库和推理机专家系统的两大核心模块，推理机依据知识库提供的专业领域知识进行自动的演绎或归纳推理，获得所需推理结论实现人工智能。显然，知识库的构建是实现专家系统的关键要点。

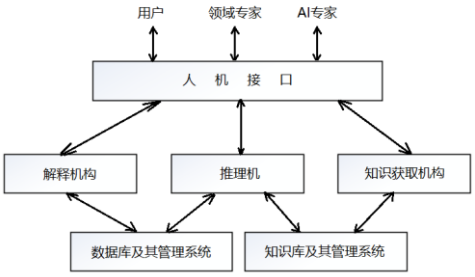


图 1-14 专家系统的基本结构

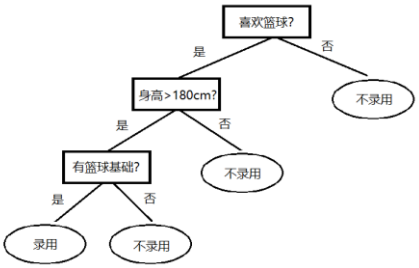


图 1-15 决策树模型示例

早期专家系统知识库中的知识为通过人工方式获取，经过专业训练的知识工程师与领域专家进行人工交互获得专家知识并将其整理成适当的数据结构存入知识库。这种人工构建知识库的方式存在如下弊端：

第一、知识库的普适性差，很多情况下需要针对特定的具体任务构建相应的知识库，需要频繁地人工改变知识库以适应任务的变化，例如用于识别猫的知识库不能用于识别狗，如果想用一个识别猫的专家系统来识别狗，就必须手工修改知识库。

第二、专家对事物的认识有时候具有一定的主观性，甚至会有一些错误，而且不同的专家对同一个事务给出的知识有时会有一些分歧，如何消除专家知识的主观错误或分歧有时候是一个非常困难的事情。

上述弊端很快成为制约专家系统和知识工程进一步发展的瓶颈。二十世纪八十年代，人们开始意识到让机器自己从样本数据中学习所需知识的重要性，并从符号主义的基本理论出发，比较系统地研究机器学习的基本理论和方法，逐步形成一套基于符号计算的机器学习理论和方法，称之为**符号学习**。

符号学习分为记忆学习、演绎学习和归纳学习这三种基本类型。**记忆学习**是一种最基本的学习方式，有时亦称之为**死记硬背式学习**，这种学习只需将知识记录下来，需要时再做原样输出，在本质上对信息进行存储和检索。**演绎学习**是一种以演绎推理为基础的学习方式，即从现有知识当中通过由一般到特殊的演绎推理获得并保存推理结论。**归纳学习**是一种以归纳推理为基础的学习方式，试图从具体示例或样例中归纳出一般规律。

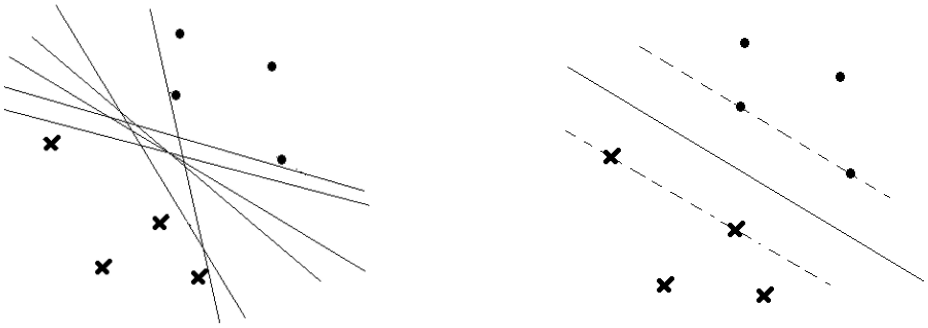
归纳学习是最重要的一种符号学习方式，研究成果也较为突出。作为符号学习代表性成果的决策树模型就是基于归纳学习。决策树模型是一个树形结构，包含了一个根结点、若干内部结点和若干叶子结点。该模型主要用于表示某种级联判断或决策，例如图 1-15 表示一个用于挑选篮球运动员的决策树模型，其中每个结点对应一次判断或决策，叶子结点表示判断或决策的最终结果。可以通过ID3、C4.5和CART等机器学习算法对数据样本自动构造决策树模型，构建用于符号推理的知识库。1995 年布瑞曼等人在决策树基础上进一步提出了随机森林模型及相关的机器学习算法，该模型通过构造多棵决策树并将它们各自的输出进行组合使得模型输出更具稳定性。

符号学习在机器学习历史上长期占据主导地位，基于符号学习方法的专家系统在多个领

域获得成功应用，例如辅助医疗诊断专家系统MYCIN、工业指导专家系统CONPHYDE等。直到二十世纪九十年代，基于概率统计理论的**统计学习**方法逐渐走向成熟，并凭借其理论的完备性和实际应用的卓越表现取代符号学习成为机器学习的主流方式。

统计学习源于符号学习中的归纳学习，继承了归纳学习通过分析数据获得一般化规律的思想，由基于概率统计的学习理论指导其归纳推理过程。统计学习的目标是理解样本数据，从样本数据中发现其内在规律，并利用这些规律去进行预测分析。统计学习的基本策略是假设是同一类型数据满足一定的统计学规律，并据此使用概率统计工具分析处理数据。

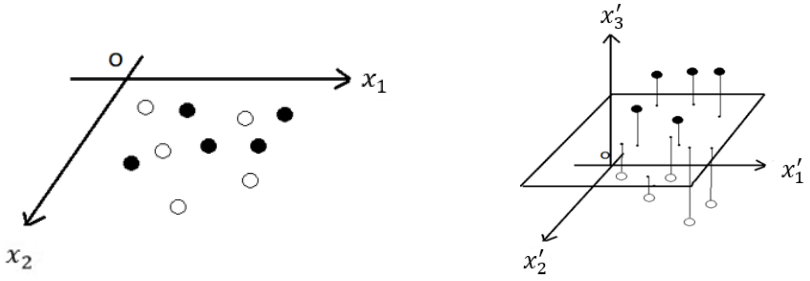
统计学习中最具代表性的成果是支持向量机模型及相关的统计学习算法。该模型由万普尼克和凯尔特斯在 1995 年正式提出，是一种基于小样本统计学习的二值分类器模型，目前已广泛应用于模式识别、自然语言处理等实际问题并取得较好的实践效果。



（a）分离超平面　　（b）间隔最大的分离超平面

图 1-16 分离超平面与最大间隔分离超平面

事实上，万普尼克早在 20 世纪 70 年代后期便提出了支持向量机的两个核心思想，即最大间隔和核方法。所谓**最大间隔**，是指正负两类样本与分离超平面之间的距离最大。如图 1-16（a）所示，对于一个可用超平面进行样本分类的线性可分任务，一般都存在无数个分离超平面，其中与两类样本的几何间隔最大的分离超平面具有最强的泛化能力，故而支持向量机所要寻找的分离超平面便是两类样本的几何间隔最大的分离超平面，如图 1-16(b)所示，其中虚线上面的点表示支持向量，实线表示间隔最大分离超平面。



（a） 二维特征空间的数据分布　　（b）三维特征空间的数据分布

图 1-17 核方法示意图

最大间隔分离超平面虽然是线性可分任务的最优解，但大部分的二分类任务都是线性不可分的，即不存在任意一个超平面可以将两类数据完美分隔开来，对于此类任务，核方法便是解决问题的关键所在。核方法的基本思想是将低维特征空间当中线性不可分的数据映射到高维特征空间当中，使得这些数据在高维特征空间当中线性可分，如图 1-17（a）所示，两类数据在二维平面当中线性不可分，但若利用某一映射将其转变为图 1-17（b）中所示的情

况，原本线性不可分的分类任务便被转化为了线性可分的任务，再通过最大间隔思想便可求得泛化性能最佳的分离超平面。

支持向量机具有一套比较完整的理论支撑，已被理论证明具有以下两方面的优势：

（1）支持向量机使用的最大间隔思想使得分类器模型只取决于支持向量，模型计算复杂度只与支持向量数目有关，有效避免了维数灾难问题并使得支持向量机对训练样本的变化具有较强的鲁棒性；

（2）支持向量机的核方法在一定程度上避免了直接在高维空间中处理问题，有效降低了问题求解的难度。

二十世纪九十年代以来，以支持向量机为代表的统计机器学习理论和方法得到蓬勃发展并取代符号学习成为机器学习的主流。在统计机器学习的带动下，机器学习进入快速发展的阶段，研究成果大量涌现、精彩纷呈，监督学习、无监督学习、半监督学习、强化学习、集成学习、迁移学习等机器学习方法不断得到发展和完善。这些机器学习方法在计算机视觉、自然语言理解、数据挖掘、经济决策与预测等多个领域的成功应用使得机器学习在人工智能重要性逐步显现，并将人工智能的发展带入机器学习时代。

