高等学校金融学专业系列教材

The Advanced Course of Programming Trading

程序化交易高级教程

——机器学习在量化交易中的应用

（国信iQuant量化平台）

主编　陈学彬

高等教育出版社

货币银行学（第四版）（普通高等教育“十一五”国家级规划教材、国家精品课程教材、第六届高等教育国家级教学成果奖、教育部推荐教材、全国普通高等学校优秀教材）	戴国强
货币银行学习题精编	曹　啸 戴国强
金融学（第四版）（“十二五”普通高等教育本科国家级规划教材、普通高等教育“十一五”国家级规划教材、上海市优秀教材）	陈学彬
金融学简明教程	陈学彬
金融学（第二版）（教育科学国家规划课题研究成果）	马小南
金融学	李文哲
金融学	陈　平 何太胜
互联网金融	陈中放 胡军辉
金融投资工具比较与应用	黄文平 陈中放
企业融资模式与应用	李　忠 陈中放
程序化交易高级教程——机器学习在量化交易中的应用	陈学彬
程序化交易初级教程	陈学彬
国际金融	闫　屹
国际金融（省级精品课程教材）	闫　福
证券投资学（第五版）（教育部推荐教材）	霍文文
证券投资学简明教程（第二版）	霍文文
证券投资学（第二版）	魏建国
商业银行经营管理（第二版）	黄亚钧
商业银行经营管理（第二版）（普通高等教育“十一五”国家级规划教材）	吴念鲁
投资银行理论与实务（第二版）（教育部推荐教材）	黄亚钧
金融工程学（第二版）（国家精品课程教材、上海市优秀教材）	吴冲锋

ISBN 978-7-04-051278-6
9 787040 512786 >

定价：39.00元

高等学校金融学专业系列教材

CHENGXUHUA JIAOYI GAOJI JIAOCHENG

程序化交易高级教程

——机器学习在量化交易中的应用

（国信iQuant量化平台）

主编　陈学彬

高等教育出版社·北京

内容提要

本书是高等学校金融学专业系列教材之一。本书主要内容包括：导论、机器学习基础、Python 编程基础、基于 Python 的机器学习软件包、国信 iQuant 量化交易平台、交易策略学习模型的数据准备、线性回归估值选股模型、逻辑回归收益率预测选股模型、决策树分类择时模型、朴素贝叶斯分类择时模型、支持向量机分类择时模型、K 均值聚类分析选股模型、Apriori 股票关联分析模型、BP 神经网络择时模型、循环神经网络择时模型、长短期记忆择时交易模型、卷积神经网络择时交易模型、结语。

本书由浅入深，结合具体案例，将机器学习的理论模型应用于程序化交易，强调程序化交易策略的实用性。本书适合作为高等学校金融学专业相关课程教材，也可作为程序化交易的深化读物。

图书在版编目(CIP)数据

程序化交易高级教程：机器学习在量化交易中的应用：国信 iQuant 量化平台 / 陈学彬主编. —北京：高等教育出版社，2019.1

ISBN 978-7-04-051278-6

Ⅰ.①程… Ⅱ.①陈… Ⅲ.①期货交易—应用软件—高等学校—教材 Ⅳ.①F713.35—39

中国版本图书馆 CIP 数据核字（2019）第 011415 号

策划编辑 熊柏根 **责任编辑** 熊柏根 刘自挥 **封面设计** 张文豪 **责任印制** 高忠富

出版发行	高等教育出版社	**网　　址**	http://www.hep.edu.cn
社　　址	北京市西城区德外大街 4 号		http://www.hep.com.cn
邮政编码	100120		http://www.hep.com.cn/shanghai
印　　刷	上海华教印务有限公司	**网上订购**	http://www.hepmall.com.cn
开　　本	787mm × 1092mm　1/16		http://www.hepmall.com
印　　张	18.5		http://www.hepmall.cn
字　　数	433 千字	**版　　次**	2019 年 1 月第 1 版
购书热线	010-58581118	**印　　次**	2019 年 1 月第 1 次印刷
咨询电话	400-810-0598	**定　　价**	39.00 元

