北大汇丰

PHBS FINANCIAL FRONTIER DIALOGUE

金融前沿对话

2019 年第 18 期　总第 29 期

PHBS HFRI
北京大学汇丰金融研究院

主办单位：北京大学汇丰金融研究院
院长：海闻
执行院长：巴曙松
秘书长：本力
编辑：都闻心（执行）曹明明 何清颖 方堉豪 朱伟豪

北京大学汇丰金融研究院简介

北京大学汇丰金融研究院（The HSBC Financial Research Institute at Peking University，缩写 HFRI）成立于 2008 年 12 月，研究院接受汇丰银行慈善基金会资助，致力于加强国内外著名高校、金融研究机构，以及知名金融学者之间的交流，构建开放的学术、政策交流平台，旨在提高中国金融理论与政策的研究水平，促进学术繁荣与发展，加强与政府决策部门的联系与合作，为政府决策提供参考意见，为国际金融体制改革和中国金融业的发展做出贡献。

北京大学汇丰金融研究院院长为北京大学校务委员会副主任、北京大学汇丰商学院院长海闻教授，执行院长为中国银行业协会首席经济学家、中国宏观经济学会副会长巴曙松教授。

机器学习应用于投资

【对话主持】

北京大学汇丰金融研究院执行院长、中国银行业协会首席经济学家、中国宏观经济学会副会长巴曙松

【特邀嘉宾】

MacKay Shields 董事总经理朱宁

一、机器学习投资模型的出发点

今天主要是从投资者的角度谈谈对机器学习的理解，在了解机器学习的背景，为什么要机器学习，机器学习投资模型的出发点是什么之前，先讲一个小故事。2000 年，巴菲特给哥伦比亚大学 MBA 的学生做了一次职业辅导，他说要做一个好的职业投资者，必须做足功课，每天读 500 页的报告。我们知道巴菲特的投资很重视基本面，投资方法论毋庸置疑，每天 500 页报告作为一个投资策略，有两个方面含义，一是书中自有颜如玉，他认为制造价值的策略就蕴含在这些报告里，二是报告数量要足够多达 500 页。以上两点实际上就是今天用机器学习做投资的主要出发点，首先数据包含着足够的信息，包含着创造价值的投资策略，其次是数据量足够大，巴菲特老先生很勤奋，每天 500 页恐怕是极限，但在今天看来也是远远不够的，SEC 10-K、10-Q、卖方报告不计其数，有必要使用计算机辅助。还有一点，巴菲特推荐

给 MBA 的材料从技术上讲都是传统量化投资理论难以驾驭的所谓非结构化数据，非结构化数据是指数据结构不完整、不规则、没有预定义的数据模型，不方便使用数据库或者二维逻辑来表现的数据，基本上让 quant 无法随便 touch，包括有格式的办公文档、文本图片、推特、图像视频等。非结构化数据的格式多种多样，标准也多样，在技术上非结构化数据的信息量比结构化数据信息更难以标准化和理解，这也是今天机器学习的主要对象。

谈到机器学习近十年来的迅速发展，数据形式的演进是主要因素，也是最深刻的经济背景。2010 年以前，非结构化数据还比较少，2010 年是个很重要的年份，为什么这一年非结构化数据开始几何式暴涨，主要原因是 4G 技术开始推广，在 2010 年 6 月 4 日，美国 spring 率先推出了 4G 手机，4G 比 3G 快了至少是十几、二十几倍，使得非结构化数据成为主流，在 2015 年以前，社交媒体几乎不存在，使用 Youtube、Facebook 的人非常少。可以想象，当全球进入 5G 模式之后，数据传输速度再次提高几十倍，甚至 100 倍，非结构化数据的传输将进一步得到飞升，可以预计 5G 将使得机器学习进入人们生活的每一个角落，这是不可避免要应对的。

机器学习是一门以数据为中心、多领域的交叉学科，涉及概率论、统计学、算法复杂度多门学科。方法论上主要是用归纳作为综合而不是演绎，用以往的经验和知识数据，特别是非结构化数据，重新组织已有的知识结构，使之不断改进自身的性能，进一步获取新的知识和