1.2.3 连接学习的兴起

进入二十一世纪，计算机硬件计算能力获得飞跃发展。特别是英伟达公司在 2007 年推出基于 CUDA 的通用 GPU 版大大增强了 GPU 的开放性和通用性，吸引了大量使用各种编程语言的工程师纷纷使用 GPU 进行系统开发。人们在 2009 年开始尝试使用 GPU 训练人工神经网络，以有效降低多层神经网络的训练时间。2010 年推出的 NVIDIA-480 GPU 芯片已经达到每秒 1.3 万亿次浮点运算，能够很好地满足多层神经网络训练的高速度、大规模矩阵运算需要，使得连接学习训练困难的问题得到了很好的解决，为以深度学习为代表的连接学习兴起奠定了良好的硬件算力基础。

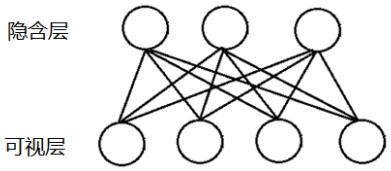


图 1-18 受限玻尔兹曼机示意图

与此同时，连接学习算法和理论研究取得了重要突破。2006 年，辛顿使用逐层学习策略对样本数据进行训练，获得了一个效果较好的深层神经网络——深度信念网络，打破了深层网络难以被训练的局面。逐层学习策略首先将深层神经网络拆分成若干相对独立的浅层的自编码网络，各个自编码网络可以根据其输入与输出一致的特点进行无监督学习，由此计算出连接权重；然后通过将多个训练好的自编码网络进行堆叠的方式获得一个参数较优的深层神经网络；最后，通过少量带标注的样本对网络进行微调便可获得一种性能优良的深层神经网络，即深度信念网络。

深度信念网络以受限玻尔兹曼机（RBM）为基本构件堆叠组建而成。RBM是一种自编码网络，其结构如图 1-18 所示，包含可视层和隐藏层。图 1-19 表示深度信念网络的基本结构，由图 1-19 可知，深度信念网络通过堆叠受限玻尔兹曼机的方式构建，前一个训练完成的RBM的隐藏层作为后一个RBM的可视层，层层堆叠，由此形成深度信念网络。

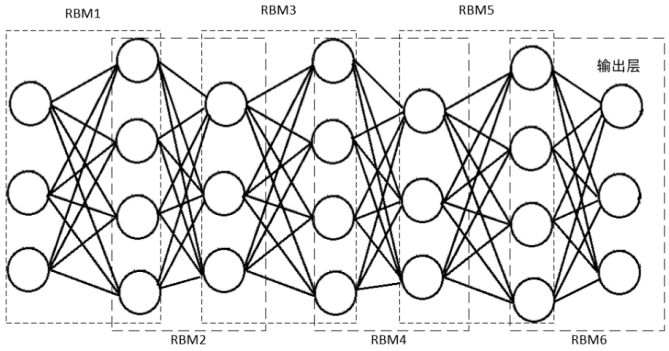


图 1-19多个 RBM 堆叠示意图

用于堆叠的受限玻尔兹曼机通过逐层训练的方式已经获得了较优的权重设置，这使得深度信念网络的初始权重较优，只需利用反向传播算法对连接权重进行微调即可完成训练。深度信念网络在效果上要优于支持向量机，这使得人们的目光再次回到连接学习上。为此次连接学习复兴做出重要贡献的辛顿将深层次神经网络的训练构造过程命名为**深度学习**，此后连接学习的理论和应用研究便在深度学习的名号下如火如荼地展开。

在 2012 年，基于深度卷积神经结构的Alexnet图像分类模型利用分布式 GPU 完成了 ImageNet 数据集中海量图像分类样本的训练，在合适的训练时间长度内取得了较好的训练效果，赢得了 2012 届图像识别大赛的冠军并实现了识别准确率高达 10.8%的提升。ImageNet 数据集在普林斯顿大学李飞飞教授主导下，通过众包平台 Mechanical Turk 历时两年时间创建，由 1500 万个标记图像组成，分为 22000 个类别，为当时最大的图像分类开源数据集，也是各种机器视觉算法的最有力的检测工具。

卷积神经网络是一个特殊的神经网络，它利用卷积操作使得网络层与层之间采用局部连接方式，这种连接方式不仅可以减少网络参数，并且更加符合生物神经系统工作的感受野机制。除此之外，卷积神经网络各层中对不同感受野进行处理时共享同一组参数，这进一步减少了模型参数数量。Alexnet图像分类模型的成功表明，通过大量样本训练获得的深层次的卷积神经网络可以有效解决过拟合问题。这项研究成果使得卷积神经网络迅速成为模式识别与计算机视觉研究领域的新宠，并涌现大批优秀研究成果，例如用于图像分类任务的GoogleNet, VGG, ResNet等深度卷积网络模型。目前，面向图像分类的深度卷积网络模型已呈百花齐放的发展态势。

近几年，人们进一步将深度卷积网络等深度网络模型用于图像中目标的自动检测，并取得了丰硕的研究成果。目前，基于深度网络模型的目标检测的算法主要分为两大类：第一类是两阶段检测算法，这类算法将检测问题划分为两个阶段，即首先产生候选区域，然后对候选区域进行分类，主要使用 R-CNN，Fast R-CNN，Faster R-CNN 等深度卷积网络模型；第二类是单阶段检测算法，这类算法不需要产生候选区域，直接生成物体类别概率和位置坐标值，主要使用 YOLO 和 SSD 等深度卷积网络模型。这两类目标检测算法各具特色，在一般

情况下，两阶段算法准确度较高，单阶段算法则速度较快。

图像分类和目标检测这类任务的研究对象都是独立的图像样本，模型无需考虑样本之间的联系，使用深度卷积网络往往就能取得较好的效果。如果所处理的信息都是一个连续序列，例如一段音频或视频，此时卷积神经网络模型会割裂信息序列项的前后联系，为解决这一问题，人们进一步提出了可以处理序列信息的深度循环神经网络（RNN）和长短时记忆网络（LSTM），以有效解决信息序列的表示和处理问题。目前，RNN和LSTM已在视频行为分析、语音信息的识别与合成、自然语言理解与机器翻译等多个领域取得成功应用。

深度学习方法也使得强化学习领域的研究取得了长足的进展，利用基于深层神经网络模型所实现的深度价值网络和深度策略网络在强化学习领域起到了重要的作用，例如著名的AlphaGO围棋程序使用强化学习策略构建深度策略网络，将其用于根据当前盘面状态确定走棋策略，并通过构建深度Q网络模型实现对走棋策略的评估，以寻求策略和评估策略的交互的方式实现布棋并取得了很好的效果。

此外，人们还尝试使用深度学习模型模拟现实生活中的真实数据。2014 年 6 月，古德费勒等学者提出了名为生成对抗网络（GAN）的生成模型，该模型可以根据需要生成新的样本。该模型由两个子模型组成，第一个为生成器，它可以根据训练样本来生成新的样本，另一个为判别器，它的输入为训练集中的真实样本或生成器所生成的虚假样本，目标是判断输入样本是否为真实样本或虚假样本。当判别器无法判别时，就意味着生成器所生成的样本与真实样本几乎来自同一分布，从而完成了新样本的生成任务。

深度学习除了实际效果上的大幅改善之外，还能避免特征人工选择或构造方式的不足，深度学习利用网络模型自动提取的特征往往更有利于模型解决实际问题。但这种特征自动提取方式也存在不容忽视的问题：首先深度学习缺乏严格的理论基础，实现过程是个黑盒，即深度学习模型所用的特征所表达的信息往往难以理解；其次，深度学习模型拥有大量参数，通常需要海量训练样本，这无疑增加了训练的难度。如何在保证效果的基础上减小模型、减少参数或者实现大模型的小样本训练仍然是一个具有挑战性的问题。

1.3 机器学习的基本问题

机器学习的基本思想是通过从样本数据中提取所需信息构建一个有效的机器学习模型，并根据所建模型完成分类、回归、聚类等具体的机器学习任务。使用机器学习方法求解具体问题时需要面临一些基本问题。首先是样本特征的提取问题：样本数据所包含的信息是多种多样的，不同的机器学习模型和任务所需的样本信息通常也各具特色，样本特征的提取问题要解决的是如何从样本数据中获取适当的信息满足模型构建和完成机器学习任务的需要。其次是机器学习规则构建问题：规则是机器学习模型的基本构件或具体表现形式，不同的机器学习方式采用与之相适应的不同类型的规则。演绎学习使用逻辑规则，归纳学习使用关联规则，统计学习和连接学习使用映射规则，要实现机器学习模型的构建就必须解决机器学习规则构建问题。最后是模型评估问题：对于已建的机器学习模型，必须对其进行性能评估以判定是否满足任务需求。因此，机器学习的基本问题主要是特征提取、规则构建和模型评估。本节主要讨论这三个基本问题及解决方法。

1.3.1 特征提取

机器学习中的样本信息由一组表征数据描述，例如一幅彩色图像可用三个分别表示红、绿、蓝幅度值的矩阵进行表示，这三个矩阵就是这幅彩色图像的表征数据。样本表征数据虽然包含样本的所有信息，但数据量往往较大且存在一定的冗余，不便直接处理。因此，需要对样本的表征数据进行适当处理获得机器学习和实际问题求解所需要的特定信息，这种处理过程通常称之为样本的**特征提取**。具体地说，特征提取是对机器学习任务所涉及的原始数据表征进行处理，得到一组具有特定意义的特征数据作为样本数据的优化描述，并以尽可能少的特征数据表达出尽可能多的信息，以方便后续的模型优化和任务求解。

