第零章 智能算法程序设计引言

智能是一种通过学习获得的能力

1.1 人脑智能与人工智能

近年来，随着各种智能产品和智能应用的普及，“**智能**”逐渐成为人们谈论的热词。可是，如果要认真的问一问“什么是智能”这个问题，既使是从事这方面研究的专家也觉得很难回答。

如果我们以人类为中心，相信人类的“智能”是唯一的“智能”，那么我们可以把智能简单的定义为“**人脑智能**”，从而模仿“**人脑智能**”的机器所具有的智能就可以相应定义为“**人工智能**”——人工建造的智能。我们且不说是否人类的“智能”就是唯一的“智能”，上面的定义还是没有正面回答“什么是智能”这个问题!

文献【8】给出了这样一个定义：智能是系统通过获取和加工信息而获得的**能力**。这个定义把智能定义为一种能力，而且这种能力可以通过获取和加工信息而获得，就如同人通过学习（学习就是获取和加工信息的过程）获得新的能力。能力有高有低，既然智能是一种能力，那么智能也有低级智能、高级智能之分。以“**人脑智能**”为例，刚出生的婴儿几乎是一张白纸，只具有一些最基本的低级智能（比如吃奶、哭闹），到 1 岁左右就能学会很多更高级的智能（比如走路、表达丰富的情绪、开始伢伢学语），到三岁左右开始记事（记忆能力），然后上学，这个过程一直持续到成年，不断具备更多更高级的智能（比如社会交往、逻辑分析、艺术创作）。相应的，“**人工智能**”有一个类似的典型例子——自动驾驶，工业界将其划分为从 L0 到 L5 这六个从低到高的级别。最低的 L0 级，完全人类驾驶，

机器可以提供周围环境的警报。最高的 L5 级就是完全自动驾驶，无需人类驾驶员。

如此看来，把智能定义为一种能力是合适的，也是目前具有最多共识的。既然智能是一种能力，那么很自然这种能力就可以通过“**学习**”来获得，这就解决了智能从何而来的关键问题——学习是智能的关键所在。如此，笔者给出智能的定义：**智能是一种通过学习获得的能力**。

那么，学习又是什么呢？

1.2 什么是学习？

按照上面文献【8】给出的定义，学习就是获取和加工信息的过程。这个定义不太令人满意，太宽泛，因为几乎所有过程（比如我们做一个决策，计算机操作系统调度一个进程）都可以视为获取和加工信息的过程，显然并不是所有这些过程都获得了一种能力。所以，我们需要从**是否获得能力**这个角度来定义“学习”，抓住学习的本质。

文献【9】给出了这样一个定义：如果一个系统能够通过执行某个过程改进它的性能，这就是学习。这个定义的好处是，突出了“改进”——并不是所有的过程都是学习，只有达到了“改进”的过程才是学习。为了与上面“智能是一种能力”相一致，笔者将此定义稍加修改：**如果一个系统能够通过执行某个过程提升它的能力，这就是学习**。

下一个问题，如何学习呢？

先看看我们人是如何学的。婴儿蹒跚学步，是自己在试错中学习。上学后，我们在老师的指导和帮助下，逐渐学会用符号（比如文字符号、数学符号）和逻

辑（比如恒等变换、反证法）进行推理和学习。而脑和神经科学家会说，人的学习就是人脑中巨量的神经元连接而成的网络的塑造过程——神经元的可塑性。“**人工智能**”的三个主要流派可以看作是对这三种学习方式的模仿。如图 1 所示，婴儿在试错中学会走路被称为强化学习方式，属于行为主义这个流派，其基本思路是：智能体通过与环境的交互进行学习。基于符号和逻辑的学习方式属于符号主义或逻辑主义流派，其基本思路是：智能可以形式化为符号和逻辑规则，进而表达和学习新的知识、获得新的能力。比如我们都熟悉的基于公理和定义推出定理。最后，连接主义流派对应脑和神经科学家的观点，用神经元构造神经网络，进而产生智能。当前红透学术界和工业界的**神经网络与深度学习**就是连接主义流派的代表作，其三位代表人物一起获得了 2018 年的计算机领域最高奖——ACM 图灵奖，可谓实至名归。

