关于机器学习的领悟与反思

张志华
北京大学数学学院

近年来，人工智能的强势崛起，特别是去年AlphaGo和韩国九段棋手李世石的人机大战，让我们深刻地领略到了人工智能技术的巨大潜力。数据是载体，智能是目标，而机器学习是从数据通往智能的技术、方法途径。因此，机器学习是数据科学的核心，是现代人工智能的本质。

通俗地说，机器学习就是从数据中挖掘出有价值的信息。数据本身是无意识的，它不能自动呈现出有用的信息。怎样才能找出有价值的东西呢？第一步要给数据一个抽象的表示；接着基于表示进行建模；然后估计模型的参数，也就是计算；为了应对大规模的数据所带来的问题，我们还需要设计一些高效的实现手段，包括硬件层面和算法层面。统计是建模的主要工具和途径，而模型求解大多被定义为一个优化问题或后验抽样问题，具体地，频率派方法其实就是一个优化问题。而贝叶斯模型的计算则往往牵涉蒙特卡罗(Monte Carlo) 随机抽样方法。因此，机器学习是计算机科学和统计学的交叉学科。

借鉴计算机视觉理论创始人马尔 (Marr) 的关于计算机视觉的三级论定义，我把机器学习也分为三个层次：初级、中级和高级。初级阶段是数据获取以及特征的提取。中级阶段是数据处理与分析，它又包含三个方面：首先是应用问题导向，简单地说，它主要应用已有的模型和方法解决一些实际问题，这可以理解为数据挖掘；其次，根据应用问题的需要，提出和发展模型、方法和算法以及研究支撑它们的数学原理或理论基础等，这则是机器学习学科的核心内容；第三，通过推理达到某种智能。高级阶段是智能与认知，即实现智能的目标。数据挖掘和机器学习本质上是一样的，其区别是数据挖掘更接近于数据端，而机器学习则更接近于智能端。

统计与计算

卡内基梅隆大学统计系教授沃塞曼 (Larry Wasserman) 写了一本名字非常霸道的书：《统计学完全教程》(All of Statistics)。这本书的引言部分有一个关于统计学与机器学习非常耐人寻味的描述。沃塞曼认为，原来统计是在统计系，计算机是在计算机系，这两者是不相来往的，而且互相都不认同对方的价值。计算机学家认为那些统计理论没有用，不解决问题，而统计学家则认为计算机学家只是在“重新发明轮子”，没有新意。然而，他认为现在情况改变了，统计学家认识到计算机学家正在做出的贡献，而计算机学家也认识到统计的理论和方法论的普遍性意义。所以，沃塞曼写了这本书，可以说这是一本为统计学者写的计算机领域的书，为计算机学者写的统计领域的书。

现在大家达成了一个共识：如果你在用一个机器学习方法，而不懂其基础原理，这是一件非常可怕的事情。正是由于这个原因，目前学术界对深度学习还是有些疑虑的。尽管深度学习已经在实际应用中展示出其强大的能力，但其中的数学原理目前大家还不是太清楚。

让我们具体讨论计算机与统计学之间的关系。计算机学家通常具有强大的计算能力和解决问题的直觉，而统计学家擅长于理论分析和问题建模，因此，两者具有很好的互补性和合作空间。例如，数学家瓦普尼克 (Vapnik) 等人早在20 世纪60 年代就提出了支持向量机的理论，但直到计算机界于90 年代末发明了非常有效的求解算法，并随着后续大量实现代码的开源，支持向量机现在成为了分类算法的一个基准模型。再比如，核主成分分析(Kernel Principal Component Analysis, KPCA) 是由计算机学家提出的一个非线性降维方法，其实它等价于经典多维尺度分析(Multi- Dimensional Scaling, MDS)。而后者在统计界是很早就存在的，但如果没有计算机界重新发现，有些好的东西可能就被埋没了。

计算机界和统计界的通力合作，成就了机器学习从20世纪90年代中期到21世纪00年代中期的黄金发展时期，主要标志是学术界涌现出一批重要成果，比如，基于统计学习理论的支持向量机，随机森林和Boosting等集成分类方法，概率图模型，基于再生核理论的非线性数据分析与处理方法，非参数贝叶斯方法，基于正则化理论的稀疏学习模型及应用等等。这些成果奠定了统计机器学习的理论基础和框架。

机器学习现在已成为统计学的一个主流方向，许多著名大学的统计系纷纷从机器学习领域招聘教授，比如斯坦福大学统计系新进的两位助理教授来自机器学习专业。计算在统计领域已经变得越来越重要，传统多元统计分析是以矩阵分解为计算工具，现代高维统计则是以数值优化为计算工具。