特征提取一般包括如下两个基本步骤：

（1）构造出一组用于对样本数据进行描述的特征，即特征构造；

（2）对构造好的这组特征进行筛选或变换，使得最终的特征集合具有尽可能少的特征数目且包含尽可能多的所需样本信息。

对于任意给定的一个样本，其所有特征值构成的向量称为该样本的**特征向量。**对于经过特征提取的样本，通常使用特征向量代替表征向量以获得对样本数据的一种优化描述。

可针对不同的机器学习任务构造出相应的特征。例如在自然语言理解领域通常用词袋特征或词频特征抽象地表达具体本文特征。词袋特征忽略了文本中词语之间的上下文联系，而只考虑每个词语出现的次数，例如句子 “我和哥哥都喜欢看电视剧。”和“哥哥还喜欢看篮球比赛。”可按照其中出现的词语构建如下词典：

{我：1；和：2；哥哥：3；都：4；喜欢看：5；电视剧：6；还：7；篮球比赛：8}

该词典中包含 8 个词语，每个词语都有唯一的索引，根据该词典可分别将上述两个句子转化为一个 8 维向量，即：

$$(1,1,1,1,1,1,0,0);(0,0,1,0,1,0,1,1)$$

向量中每个分量值表示该分量所对应词语在文本中出现的次数ω。例如，若将上述两句合并为一个文本“我和哥哥都喜欢看电视剧，哥哥还喜欢看篮球比赛。”，则该文本对应的词袋特征可表示为：$(1,1,2,1,2,1,1,1)$。

对于任意给定的一个文本，词典中各个词语该文本中出现的次数及分布情况通常与该文本具体内容有着非常密切的关系，故可根据词袋特征对文本进行分类、聚类等处理。例如，对于如下三段文本：

文本 1:“在非洁净的环境下生产的半导体工业产品合格率仅为 10%—15%左右。显示屏生产线厂房项目总建筑面积 178 万平方米，5 个核心厂房均各设一个洁净区，中建一局负责的切割与偏贴厂房洁净区最高洁净度要求为 1000 级，即要求洁净区域内每立方英尺存在 0.5μm 的微尘粒子不能超过 1000 颗。建成后，屏幕生产所用的切割片、涂色和背光板组装等系列设备将安放在此。而屏幕生产的最核心设备曝光机所在的阵列厂房洁净区对洁净等级要求为 100 级甚至 10 级。此外，生产线车间对空气洁净度、温度、湿度、防静电、微振、光照度、噪声等都有严格的参数要求。”

文本 2:“不同于其他厂房建设，所有进入洁净室的施工人员都要穿着洁净服，进入前需在更衣室吹掉灰尘，随时随地有专业人员打扫擦拭。洁净区建成后还要擦拭两遍，竣工前经十余项检测通过后才可搬入生产设备。” 目前施工区域正在进行第一阶段分层级地面打磨环氧和水电风管道安装工作。“地面如果高低不平，会产生大量灰尘微粒积压，一个细小的微粒会直接导致产品报废，特别是洁净区域对平整度的要求近乎苛刻，1 平方米内地面高差最多 2 毫米。”

文本 3:“公民科学素质水平是决定国家整体素质的重要指标，至少 10%的公民具备科学素养是该国家成为创新型国家的重要节点。发布于 2015 年的第九次中国公民科学素质调查结果显示，我国具备科学素质的公民比例达到了 6.20%。白希介绍说，改革开放以来，中国公民的科学素质稳步提升，但数据也反映出我国公众科学素质发展中一些不平衡、不充分的情况，下一步将进一步缩小差距，力争尽快赶上世界先进水平。”

可分别对这些文本提取词袋特征，得到如图 1-20 所示的词袋特征分布图。图中横坐标中每个点对应一个词语，纵坐标表示词语出现的次数，Para1、Para2、Para3分别为文本 1、文本 2 和文本 3 的词袋特征分布。这里只统计文本中包含具体含义的实词部分，且将相同词根的词语统计为同一个词语，语气词和系动词等不含具体含义的虚词不纳入统计范围。有了图 1-20 所示的文本词袋特征，就可通过适当的聚类算法进行文本聚类，例如用$k-$均值聚类算法可得到聚类结果为：文本 1 和文本 2 聚为一类，文本 3 单独作为一类。

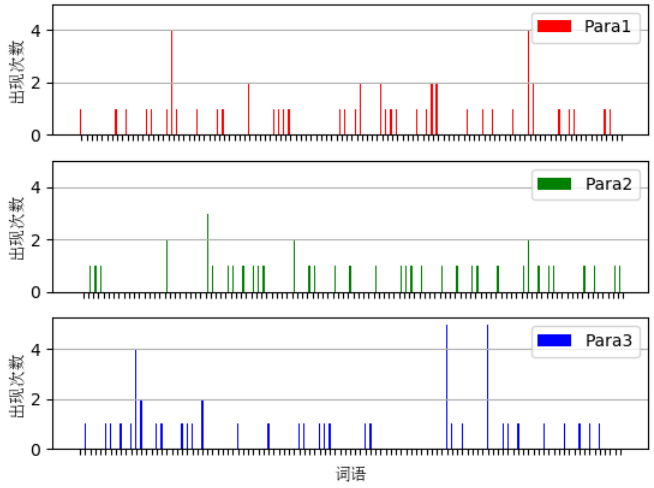


图 1-20 三段文本的词袋特征分布

计算机视觉和视频图像处理领域常用特征有 LBP 特征、Canny 特征等。LBP 特征是一种图像纹理特征，表达的是物体表面具有缓慢或周期性变化的结构组织排列属性。LBP 特征算子定义在3×3像素网格窗口中，设某个3×3窗口中像素灰度取值如图 1-21 中左半部分所示，则该窗口中心像素点的 LBP 值计算过程如下：

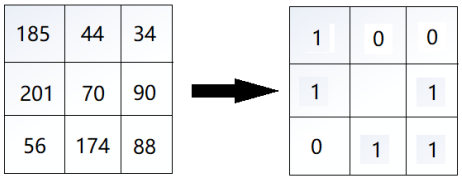


图 1-21 LBP 二进制编码意图

第一步：将周围像素点的取值与中心像素点取值大小进行比较，若边缘像素点的取值大于中心像素点位置，则该点取值为 1，否则取值为 0，经过这一操作可将周围像素值转化为二进制编码，计算结果如图 1-21 中右边部分所示。

第二步：从左上角像素开始，按顺时针方向将 LBP 二进制编码组合成一个八位二进制数并将其转化为一个十进制数，即得到 LBP 码，例如图 1-21 中窗口 LBP 码转化结果为：

$$10011101_2 = 157$$

即该3×3窗口中心像素的 LBP 码为 157。

第三步：重复上述步骤求得图像所有像素的 LBP 码，并将其作为 LBP 特征图。图 1-22（b）为图 1-22（a）所示图像的 LBP 特征图。

(a)原始图像 (b) LBP 特征图 （c）Canny 边缘特征

图 1-22 图像的特征提取

LBP 特征是通过计算单个像素与其相邻像素之间的灰度关系得到，由于相邻像素之间通常存在一定的相关性，故整幅图像各像素的 LBP 码之间也会存在一定的相关性，这使得 LBP 特征图能够表现出一定的全局纹理特征。

Canny 特征主要表达图像的边缘信息，可用于确定图像中目标的轮廓和位置。由于图像中目标边缘的亮度有一定变化，边缘像素的梯度通常较大，Canny 特征提取主要根据梯度值和方向寻找边缘像素。图 1-22（c）表示图 1-22（a）所示图像的 Canny 边缘特征。

除了上述 LBP 纹理特征和 Canny 边缘特征之外，计算机视觉领域还可使用很多其他特征，例如颜色直方图、Haar 特征、SIFT 特征等。这些特征都是人们基于对实际问题的分析而人工构造出来的。随着深度学习技术的不断发展，人们逐渐开始尝试让计算机根据任务的实际需求自动进行特征提取，深度卷积神经网络是一种最常用的特征自动提取模型。卷积神经网络是一种以层级连接方式构建的网络模型，该模型从第一个卷积层开始逐层对样本进行特征提取，模型后一层的特征提取基于前一层所提取特征。图 1-23 表示卷积神经网络 LeNet-5 模型的特征提取过程，图中最下方的图片为模型的原始输入，越往上表示由越深的网络层所提取到的较高层特征。