本书如有缺页、倒页、脱页等质量问题，请到所购图书销售部门联系调换

物 料 号　51278-00

前　言

在计算机、互联网、大数据和人工智能技术迅速发展的今天，金融市场决策和交易技术也在快速发展。借助于大数据、计算机高速度的处理能力和各种统计分析方法及优化决策模型，对海量高频数据进行处理和信息挖掘，从而进行投资决策的量化交易(quantitative trading)，以及利用计算机的高速处理能力和互联网络的高速传输能力及优化算法对交易指令进行优化交易的算法交易(algorithmic trading)都得到了极为迅速的发展。而将投资决策和交易执行优化整合在一起并高频率地进行交易，以捕获各种转瞬即逝的投资盈利机会的高频交易(high-frequency trading)，更是将这种计算机系统交易技术发挥到了极致。程序化交易是指利用计算机系统程序进行的自动化交易。[①]各种新的程序化交易方法、策略正在不断地大量涌现和更新。显然，程序化交易正在成为全球金融市场交易的主流方式和发展趋势。特别是近年迅速发展的人工智能技术，通过机器学习对金融大数据处理，并自我学习、自我完善而形成的智能交易系统正在将金融程序化交易的自动化交易推向更加高级的智能化交易阶段。

程序化交易之所以能够迅速地发展，主要原因不仅在于其具备有效地利用计算机和网络技术处理海量数据的能力和对市场信息快捷、高速的应对能力，更重要的是它可以有效地克服人工交易时人类恐惧、贪婪的天性和犹豫不决、感情冲动等人性弱点，坚定地、严格地按照规则和程序进行交易决策并执行。这对于有效地把握金融市场投资机会和控制风险都是极为重要的。当然，传统的程序化交易也有其明显的缺陷：程序化交易系统是由人开发的，它对市场的适用性和稳健性，取决于开发人员对市场规律的认识和在交易系统上的实现；而金融市场又是一个极其复杂多变的市场，人的认知能力的有限性和市场的多变性可能制约交易系统的适用性和稳健性。此外，大量的程序化交易系统的批量快速反应在市场剧烈波动时引发的羊群效应可能导致金融市场系统短期出现崩溃。程序化交易系统故障对金融市场带来冲击，以及给投资者带来巨大损失屡见不鲜。但是，程序化交易的缺陷并不能阻止其快速地发展，反而吸引了更多的机构和个人投入程序化交易策略系统的研究、开发和应用之中。它将有利于克服程序化交易系统开发人员认知能力的有限性和金融市场多变性的矛盾。而正在迅速发展的具有自动学习、自我完善功能的机器学习技术在交易系统中的引入，将推动程序化交易从简单地执行开发人员制定的交易策略向自动总结市场规律，不断适应市场变化的智能化方向发展。像智能机器人和无人驾驶

① 不管是量化交易、算法交易、高频交易还是程序化交易，都是利用计算机程序进行的自动交易，但它们各自的重点又有所差异。

汽车在它们的应用领域将逐步取代人的操作一样，智能交易系统必将在未来逐步地取代金融市场大量的人工交易。未来金融市场的竞争将更多的是智能交易系统之间的竞争。

我国的金融市场虽然起步较晚，但是，近十几年随着计算机技术和网络技术的快速发展，各种计算机投资分析系统和交易系统大量涌现。其中大多数系统的决策和交易仍然互相分离，且主要依靠人工进行，个人投资者开发和使用程序化交易策略系统的寥寥无几。近几年，机构投资者开发使用程序化交易策略系统的逐步增多，但大多数还是由海外回来的人员推动。我国自己培养的程序化交易策略系统开发人员的缺乏和普通投资者对程序化交易策略系统开发和使用知识的缺乏成为制约我国程序化交易发展的重要障碍。特别是在2015年的股市危机中，一些机构和个人投资者对配资系统和高频交易的滥用，加上许多人对程序化交易功能的错误理解，使得程序化交易被视为股市危机的罪魁祸首而受到限制，从而严重制约其发展。但是，科技进步的浪潮是任何力量也阻挡不住的，量化交易、程序化交易、智能交易是全球金融交易的发展趋势，只有尽快融入这个发展趋势，才能不断前行。

尽管程序化交易在国际金融市场交易中已经获得充分发展并正在向智能交易阶段迈进，但国内高校金融专业程序化交易人才培养工作仍然显著落后于市场的需求。满足专业人才培养需要的、较为系统深入讨论程序化交易策略系统的开发和使用的理论和应用书籍仍然较为缺乏。

