# 第一章　绪论

**张皓**

https://haomood.github.io/homepage/

zhangh0214@gmail.com

## 摘要

本章简要介绍人工智能及机器学习的研究内容和发展历史, 以及未来可能的发展趋势. 最后介绍全书采用的主要符号和术语. 人工智能领域的六位图灵奖得主也在本章中得以简要提及.

本系列文章有以下特点: *(a).* 为了减轻读者的负担并能使尽可能多的读者从中收益, 本文试图尽可能少地使用数学知识, 只要求读者有基本的微积分、线性代数和概率论基础, 并在第一节对关键的数学知识进行回顾和介绍. *(b).* 本文不省略任何推导步骤, 适时补充背景知识, 力图使本节内容是自足的, 使机器学习的初学者也能理解本文内容. *(c).* 机器学习近年来发展极其迅速, 已成为一个非常广袤的领域. 本文无法涵盖机器学习领域的方方面面, 仅就一些关键的机器学习流派的方法进行介绍. *(d).* 为了帮助读者巩固本文内容, 或引导读者扩展相关知识, 文中穿插了许多问题, 并在最后一节进行问题的 "快问快答".

## 1　人工智能的发展历程

### 1.1　人工智能是什么

**定义 1** (人工智能 (artifical intelligence, AI))**.** 让机器完成那些让人来做则需要智能的科学.

什么是智能? 计算机在很多方面 (例如 IQ、记忆力和计算能力) 都超过了一般意义上的人类, 但是我们不认为这些计算机具有智能. 我们能够列举出若干的行为 (或表现) 是智能的, 但是无法定义什么是智能. 图灵 (Turing)[*]测试有助于躲避试图定义 "智能" 的哲学泥沼.

**定义 2** (图灵测试 (Turing test))**.** 如果一个人 (代号 $C$) 使用测试对象皆理解的语言去询问两个他不能看见的对象任意一串问题. 对象为一个正常思维的人 (代号 $B$) 和一个机器 (代号 $A$). 如果经过若干询问以后, $C$ 不能得出实质的区别来分辨 $A$ 与 $B$ 的不同, 则此机器 $A$ 通过图灵测试.

能否认为通过图灵测试就具有了智能? 图灵测试的本质是从表现评判智能. Searle 中文屋子实验说明即使通过测试, 也未必具有智能 (例如计算机下棋能力很强, 我们不认为其有智能). 另一方面, 图灵测试存在 "人为陷阱" (例如问大英百科全书第 1783 页第 178 行写的是什么), 计算机还需要知道人类的智能水平, 这已经超过了数学和计算机科学的研究范围. 总之, 通过图灵测试对于严肃的人工智能来说不是一个明智的研究和发展目标 [3] : (1). 即使能制造这样的机器, 还会认为这不是智能的. (2). 这个目标本身能否通过还有待哲学家进一步思考.

**定义 3** (中文屋子实验 (Chinese room argument))**.** 西尔勒博士 (扮演计算机中的 CPU) 在一个封闭的房子里, 有输入和输出缝隙与外部相通. 输入的是中文问题, 而他对中文一窍不通. 房子内有一本英语的指令手册 (相当于程序), 从中可以找到相应的规则. 他按照规则办事, 把作为答案的中文符号写在纸 (相当于存储器) 上, 并输出到屋子外面. 这样, 看起来他能处理输入的中

[*]计算机科学之父. 他的三大贡献是图灵机 (奠定计算机科学的基础)、破译德军 Enigma 密码机 (对第二次世界大战产生重要影响)、图灵测试.

文问题, 并给出正确答案 (如同一台计算机通过了图灵测试). 但是, 他实际上对那些问题毫无理解, 不理解其中的任何一个词.