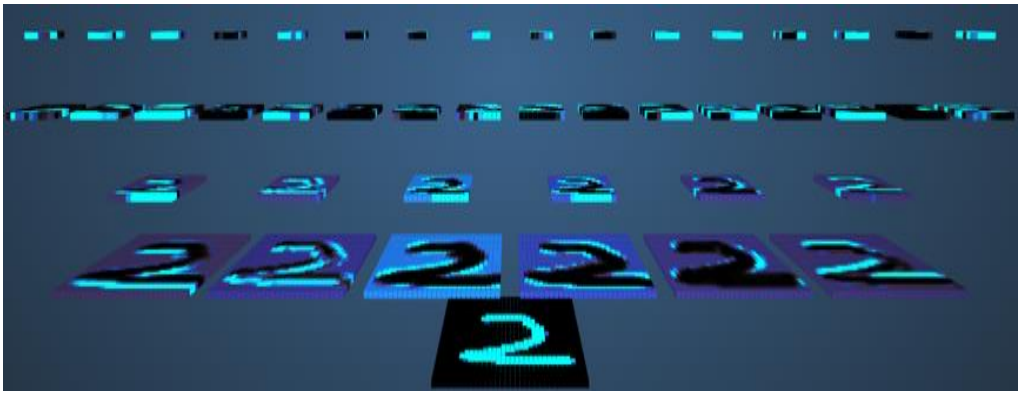


图 1-23 卷积神经网络自动提取特征示意图

与人工构造特征相比，机器自动提取的特征所包含的信息往往难以被人们理解，越高层的特征越抽象，但由于计算机自动提取的特征包含更多模型所需信息，故此使用这些自动提取的特征来解决实际问题通常能够得到更好的效果。

在特征构造完成之后有时还需要对这些特征进行进一步的筛选或融合，剔除可能存在的

与实际任务无关的信息或冗余信息。对特征进行筛选的过程称之为**特征选择**。具体而言，对于构造好的特征构成的特征集合D,特征选择的目标是寻找到D的某个子集D',使得基于D'所建模型的性能与基于D所建模型性能相当，并具有较低的模型优化计算量。通常使用子集搜索或相关性评估方式实现对特征的选择。

子集搜索就是通过搜索特征集合D的所有子集选择效果最好的子集作为最优特征子集。例如，某机器学习任务的特征集合为：

{Canny 边缘特征，颜色直方图，Laplacian 边缘特征，LBP 纹理特征}

分别用该特征集合每个子集进行模型训练，如果发现使用

{LBP 纹理特征，颜色直方图，Laplacian 边缘特征}

这一特征子集训练获得的模型性能最优，则该特征子集便为最优特征子集。

如果特征集合D中元素个数较多，子集搜索方法显然不可行。此时可用特征相关性评估的方式确定最优子集。相关性评估方法的基本思想是使用某个统计量来评估单个特征与样本真实标记之间的相关性，并选择相关性较大的几个特征来构成近似的最优特征子集。

显然，与样本真实标记有较大相关性的特征通常能更好地帮助模型对样本的预测，而与样本真实标记相关性不大的特征通常只能提供较少的参考信息。相关性评估常用的统计量有χ^2统计量、信息熵等，下面以χ^2统计量为例简要介绍相关性评估的具体做法。

对于任意给定的某个样本特征，相关性评估方法首先假设某一特征与样本的真实标记值无关，然后对该假设进行假设检验，判断该假设是否成立。使用χ^2统计量进行假设检验的方式被称为χ^2检验。例如假设H为某一特征与样本的真实标记值无关，则可通过样本数据的实际值A和在假设成立条件下的理论值T计算出如下χ^2统计量：

$$\chi^2 = \sum \frac{(A-T)^2}{T} \qquad (1\text{-}11)$$

在假设H成立的条件下，实际值A和理论值T之间的差别应该较小，即χ^2是一个较小的数，故当χ^2的值超过某一阈值时，则可以拒绝假设，认为该特征与样本的真实标记值相关。

【例题 1.2】对于一个二分类问题，其特征集合为：

D = {Canny 边缘特征，颜色直方图，Laplacian 边缘特征，LBP 纹理特征}

将D中各特征离散化为 0 或 1 这两种取值状态，特征与样本真实标记y之间的关系分别如表 1-1 至表 1-4 所示，试选出包含三个特征的最优特征子集。

表 1-1 Canny 边缘特征取值情况与真实标记y取值关系表

	$y=-1$的样本数	$y=+1$的样本数	合计
Canny 边缘特征 = 0	60	59	119
Canny 边缘特征 = 1	10	122	132
合计	70	181	251

表 1-2 颜色直方图取值情况与真实标记y取值关系表

	$y=-1$的样本数	$y=+1$的样本数	合计
颜色直方图 = 0	31	70	101
颜色直方图 = 1	39	111	150
合计	70	181	251

表 1-3 Laplacian 边缘特征取值情况与真实标记$\boldsymbol{y}$取值关系表

	$y=-1$的样本数	$y=+1$的样本数	合计
Laplacian 特征 $=0$	27	82	109
Laplacian 特征 $=1$	43	99	142
合计	70	181	251

表 1-4 LBP 纹理特征取值情况与真实标记$\boldsymbol{y}$取值关系表

	$y=-1$的样本数	$y=+1$的样本数	合计
LBP 纹理特征 $=0$	18	86	104
LBP 纹理特征 $=1$	52	95	147
合计	70	181	251

【解】(1)假设H_0: Canny 边缘特征与y不相关。在假设H_0成立的条件下，即 Canny 边缘特征与y不相关时，理论上$y=-1$的样本占总样本的比例约为 27.89%，则理论上 Canny 边缘特征与y取值之间的关系应该如表 1-5 所示，

表 1-5 理论上特征取值与真实标记$\boldsymbol{y}$取值关系表

	$y=-1$的样本数	$y=+1$的样本数	合计
$f=0$	30.4001	78.5999	109
$f=1$	39.6038	102.3962	142

则可算出相应的χ^2统计量：

$$\chi^2=\sum\frac{(A-T)^2}{T}\approx 59.59$$

假设H_1:颜色直方图与y不相关。在假设H_1成立的条件下，理论上颜色直方图与y取值之间也应该满足表 1-5 中的关系，由此可算出相应的χ^2统计量：

$$\chi^2=\sum\frac{(A-T)^2}{T}\approx 2.03$$

假设H_2:Laplacian 边缘特征与y不相关。在假设H_2成立的条件下，理论上 Laplacian 边缘特征与y取值之间也应该满足表 1-5 中的关系，可算出χ^2统计量：

$$\chi^2=\sum\frac{(A-T)^2}{T}\approx 0.93$$

假设H_3:LBP 纹理特征与y不相关。在假设H_3成立的条件下，理论上 LBP 纹理特征与y取值之间也应该满足表 1-5 中的关系，可算出χ^2统计量：

$$\chi^2=\sum\frac{(A-T)^2}{T}\approx 5.30$$

根据上述统计量值的排序选择三个χ^2统计量较小的特征构成所需的最优特征子集D'，即有：$D'=\{$Canny 边缘特征，LBP 纹理特征，颜色直方图$\}$。□

特征选择方式仅仅从原特征集合当中选择了几个特征组成特征子集，却并未改变其中的特征。事实上，除了特征选择之外，还可通过特征变换方式排除或减少特征集合中特征所包含的无关或冗余信息，例如通过某种投影映射方式对特征数据进行适当降维等，具体可参考有关 LDA 和 PCA 算法的相关内容。

1.3.2 规则构建

机器学习模型通常根据样本的表征或特征信息实现回归、分类、聚类等问题的求解，因此需要在样本表征或特征信息与模型输出之间建立一定的联系。机器学习模型通常以规则的形式表达这种联系，可将规则看成是机器学习模型的基本构件或具体表现形式。因此，规则构建是建立机器学习模型所必须解决的一个基本问题。不同的机器学习方式通常采用与之相适应的不同类型的规则。这些规则主要有用于演绎学习的逻辑规则、用于归纳学习的关联规则以及用于统计学习和连接学习的映射规则。

演绎学习主要通过命题逻辑和谓词逻辑的演绎推理进行学习，使用假言三段论、排中律、矛盾律等逻辑规则进行演绎推理。演绎学习的理论基础完备、严谨，学习过程语义清晰、易于理解，但难以处理不确定性信息，对复杂问题的求解会出现难以解决的组合爆炸问题。因此，对于演绎学习，其规则构建主要目标和难点是建立一套能有效处理不确定性信息的逻辑规则。对模糊性、随机性等不确定信息的处理局限是制约演绎学习发展的主要瓶颈。事实上，机器学习目前使用的基本策略是从样本或样例中学习，归纳学习、统计学习和连接学习已经成为机器学习的主流方式。下面着重分析讨论关联规则和映射规则的构建方法。

所谓关联规则，是指一类已指明条件蕴含关系的规则，故亦称为$\text{if}-\text{then}$规则。归纳学习的目标是采用适当方式从若干样例或样本中归纳总结出一组具有较好普适性的关联规则。这组关联规则的普适性主要表现为既符合已知样例或样本的性质，又能给新的示例或样本赋予较为合理的逻辑判断输出。

可用命题逻辑的蕴含式$X \to Y$表示一个具体的关联规则，意为如果命题X成立，则命题Y成立。其中X称为前件或条件，Y称为后件或结论。由于关联规则具有明确的因果蕴含关系，故用关联规则构造的模型通常都具有很好的可解释性，便于分析和理解。