作为专业知识来讲，程序化交易既是一门理论性很强的学科，需要很多复杂而高深的理论支持；更是一门应用性很强的学科，需要将各种复杂的理论方法应用于实际的市场交易决策和执行之中，并直接接受市场运行的检验。程序化交易的学习和应用都需要与市场实际紧密结合。一些量化分析软件虽然提供了许多十分有用的量化投资分析工具，但由于它们缺乏与市场直接连接获取实时数据和下达交易指令的接口而使其应用受限。而要自己开发一个专用的优秀交易平台并非易事，只有大型金融机构能够做到。这不仅是个人投资者不可及，即使是一些颇具规模的中小型金融专业投资机构也无法做到。能较好解决这些问题的公用程序化交易平台应运而生，并受到许多个人投资者和中小型专业投资机构的欢迎。程序化交易课程的学习正好建立在这些交易平台之上，使学生的学习和实训都与金融市场的交易实际紧密结合。

我在前几年为复旦大学金融学专业研究生开设程序化交易课程的基础上，2015年编写出版了两本关于程序化交易的教材《程序化交易》和《期权策略程序化交易》，对此领域的教材和教学进行了一些初步的尝试和探索。2017年在国信证券的大力支持下，进一步编写出版了以国信TradeStation平台为基础的《程序化交易初级教程》和《程序化交易中级教程》。该系列教材出版以来，受到一些高校和业界人士的欢迎，为推动我国程序化交易人才的培养尽了绵薄之力。随着程序化交易技术的发展，程序化交易平台在不断完善和发展，程序化交易的应用领域也在逐步拓展。几年前，我国程序化交易主要集中在期货交易领域，而近几年已经开始逐步向股票、债券、基金和期权等领域拓展。特别是随着我

国金融市场的全方位对外开放，程序化交易在我国金融市场交易的全方位使用是其发展的必然趋势。为了适应这种发展趋势的需要，有必要以具有更强功能的、能够覆盖更多市场领域的程序化交易平台为基础编写教材和培养学生。

由于程序化交易技术涉及的专业知识较多，不可能在一本书里全部介绍和讨论，学生也不可能在一个学期里全面深入地学习和掌握。较好的选择是划分为初级、中级和高级系列教程，供入门者、进阶者和深化者学习讨论使用。由于面向对象编程对于初学者具有一定的难度，因此，初级教程主要介绍 TradeStation 量化平台中面向用户的 EasyLanguage 的基础功能和应用，而将面向对象编程的组件带来的拓展功能放在中级教程中介绍和讨论。对于将机器学习、人工智能方法引入程序化交易的方法则放在高级教程中介绍和讨论。

本书是该系列教材中的高级教程，作为程序化交易的深化读物，在读者已经掌握程序化交易策略系统的开发、测试和应用知识的基础之上，解决怎样更好地利用机器学习方法自动地寻找最优交易策略的程序化交易问题。在传统的程序化交易平台中，各种交易策略都是完全由策略开发者进行开发的，包括提出策略思想、编制策略程序、进行历史回测、参数优化后再投入实盘交易的整个开发过程。策略的优劣与策略开发者的知识和经验密切相关；而且金融市场变化万千，策略开发者根据过去的经验开发的策略在回测和优化中表现良好，但却很难适应市场的变化。而机器学习不仅可以从历史数据中学习总结金融市场规律，用于指导当前和未来的交易，而且可以根据市场的变化，通过自身的不断学习完善其交易规则。因此，以机器学习为基础的智能交易正在成为金融交易发展的必然趋势。作为程序化交易高级教程，自然以探讨机器学习的金融交易应用为目的。

本书具有以下几个特点。第一，继承性，本书作为程序化交易的高级教程，既是在程序化交易的基本理论和方法基础上展开的，也是在机器学习方法和相应的计算机编程实现的基础上展开的。对于初级教程和中级教程里已经讨论过的程序化交易的基础原理，在此并不重复地专门讨论。但高级教程在许多方面的分析却都是建立在初中级教程中已经分析的基本原理基础之上的，是对初中级教程学习的一种继承。而机器学习和 Python 编程的基本原理也是对相关理论的继承。第二，拓展性，重点讨论利用主流的机器学习编程语言 Python 在程序化交易平台的应用，将机器学习方法从理论模型连接到金融市场的程序化交易应用。第三，实用性，强调程序化交易策略的实用性，本书将讨论怎样利用国信 iQuant 平台，进行机器学习选股策略、择时交易策略的模型开发、数据准备、模型训练、历史回测和模拟交易等策略开发和应用过程，为将机器学习方法成功地应用到实盘交易奠定坚实基础。第四，启发性，本书对机器学习方法在程序化交易应用的讨论并不局限于所讨论的几种方法和策略的具体使用，而更着眼于启发读者去思考、理解拓展机器学习的原理及其在程序化交易策略的应用方法，从而为读者开发自己的策略奠定较好的基础。第五，具体化，每一种机器学习方法的讨论都结合具体的案例进行分析，而非抽象地讨论问题，便于读者的学习和理解。

