2.机器学习应用现状

在过去二十年中,**人类收集、存储、传输、处理数据的能力取得了飞速提升**,人类社会的各个角落都积累了大量数据，亟需能有效地对数据进行分析利用的计算机算法，而机器学习恰顺应了大时代的这个迫切需求，因此该学科领域很自然地取得巨大发展、受到广泛关注.

今天,在计算机科学的诸多分支学科领域中,无论是**多媒体、图形学,还是网络通信、软件工程,乃至体系结构、芯片设计，都能找到机器学习技术**的身影,尤其是在计算机视觉、自然语言处理等“计算机应用技术”领域，机器学习已成为最重要的技术进步源泉之一.

机器学习还为许多交叉学科提供了重要的技术支撑，例如，“生物信息学”试图利用信息技术来研究生命现象和规律，而基因组计划的实施和基因药物的美好愿景让人们为之心潮澎湃.**生物信息学研究涉及从“生命现象”到“规律发现”的整个过程**，其间必然包括数据获取、数据管理、数据分析、仿真实验等环节,而“数据分析”恰是机器学习技术的舞台,各种机器学习技术已经在这个舞台上大放异彩.

事实上，随着科学研究的基本手段从传统的“**理论+实验”**走向现在的“**理论+实验+计算**”,乃至出现“数据科学”这样的提法,机器学习的重要性日趋显著,因为“计算”的目的往往是数据分析,而数据科学的核心也恰是通过分析数据来获得价值.若要列出目前计算机科学技术中最活跃、最受瞩目的研究分支,那么机器学习必居其中。2001年，美国NASA-JPL的科学家在 Science 杂志上专门撰文Mjolsness and DeCoste,2001]指出，机器学习对科学研究的整个过程正起到越来越大的支撑作用,其进展对科技发展意义重大，2003年,DARPA启动PAL计划,将机器学习的重要性上升到美国国家安全的高度来考虑.众所周知,美国最尖端科技的研究通常是由NASA和DARPA推进的,而这两大机构不约而同地强调机器学习的重要性,其意义不言而喻.

2006年,卡耐基梅隆大学宣告成立世界上第一个“机器学习系”，机器学习领域奠基人之一T.Mitchell教授出任首任系主任，2012年3月,美国奥巴马政府启动“大数据研究与发展计划”，美国国家科学基金会旋即在加州大学伯克利分校启动加强计划，强调要深入研究和整合大数据时代的三大关键技术:机器学习、云计算、众包(crowdsourcing).显然,机器学习在大数据时代是必不可少的核心技术,道理很简单:**收集、存储、传输、管理大数据的目的,是为了“利用”大数据，而如果没有机器学习技术分析数据，则“利用”无从谈起**.

谈到对数据进行分析利用，很多人会想到“**数据挖掘**”(data mining)，这里简单探讨一下数据挖掘与机器学习的联系.数据挖掘领域在二十世纪九十年代形成,它受到很多学科领域的影响,其中数据库、机器学习、统计学无疑影响最大[Zhou,2003].**数据挖掘是从海量数据中发掘知识,这就必然涉及对“海量数据”的管理和分析**.大体来说,**数据库领域的研究为数据挖掘提供数据管理技术，而机器学习和统计学的研究为数据挖掘提供数据分析技术，由于统计学界的研究成果通常需要经由机器学习研究来形成有效的学习算法,之后再进入数据挖掘领域**,因此从这个意义上说，统计学主要是通过机器学习对数据挖掘发挥影响,而机器学习领域和数据库领域则是数据挖掘的两大支撑.

今天，机器学习已经与普通人的生活密切相关.例如在**天气预报、能源勘探、环境监测等方面,有效地利用机器学习技术对卫星和传感器发回的数据进行分析,是提高预报和检测准确性的重要途径;在商业营销中,有效地利用机器学习技术对销售数据、客户信息进行分析,不仅可帮助商家优化库存降低成本，还有助于针对用户群设计特殊营销策略**;……下面再举几例:

众所周知，谷歌、百度等互联网搜索引擎已开始改变人类的生活方式，例如很多人已习惯于在出行前通过互联网搜索来了解目的地信息、寻找合适的酒店、餐馆等.美国《新闻周刊》曾对谷歌有一句话评论:“**它使任何人离任何问题的答案间的距离变得只有点击一下鼠标这么远.**”显然,互联网搜索是通过分析网络上的数据来找到用户所需的信息,在这个过程中，用户查询是输入、搜索结果是输出，而要建立输入与输出之间的联系，内核必然需要机器学习技术，事实上,互联网搜索发展至今，机器学习技术的支撑居功至伟.到了今天,搜索的对象、内容日趋复杂,机器学习技术的影响更为明显，**例如在进行“图片搜索”时,无论谷歌还是百度都在使用最新潮的机器学习技术**.谷歌、百度、脸书、雅虎等公司纷纷成立专攻机器学习技术的研究团队,甚至直接以机器学习技术命名的研究院,充分体现出机器学习技术的发展和应用,甚至在一定程度上影响了互联网产业的走向，