对所有已知样例或样本做出正确推断是对归纳学习所得关联规则的基本要求。也就是说，要求关联规则的正确推断涵盖或覆盖所有训练样本数据。由此设计而的关联规则构造算法称为**序列覆盖算法**。具体地说，序列覆盖算法递归地归纳出单条关联规则去逐步覆盖训练样本集中的正样例，当训练样本集中所有正例均被已归纳的关联规则所覆盖时，此时对应的关联规则集就是所求规则集。然后按适当标准对所求规则进行排序，确定规则使用的优先级。

表 1-6 小张户外运动情况记录表

编号	天气	气温	事务	户外运动
1	晴朗	不舒适	有	否
2	晴朗	舒适	无	是
3	阴雨	不舒适	无	否
4	阴雨	舒适	无	是
5	晴朗	不舒适	无	是

序列覆盖算法的关联规则归纳构建主要通过对假设空间的搜索完成。下面结合表 1-6 所示数据简要介绍基于假设空间的关联规则搜索过程。表 1-6 是关于小张是否进行运动的相关数据集。若要根据该数据集归纳出小张是否进行户外运动的规则，最容易想到的方法是从最一般的规则前件$\emptyset$开始对假设空间进行遍历搜索，逐步特化规则，最终获得所需规则，具

体搜索过程如下：

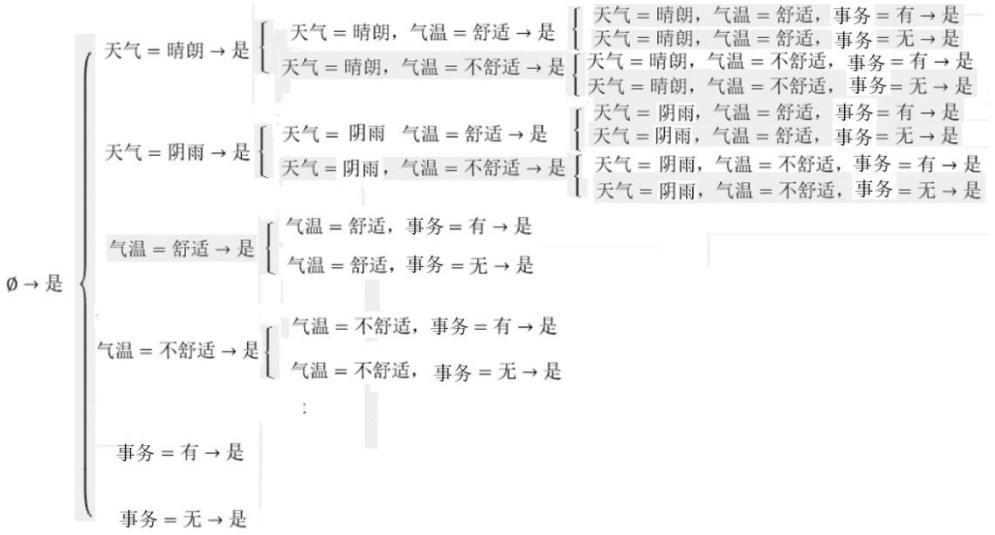


图 1-24 关联规则搜索空间

选择关联规则的后件为“是”，则搜索起始位置为∅ → 是，搜索空间如图 1-24 所示。首先搜索到的规则为“天气 = 晴朗 → 是”，则 1、2、5 号样例的属性取值与其前件一致，但 1 号样例不符合该规则的推断，故忽略该规则继续向后搜索。当搜索到关联规则“气温 = 舒适 → 是”时，2、4 号样本的特征取值于其前件一致且均满足该规则的推断，故 2、4 号样例被这条关联规则所覆盖。将这条规则记录下来并删除 2、4 号样例，由此完成一条关联规则的归纳构建。递归上述关联规则的归纳构建过程可得两条关联规则，并且它们可以覆盖训练样本集中所有正例。因此，这两条关联规则即为所求关联规则。最后，根据关联规则所覆盖的样例数目的大小对其进行排序，得到序列覆盖算法的计算结果：

气温 = 舒适 → 是

天气 = 晴朗，气温 = 不舒适，事务 = 无 → 是

假设某天的情况为（天气 = 阴雨，气温 = 舒适，事务 = 有），则可根据上述两条关联规则推断出小张会进行户外运动的结论。

当样本属性个数或属性取值较多时，关联规则的搜索空间可能会变得很大，遍历搜索算法会因计算复杂度大幅上升而变得不可行。此时可用贪心搜索算法求得近似解，即每一次搜索都只朝着当前最优方向进行，从而有效地缩小搜索空间。基于贪心搜索的关联规则学习算法有很多，其中最具代表性的主要有CN2算法、AQ算法等，不再赘述。

在统计学习和连接学习领域，样本数据表征或特征与模型输出之间的关系通常表现为映射规则，即从输入空间到输出空间的映射函数，映射规则的构建过程其实就是确定映射函数的过程，而映射函数确定问题则可转化为求解目标函数最值的优化问题。因此，映射规则的构建主要是通过对目标函数进行优化的方式实现，基本步骤如下：

（1）根据求解问题的具体要求确定机器学习模型的基本类型或映射函数的基本结构；

（2）根据样本数据的具体形式和模型特点确定合适模型优化标准，如经验风险最小化、结构风险最小化、类内距离最小化、类间距离最大化等；

（3）设计构造模型优化的目标函数；

（4）通过对目标函数进行最值优化获得所需映射函数，完成映射规则构建。

下面结合实例介绍统计学习映射规则构建的具体过程。

【例题 1.3】表 1-7 为某公司部分职员的年龄（岁）和薪资（千元/月）数据，这些职员由公司管理人员和普通员工组成。试根据表中信息大致判断哪些职员为管理人员，哪些职员为普通员工。

【分析】虽然表 1-7 中数据没有标注职位信息但管理层职员通常年龄相对较大且薪资相对较高，故可用聚类方式求解。聚类的基本思想通过聚类算法将不带标注的样本聚合成适当的族群。由于聚类的学习对象是不带标注的样本，无法使用基于误差的经验风险最小化或结构风险最小化原则，故聚类算法一般首先使用类内距离最小化原则构建目标函数，然后对目标函数进行优化确定映射规则。

表 1-7 某公司职员年龄及薪资情况表

编号	1	2	3	4	5	6	7	8	9
年龄	27	47	31	44	21	23	50	23	21
薪资	4.3	6.3	5.2	7.1	4.0	3.9	6.7	4.4	3.9
编号	10	11	12	13	14	15	16	17	18
年龄	32	48	54	29	34	51	56	40	27
薪资	4.6	6.3	7.2	5.0	4.7	6.8	7.2	6.1	4.1

【解】使用对表 1-7 中样本数据进行聚类的方式求解。令$X_i = \{a_i, e_i\}$表示表 1-7 中第i个职员的样本数据，其中a_i表示年龄，e_i表示薪资，由此建立如下聚类映射规则基本结构：

$$f(X_i) = \begin{cases} C_1, & i \in \Delta_1 \\ C_2, & i \in \Delta_2 \end{cases}$$

其中C_j表示第j个簇群，Δ_1和Δ_2待定的样本数据编号集合。

根据类内距离最小化原则确定如下目标函数：

$$D = \sum_{j=1}^{2} \sum_{X_i \in C_j} \sqrt{(a_i - a_{u_j})^2 + (e_i - e_{u_j})^2} \qquad (1-12)$$

其中：u_j表示第j个簇群的聚类中心，a_{u_j}表示u_j的年龄值，e_{u_j}表示u_j的薪资值。

D为所有数据点与其所在族聚类中心的距离之和，可以很好地表示聚类的类内距离。下面使用$k-$均值聚类算法对目标函数D进行优化，其中k表示族群个数，本例中$k = 2$。$k-$均值聚类的基本思路为：首先任选两个点分别作为两个族群的初始聚类中心，然后将剩余数据根据其与聚类中心的距离划分到对应的簇中并根据所聚数据的均值更新聚类中心，递归上述过程直至聚类中心的位置不再变化（聚类中心收敛）即完成对目标函数的优化，最后根据收敛的聚类中心生成聚类映射规则的具体形式。

表 1-8 第一轮类内距离计算

	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
u_1	4.1	17.2	6.0	4.0	23.1	4.0	6.0	5.0
u_2	16.0	3.1	26.1	24.1	3.0	24.0	26.1	15.1
	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}
u_1	31.0	37.2	2.1	7.0	24.1	29.1	13.1	0.2

u_2	1.0	7.1	18.1	13.1	4.0	9.0	7.0	20.1

令$u_1 = X_1 = \{27，4.3\}$和$u_2 = X_2 = \{47，6.3\}$，分别计算数据$X_3, X_4, \ldots, X_{18}$到数据X_1, X_2之间的欧式距离，表 1-8 中数据为计算结果。根据表 1-8 中的计算结果将每个数据分别划入其与聚类中心距离较小的簇群中，可得如下划分：

$$C_1：X_1、X_3、X_5、X_6、X_8、X_9、X_{10}、X_{13}、X_{14}、X_{18}$$

$$C_2：X_2、X_4、X_7、X_{11}、X_{12}、X_{15}、X_{16}、X_{17}$$

计算C_j中数据均值，并将该均值作为簇C_j新的聚类中心，即将聚类中心更新为：

$$u_1 = \{26.8，4.41\}，u_2 = \{48.75，6.7125\}$$

计算样本数据与上述聚类中心的距离，计算结果如表 1-9 所示。

表 1-9 第二轮类内距离计算

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
u_1	0.23	20.3	4.3	17.4	5.8	3.8	23.3	3.8	5.8
u_2	21.9	1.8	17.8	4.8	27.9	25.9	1.3	25.9	27.9
	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}
u_1	5.2	21.3	27.3	2.3	7.2	24.3	19.3	13.3	0.4
u_2	16.9	0.9	5.3	19.8	14.9	2.3	7.3	8.8	21.9