本书将国信 iQuant 量化平台作为本书案例分析的交易平台和策略开发平台。重点讨论机器学习方法在程序化交易中的应用。由于各种支持 Python 语言编程的程序化交易平台的基本原理是一致的，在该平台上讨论的机器学习程序化交易策略的基本原理和方法同样适用于其他的交易平台。但每一种交易平台又具有自己的特点，有些特殊性的东西在其他平台可能并不适用。读者需要注意。

本书作为程序化交易的深化读物，适宜程序化交易的进阶者和深化者，包括大学金融学专业硕士研究生、金融个人投资者和金融机构的相关人员。需要的预备知识是金融投资的基础知识、机器学习和 Python 编程的基础知识以及程序化交易的基础知识。有金融投资交易的丰富实战经验的人员阅读和使用本书的收获将更丰。缺乏程序化交易基础知识的读者，建议先阅读本系列教材的初级和中级教程。对于缺乏机器学习和计算机编程基础的读者，可以先阅读相关书籍，也可以在阅读本书中需要的地方再查阅相关资料。

本书由复旦大学金融研究院教授陈学彬和国信证券公司的王燕华、张治、徐睿、李文洋、曾健勇、吴文娟和陈俊杰等编写。具体分工为：陈学彬负责第一至第四章和第十四至第十八章；徐睿负责第五至第八章；李文洋负责第九至第十一章；曾健勇负责第十二至第十三章的初稿。陈学彬、王燕华、张治、吴文娟、陈俊杰负责书稿编写大纲的确定和书稿的审定，陈学彬最后修改定稿。

本书的编写和出版得到上海市高峰学科建设项目、国信证券公司、复旦大学经济学院和高等教育出版社的大力支持和资助。国信证券公司不仅提供了资金的支持，而且派出强大的专家团队参与到该系列教材的编写、程序的调试、策略的回测优化等过程，对其中出现的各种问题及时地给予解答和解决。在此，谨向本书的编写和出版给予大力支持和资助的机构和个人表示衷心的感谢！

此外，还要感谢我的妻子廖玉英女士！我几十年的学习和工作上的进步都离不开她的大力支持。正是在她的无微不至的关怀、照料和帮助下，我才能在退休后继续写作，并能够在 2018 年暑期完成了本书的写作。在此，也借此书的出版向她表示最衷心的感谢！

本书是作者根据自己在程序化交易策略的开发和教学中的经验，并参考机器学习、Python 编程的大量资料、文章和国信 iQuant 的使用说明书等资料的基础上编写而成。在此，也向这些作者表示衷心感谢！我们的知识和经验的不足，可能导致书中难免存在一些错误，敬请各位读者批评指正。另外，本书讨论的各种策略及其程序仅供学习理解其原理和方法之用，并非向读者推荐这些策略程序，实际应用需要读者根据自己的经验修善，不要盲目照搬，否则可能给你带来不必要的损失。希望通过对本书的阅读和相关的应用练习，能够为你进入机器学习程序化交易的大门并在该领域取得更大的发展有所帮助！

陈学彬

2019 年 1 月

于复旦大学金融研究院

目　录

第一篇　机器学习交易基础

第二篇 机器学习回归分析

第三篇 机器学习分类模型

第四篇 机器学习聚类和关联分析

第五篇 神经网络学习

第一章
导　论

第一节　机器学习导论

一、什么是机器学习

机器学习(machine learning, ML)是专门研究计算机怎样模拟或实现人类的学习行为,以获取新的知识或技能,重新组织已有的知识结构使之不断改善自身性能的科学理论和方法。它是一门多领域交叉学科,涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。它是人工智能的核心,是使计算机具有智能的根本途径,其应用遍及人工智能的各个领域。

学习是人类具有的一种重要智能行为,机器能否像人类一样具有学习能力呢? 1959年美国的塞缪尔(Samuel)设计了一个下棋程序,这个程序具有学习能力,它可以在不断的对弈中改善自己的棋艺。4年后,这个程序战胜了设计者本人。又过了3年,这个程序战胜了美国一个保持8年之久的常胜不败的冠军。这个程序向人们展示了机器学习的能力,提出了许多令人深思的社会问题与哲学问题。

机器的能力能否超过人? 很多持否定意见的人的一个主要论据是:机器是人造的,其性能和动作完全是由设计者规定的,因此无论如何其能力也不会超过设计者本人。这种意见对不具备学习能力的机器来说的确是对的,可是对具备学习能力的机器来说就值得考虑了,因为这种机器的能力在应用中不断地提高,过一段时间之后,设计者本人也不知它的能力到了何种水平。