再举一例.车祸是人类最凶险的杀手之一,全世界每年有上百万人丧生车轮,仅我国每年就有约十万人死于车祸.由计算机来实现自动汽车驾驶是一个理想的方案,**因为机器上路时可以确保不是新手驾驶、不会疲劳驾驶，更不会酒后驾驶，**而且还有重要的军事用途.美国在二十世纪八十年代就开始进行这方面研究.这里最大的困难是无法在汽车厂里事先把汽车上路后所会遇到的所有情况都考虑到、设计出处理规则并加以编程实现，**而只能根据上路时遇到的情况即时处理.若把车载传感器接收到的信息作为输入,把方向、刹车、油门的控制行为作为输出**，则这里的关键问题恰可抽象为一个机器学习任务.2004年3月,在美国DARPA组织的自动驾驶车比赛中,斯坦福大学机器学习专家 S.Thrun的小组研制的参赛车用6小时53分钟成功走完了132英里赛程获得冠军.比赛路段是在内华达州西南部的山区和沙漠中,路况相当复杂,在这样的路段上行车即使对经验丰富的人类司机来说也是一个挑战.S.Thrun 后来到谷歌领导自动驾驶车项目团队，值得一提的是，自动驾驶车在近几年取得了飞跃式发展,除谷歌外,通用、奥迪、大众、宝马等传统汽车公司均投入巨资进行研发，目前已开始有产品进入市场.2011年6月，美国内华达州议会通过法案，成为美国第一个认可自动驾驶车的州,此后,夏威夷州和佛罗里达州也先后通过类似法案.自动驾驶汽车可望在不久的将来出现在普通人的生活中，而机器学习技术则起到了“司机”作用.

机器学习技术甚至已影响到人类社会政治生活。**2012年美国大选期间**，奥巴马麾下有一支机器学习团队，他们对各类选情数据进行分析，为奥巴马提示下一步竞选行动.例**如他们使用机器学习技术分析社交网络数据,判断出在总统候选人第一次辩论之后哪些选民会倒戈,并根据分析的结果开发出个性化宣传策略，能为每位选民找出一个最有说服力的挽留理由**;他们基于机器学习模型的分析结果提示奥巴马应去何处开展拉票活动,有些建议甚至让专业竞选顾问大吃一惊，而结果表明去这些地方大有收获.总统选举需要大量金钱，机器学习技术在这方面发挥了奇效.例如，**机器学习模型分析出,某电影明星对某地区某年龄段的特定人群很有吸引力,而这个群体很愿意出高价与该明星及奥巴马共进晚餐……果然,这样一次筹资晚宴成功募集到1500万美元**;最终，借助机器学习模型，奥巴马筹到了创纪录的10亿美元竞选经费.机器学习技术不仅有助于竞选经费“开源”,还可帮助“节流”，例如机器学习模型通过对不同群体选民进行分析，建议购买了一些冷门节目的广告时段，而没有采用在昂贵的黄金时段购买广告的传统做法，使得广告资金效率相比2008年竞选提高了14%;……胜选后，《时代》周刊专门报道了这个被奥巴马称为“竞选核武器”、由半监督学习研究专家R.Ghani领导的团队.

值得一提的是,机器学习备受瞩目当然是由于它已成为智能数据分析技术的创新源泉，但机器学习研究还有另一个不可忽视的意义，即通过建立一些关于学习的计算模型来促进我们理解“人类如何学习”，例如,P.Kanerva在二十世纪八十年代中期提出SDM(Sparse Distributed Memory)模型[Kanerva,1988]时并没有刻意模仿脑生理结构，但后来神经科学的研究发现,SDM的稀疏编码机制在视觉、听觉、嗅觉功能的脑皮层中广泛存在，从而为理解脑的某些功能提供了一定的启发.自然科学研究的驱动力归结起来无外是人类对宇宙本源、万物本质、生命本性、自我本识的好奇,而“人类如何学习”无疑是一个有关自我本识的重大问题.从这个意义上说，机器学习不仅在信息科学中占有重要地位,还具有一定的自然科学探索色彩。