人工智能的目标是什么? 人工智能的研究大致可分为两类: (1). 强人工智能 (general AI): 研制出达到甚至超越人类智慧水平的机器, 具有心智和意识、能根据自己的意图开展行动. (2). 弱人工智能 (weak AI): 借鉴人类的智能行为, 研制出更好的工具以减轻人类智力劳动, 让机器做事情时聪明一点. 国际主流人工智能学界所持的目标是弱人工智能. 人工智能技术现在所取得的进展和成功, 是缘于弱人工智能而不是强人工智能的研究. 弱人工智能研究的主要目的并不是人造智能, 而是受智能启发的计算 (intelligence-inspired computing). 事实上, 强人工智能还涉及科学研究的伦理问题, 这也是主流人工智能学界不往这个方向努力的原因之一. 强人工智能具有自主意识, 能力全面超越人类, 将不再是能被人类控制的工具, 无法保证它的 "利益" 与人类一致.

是否应该担心所谓的人工智能的奇点和人工智能威胁论? 现在虽然人工智能在图像识别、语音识别、棋类游戏等一些领域达到甚至超越人类水平, 但它们都是在考虑特定类型的智能行为, 它们能并且只能完成这些特定类型的智能行为, 不具有也无需考虑自主心智、独立意识、甚至情感. 现在宣传人工智能的奇点和人工智能威胁论的人中, 有人因为不了解、有人因为个人名望、而有人故意宣传. 不过, 现在的人工智能技术还不够强大和可靠 (参见第 2.3 节). 过早地将自主驾驶车辆、电网管理、自主武器系统等高风险决策交予计算机将存在隐患, 只要机器学习组件中存在一点错误, 人工智能技术的过早部署都将可能导致重大的生命损失.

## 1.2 人工智能发展的三个时期

1956 年夏美国达特茅斯会议标志人工智能学科诞生, 这个会议参加者包括 McCarthy[†]、Minsky[‡]、香农 (Shannon)[§]等. 随后, 人工智能的发展经历了三个时期 [17] , 并且中间穿插两次寒冬期.

**推理期**.

- 时期. 二十世纪五十年代至七十年代初.
- 思路. 只要赋予机器逻辑推理的能力, 机器就具有智能.
- 代表工作. Newell 和 Simon[¶]的 Logic Theorist 和 General Problem Solving. 例如, Logic Theorist 在 1963 年证明了 Whitehead 和 Russell 的 "Principia Mathematica" 的全部 52 条定理, 其中定理 2.85 甚至比证明得比原著更加巧妙.
- 局限. 仅具有逻辑推理能力远远实现不了人工智能.

**第一次寒冬**.

- 时期. 二十世纪七十年代中期至八十年代初.
- 原因. 人们对人工智能过于乐观, 但发现了一些推理期面临的障碍, 导致对人工智能的资金支持减少, 也影响了人们在这个领域研究的积极性.

**知识期**.

- 时期. 二十世纪七十年代中期至八十年代末.
- 思路. 要使机器具有智能, 必须设法使机器拥有知识.
- 代表工作. Feigenbaum[‖] 的 DENDRAL 专家系统.
- 局限. (1). 由人把知识总结出来再教给计算机相当困难. (2). 有的专家不愿意分享知识.

**第二次寒冬**.

- 时期. 二十世纪八十年代末至九十年代初.
- 原因. 专家系统虽然有用, 但它的应用领域过于狭窄, 而且更新和维护成本高. 日本第五代计算机计划失败.

**学习期**.

- 时期. 二十世纪九十年代至今.
- 思路. 由机器自己学习知识, 解决知识工程瓶颈.
- 代表工作. 从样例中学习 (广义的归纳学习).
- 局限. 无法学得泛化性能好且语义明确的概念, 今天的机器学习得到的大多数是黑箱.

人工智能会出现第三次寒冬吗? 在计算机科学的所有分支中, 人工智能可能是公众最感兴趣的. 但由于科普程度不够, 在每一次大发展之后, 社会上往往对其抱有不切实际的幻想. 当那些超前的设想在技术上遇到困难时, 过高的期望转变为普遍的质疑, 整个研究领域也

[†]因对人工智能的贡献获得 1971 年图灵奖. 被誉为人工智能之父.