机器学习已经有了十分广泛的应用,例如:数据挖掘、计算机视觉、自然语言处理、生物特征识别、搜索引擎、医学诊断、检测信用卡欺诈、证券市场分析、金融投资交易、风险管理、DNA序列测序、语音和手写识别、战略游戏和机器人运用等,各种新的运用正在迅速涌现,未来前景无限。

二、机器学习的发展演变

机器学习是人工智能研究较为年轻的分支,它的发展过程大体上可分为四个阶段。

第一阶段是在20世纪50年代中叶至60年代中叶，属于热烈时期。

第二阶段是在20世纪60年代中叶至70年代中叶，被称为机器学习的冷静时期。

第三阶段是从20世纪70年代中叶至80年代中叶，被称为复兴时期。

第四阶段始于1986年。机器学习进入新阶段的重要表现在下列诸方面：

(1) 机器学习已成为新的边缘学科并在高校形成一门课程。它综合应用心理学、生物学和神经生理学以及数学、自动化和计算机科学形成机器学习理论基础。

(2) 结合各种学习方法，取长补短的多种形式的集成学习系统研究正在兴起。特别是联结学习符号学习的耦合可以更好地解决连续性信号处理中知识与技能的获取与求精问题而受到重视。

(3) 机器学习与人工智能各种基础问题的统一性观点正在形成。例如，学习与问题求解结合进行、知识表达便于学习的观点产生了通用智能系统SOAR的组块学习。基于案例分析的类比学习与问题求解结合的方法已成为经验学习的重要方向。

(4) 各种学习方法的应用范围不断扩大，一部分已形成商品。归纳学习的知识获取工具已在诊断分类型专家系统中广泛使用。联结学习在声音图文识别中占优势，分析学习已用于设计综合型专家系统，遗传算法与强化学习在工程控制中有很好的应用前景，与符号系统耦合的神经网络联结学习在企业的智能管理与智能机器人运动规划中发挥重要作用。

(5) 与机器学习有关的学术活动空前活跃。国际上除每年一次的机器学习研讨会外，还有计算机学习理论会议、机器学习算法研讨会议以及机器学习在各个运用领域的研讨会议等。

第二节 金融交易如何使用机器学习方法

一、人工智能技术在金融交易领域的应用

交易是金融市场的基本功能。金融投资、融资和风险管理都必须通过金融交易来实现。有效率的金融交易就成为制约金融投融资和风险管理效率的重要环节。早期的金融交易与投资分析和决策行为基本上是融为一体的，由投资者直接实施。以后，随着金融市场的发展、金融交易活动逐步与投资分析决策分离，投资分析和决策由投资者或者代理人进行，发出交易指令后由交易经纪人在金融交易所执行。金融交易活动逐步演变成为仅是交易指令的执行实施活动并成为一种职业。但随着计算机技术和计算机网络的发展，一方面，金融信息的传播更加迅速，金融分析决策需要处理的信息量呈几何级数的爆发式增长；另一方面，金融市场的复杂多变和激烈的竞争要求完成投资分析决策到下达交易指令到交易所占用的时间越来越短，交易决策和执行的竞争已经不是“争分夺秒”而是“争毫秒夺微秒”。金融投资分析、决策、交易执行行为的分离已经不能满足金融市场激烈竞争的需要。因此，集金融信息收集处理、分析决策与交易执行和管理多重任务于一身的现代金融程序化(量化)交易系统应运而生。

19世纪80年代，美国已有记载的传统技术分析方法出现，但使用并不普遍。20世纪70年代起，由于计算机技术进步和应用的推动，需要大量计算的量化分析方法相继问世。将两者结合的程序化交易最早起源于美国1975年出现的“股票组合转让与交易”。纽约

股票交易所要求程序化交易须达到15只股票以上，交易金额100万美元以上。80年代，程序化化交易发展迅猛，交易量急剧增加。但是，由于美国1987年10月19日的“黑色星期一”股灾的发生，股市崩盘，跌幅超过22%。程序化交易一度成为替罪羊，其发展也一度处于停滞状态。众多的研究发现：程序化交易与股票市场的价格波动并没有必然的联系，同样，也没有证据显示指数套利加剧了股票市场价格的波动。20世纪90年代以后，程序化交易的发展跃入一个新的台阶。21世纪初叶，随着大数据、互联网和人工智能技术的迅速发展，程序化交易的智能化正在快速地兴起和发展。