图 1 人工智能的三个主要流派（来自文献【8】）

1.3 人脑智能与人工智能的关系

前面我们把“**人工智能**”定义为对“**人脑智能**”的模仿。这种模仿可以是功能层面的（比如能够推导公式、能够自主行走），也可以是结构层面的（比如用人工

神经元构造人工神经网络)。回顾历史，Marr（文献【9】）认为我们不应该模仿结构，而应该模仿功能。他说，研究鸟如何飞行不是去研究鸟的羽毛是如何构成的，而应该是弄清楚其背后的空气动力学原理。Marr 的这个观点很长时间以来占据主导地位。文献【8】对 Marr 的观点提出了质疑，认为结构是功能的基础，结构的模仿是必要的。文献【8】指出，最早的飞机就是从结构上模仿鸟的飞行，而空气动力学原理的建立是几十年后的事情了。所以，类比过来，“**人工智能**”应该在结构和功能两个层面对“**人脑智能**”进行模仿。未来的人工智能应该是在结构上模仿脑，在功能上模仿和超越脑。

目前，欧盟的脑计划就是走的结构仿脑的路线。问题是，假设做到了结构仿脑，是否自然就能够在功能上做到模仿和超越脑呢？这里要画上一个大大的问号！为了回答这个问题，研究者们一直都在坚持不懈的做出艰苦的努力。

从人工智能的发展历史来看，结构仿脑和功能仿脑两条路线始终在并行展开，且呈现相互促进、相互融合的局面。1943 年，沃伦·麦卡洛克（Warren McCulloch）和沃尔特·皮茨（Walter Pitts）就提出了沿用至今的 **M-P 神经元模型**。这个模型是对生物神经元十分简化的模拟，其中既有结构上的模拟（输入对应树突、计算对应细胞核、输出对应突触），也有功能上的模拟（求和、非线性变换）。

1950 年代，在此基础上发展出了仅有两层的人工神经网络——感知机。虽然只有输入输出两层，只能解决线性问题，感知机提出的**学习算法**却是具有开创性的，是功能仿脑的典范。1980 到 1990 年代发展出了多层神经网络，其中卷积神经网络成功应用到了手写数字识别。理论方面，单隐层神经网络的万能逼近定理得到严格证明，尽管这个定理对如何设计神经网络并没有实际的指导作用。这一波的另外一个重大关键进展是基于梯度下降的反向传播算法的提出和应用。梯

度下降是基于分析数学（微积分）的一阶优化方法，反向传播算法也是基于分析数学（复合函数的链式求导规则），这都是功能仿脑。

2012 年以来深度神经网络迎来爆发式发展，在诸多评测数据集和实际应用中**超越**了人类。这一波的发展主要是大数据和大规模并行计算（比如 GPU）带来的，在结构仿脑层面只有一些小的细节上的进展（比如横向连接、反馈、短路），在算法和功能仿脑层面也没有根本上的进展。比较有**亮点**的是，人们开始关注深度神经网络的理论解释，数学家也加入其中，希望破解深度神经网络的黑盒特性，指导实际的网络设计。可以说，这一波的发展是因为算料（数据）、算力（GPU）、算法三个必备条件都已具备，水到渠成。这里面，结构仿脑（人工神经元和多层神经网络）和功能仿脑（线性和非线性运算、优化方法）相互促进、相互融合，实现了在结构上模仿脑、在功能上模仿和超越脑。

值得特别注意的是，在这个发展过程中，**类脑计算**的概念及其生态体系逐渐形成，简单的说就是要在新的非冯.洛伊曼的类脑计算系统上实现人工智能。比如，对生物神经元更精细更精确的建模，对大脑学习过程与机制的深入探索和模仿。也就是说，类脑计算在基本理念上**更加强调结构仿脑**，希望借此实现在功能上更好的模仿和超越脑（比如低功耗、高效率、高可靠）。类脑计算面临的挑战很多，硬件层面上要解决器件、系统的设计实现，软件层面上要解决学习算法的设计实现，应用上要能体现出类脑计算相比现有模式具有不可替代的优势。

路漫漫其修远兮，唯有上下而求索！

1.4 模仿脑的重要途径——统计机器学习

前面我们谈到了什么是学习，笔者给出了定义：**如果一个系统能够通过执行**

某个过程提升它的能力，这就是学习。我们谈到了三种学习方式，分别对应三种基本理念。不同的理念采用的数学方法有所不同，本课程重点关注非符号主义理念，采用概率与统计（包括信息论）方法。

实际上，大概从 1980 年代开始，一门称为机器学习或统计学习的学科悄然发展起来，我们不妨将其合称为**统计机器学习**。这个学科建立在概率与统计（包括信息论）方法基础之上，综合运用数值最优化、正则化等方法，与计算机科学、模式识别、数据挖掘、脑与神经科学等学科充分交叉。比如，概率与统计为我们提供建模的思想来源，数值最优化和正则化帮助我们找到最优解（全局或局部最优），计算机科学帮助我们设计、分析、实现学习算法，模式识别和数据挖掘提供应用领域，脑与神经科学为我们仿脑提供思想来源（比如神经网络与深度学习）。