依据表 1-9 中的计算结果，得到如下划分：

$$C_1：X_1、X_3、X_5、X_6、X_8、X_9、X_{10}、X_{13}、X_{14}、X_{18}$$

$$C_2：X_2、X_4、X_7、X_{11}、X_{12}、X_{15}、X_{16}、X_{17}$$

此次划分与前次划分相同，聚类中心收敛，故算法结束，获得下标集合：

$$\Delta_1 = \{1,3,5,6,8,9,10,13,14,18\}；\Delta_2 = \{2,4,7,11,12,15,16,17\}$$

由此得到聚类映射规则为：

$$f(X_i) = \begin{cases} C_1, & i \in \{1,3,5,6,8,9,10,13,14,18\} \\ C_2, & i \in \{2,4,7,11,12,15,16,17\} \end{cases}$$

即第 1、3、5、6、8、9、10、13、14、18 号职员为普通员工，第 2、4、7、11、12、15、16、17 号职员为管理人员。□

对于统计学习分类任务映射规则的构建，其基本流程与上述聚类任务类似，只是在优化标准和目标函数的设计上有所差异。下面以线性可分的二分类任务为例简要介绍分类映射规则构建的具体过程。设有训练样本集$S = \{(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)\}$，其中每个样本由$t$个特征描述，分别为$x_1, x_2 \ldots, x_t$，$y_i$为样本的标注值且$y_i \in Y = \{+1, -1\}, i = 1,2, \ldots, n$。现对上述分类任务构建映射规则，具体步骤如下：

第一步：确定机器学习模型的基本类型和映射函数的基本结构。使用支持向量机模型实现线性可分的二分类任务。支持向量机通过如下超平面实现样本数据的二分类：

$$\boldsymbol{w}^T X + b = 0 \qquad (1-13)$$

其中$\boldsymbol{w} = (w_1, w_2, \ldots，w_t)^T$为参数向量，$X = (x_1, x_2, \ldots，x_t)^T$为特征向量，$b$为偏置项。支持向量机的分类目标将分类数据分置超平面的两侧，由此可得分类映射规则基本形式为：

$$f(X) = \mathrm{sgn}(\boldsymbol{w}^T X + b) \qquad (1-14)$$

其中$\mathrm{sgn}(t)$为阶跃函数。

第二步：确定合适模型优化标准。支持向量机采用硬间隔最大化原则进行学习，即使得两类数据到分离超平面的距离最远。例如对于图 1-25 中用实线和虚线表示的两个分离超平面，由于数据点距离实线超平面较远，故实线超平面的硬间隔较大。

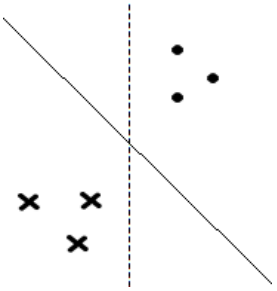


图 1-25 硬间距大小示意图

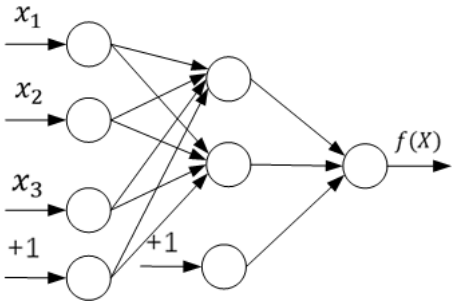


图 1-26 一个隐层的多层感知机

第三步：设计构造模型优化的目标函数。根据硬间隔最大化原则，需要构建一个分离超平面，使得样本点到超平面的距离最大。令$X_i=(x_{1i},x_{2i},\ldots,x_{ti})^T$表示训练样本集S中任意给定的第$i$个示例，则不难得到$X_i$到分离超平面$\boldsymbol{w}^TX+b=0$的几何间隔$d_i$为：

$$d_i=\frac{1}{||\boldsymbol{w}^T||}y_i(\boldsymbol{w}^TX_i+b) \qquad (1-15)$$

令$d=\min\{d_1,d_2,\ldots,d_n\}$，则使得$d$最大的分离超平面即为所求。由此可得如下目标函数：

$$\max_{\boldsymbol{w}^T,b}\ d=\frac{1}{||\boldsymbol{w}^T||}Y(\boldsymbol{w}^TX+b);\quad \text{s.t.}\quad \frac{1}{||\boldsymbol{w}^T||}y_i(\boldsymbol{w}^TX_i+b)\geq d$$

显然，若按比例缩放$\boldsymbol{w}^T$中元素和偏置b，则$Y(\boldsymbol{w}^TX+b)$值也会同比例缩放，这并不影响优化结果。故可令$Y(\boldsymbol{w}^TX+b)=1$，则有$d=1/||\boldsymbol{w}^T||$。由此可将约束条件简化为：

$$y_i(\boldsymbol{w}^TX_i+b)-1\geq 0$$

另外，由于最大化$1/||\boldsymbol{w}^T||$与最小化$||\boldsymbol{w}^T||^2$等价，故可将目标函数转化为：

$$\min_{\boldsymbol{w}^T,b}\ ||\boldsymbol{w}^T||^2;\ \text{s.t.}\quad y_i(\boldsymbol{w}^TX_i+b)-1\geq 0 \qquad (1-16)$$

第四步：采用适当优化计算方法对上述目标函数进行优化，求得最优参数$\boldsymbol{w}^{*T}$和偏置b*项，得到所求的分类映射规则$f(\boldsymbol{X})=\mathrm{sgn}(\boldsymbol{w}^{*T}\boldsymbol{x}+\mathrm{b}^*)$，完成映射规则的构建。

连接学习映射规则的构建也遵从上述步骤，不过在选择模型时会选用连接学习模型，即神经网络模型。下面以 3 维特征向量的二分类问题为例，讨论其映射规则的构建过程：

第一步：确定机器学习模型的基本类型和映射函数的基本结构。选择包含一个隐层的多层感知机作为分类模型，其网络结构如图 1-26 所示，其中每个圆圈代表一个神经元。由于特征向量维数为 3，故输入层仅含三个输入神经元。

令w_{ki}^l为第l层第i个神经元与第$l+1$层第k个神经元之间的连接权重为w_{ki}^l，b_i^l为第l层第i个节点的偏置项，各层激活函数均为φ，则对于样本输入$X=(x_1,x_2,x_3)^T$，该模型两个隐层节点的输出h_1,h_2分别为：

$$h_1=\varphi\left(\sum_{i=1}^{n}w_{1i}^1x_i+b_1^1\right);h_2=\varphi(\sum_{i=1}^{n}w_{2i}^1x_i+b_2^1)$$

将上式表示为矩阵形式, 则有：

$$\boldsymbol{h}=\begin{bmatrix}h_1\\h_2\end{bmatrix}=\varphi\left[\begin{pmatrix}w_{11}^1 & w_{12}^1 & w_{13}^1\\w_{21}^1 & w_{22}^1 & w_{23}^1\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}+\begin{pmatrix}b_1^2\\b_2^2\end{pmatrix}\right]=\varphi(\boldsymbol{w}^1X+\boldsymbol{b}^2)$$

其中$\boldsymbol{w}^1$为输入层到隐含层的连接权重矩阵，$\boldsymbol{b}^2$为隐含层的偏置向量。同理可得该模型的输出$f(X)$为：

$$f(X)=\varphi(w_{11}^2h_1+w_{21}^2h_2+b_1^3)=\varphi\left(\boldsymbol{w}^{2^T}\boldsymbol{h}+b_1^3\right)$$

其中$\boldsymbol{w}^2$为隐含层到输出层的连接权重向量，$\boldsymbol{h}$为隐含层的输出向量。

第二步：采用结构风险最小化原则模型优化标准，该原则是对经验风险最小化原则的一种改进。模型F在m元训练集$G=\{(X_1,y_1),(X_2,y_2),\dots,(X_m,y_m)\}$上的结构风险定义为：

$$R_{\text{srm}}(F)=\frac{1}{m}\sum_{k=1}^{m}L(y_k,F(X_k))+\lambda K(F)\qquad(1-17)$$

其中L为损失函数，$K(\text{F})$为模型复杂程度，λ为非负系数；

第三步：根据结构风险最小化原则确定目标函数，使用 0-1 损失函数和L^1范数惩罚项构造如下目标函数：

$$R_{\text{srm}}(f)=\frac{1}{n}\sum_{k=1}^{n}I(f(X_i)\neq y_i)+\lambda\left\|\boldsymbol{w}^l\right\|_1\qquad(1-18)$$