技能。它是当前的人工智能的核心，已经有了很多的应用。影响学习系统最重要的因素是环境向系统提供的信息，或者说数据。信息库里存放的那些数据，一般来说如果信息的质量比较高，在学习的部分就比较容易处理，如果提供的信息杂乱无章，比如非结构化数据，学习系统需要获得足够的数据之后，删除不必要的噪音。虽然机器学习本身对噪音比较 robust，但是学习部分的任务也会比较繁重，设计起来比较困难，所以在机器学习时大家都考虑到先进行 noise reduction。

上世纪 90 年代的时候，机器学习这个词就存在，但是并没有引起特别的注意，近十年才变得流行起来，这门学科起始于上世纪 50 年代，六七十年代沉寂了一段时间，1986 年神经网络再次兴起，成为高校的一门课程。经过过去几十年的发展，各种学习方法的应用范围不断扩大，一部分已经成为生活中我们经常使用的产品，比如目前主要的 email 系统里的 spam filter 多多少少都用到机器学习的技术。近十年来机器学习的主要方向是深度学习，深度学习是指训练大型的神经网络。深度学习的兴起，一方面得益于 4G 技术带来的数据爆炸，social media 迅速增长，另一方面是一些技术的突破，包括硬件、软件，使得深度学习得以在专家系统、人脸识别、企业的智能管理还有智能机器人的运动规划中得到广泛应用。

二、机器学习的局限性和应用

目前机器学习和金融有关的一些成功应用，主要就是信用卡的欺诈、保险灾害的损失评估，在投资领域还没有看到特别成功的应用，

机器学习用于投资的一些条件虽然已经具备，特别是数据已经足够，像无人驾驶汽车技术都是机器学习、深度学习专注的研究题目，但是使用自动交易业界花了好多功夫，至今也缺乏突破。从一个业界人士角度来看，这里既有机器学习技术本身的局限性，也有行业资源的分配问题。

首先是机器学习技术本身的局限性，住在北京的朋友可能去过地坛庙会，有一项节目叫叼旗，游客手里拿着一面小旗或者钢蹦，师傅放出一只麻雀飞过来把旗或者钢镚叼走，这种鸟一般都是在刚孵化的时候从鸟巢里偷出来驯养的。我曾好奇新孵化的雏鸟怎么会飞，其实鸟能飞是天性本能，根本不需要人教，不需要学习流体力学，训练无非就是经常要用食物激发鸟儿飞，或者带着鸟儿骑自行车让它感受速度，鸟的另外一个先天本能就是直觉，它知道什么时候上，什么时候下，什么是危险，能够主动避开障碍，进了死胡同掉头。但是这些在机器学习的身上，并不是 common sense，稍微一个小拐弯就得推倒设计重新做，鸟儿不存在这种情况。

所谓机器学习，我们期望用计算机将所有信息归纳成基本的模式，再用模式来推断未来。假设数据中蕴含的信息可以帮助产生问题的解决方案，但实际情况是怎样，机器并不知道，特别是金融数据，几年前的数据有没有用很难说。所以一只鸟儿学会飞翔经过几百万年的进化，才有很多直觉。鸟儿面临的挑战是发现所有问题的模式，然后用直觉来限制自己的猜测，但是计算机没有这种直觉，从计算机的角度

来说，模式识别中只有一个模式，要么是正确的，要么就是错误的，所以计算机在学习中并没有真正理解问题，更谈不上问问题。想象一下不会问问题，但是可以回答问题的系统，对于简单的不用做太多学习的工作可以应付，但是复杂的系统无法做到。

那么，有没有可能做成一种 Universal learner，它能够主动问问题，主动地学习？1996 年数学家 David Wolpert 证明不存在这样的数学学习，这就是著名的“no free lunch”定理，当一个学习器善于某种学习模式，就有它不善于学习的模式，没有学习器能够擅长所有的学习，也就是说机器学习很难达到全方位地代替人脑。即使将来的计算机或者系统有更多的内存、更强的速度，机器学习作为一种学习方式是不完备的。