[‡]因创立和促进人工智能领域获得 1969 年图灵奖.

[§]被誉为信息论之父.

[¶]两人因对人工智能、人类认知心理学和列表处理 (list processing) 的贡献同时获得 1975 年图灵奖. Simon 因决策理论的贡献获得 1978 年诺贝尔经济学奖.

[‖]和 Reddy 两人因研制大规模人工智能系统, 展示人工智能的实际价值和潜在商业应用同时获得 1994 年图灵奖.

随之遇冷. 前两次人工智能遇到寒冬的原因大致如此. 蒸汽机和电力技术一定程度上将人类从繁重的体力劳动中解放出来从而集中于智力劳动, 人工智能目标将人类从重复性强的简单智力劳动中解放出来, 从事更高层次的智力活动. 未来可能人工智能不像现在这么举世瞩目, 但这个趋势是不可逆转的.

**定义 4** (机器学习 (machine learning, ML))**.** 通过计算的手段, 利用经验 (数据) 来改善系统自身的性能 [6] .

*机器学习目前成为热潮的原因?* 现在人类社会的各个角落都积累了大量的数据, 亟需能有效地对数据进行分析利用的计算机算法. 机器学习恰好顺应了大时代的这个迫切需求. 机器学习本身作为 "突破知识工程瓶颈" 利器出现, 但发展到今天主要研究智能数据分析 (利用计算机分析数据) 的理论和算法, 其影响力和意义可能已经超过人工智能本身. 大数据时代的三大关键技术: 机器学习 (数据分析)、云计算 (数据处理)、众包 (数据标记).

*什么场景不适合机器学习?* 太 "简单" 或太 "困难" 的问题不适合用机器学习. 太 "简单" 的问题包括: (1). 可以手工计算出或利用计算机得出问题的解. 例如判断图中是否包含环. 太 "困难" 的问题包括: (1). 特征信息不充分. 例如重要特征信息没有获得. (2). 样本信息不充分. 例如仅有很少的数据样本. (3). 问题本身不存在潜在规律.

### 1.3 机器学习与相关领域

*机器学习和统计、统计学习的区别和联系?* 传统统计学往往醉心于理论的优美而忽略实际效用, 统计学的研究成果通常需要经由机器学习的研究来形成有效的学习算法. 机器学习是人工智能中的一个领域, 统计学习只是只是利用机器学习解决人工智能问题的一种手段. 另一方面, 统计学习中一部分研究者是统计学家, 他们关注的是如何把统计学中的理论和方法变成可以在计算机上有效实现的算法. 而并不关心这样的算法对人工智能中的什么问题有用.

**定义 5** (数据挖掘 (data mining))**.** 从海量数据中发现知识.

*机器学习和数据库的区别和联系?* 数据库为数据挖掘提供数据管理技术, 而机器学习为数据挖掘提供数据分析技术.

**定义 6** (模式识别 (pattern recognition))**.** 使用计算机算法来自动地发现数据中的规律, 并利用这些规律进行决策 [2] .

*机器学习和模式识别的区别和联系?* 传统机器学习很少关注数据采集过程, 而通过对数据进行数学上的假设和约束来设计学习算法. 模式识别关注从数据采集、特征提取/学习、模型学习、评估和部署的全过程, 其中机器学习为模式识别提供数据分析技术 [16] .

## 2 机器学习的发展历程

### 2.1 机器学习的三个流派

**符号主义** (symbolism).

- *时期.* 二十世纪八十年代至九十年代中期.
- *思路.* 符号知识表示.
- *代表工作.* 决策树 (以信息论为基础, 以信息熵最小化为目标, 直接模拟人类对概念进行判定的树形流程)、归纳逻辑程序设计 ILP (使用一阶逻辑即逻辑谓词进行知识表示, 通过修改和扩充逻辑表达式例如 Prolog 表达式完成对数据的归纳).
- *优点.* 符号主义可以学习产生明确的概念表示. 决策树简单易用. ILP 有强知识表示能力, 可以较容易地表达出复杂数据关系, 而且领域知识通常可方便地通过逻辑表达式进行描述, ILP 不仅可利用领域知识辅助学习, 还可以通过学习对领域知识进行精化和增强.
- *局限.* ILP 表示能力太强, 直接导致学习过程面临的假设空间太大、复杂度极高, 因此问题规模稍大就难以有效进行学习.