最近有一本尚未出版的书《数据科学基础》(Foundation of Data Science)，作者之一霍普克洛夫特 (John Hopcroft) 是图灵奖得主。在这本书前言部分，提到了计算机科学的发展可以分为三个阶段：早期、中期和当今。早期就是让计算机可以运行起来，其重点在于开发程序语言、编译技术、操作系统，以及研究支撑它们的数学理论。中期是让计算机变得有用，变得高效，重点在于研究算法和数据结构。第三个阶段是让计算机具有更广泛的应用，发展重点从离散类数学转到概率和统计。我曾经和霍普克洛夫特教授交谈过几次，他认为计算机科学发展到今天，机器学习是核心。而且他正致力于机器学习和深度学习的研究和教学。

现在计算机界戏称机器学习为“全能学科”，它无所不在。除了有其自身的学科体系外，机器学习还有两个重要的辐射功能。一是为应用学科提供解决问题的方法与途径。对于一个应用学科来说，机器学习的目的就是把一些难懂的数学翻译成让工程师能够写出程序的伪代码。二是为一些传统学科，比如统计、理论计算机科学、运筹优化等找到新的研究问题。因此，大多数世界著名大学的计算机学科把机器学习或人工智能列为核心方向，扩大机器学习领域的教师规模，而且至少要保持两、三个机器学习研究方向具有一流竞争力。有些计算机专业有1/3甚至1/2 的研究生选修机器学习或人工智能。

然而，机器学习是一门应用学科，它需要在工业界发挥作用，能为他们解决实际问题。幸运的是，机器学习切实能被用来帮助工业界解决问题。特别是当下的热点，比如说深度学习、AlphaGo、无人驾驶汽车、人工智能助理等对工业界的巨大影响。当今 IT 的发展已从传统的微软模式转变到谷歌模式。传统的微软模式可以理解为制造业，而谷歌模式则是服务业。谷歌的搜索完全是免费服务社会，他们的搜索技术做得越来越极致，同时创造的财富也越来越丰厚。

财富蕴藏在数据中，而挖掘财富的核心技术则是机器学习，因此谷歌认为自

己是一家机器学习公司。深度学习作为当今最有活力的机器学习方向，在计算机视觉、自然语言理解、语音识别、智力游戏等领域的颠覆性成就，造就了一批新兴的创业公司。工业界对机器学习领域的人才有大量的需求。不仅仅需要代码能力强的工程师，也需要有数学建模和解决问题的科学家。

机器学习发展启示

机器学习的发展历程告诉我们：发展一个学科需要一个务实的态度。时髦的概念和名字无疑对学科的普及有一定的推动作用，但学科的根本还是所研究的问题、方法、技术和支撑的基础等，以及为社会产生的价值。

“机器学习”是个很酷的名字，简单地按照字面理解，它的目的是让机器能像人一样具有学习能力。但在其十年的黄金发展期，机器学习界并没有过多地炒作“智能”或者“认知”，而是关注于引入统计学等来建立学科的理论基础，面向数据分析与处理，以无监督学习和有监督学习为两大主要的研究问题，提出和开发了一系列模型、方法和计算算法等，切实地解决了工业界所面临的一些实际问题。近几年，因为大数据的驱动和计算能力的极大提升，一批面向机器学习的底层架构先后被开发出来。神经网络其实在20世纪80年代末或90年代初就被广泛研究，但后来沉寂了。近几年，基于深度学习的神经网络强势崛起，给工业界带来了深刻的变革和机遇。深度学习的成功不是源自脑科学或认知科学的进展，而是因为大数据的驱动和计算能力的极大提升。

机器学习的发展诠释了多学科交叉的重要性和必要性。然而这种交叉不是简单地彼此知道几个名词或概念就可以的，是需要真正的融会贯通。已故的布莱曼(Leo Breiman) 教授是统计机器学习的主要奠基人，他是众多统计学习方法的主要贡献者，比如Bagging、分类回归树(CART)、随机森林以及非负garrote稀疏模型等。布莱曼教授经历传奇，他从学术界转到工业界从事统计的实际应用十多年，然后又回到学术界。布莱曼是乔丹(Michael Jordan) 教授的伯乐，当初是他力主把乔丹从麻省理工学院引进到伯克利分校的。乔丹教授既是一流的计算机学家，又是一流的统计学家，而他的博士专业为心理学，他能够承担起建立统计机器学习的重任，为机器学习领域培养了一大批优秀的学者。