机器学习技术的局限性，加上投资行为的复杂性，造成了机器学习在投资领域并不是那么好用，至今没有很成熟突出的应用。金融的分析属于所谓非实验性科学，无法进行对照实验，这意味着虽然存在大量的金融交易数据，但是我们没有方法通过设计实验来控制自变量的变化，没法通过重复性实验来检验提出的假设。比如机器学习发现了某种选股模型，这些数据结构得到的分析，大部分看起来实际上是一种所谓的欺骗性模式，尤其是对于样本外数据，我们很有可能被所谓的 involve data snooping 欺骗，从数据中发现的模式实际并不存在。Data snooping 存在于所有的非实验性研究中，尤其当我们把复杂的机器学习算法用于投资时，这种问题尤甚。因为复杂的非线性算法中包

含着很多参数，通过这些参数的配合，总能发现一些让人无法理解的可以获得超额收益的模式或模型，如果不能够认真理解它的 economic significance，所谓的 data snooping 就会使复杂的机器学习算法成为从历史数据中发现无效巧合的高效工具。

目前机器学习在投资领域的应用开发主要分成两块，第一是由 FinTech Community 开发周边的服务，FinTech Community 在机器学习技术方面是强项，有很多在微软、谷歌工作过的人，但对投资比较陌生，所以他们开发了一大批所谓的“high data velocity，low decision dimension”的应用，这都不是核心的投资服务。投资服务的核心业务是指怎么看宏观、怎么看微观、风险管理、资产配置等，这些目前还主要是华尔街来做。但是华尔街目前做机器学习的大部分都是 quant，这些人对陌生的前沿技术不是很灵光，比如我本来是研究神经网络，在华尔街干了二十几年，专业已经放下了，所以个人认为，目前一些 high profile 基于模型的深度学习策略基本上都是忽悠。

三、机器学习如何用于投资

机器学习主要分三类，每一类有一些小分枝，若干应用场景。首先是交易机器人，Chatbot 是对话机器人，比较著名的就是 Facebook 的 Chatbot。Trading Chatbot 是基于自然语言识别和机器学习技术的人机交互方案的统称，主要的目的是帮助交易员将杂乱无章的市场信息转换成一些精炼的信息、报告，并且筛选出某些投资机会。举例来说，有个小型的金融私有公司 ACE Cash Express，只有 3.15 亿美元的

短债（secured bond），另外还有 1.15 亿的抵押贷款，没有股票，7 月 15 日收盘后，这家公司发布了一个 tender offer，要在 7 月 24 日以 100 美元的价格收购所有债券，这支 bond 在 offer 公布前的价格是 90 美元，所以理论上 7 月 16 号只要这支债券的交易价格在 100 美元以下就有交易机会，买进之后在 7 月 24 号之前 consent tender offer 就可以赚取无风险差价。实际情况是 7 月 16 号的市场一开始没反应过来，直到尾盘某个券商以 96.75 美元的价格收购了一笔，很快以 97 块卖出，赚了两毛五的小 profit，17 日对冲基金才反应过来，追逐剩余的 offer。像这种小公司债券的流动性很差，大的基金一般都没有专人的 coverage，所以在券商普遍没有 balance sheet 的情况下，这种投资机会就留给了对冲基金和一些比较小的金融机构。这样的交易机会一直存在，可以轻易地被所谓的交易机器人用机器学习技术辨别出来，而且要用自然语言再加上一些逻辑。目前华尔街已经有很多公司在做这种交易的机器人，有的做得也很不错。