**统计学习** (statistical learning).

- *时期.* 二十世纪九十年代中期至二十一世纪初.
- *思路.*
- *代表工作.* 支持向量机、核方法、Valiant[**]等人开创的统计学习理论 [11] .
- *优点.* 有理论直接支撑.
- *局限.*

**连接主义** (connectionism).

- *时期.* 二十世纪五十年代中后至九十年代中期、二十一世纪初至今.

[**]因对计算学习理论的贡献获得 2010 年图灵奖.

- 思路. 神经网络.
- 代表工作. M-P 神经元模型 [5] 、误差逆传播 (error back-propagation, BP) 算法、深度学习 (deep learning).
- 优点. (1). 对使用者要求不高. (2). 模型复杂度高、容量大、学习能力强, 可以在很多现实问题 (尤其是设计语音、图像等复杂对象的应用) 上发挥作用.
- 局限. (1). 连接主义学习产生的是黑箱模型. (2). 缺乏严格的理论基础. (3). 学习过程设计大量超参数, 而参数设置缺乏指导, 主要靠手工调参. (4). 模型需要连续可微, 难以处理符号化、离散数据.

## 2.2 深度学习

人工神经网络和生物神经网络的联系和区别? 早期的人工神经网络设计确实受到生物神经网络的启发, 但有效的人工神经网络学习算法大多以数学证明为支撑. 可以将一个神经网络视为包含了许多参数的数学模型, 这个模型是若干个函数相互嵌套代入而得. 就像研制飞机需要的是物理中的空气动力学而不是鸟的生物学知识, 人工神经网络的研究更需要计算机科学和数学知识, 而不是认知科学和神经科学知识.

人工神经网络的可塑性-稳定性窘境是什么? 可塑性 (plasticity) 是指神经网络要有学习新知识的能力, 稳定性 (stability) 是指神经网络在学习新知识时要保持对旧知识的记忆.

人工神经网络发展的几起几落及原因?

- 第一次高潮. 二十世纪四十年代的 M-P 神经元模型、Hebb 学习律, 五十年代的感知机、Adaline.
- 第一次低谷. 1969 年 Minsky 与 Papert 在 "Perceptrons" 一书中指出, 单层神经网络无法解决非线性问题, 而多层网络的训练算法尚看不到希望. 这一论断使美国和苏联均停止了对神经网络研究的资助.
- 第二次高潮. 1983 年 Hopfield 利用神经网络在旅行商这个 NP 完全问题的求解上获得当时最好结果、Rumelhart 等人 [8] 重新发明了 Werbos 在 1974 年发明的 BP 算法 [12] .
- 第二次低谷. 二十世纪九十年代中期, 随着统计学习理论和支持向量机的兴起, 而神经网络学习的理论性质不够清楚、试错性强、在使用中充斥大量技巧 (trick), 这使得神经网络研究又陷入低谷.
- 第三次高潮. 2010 年前后, 随着计算能力的迅猛提升和大数据的涌现, 神经网络研究在深度学习的名义下又重新崛起.

事实上, 神经网络的几次高潮和数据储量和计算设备的发展有很大关系. 神经网络第一次高潮得益于二十世纪五十年代现代电子计算机的广泛应用, 第二次高潮得益于 Intel x86 系列微处理器与内存条技术的广泛应用, 而第三次高潮得益于图形计算单元 (graphics processing units, GPU) 的广泛应用. 相比之下, 神经网络是相对最容易利用新增计算能力的机器学习方法.