其中$I(C)$在满足条件C时取 1，否则取 0。

第四步：采用适当优化计算方法对上述目标函数进行优化，解得$\boldsymbol{w}^1$，$\boldsymbol{w}^2$，$\boldsymbol{b}^2$，b_1^3的最优值$\boldsymbol{w}^{*1}$，$\boldsymbol{w}^{*2}$，$\boldsymbol{b}^{*2}$，b_1^{*3}，由此得到所求的分类映射规则：

$$f(\boldsymbol{X})=\varphi[\boldsymbol{w}^{*2^T}\varphi(\boldsymbol{w}^{*1}X+\boldsymbol{b}^{*2})+b^{*3}]\qquad(1-19)$$

目标函数的优化方法有很多，当目标函数较为简单时，可通过参数估计方式直接估计目标函数最小时所对应参数值，当目标函数较为复杂而无法直接估计参数时，则可通过迭代逼近方式逐渐优化目标函数并确定参数，这里不再赘述。

1.3.3 模型评估

机器学习的目的是提升模型性能以满足学习任务的需求，对于训练完成的模型需要对其进行性能评估以评判机器学习是否实现了目标或目标实现的程度。因此，对已训练模型进行有效的性能评估是机器学习必须面对和解决的一个基本问题。如前所述，机器学习模型性能的优劣主要取决于其泛化性能，模型评估的基本策略是设法估算出模型的泛化误差并通过泛化误差评估模型泛化性能。直接计算模型泛化误差通常是一件非常困难的事情，故具体的模型评估实施过程中一般使用测试误差近似代替泛化误差，即在测试样本集上计算模型误差并将其作为泛化误差的近似替代。下面具体讨论模型性能评估的基本方法。

要实现对模型的有效评估，首先必须确定能够对模型性能进行有效度量的指标。假设f是一个任意给定的分类模型，$T=\{(X_1,y_1),(X_2,y_2),\dots,(X_k,y_k)\}$是用于测试模型$f$的测试样本集，模型$f$对样本$X_i$的输出为$\hat{y}_i=f(X_i)$。如果$T$的$k$个样本中有$m$个样本的模型输出与其标

记值不一致，则可用 0-1 损失函数算得该模型的测试误差为：

$$R_{\text{test}}(f)=\frac{1}{k}\sum_{i=1}^{k}L(y_i,f(X_i))=\frac{m}{k}$$

即为模型输出值与标记不一致样本数占测试样本总数的比，通常亦称之为**错误率**e，即有：

$$e=\frac{1}{k}\sum_{i=1}^{k}I(f(X_i)\neq y_i) \qquad (1-20)$$

其中$I(C)$为条件函数，即$I(C)$的值在满足条件C时取 1，否则取 0。

正确率a则是分类正确的样本数与占测试样本总数的比，即有：

$$a=\frac{1}{k}\sum_{i=1}^{k}I(f(X_i)=y_i) \qquad (1-21)$$

正确率与错误率是面向分类任务模型最常用的两种性能度量标准。显然，对于任意一个分类模型，其正确率与错误率之和恒为 1。

对于很多分类问题，仅用正确率和错误率对分类模型评估是不全面的。例如在计算机辅助诊断应用领域，对于输出空间$Y=\{$患有癌症，不患癌症$\}$，假设只有1%的人患有癌症，若直接设置模型输出全部为不患癌症，则可将错误率控制在1%，但是这显然不是一个有效模型。因为该模型的主要目的是将尽可能患有癌症的示例找出来，将患有癌症的示例分类为不患癌症将会产生非常严重的后果。可用查准率和查全率来评价这类模型。

对于任意给定的一个二分类任务，通常会将其中某一类指定为正类，将另外一类指定为负类。令f为完成该二分类任务的分类模型，则可得如下四项基本统计指标：真正例数$\text{TP}(f)$、假正例数$\text{FP}(f)$、真反例数$\text{TN}(f)$和假反例数$\text{FN}(f)$。它们分别表示预测为正类且实际为正类的样例数、预测为正类且实际为负类的样例数、预测为负类且实际为负类的样例数和预测为负类且实际为正类的样例数。可根据这些指标算出模型f的查准率$P(f)$与查全率$W(f)$，具体计算公式如下：

$$P(f)=\frac{\text{TP}(f)}{\text{TP}(f)+FP(f)};\qquad W(f)=\frac{\text{TP}(f)}{\text{TP}(f)+\text{FN}(f)}$$

根据查全率与查准率计算公式不难看出$P(f)$和$W(f)$的取值有一定制约关系。例如，若想得到尽可能高的查准率$P(f)$，则应尽可能地减少假正例。最简单的办法是将正例概率很高的样本作为预测正例且将其它样本均预测为反例，但会提高假反例的数目，导致查全率$W(f)$降低。同理，若想提高查全率$W(f)$，则有可能降低查准率$P(f)$。

利用查全率与查准率对模型进行性能度量时，难免会出现某个模型查全率高但查准率低而另一个模型查全率低但查准率高的情况，此时难以使用查全率与查准率指标对这两个模型进行性能对比。为此引入一个名为F_1值的指标解决这个问题。所谓模型f的$F_1(f)$值，是指该模型查全率与查准率的调和均值，即有：

$$\frac{2}{F_1(f)}=\frac{1}{P(f)}+\frac{1}{W(f)}$$

由此可得：

$$F_1(f)=\frac{2\mathrm{TP}(f)}{2\mathrm{TP}(f)+\mathrm{FP}(f)+\mathrm{FN}(f)} \qquad (1-22)$$

$F_1(f)$值综合了查全率与查准率，当查全率与查准率都较高时$F_1(f)$值也较高。因此，如果模型的F_1值越大，则认为该模型的性能越优良。

除了F_1值之外，还可以通过一种名为**ROC曲线**的函数图像直观表示两个模型的性能对比。对于任意给定的分类模型f，其ROC曲线表示该模型真正例率$\mathrm{TPR}(f)$和假正例率$\mathrm{FPR}(f)$这两个变量之间的函数关系，其中$\mathrm{TPR}(f)$为函数因变量，表示真正例数$\mathrm{TP}(f)$占测试样本集中全部正例数的比（即查全率），$\mathrm{FPR}(f)$函数自变量，表示假正例数$\mathrm{FP}(f)$占测试样本集中全部假例数的比，即有：

$$\mathrm{TPR}(f)=\frac{\mathrm{TP}(f)}{\mathrm{TP}(f)+FN(f)};\qquad \mathrm{FPR}(f)=\frac{\mathrm{FP}(f)}{\mathrm{FP}(f)+\mathrm{TN}(f)}$$

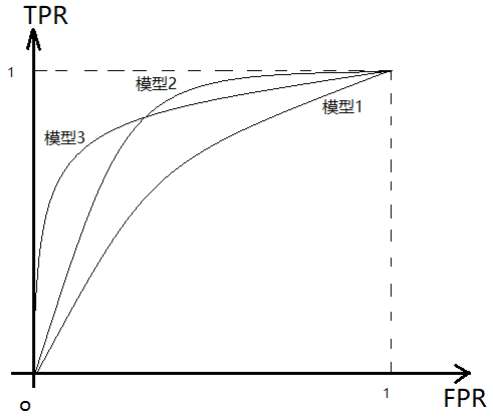


图 1-27 ROC曲线示意图

显然，如果分类模型的真正例率$\mathrm{TPR}(f)$接近于 1 且假正例率$\mathrm{FPR}(f)$接近于 0，则该模型具有比较好的分类性能。从图形上看，模型ROC曲线越靠近左上方则该模型的性能就越好。图 1-26 表示三个不同模型的ROC曲线，由ROC曲线分布特点可知模型 2 和模型 3 的性能优于模型 1。模型 2 和模型 3 的性能在不同情况下各有优劣，可通过ROC曲线下方面积指标进一步比较它们的平均性能。ROC曲线下方面积指标称为**AUC指标**。在一般情况下，模型所对应的AUC值越大，则该模型的平均性能就越好。

上述度量标准主要针对分类模型，下面进一步考察回归模型的性能度量标准。令f是任意给定的一个回归模型，$T=\{(X_1,y_1),(X_2,y_2),\ldots,(X_s,y_s)\}$是用于测试模型$f$的测试样本集合，$\hat{y}_i=f(X_i)$是模型$f$对于样本$X_i$的输出。若用平方损失函数度量单个样本的回归误差，则可用均方误差（MSE）度量模型f关于测试集T的测试误差，即有：

$$\mathrm{MSE}(f)=\frac{1}{s}\sum_{i=1}^{s}(f(X_i)-y_i)^2 \qquad (1-23)$$

另一个常用回归模型性能度量标准为**决定系数**。令$\bar{y}$为测试样本集T中所有样本标记的均值，则模型f在T上的决定系数R^2定义为：

$$R^2=1-\frac{\sum_{i=1}^{s}\big(y_i-f(X_i)\big)^2}{\sum_{i=1}^{s}(y_i-\bar{y})^2} \qquad (1-24)$$