分析员通过编写简单函数，设计一些指标，观察数据分布，而这些仅仅把计算机当做一个大型计算器来使用。直到近十多年机器学习的崛起，数据可以快速海量地进行分析、拟合、预测，人们逐渐把人工智能与程序化交易联系得愈加紧密。我们可以把程序化交易按照人工智能的子领域（机器学习、自然语言处理、知识图谱）分为三个子领域。

（一）机器学习

建模分析员们对财务、交易数据进行建模，分析其显著特征，利用回归分析等传统机器学习算法预测市场走势，制定交易策略。这种方式的优点在于，建模方便，在具有大量交易数据和财务数据的情况下，通过机器学习，可以迅速地提取市场特征，并训练具有较强预测功能的市场模型，并可以较好地用于交易实际。但这种方式也有两个主要弊端，其一是数据不够丰富，仅限于交易数据和财务数据；更重要的是它受限于特征的选取与组合（feature engineering），模型的好坏取决于建模分析员对数据的敏感程度。

此外一种做法是，模仿专家的行为，选择某一领域的特定专家，复制他们的决策过程，并导入可重复的计算框架。

其代表公司有：纽约的Rebellion Research，其在2007年推出了第一个纯人工智能（AI）投资基金。该公司的交易系统基于贝叶斯机器学习，结合预测算法，响应新的信息和历史经验从而不断演化，有效地通过自学习完成在全球44个国家的股票、债券、大宗商品和外汇的交易。

日本的初创公司Alpaca，他们的交易平台Capitalico利用基于图像识别的深度学习技术，允许用户很容易地从存档里找到外汇交易图表并帮助做好分析，使普通投资者就能知道明星交易员是如何交易的，从他们的经验中学习并作出更准确的交易。

伦敦的对冲基金机构Castilium，由金融领域权威与计算机科学家一同创建，包括前德意志银行衍生品专家、花旗集团前董事长兼首席执行官和麻省理工的教授。他们采访了大量交易员和基金经理，复制分析师、交易员和风险经理们的推演和决策过程，并将它们纳入算法中。

中国香港的Aidyia，致力于用人工智能分析美股市场，依赖于多种AI的混合，包括遗传算法（genetic evolution）、概率逻辑（probabilistic logic），系统分析大盘行情以及宏观经济数据之后做出自己的市场预测，并对最后的行动进行表决。

全球最大的对冲基金桥水联合（Bridgewater Asspcoates），使用一种基于历史数据与统计概率的交易算法，让系统能够自动学习市场变化并适应新的信息。与其类似的公司还有Point72 Asset Management，Renaissance Technologies，Two Sigma等。

（二）自然语言处理

在利用机器学习进行金融交易中，人们发现仅仅从数字训练模型是不够的，他们开始

考虑引入新闻、政策、社交网络中的丰富文本并运用自然语言处理技术进行分析，将非结构化数据进行结构化处理，从中探寻影响市场变动的线索。

这方面直接用于投资交易的并不多，更多是用于风控与征信。通过抓取个人及企业在其主页、社交媒体等地方的数据，首先可以判断企业或其产品在社会中的影响力，比如观测 App 下载量，微博中提及产品的次数，在知乎上对其产品的评价；此外，将数据结构化后，也可推测投资的风险点。

这方面国内的很多互联网贷款，征信公司都在大量使用自然语言处理技术，如宜信、闪银等。另外一些公司则利用这些技术进行 B 端潜在客户的搜寻，如 EverString，并将信息出售给其上游公司。代表性公司有：

① CommEq。其是 2016 年 6 月份在伦敦新设的一家基于人工智能(AI)的对冲基金。CommEq 的投资方法结合了定量模型与自然语言处理(NLP)，使计算机能够如人类一样通过推断和逻辑演绎理解不完整和非结构化的信息。

② 由李嘉诚与塔塔通信投资的 Sentient Technologies。其运用自然语言处理、深度学习(deep learning)等多种 AI 技术，进行量化交易模型的建立。

③ Kensho。其是美国一家基于云计算的智能计算机系统先锋公司。它综合使用自然语言搜索，图形化用户界面和云计算，为金融市场的投资者提供一套全新的数据分析工具——Warren。Warren 能够回答复杂的金融市场问题，如各种数据、股票走向等，可回答约 100 万种关于全球事件对股价影响的英文问题。