人工智能、机器学习、深度学习之间的关系. 机器学习是人工智能的一个分支, 而深度学习是机器学习的一个分支.

深度学习目前成为热潮的原因? (1). 大量的数据. 深度学习模型容量大、过拟合风险高. 大数据是最简单有效的缓解过拟合的方法. (2). 强力的计算设备 (例如 GPU). 能够训练如此复杂的模型. (3). 有效的算法 (例如使用修正线性函数 (rectified linear unit, ReLU) 替代 Sigmoid 激活函数、合适的参数初始化等). 事实上, 因果关系应该是反过来, 我们要用深度学习, 所以才会考虑上面这些因素.

深度学习为什么要深? 增加隐层数目是提高神经网络容量的一个简单方法. 虽然通过单纯增加神经元数目也可以增加模型复杂度, 但实际上加深网络 (增加隐层数目) 比加宽网络 (增加隐层神经元数目) 更有效. 这是因为加深网络不仅增加了拥有激活函数的神经元数目, 还增加了激活函数嵌套的层数. 在现实任务中如何设计隐层神经元个数仍是个未决问题, 通常靠试错法 (trial-by-error), 我们通常会用一个非常深的神经网络来保证足够的网络容量.

**引理 1** (人工神经网络的万有逼近定理 [4] ). *当隐层神经元个数足够多时, 单隐层前馈神经网络能以任意精度逼近任意复杂度的连续函数.*

深度学习有效的原因? 其表示学习 (representation learning) 的能力 [1] . 深度学习的多隐层堆叠、每层对上一层的输出进行处理的机制, 可以看做是对输入样本进行加工, 从而把初始的、与输出标记之间联系不太密切的输入表示, 转化成与输出目标联系更密切的表示, 之后用简单模型即可完成复杂的学习任务. 以往机器学习用于现实任务时, 描述样本的特征通常需由人类专家来设计, 这称为特征工程 (feature engineering). 特征的

好坏对泛化性能有至关重要的影响, 人类专家设计出好特征也并非易事. 特征学习则通过机器学习技术自身来产生好特征, 这使得机器学习向全自动数据分析又前进了一步. 简而言之, 深度学习能有效进行表示学习的原因在于: 逐层处理、特征变换、足够的模型复杂度.

端到端学习是否应该是我们追求的目标? 不应该. 是否要采用端到端学习 (end-to-end learning) 的一个重要考虑因素是: 有没有足够的数据对应端到端的过程, 以及我们有没有一些领域知识能够用于整个系统中的一些模块.

### 2.3 机器学习的未来趋势

**多样化的深度学习算法**.

深度学习的成功在于其表示学习能力, 而深度神经网络是目前几乎唯一行之有效的深度学习解决方案. 但是, 深度神经网络要求计算单元是连续可微的, 这样才能使用误差反向传播算法基于梯度进行学习. 但是, 现实任务中的数据并不都是实值的, 如何利用深度学习处理符号数据、离散数据, 是未来的一个研究方向. 例如, 套现欺诈检测问题中存在很多离散属性, 并且特征十分高维和稀疏 [14] . 目前深度神经网络通常将离散属性转化为连续值后进行使用, 但这个过程会额外地引入距离, 带来信息偏差. 图神经网络 (graph neuron networks, GNN) [13] 可以对具有图结构的数据进行分析. 图可以对不具有欧式距离的数据进行建模, 并可以表示目标之间的依赖关系.

**弱监督条件下的学习**.

目前, 机器学习 (尤其是深度学习) 取得的成功高度依赖大规模有标记数据, 这甚至催生出数据标记产业. 但是, 在许多现实任务中, 既缺乏大量有标记数据, 又难以通过无成本探索获得大量训练样本 (例如 AlphaGo Zero [9] 可以通过探索获得大量围棋对弈样本, 但围棋本身的胜负规则是很强的监督信息). 完全无监督学习还很渺茫, 因此弱监督学习是亟待研究的一个方向.