决定系数表达式后项的分子为模型f关于测试集T的误差平方和，分母表示训练数据标记值的波动程度，二者相除可消除测试样本标记值的波动对模型性能的影响。对于给定的测试样本集T，如果模型f关于T的误差平方和越小，则其拟合效果就越好。故模型的决定系数越接近于 1，则该模型的性能就越好。

有了模型的性能度量指标，就可采用适当方法对模型进行性能评估。模型评估需要使用测试样本集估计模型的泛化误差。如果将参与模型训练的样本作为测试样本用于对模型性能的评价，则会降低对模型泛化性能估计的准确性。因此，测试样本一般不能用于对模型的训练，需将整个数据样本集划分为互斥的训练样本集和测试样本集。基于对样本集的不同划分策略，形成留出法、交叉验证法和自助法等多种模型评估方法。

对于样本数据集D，最简单的划分方法是直接从D中随机划分出部分数据组成训练样本集S，剩下部分作为测试样本集T用于估计模型的泛化误差，这种方法称之为**留出法**。为了保证留出法评估结果的可信度，通常要求训练集和测试集中的样本分布大致相同，从而避免划分不当带来偏差对模型评估结果的影响。

划分的随机性显然会给留出法的评估结果带来一定的波动，故仅用一次留出法的评估结果作为模型评估的最终结果是一种比较片面做法，一般需要多次使用留出法对模型进行评估，并将这些评估结果的均值作为最终的评估结果。在实际的模型训练与测试过程中，训练样本数通常占数据集总样本数的$2/3$至$4/5$，其余样本组成测试集。这意味着对于一次性留出法评估，数据集D中有部分样本未能参与到训练过程当中。因此，一次性留出法的评估结果与直接使用D中全部样本进行训练的真实模型性能存在一定差别。显然，当训练样本数占全部样本数的比例越高，这种差别就越小。

基于以上分析，K折交叉验证法将绝大多数样本用于训练。K折交叉验证法的基本思路为：首先将数据集D等分为K子集$D_i(i = 1,2, \dots, K)$，然后依次保留其中一个子集作为测试集T，而将其余$K-1$个子集合进行合并后作为训练集S。令$R_i(i = 1,2, \dots, K)$表示第i次模型评估结果，则各次评估结果R_i的均值就是K折交叉验证法对模型的最终评估结果。

显然，在使用K折交叉验证法进行模型评估的过程中，数据集D的每个样本仅参与了一次测试过程且参与了$K-1$次模型训练过程。这意味着每次训练的训练集绝大部分是重叠的，当K值越大时，参与模型训练的样本数越多，得到的模型性能越接近于使用数据集D中所有样本进行训练所得到的模型性能，但此时用于测试的样本数则较少，测试结果难以真实反映模型的实际的泛化性能。

留一法将每个样本单独作为一个划分，然后采样交叉验证的方式进行模型评估，是K折交叉验证法的一个特例，也是一种经典的交叉验证法。假设数据集D包含n个样本，留一法则将D进行n等分，依据交叉验证的规则分别进行n次模型训练和测试，每次有$n-1$个样本参与模型训练且1个样本参与模型测试。最终评估结果亦为各次评估结果的均值。在D中样本数较多的情况下，使用留一法进行模型评估的计算成本较高。

5×2交叉验证法是另一种经典的交叉验证法，该方法使用基数相等的训练样本集和测试样本集进行模型评估，主要包括对样本数据的随机等分和对折这两种操作。所谓随机等分，就是将整个样本数据集合D随机地切分成样本数目相等的两个子集合，并将其中一个作为训练样本集，另外一个作为测试样本集。对折则是将样本数目相等的训练样本集和测试样本集

进行性质对换，即将原来的训练样本集变成测试样本集、原来的测试训练样本集变成训练样本集。5×2交叉验证法对数据集合D进行五次随机等分和对折操作。令D_i^j为第i次随机等分的第j个样本集合，则第一次随机划分可得到D_1^1、D_1^2这两个样本子集，选择其中之一作为训练集，另一个作为测试集，由此得到了第一对训练集S_1和测试集T_1，对S_1和T_1进行对折操作便可得到第二对训练集S_2和测试集T_2，即有：

$$S_1 = D_1^1, T_1 = D_1^2;\quad S_2 = D_1^2, T_2 = D_1^1$$

由以上分析可知，5×2交叉验证法对数据集合D的一次随机等分和对折可以得到两对训练集和测试集，进行五次随机等分和对折则可得到十对训练集和测试集，表 1-10 给出了这十对训练集和测试集的具体产生过程。

表 1-10 $\mathbf{5\times2}$交叉验证法训练集和测试集

第 1 次随机等分	第 2 次随机等分	第 3 次随机等分	第 4 次随机等分	第 5 次随机等分
$S_1 = D_1^1$, $T_1 = D_1^2$	$S_3 = D_2^1$, $T_3 = D_2^2$	$S_5 = D_3^1$, $T_5 = D_3^2$	$S_7 = D_4^1$, $T_7 = D_4^2$	$S_9 = D_5^1$, $T_9 = D_5^2$
$S_2 = D_1^2$, $T_2 = D_1^1$	$S_4 = D_2^2$, $T_4 = D_2^1$	$S_6 = D_3^2$, $T_6 = D_3^1$	$S_8 = D_4^2$, $T_8 = D_4^1$	$S_{10} = D_5^2$, $T_{10} = D_5^1$

显然可进行更多次随机等分和对折操作获得更多的训练集和测试集，但次数过多的随机等分会使得样本子集之间的具有很强的相关性，无法提供新的有助于模型评估的信息，却徒增模型评估的成本。数据集D中样本数量较多的情况下，可以进行多次划分建立多组相关性较小的数据集。因此，留出法和交叉验证法通常比较适用于样本数量较多情形。

当D中样本数量较少时，可以采用一种名为自助法的方式构造训练集和测试集。自助法主要通过对D中样本进行可重复随机采样的方式构造训练集和测试集。具体地说，假设数据集D中包含n个样本，自助法对数据集D中样本进行n次有放回的采样，并将采样得到的样本作为训练样本生成一个含有n个样本的训练样本集S，所有未被得到的样本则作为测试样本构成测试集T。对于D中的任一样本，该样本在自助采样中未被采样的概率为：

$$\lim_{n\to\infty}(1-\frac{1}{n})^n = \frac{1}{e} \approx 0.368$$

因此，在数据集D中样本数足够大的情况下，测试样本数约占总样本数的36.8%。值得注意的是，通过自助法构造的训练集中样本数量与整个数据集的样本数量相同，但其中可能包含重复样本，因此所获得的训练集的样本分布与整个数据集的样本分布不同，因此，通常只有在样本量较小，难以对数据集进行有效划分时才使用自助法进行模型评估。

综上所述，在给定数据集D的情况下对机器学习模型进行模型评估，首先需要选择合适的性能度量指标对模型的性能进行度量，然后通过适当的模型评估方法计算出模型所对应的性能度量指标的具体取值，最后通过综合考察机器学习任务的具体特点和性能度量指标值判定所训练模型是否满足机器学习任务的需求。

1.4 习 题

（1）试描述人工智能、机器学习、深度学习的含义，并说明它们之间的关系。

（2）列举在实际生活中有关机器学习的具体应用实例及相关的机器学习方法。

（3）阐述过拟合和欠拟合现象的产生原因及常用的解决办法。

（4）假设现对消费者的消费行为进行分析，要找出某类商品定价 A 和其销量 B 之间的关系，这属于机器学习中的何种任务？此类任务有何特点？

（5）简述方差和偏差的概念，并说明二者的区别和联系。

（6）对于样本集$S=\{(2,1),(5,4),(3,3),(7,5),(8,9)\}$,模型$f(x)=4x+1$,试求模型$f(x)$在$S$上的整体误差$R_S(f)$。

（7）机器学习的发展历程有哪些阶段？试说明每个阶段机器学习方法的基本理论。

（8）如何训练一个感知机？该训练过程的理论依据是什么？

（9）简述专家系统的组成部分，并简要说明构建一个专家系统的过程。

（10）什么是特征空间？什么是特征向量？

（11）机器学习中为什么要进行特征提取？卷积神经网络的特征自动提取有何特点？

（12）结合具体实例来说明正确率、错误率和 ROC 曲线之间区别与联系。

（13）在简要介绍模型性能度量中真正例率、假正例率、查准率、查全率的概念，并讨论它们之间的区别与联系。

（14）要确定某基因X与某种疾病y之间是否存在联系，现对 260 人进行检查，得到体内是否含义该基因与是否患病之间的关系如表 1-11 所示，若χ^2统计量的阈值设为 7.2，$X=0$表示体内不含该基因，$X=1$表示含有该基因，$y=-1$表示不患病，$y=+1$表示患病。试根据表中数据确定该基因是否与该疾病存在某种联系。

表 1-11 是否含有该基因与是否患病关系表

	$y=-1$的人数	$y=+1$的人数	合计
$X=0$	99	53	152
$X=1$	41	67	108
合计	140	120	260

（15）数据集 D 中包含 100 个样本，其中正例 50 个，反例 50 个，若使用自助法进行模型评估，理论上训练集和测试集中各包含多少个样本？