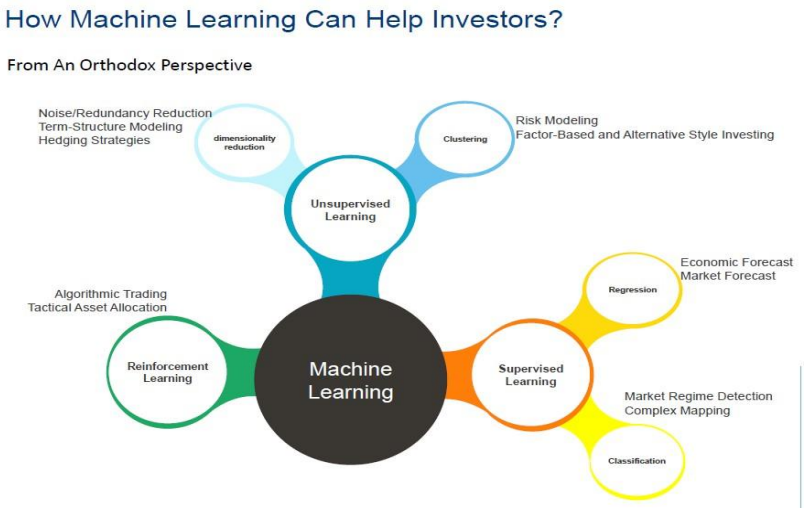

图 1 机器学习

上周在 TCF 的 FinTech 讲座上，一位 AI 大咖说他也在做机器学习，并且目标很高，要模拟交易员的行为、自动学习。当时我很受震动，感觉华尔街和 Fintech 在交易机器人的设计方向和理念上有很大不同，总结如图 2。

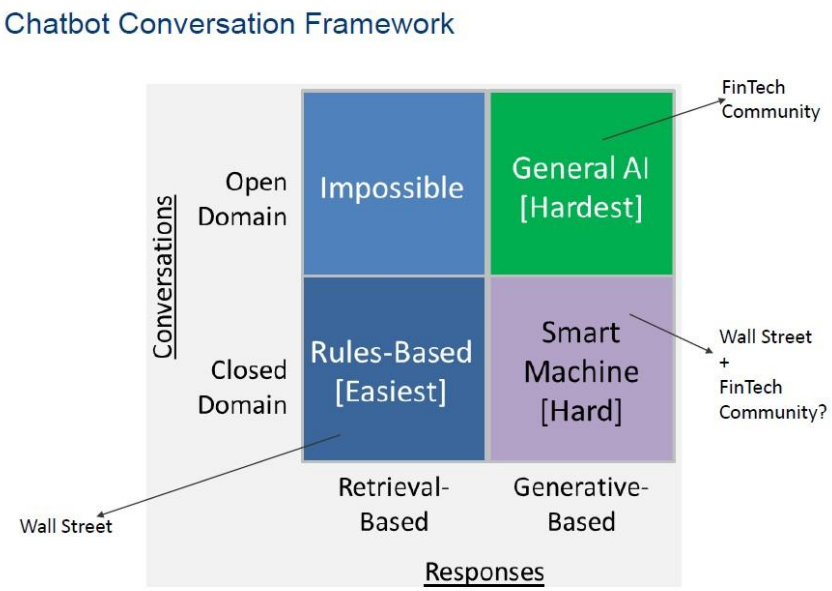

图 2 对话机器人交流模型

图 2 纵坐标是对专业知识信息的把握，即 domain knowledge，Open domain 是机器人可以理解的广谱信息，新的信息也懂，Close domain 是有限的领域，比如说企业债、银行、tender offer、call 等。横坐标是反馈，基本就是反映出来的策略，retrieval-based 是一个事先定义好的策略集，Generative-based 指的是动态产生反馈、策略。目前华尔街处在象限的左下角，华尔街的优势是有专业知识，有 domain knowledge，机器学习和开发能力稍弱，所以目前的 focus 是在同一个 Playbook 里提取反馈,但是有限的 domain 可能有无限的

scenario，有很多事情发生，所以最优的 response 不一定能够提前预定。 另一方面 Fintech community 占据右上角，他们的产品貌似高大上，但是缺乏投资经验，做出东西不一定实用，如果两方面合作开发出有强劲学习能力，并且有专业知识的实用产品，可能会更容易获得成功，目前还没有看到在象限右下角的产品。

开发基于机器学习的投资模型很难，目前成功的案例很少，都是信息比较充分的投资决定，逻辑比较简单直接的应用，前面的企业债交易机器人就是一例。怎样结合市场信息和宏观环境推导出可靠的结论，还要做很多研究，即使如此将来做出来的模型也肯定非常复杂，特别是在目前全球低增长低通胀的环境下，投资可能没有一定之规，要看市场风格的变化。