(三) 知识图谱

机器学习擅长发现数据间的相关性而非因果性。这就需要根据经济金融理论和专家经验建立的知识库(规则)来避免各种虚假相关性的发生。知识图谱本质上是语义网络，是一种基于图的数据结构，根据专家设计的规则与不同种类的实体连接所组成的关系网络。知识图谱提供了从“关系”的角度去分析问题的能力。

就金融领域来说，规则可以是经济理论和专家对行业的理解、投资的逻辑、风控的把握，关系可以是企业的上下游、合作、竞争对手、子母公司、投资、对标等关系，可以是高管与企业的任职等关系，也可以是行业间的逻辑关系，实体则是投资机构、投资人、企业等，把它们以知识图谱表示出来，从而进行更深入的知识推理，从而提高金融决策的精准性。代表性公司有：

① Garlik。其是知识图谱在金融最早的应用代表之一。该公司 2005 年成立于英国，核心成员来自南安普顿大学，是语义网的核心研究机构之一，主要业务是在线个人信息监控。他们收集网络和社交媒体上的个人信息，当发生个人信息盗窃时 Garlik 会及时报警。2011 年他们被美国的三大个人信用记录公司之一 Experian 收购，其技术被用于个人信用记录、信用盗窃的分析。Garlik 的核心技术之一是大规模语义数据库，前后开源发布了 3store、4store、5store 等高性能数据库。

② Palantir。其是估值仅次于 Uber 的科技创业公司。他们有一个基于知识图谱的金融数据分析平台：Palantir Metropolis，可以整合多源的量化资料，并提供一套方便易用的分析工具来满足复杂的研究需求，其中的组件能够进行复杂搜索、可视化编辑与分析，有非常丰富的人机交互能力。

③ 全球首只号称使用人工智能选股的 AI Powered Equity ETF(AIEQ)。其表现不

俗。2017 年 10 月 18 日，在纽交所推出，与传统的追踪指数的被动型 ETF 不同，该基金为主动型管理 ETF，追求资产增值与超额收益。利用人工智能技术，AIEQ 背后的量化模型每天可处理超过一百万条信息，用于对美国逾 6 000 家公司搭建预测模型。量化模型还会根据每天的经济情况、市场趋势以及市场主要事件，对预测模型的结果的可能性进行调整。AIEQ 分析美国股票和房地产投资信托基金（REITs）最近十年的历史数据，挑选 30～70 家在未来 12 个月具有超额收益机会的公司进行投资。目标是在风险对标美股市场的前提下最大化收益。截至 2018 年 6 月 13 日，AIEQ 的资产管理规模不到 8 个月就超过 1.4 亿美元，获得投资收益 13.68%，年化波动率 15.48%，最大回撤 9.75%。同期标普 500 指数的收益为 9.77%，年化波动率 14.52%，最大回撤为 10.10%。