弱监督信息大致包括以下三类 [15] :

- 监督信息不完全. 例如医学图像中, 少部分得到了专家标注, 而大部分没有标记. 这种情况下可以进行半监督学习和主动学习.
- 监督信息不具体. 例如医学图像中, 某图像被标记出有病灶, 但是未具体标出病灶在哪. 这种情况下可以进行多示例学习.
- 监督信息不精确. 例如医学图像中, 专家由于疲劳、疏忽等原因出现标记错误. 这种情况下可以进行带噪声学习和众包学习.

**开放动态环境下的学习**.

现在的机器学习算法通常工作于封闭静态环境中, 需要假设数据分布 $\mathcal{D}$ 恒定、样本类别恒定、样本属性恒定、评价目标恒定等. 但是, 在许多现实任务中, 经常会遇到开放动态环境, 高风险应用. 此时, 一个根本的要求是学得模型需要有很高的稳健性 (robust).

**因果学习**.

现在的机器学习算法大多只能推断相关性 (correlation), 而不能得到因果 (casuality). Pearl[††]提出的贝叶斯网的动机是因果推断 [7] . 在推出因果关系之后, 我们可以干预结果. 例如, 我们发现游泳池的游泳人数和甜品店的冰激凌销量有很强的相关性, 但如果甜品店老板通过鼓励大家多去游泳池游泳以希望提升冰激凌销量显然是不可行的, 真正的因果关系在于温度, 气温高了之后游泳的人多了, 吃冰激凌的人也多了. 此外, 因果学习得到的模型具有很好的可解释性.

## 3 机器学习基本术语

### 3.1 基本术语与符号

全书采用的主要符号和术语如表 1 和表 3 所示, 在后续的章节中会陆续介绍这些术语和符号的含义. 其中, 指示函数定义为

$$\mathbb{I}(x) := \begin{cases} 1 & \text{若 } x \text{ 为真}; \\ 0 & \text{若 } x \text{ 为假}, \end{cases} \tag{1}$$

Sigmoid 函数定义为

$$\operatorname{sigm} x := \frac{1}{1+\exp(-x)}, \tag{2}$$

符号函数定义为

$$\operatorname{sign} x := \begin{cases} 1 & \text{若 } x > 0; \\ 0 & \text{若 } x = 0; \\ -1 & \text{若 } x < 0. \end{cases} \tag{3}$$

为了使书写更加简洁, 本系列文章定义如下算术优先级和结合性:

[††]因概率计算和因果推理对人工智能的贡献获得 2011 年图灵奖.

Table 1: 本书采用的主要符号和术语.