斯坦福大学教授弗莱德曼(Jerome Friedman) 早期从事物理学研究，但弗莱德曼是优化算法大师，他特别善于从优化的视角来研究统计方法，比如由此提出了多元自适应回归(Multivariate Adaptive Regression Splines, MARS) 和梯度推进机(Gradient Boosting Machines, GBM) 等经典机器学习算法。多伦多大学的辛顿教授是世界最著名的认知心理学家和计算机科学家。虽然他很早就成就斐然，在学术界久负盛名，但他依然始终活跃在一线，自己写代码。他提出的许多想法简单、可行又非常有效，被称为伟大的思想家。正是由于他的睿智和身体力行，深度学习技术迎来了革命性的突破。

总之，这些学者非常务实，从不提那些空洞无物的名词和框架。他们遵循自下而上的方式，从具体问题、模型、方法、算法等着手，一步一步实现系统化。

可以说机器学习是由学术界、工业界、创业界（或竞赛界）等合力造就的。学术界是引擎，工业界是驱动，创业界是活力和未来。学术界和工业界应该有各自的职责和分工。学术界的职责在于建立和发展机器学习学科，培养机器学习领

域的专门人才；而大项目、大工程更应该由市场来驱动，由工业界来实施和完成。

我国机器学习发展现状和出路

机器学习在我国得到了广泛的关注，也取得了一定的成绩，但我觉得大多数研究集中在数据挖掘层面，我国从事纯粹机器学习研究的学者屈指可数。在计算机学术界，理论、方法等基础性的研究没有得到足够重视，一些理论背景深厚的领域甚至被边缘化。而一些“过剩学科”、“夕阳学科”则聚集了大量的人力、财力，这使得我国在国际主流计算机领域中缺乏竞争力和影响力。

统计学在我国还是一个弱势学科，最近才被国家定为一级学科。我国统计学处于两个极端，一是它被当作数学的一个分支，主要研究概率论、随机过程以及数理统计理论等。二是它被划为经济学的分支，主要研究经济分析中的应用。而机器学习在统计学界还没有被深度地关注。统计学和计算机科学仍处于沃塞曼所说的“各自为战”阶段。

我国计算机学科的培养体系还基本停留在早期发展阶段，如今的学生从小就与计算机接触，他们的编程能力和国外学生相比没有任何劣势。但由于理论知识一直没有被充分重视，而且统计学的重要性没有被充分认识到，这些造成了学生的数学能力和国外著名高校相比差距很大。我国大多数大学计算机专业的本科生都开设了人工智能课程，研究生则开设了机器学习课程，但无论是深度、宽度还是知识结构都落后于学科的发展，不能适应时代的需要。因此，人才的培养无论是质量还是数量都无法满足工业界的迫切需求。

目前数据科学专业在我国得到了极大的关注，北京大学、复旦大学和中国人民大学等依托雄厚的统计学实力纷纷建立了数据科学专业或大数据研究院，并已经开始招收本科生和研究生。但是目前还没有一所大学开设机器学习专业。机器学习对其他应用或理论学科有辐射作用，也是连接两者的纽带。一方面它可以为理论端储备人才，另一方面可以结合不同领域问题，比如医疗数据、金融数据、图像视频数据等，为应用端输送人才。因此，我认为在计算机科学和应用数学本科专业中，增加机器学习的训练是必要的。

机器学习集技术、科学与艺术于一体，它有别于传统人工智能，是现代人工智能的核心。它牵涉到统计、优化、矩阵分析、理论计算机、编程、分布式计算等。因此，建议在已有的计算机专业本科生课程的基础上，适当加强概率、统计和矩阵分析等课程，下面是具体课程设置和相关教材的建议：

1．加强概率与统计的基础课程，建议采用莫里斯·德格鲁特(Morris H. DeGroot) 和马克·舍维什(Mark J. Schervish) 合著的第四版《概率论与数理统计》(Probability and Statistics) 为教材。

2．在线性代数课程里，加强矩阵分析的内容。教材建议使用吉尔伯特·斯特朗(Gilbert Strang) 的《线性代数导论》(Introduction to Linear Algebra)。吉尔伯特·斯特朗在麻省理工学院一直讲述线性代数，他的网上视频课程堪称经典。后续建议开设矩阵计算，采用特雷费森·劳埃德(Trefethen N. Lloyd) 和戴维·鲍(David Bau lll)