二、建立和使用机器学习交易模型的基本环节

建立和使用机器学习交易模型需要遵循一定的基本步骤或环节如图 1-1 所示。

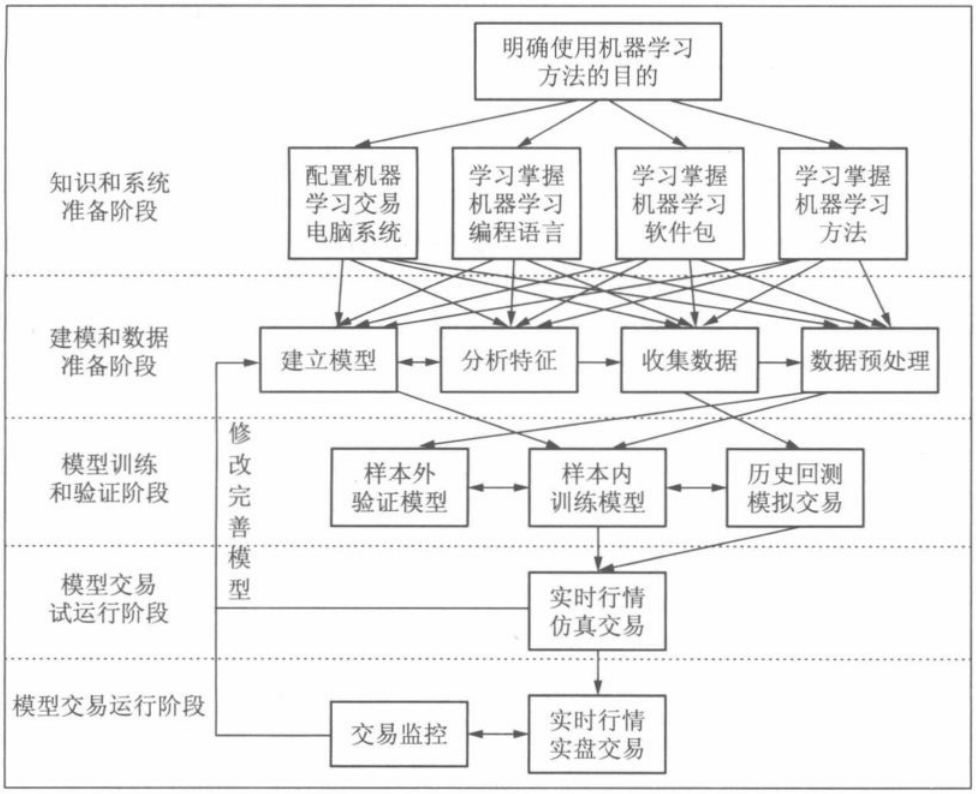

图 1-1　建立和使用机器学习交易模型的基本环节

（一）知识和系统准备阶段

1. 明确使用机器学习方法进行金融交易的目的

我们首先必须明确使用机器学习方法进行金融交易的目的，不是为使用而使用，而是提升金融交易的效率。提升金融交易效率有许多方式：精准选择投资品种（选股）、不失时机地抓住最有利的投资机会（择时）、有效地控制投资风险（风控）。机器学习方法很多，它们在选股、择时和风控方面各有所长，我们必须根据使用目的有针对性地选择机器学习方

法，从而提高其使用效率。

2. 配置使用机器学习进行交易的计算机系统

在明确使用机器学习方法的目的后，我们需要有针对性的配置使用机器学习进行金融交易的计算机系统。它包括能够进行机器学习的模型训练的系统和支持训练好的模型进行仿真交易和实盘交易的系统。

3. 学习使用机器学习的基本编程语言

要使用机器学习方法训练金融交易模型，必须利用计算机编程语言编辑机器学习模型程序。因此，使用者需要学习掌握基本的编程语言及其编程方法。目前，使用最为广泛的机器学习模型编程语言为 Python 语言。使用者需要学习掌握 Python 语言的基本编程方法。

4. 学习使用机器学习的常用软件包

由于 Python 是一种开放式的编程语言，在它的基础上，许多机构和个人开发了大量的机器学习框架和软件包，免费提供给公众使用，如 TensorFlow、Keras 等。因此，使用者不必完全由自己用 Python 语言编写整个机器学习模型的所有程序，而可以直接调用这些机器学习框架和软件包来建立和训练模型，从而大大地简化编程任务。因此，使用者需要学习如何使用这些机器学习框架和软件包，以便提高自己的编程效率。

5. 学习机器学习的基本方法

机器学习方法种类甚多，每一种方法都有自己擅长的领域和局限之处。使用者必须根据自己使用机器学习进行金融交易的主要目的有针对性地学习掌握相关的机器学习方法。这里的学习不是脱离应用实际的纯理论的学习，而是密切联系自己需要解决的问题进行学习，并结合机器学习框架和软件包的使用进行学习，其中可以借鉴许多成熟的开源的机器学习模型，从而提高自己应用机器学习方法和框架以及相关软件包建立应用模型的能力。

(二) 建模和数据准备阶段

1. 根据使用目的建立相关机器学习交易模型

在学习掌握机器学习方法和软件的基础上，使用者需要根据使用目的，尝试建立自己的机器学习交易模型。不同的机器学习模型各有所长，并非越复杂越好，需要根据自己使用机器学习方法的目的和交易对象的特点来选取。

2. 分析交易对象行情影响因素选取特征

在建立机器学习模型的同时，需要对交易对象价格的影响因素进行分析，选取能够充分刻画其价格变动的主要特征变量作为机器学习的输入变量。特征变量的选取对于机器学习模型的训练和预测结果至关重要。选取方法包括经济理论指导、相关性分析、主成分分析和专家经验等。

3. 收集整理相关原始数据

机器学习是建立在对能够反映交易对象价格变动特征的大量数据进行模型训练的基础上的。因此，收集特征变量的大量历史数据十分重要。作为金融交易模型的输入数据，不仅需要考虑历史数据的可得性，还必须考虑实时数据的可得性。因此，在模型训练完成后，当其实时交易的时候，必须能够联网进行实时数据更新。

4. 进行模型训练前的数据预处理

在收集好原始数据后，不能直接用来机器训练，还需要对它们进行必要的预处理。预处理包括特征指标的计算、特征指标的标准化(归一化)和正则化。标准化用来消除不同