| 符号 | 术语/含义 | 定义/对应章节 |
|---|---|---|
| $\boldsymbol{x} \in \mathcal{X}$ | 示例 (instance)/样本 (sample)/特征向量 (feature vector) | 对一个对象或事件的描述 |
| $x_j \in \mathbb{R}$ | 属性 (attribute)/特征 (feature) | 对象或事件在某方面的表现或性质的事项 |
| $\mathcal{X} := \mathrm{span}(\{\boldsymbol{x}_i\}_{i=1}^m) \subseteq \mathbb{R}^d$ | 属性空间/样本空间/输入空间 | 所有属性张成的空间 |
| $y \in \mathcal{Y}$ | 标记 (label) | 示例结果的信息 |
| $\mathcal{Y} := \mathrm{span}(\{y_i\}_{i=1}^m) \in \mathbb{R}$ | 标记空间 | 所有标记张成的空间 |
| $(\boldsymbol{x}, y) \sim \mathcal{D}$ | 样例 (example) | 拥有标记信息的示例 |
| $D := \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$ | 数据集 (data set) | 数据记录的集合 |
| $\mathcal{D}$ | 数据分布 | 全书广泛使用 |
| $f\colon \mathcal{X} \to \mathcal{Y}$ | 真相 (ground-truth) | 数据的潜在规律 |
| $h \in \mathcal{H}$ | 模型 (model)/假设 (hypothesis)/学习器 (learner) | 执行学习算法从数据中学习/训练的结果 |
| $\mathcal{H}$ | 假设空间 | 所有假设组成的空间 |
| $a$ | 属性的取值 | 第 7 章: 决策树 |
| $b$ | 偏置 (bias) | 全书广泛使用 |
| $C$ | 类别数 | 全书广泛使用, 使用 $c$ 进行索引 |
| $d$ | 特征维度 | 全书广泛使用, 使用 $j$ 进行索引 |
| $e$ | 泛化误差 | 第 2 章: 学习理论 |
| $\hat{e}$ | 经验误差 | 第 2 章: 学习理论 |
| $\tilde{e}$ | 带权经验误差 | 第 6 章: 集成学习 |
| $H$ | 假设的集成 (ensemble) | 第 6 章: 集成学习 |
| $\mathcal{L}(\cdot)$ | 损失函数/目标函数、拉格朗日函数 | 第 4 章: 概率方法 |
| $\ell(\cdot)$ | 经验风险 | 第 4 章: 概率方法 |
| $m$ | 样本数 | 全书广泛使用, 使用 $i$ 进行索引 |
| $\boldsymbol{w}$ | 权重 | 全书广泛使用 |
| $\alpha$ | 不等式约束的拉格朗日乘子、集成加权权重 | 第 5 章: 支持向量机、第 6 章: 集成学习 |
| $\beta$ | 等式约束的拉格朗日乘子 | 第 5 章: 支持向量机 |
| $\gamma$ | 间隔 (margin) | 第 5 章: 支持向量机 |
| $\delta$ | 置信区间 | 第 2 章: 学习理论 |
| $\epsilon$ | 很小的数 | 第 2 章: 学习理论 |
| $\varepsilon$ | 噪声 (noise) | 第 2 章: 学习理论 |
| $\eta$ | 学习率 (learning rate) | 第 4 章: 概率方法 |
| $\boldsymbol{\theta}, \boldsymbol{\Theta}$ | 学习算法的参数 | 第 4 章: 概率方法 |
| $\kappa(\cdot,\cdot)$ | 核函数 | 第 5 章: 支持向量机 |
| $\lambda$ | 经验风险和结构风险的权衡系数 | 第 4 章: 概率方法 |
| $\mu, \boldsymbol{\mu}$ | 均值 | 全书广泛使用 |
| $\xi$ | 松弛变量 | 第 5 章: 支持向量机 |
| $\Pi(\cdot)$ | 增长函数 | 第 2 章: 学习理论 |
| $\sigma, \boldsymbol{\Sigma}$ | 标准差, 协方差矩阵 | 全书广泛使用 |
| $\tau$ | 阈值 | 第 6 章: 集成学习 |
| $\phi(\cdot)$ | 特征映射 | 第 4 章: 概率方法、第 5 章: 支持向量机 |
| $\Omega(\cdot)$ | 结构风险/正则化项 | 第 4 章: 概率方法 |

Table 2: 根据监督信息划分不同的学习任务类型.

| 任务 | 监督信息 | 举例 |
|---|---|---|
| 监督学习 (supervised learning) | 完全 | 回归 (regression)、分类 (classification)、结构预测 (structure prediction) |
| 半监督学习 (semi-supervised learning) | 部分 | 纯 (pure) 半监督学习、直推 (inductive) 半监督学习 |
| 无监督学习 (unsupervsied learning) | 未知 | 聚类 (clustering)、密度估计 (density estimation)、异常检测 (outlier detection) |
| 强化学习 (reinforcement learning) | 延迟、隐式 | 有模型学习 (model-base learning)、免模型学习 (model-free learning) |

Table 3: 其他全书广泛使用的符号.