最后，做基于机器学习的投资模型，Framework 很重要，之前李强老师谈到目前模型组合的情景设定不合理，不适合机器学习发力，我个人比较同意，因为专业人士实际上是被强迫去预测波动性、收益相关性，这些是否有用都不清楚，所以我们可能需要崭新的 investment framework，不同于以往的现代资产组合理论或投资模型。作为专业人士去接受完全不同的理论可能有些困难，但是就像当年吴文俊先生平面几何证明采用完全不同的视角，有一天也许智能机器人投资系统能像 Alpha Go 横扫职业棋手那样取代华尔街的基金经理，虽然我认为这一天还比较遥远，但是更现实的途径是这些技术公司比如谷歌能够收购一些像高盛这样的企业，如果监管同意的话，那么许多的好故

事可能就会马上开始。目前，买方的机器学习方面，AQR 和 BlackRock 比较领先，卖方主要是 JP Morgan，其他公司基本上都在研究，美国主要的大学都设置了机器学习课程，有更多的人才、新的血液加入，这项新的学科将会得到长足的发展。

四、问答环节

Q1：您认为机器学习最有可能在哪些具体金融交易领域取得进展？

A1：流动性不太好的小众 asset class，比如企业债、muni bonds。卖方难以插手，信息扩散慢。

Q2：除了交易领域，在合规、反洗钱等领域有应用吗？

A2：这些早已有成功应用。关键是这些其实不是投资，基于机器学习的投资模型主要 cover macroeconomics，market forecast，asset allocation & trading，security selection。很遗憾至今还没有像人脸识别、spam filter 那样成功的基于机器学习的投资模型 。

Q3：老北京庙会的小鸟叼旗，鸟儿学飞像机器学习一样主要是本能，那鸟儿怎么会叼到正确的旗呢？

A3：叼旗本身是通过训练、食物鼓励。虽然是训鸟师傅的主要工作，但是技术上不是最复杂的部分。从工程的角度，教飞翔、躲避障碍才不可理解，靠的是鸟的 common sense，这其实才是开发机器模型的真正挑战。

Q4：机器学习难道不是用更多的数据和算法找出一些潜在规律

和关系吗？如果是这样，机器学习应该也可以用在流动性好的 asset class？

A4: 不是不可以，liquid asset 目前没有很成功的机器学习模型，有的多是 data snooping。

本文根据北京大学汇丰金融研究院执行院长巴曙松教授发起并主持的“全球市场与中国连线”第二百九十八期(2019年7月24日)内容整理而成，特邀嘉宾为 MacKay Shields 董事总经理朱宁先生。

朱宁（Michael Zhu Ning），CFA，Ph.D。朱宁先生于 2019 年 4 月加入 MacKay Shields 担任董事总经理。此前，他曾担任 Phase Capital 的首席投资官。在 2016 年 11 月加入 Phase 之前，他是 First Eagle 的多资产绝对回报和尾部对冲策略的投资组合经理。在 2013 年 4 月加入公司之前，朱宁先生是 Alliance Bernstein 的 Absolute Return Group 的定量信贷研究主管和研究主管，负责管理 Enhanced Alpha Global Macro，Tail Hedge，Unconstrained Bond 策略和信贷产品。在 2004 年加入 Alliance Bernstein 之前，他曾在花旗集团担任高级研究分析师 7 年，在信贷、利率和货币交易策略的研究，开发和管理方面拥有扎实的专业知识。此外，朱宁先生是福特汉姆大学（Fordham University）的兼职教授，教授量化金融和机器学习。他在牛津大学获得博士学位，并持有特许金融分析师（CFA）的称号。

【免责声明】

“全球市场与中国连线”为中国与全球市场间内部专业高端金融交流平台。本期报告由巴曙松教授和居姗博士共同整理，未经嘉宾本人审阅，文中观点仅代表嘉宾个人观点，不代表任何机构的意见，也不构成投资建议。

PHBS HFRI

北京大学汇丰金融研究院