简单说来，统计机器学习就是**基于数据构建概率统计模型，从而对数据进行预测与分析**。举个例子，假设我们手头有一堆数据（1 维或 2 维），我们要对这些数据的分布进行建模，并对新的数据进行预测。一种通用的方法是采用多个高斯分布来逼近，得到混合高斯分布。比如图 2，我们采用三个 1 维高斯分布的混合来逼近真实的 1 维三峰分布。再比如图 3，我们采用三个 2 维高斯分布的混合来逼近真实的 2 维三峰分布。学习的过程就是要找到三个高斯分布各自的均值、方差。均值、方差就是我们需要学习的参数。基于手头已有数据学习到均值和方差参数后，任给一个新的输入 x 值，我们就能给出其出现的概率值 p(x)。

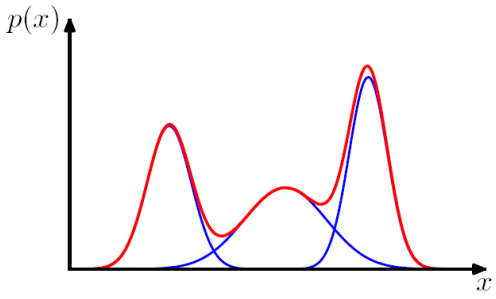

图 2　1 维混合高斯分布（来自文献【4】）

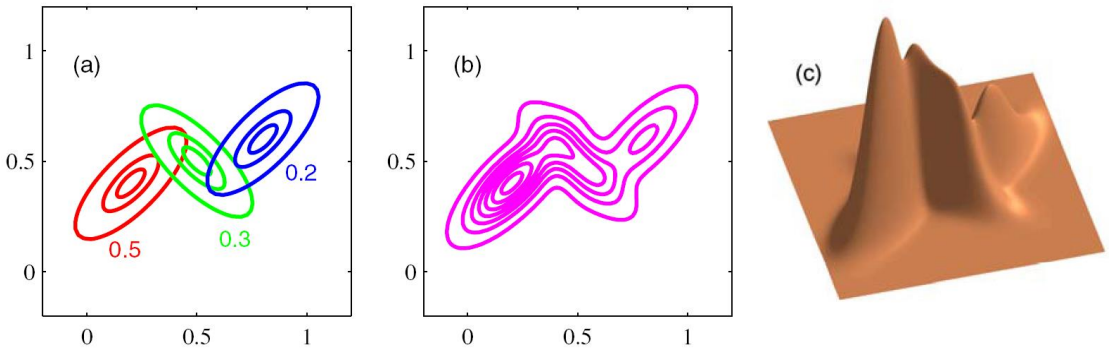

图 3　2 维混合高斯分布（来自文献【4】）

这就是统计机器学习的基本过程：收集数据，训练模型，测试模型。这个过程涉及到几个关键问题。第一个问题，如何设计模型，也就是假定数据服从什么分布。关于分布，我们有很多选择，比如上面的例子，我们假定服从混合高斯分布。通过对数据进行可视化，我们发现其呈现三个峰，于是我们进一步采用三个高斯分布进行混合，从而拟合数据的三个峰。这里读者会问，如果数据高于 2 维我们无法进行可视化怎么办呢？确实，一般情况数据都会具有**较高的维数**，这也是问题的难点所在，我们就需要采用各种办法（比如降维、流形假设等）来帮助我们理解数据本身，进而设计更合适的模型。

第二个问题，如何学习模型。不同的模型学习方法不同，同一个模型也可能采用多种学习方法，不同的学习方法在精度和效率上也有所不同，如此等等。比如上面的例子，学习三个高斯分布各自的均值、方差，最直观的方法是使得出现的数据概率最大——最大似然估计。但是，单纯的使得出现的数据概率最大对于未出现的新数据未必合适（**过拟合问题**），所以我们还有考虑了先验知识的贝叶斯方法、正则化方法。理解各种模型和各种学习方法，正确的选择和使用，这些是本课程要讲解的核心内容之一。

第三个问题，如何测试学习好的模型。同样一个模型，采用不同的测试标准，

结果可能迥异。比如有的模型**精度**高（预测对的高概率是对的），有的模型**召回**高（所有对的都能高概率预测对）。不同的应用关心的问题会有不同，需要有针对性的选择合适的测试标准。比如新冠核酸检测，更希望不能有漏检，也就是阳性不能被漏判，要求召回高。如此种种，需要具体问题具体分析，不能一概而论，不能搞“一刀切”。

这几个问题，包括没谈到的其它问题，他们相互之间实际上是相互关联、相互渗透的，学习和应用过程中要特别注意整体考虑、整体优化。比如高维问题，同样会影响到学习方法的设计与选择。再比如，测试标准同样会影响到模型的损失函数的设计与选择。