著作的《数值线性代数》(Numerical Linear Algebra) 为教科书。

3．开设机器学习课程。机器学习有许多经典的书籍，但大多不太适宜做本科生的教材。最近，麻省理工学院出版的约翰·凯莱赫(John D. Kelleher) 和布瑞恩·麦克·纳米(Brian Mac Namee) 等人著作的《机器学习基础之预测数据分析》(Fundamentals of Machine Learning for Predictive Data Analytics)，或者安得烈·韦伯 (Andrew R. Webb) 和基思·科普塞(Keith D. Copsey) 合著的第三版《统计模式识别》(Statistical Pattern Recognition) 比较适合作为本科生的教科书。同时建议课程设置实践环节，让学生尝试将机器学习方法应用到某些特定问题中。

此外，我建议设立以下课程作为本科计算机专业的提高课程或者荣誉课程。特别是，国内有些大学计算机专业设立了拔尖人才项目，我认为以下课程可以考虑列入该项目的培养计划中。事实上，上海交通大学ACM 班就开设了随机算法和统计机器学习等课程。

1．开设数值优化课程，建议参考教材乔治·诺塞达尔(Jorge Nocedal) 和史蒂芬·赖特(Stephen J. Wright) 的第二版《数值优化》(Numerical Optimization) ，或者开设数值分析，建议采用蒂莫西·索尔的《数值分析》(Numerical Analysis) 为教材。

2．加强算法课程，增加高级算法，比如随机算法，参考教材是迈克尔·米曾马克(Michael Mitzenmacher) 和伊莱·阿普法(Eli Upfal) 的《概率与计算： 随机算法与概率分析》(Probability and Computing: Randomized Algorithms and Probabilistic Analysis)。

3．在程序设计方面，增加或加强并行计算的内容。特别是在深度学习技术的执行中，通常需要GPU 加速，可以使用戴维·柯克 (David B. Kirk) 和胡文美(Wen-mei W. Hwu) 的教材 《大规模并行处理器编程实战》（第二版）(Programming Massively Parallel Processors: A Hands-on Approach, Second Edition) ；另外，还可以参考优达学城(Udacity) 上英伟达(Nvidia) 讲解CUDA 计算的公开课。

我认为以计算机科学为主导，联合统计和应用数学专业，开设机器学习研究生专业是值得考虑的。研究生专业应该围绕理论机器学习、概率与随机图模型、贝叶斯方法、大规模优化算法、深度学习等基础机器学习领域。建议开设理论机器学习、概率图模型、统计推断与贝叶斯分析、凸分析与优化、强化学习、信息论等课程。在附录我列出了一些相应书籍供参考。

结语

在 AlphaGo 和李世石九段对弈中，一个值得关注的细节是，代表 AlphaGo 方悬挂的是英国国旗。我们知道 AlphaGo 是由 deep mind 团队研发的，deep mind 是一家英国公司，但后来被 google 公司收购了。科学成果是世界人民共同拥有和分享的财富，但科学家则是有其民族情怀和归属感的。

位低不敢忘春秋大义，我深切地认为我国人工智能发展的根本出路在于教育。只有培养出一批批数理基础深厚、动手执行力极强，有真正融合交叉能力和国际视野的人才，我们才会有大作为。■

附录：参考书籍

[1] Shai Shalew-Shwartz and Shai Ben-David. Understanding Machine Learning: from Theory to Algorithms. Cambridge University Press. 2014

[2] George Casella and Roger L. Berger. Statistical Inference, second edition. The Wadsworth Group, 2002.

[3] Andrew Gelman et al. Bayesian Data Analysis, Third edition. CRC, 2014.

[4] Daphne Koller and Nir Friedman. Probabilistic Graphical Models: Principles and Techniques. MIT, 2009.

[5] Jonathan M. Borwein and Adrian S. Lewis. Convex Analysis and Nonlinear Optimization: Theory and Examples, second edition. Springer, 2006.

[6] Avrim Blum, John Hopcroft, and Ravindran Kannan. Foundation of Data Science. 2016.

[7] Richaerd S. Sutton and Andrew G. Barto. Reinforcement Learning: An Introduction. MIT, 2012.

[8] Thomas M. Cover and Joy A. Thomas. Elements of Information Theory. John Wiley & Sons, 2012.

本文是根据在统计之都微博发布的《机器学习： 统计与计算之恋》和中国计算机学会通讯发表的《机器学习的发展历程及启示》修订而成。

2017 年 1 月 9 日修订于北大静园六院