| 符号 | 含义 |
|---|---|
| $x^\star$ | (优化得到的) 最优值 |
| $\bar{x}$ | $x$ 的均值 |
| $\hat{x}$ | $x$ 的估计值 |
| $\tilde{x}$ | $a$ 变换后的结果 |
| $\mathbb{I}(\cdot)$ | 指示函数 |
| $\log x$、$\lg x$ | 以 e、2 为底的对数函数 |
| $\operatorname{sigm} x$ | Sigmod 函数 |
| $\operatorname{sign} x$ | 符号函数 |
| $\mathbf{1}$、$\mathbf{0}$ | 全 1、0 向量 |
| $\boldsymbol{I}$ | 单位矩阵 |

Table 4: 根据处理数据方式划分不同的学习任务类型.

| 任务 | 处理数据方式 |
|---|---|
| 批量学习 (batch learning) | |
| 增量学习 (incremental learning) | |
| 在线学习 (online learning) | |
| 主动学习 (active learning) | |

- 最大最小化操作符的算术优先级低于加减. 例如 $\max\ ab+cd$ 表示 $\max(ab+cd)$.
- 函数的算术优先级介于乘除和加减之间. 例如 $\log xy+z$ 表示 $\log(xy)+z$.
- 求和连乘算术优先级介于函数和加减之间. 例如 $\sum x\log y+z$ 表示 $(\sum x\log y)+z$.
- 计算采用右结合性. 例如 $\log x\log y$ 表示 $\log(x\log y)$.

### 3.2 机器学习类型

按照不同的划分标准, 机器学习可以划分成不同的类型, 如表 2 和表 4 所示. 本书主要关注批量学习下的回归和分类任务.

## 4 快问快答

人工智能、机器学习、深度学习之间的关系是什么? 答案见上文.

深度学习模型容量大了之后如何应对过拟合? 主要包括三类方法.

- 更多的数据. 当收集数据很昂贵, 或者我们拿到的是第二手数据, 数据就这么多时, 我们从现有数据中扩充生成更多数据, 用生成的“伪造”数据当作更多的真实数据进行训练, 这称为数据扩充 (data augmentation). 以图像数据为例, 数据扩充的手段例如图像水平翻转、移动一定位置、旋转一定角度、或做一点色彩变化等. 这些操作通常都不会影响这幅图像对应的标记, 并且可以尝试这些操作的组合.
- 正则化. 最常见的是 $\ell_2$ 正则化. 这倾向于使网络的权值接近 0, 这会使前一层神经元对后一层神经元的影响降低, 使网络变得简单, 降低网络的有效大小, 降低网络的拟合能力. 此外随机失活 (dropout) [10] 也是一种常见正则化手段.
- 改变网络结构.

## References

[1] Y. Bengio, A. C. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. 4

[2] C. M. Bishop. *Pattern Recognition and Machine Learning, 5th Edition*. Information Science and Statistics. Springer, 2007. 3

[3] P. J. Hayes and K. M. Ford. Turing test considered harmful. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 972–977, 1995. 1

[4] K. Hornik, M. B. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. 4

[5] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133, 1943. 4

[6] T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997. 3

[7] J. Pearl and D. Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018. 5

[8] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(9):439–458, 1986. 4

[9] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354, 2017. 5

[10] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 7

[11] V. Vapnik. *Statistical Learning Theory*. Wiley, 1998. 3

[12] P. Werbos. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. PhD thesis, Harvard University, Cambridge, MA, 1974. 4

[13] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu. A comprehensive survey on graph neural networks. *CoRR*, abs/1901.00596, 2019. 5

[14] Y. Zhang, J. Zhou, W. Zheng, J. Feng, L. Li, Z. Liu, M. Li, Z. Zhang, C. Chen, X. Li, and Z. Zhou. Distributed deep forest and its application to automatic detection of cash-out fraud. *CoRR*, abs/1805.04234, 2018. 5

[15] Z.-H. Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2018. 5

[16] 吴建鑫. 模式识别. 机械工业出版社, 2019. 3

[17] 周志华. 机器学习. 清华大学出版社, 2016. 2