证券研究报告 发布时间：2017-10-11

证券研究报告/ 策略点评报告

人工智能系列报告综述篇：人工智能发展史及算法介绍

报告摘要：

本文作为人工智能系列报告的第一篇，围绕人工智能的发展历史，人工智能主要算法以及人工智能在金融领域的应用展开综合论述，力求先为投资者勾勒出一幅人工智能全景图，后续我们将推出人工智能系列实战篇，结合多种机器学习算法，构造量化选股框架，为广大投资者提供成熟有效的策略。

人工智能发展历史波折起伏：人工智能自 1956 年正式确立以来，一直曲折发展，从产生到成为研究热点一直饱受质疑，期间经历两次发展低谷，而学科自身所迸发的生命力不断推动其走出低谷，成为引领技术革命的热点。**从诞生伊始，人工智能就有理性学派和感性学派之争**，理性学派从符号计算出发，将人脑看成信息处理器，认为任何能够以一定的逻辑规则描述的问题都可以通过人工智能程序来计算解决。感性学派简单说就是通过对脑神经的模拟来获得人工智能，随着深度学习等技术的成功，人工智能的研究热点越来越集中到感性学派。

人工智能算法发展方向不断变化：学术界早期研究重点集中在符号计算，**神经网络在人工智能发展早期被完全否定，而后逐渐被认可，再成为今天引领人工智能发展潮流的一大类算法，持续显现出生命活力。**本文着重介绍人工智能领域比较著名的 4 个算法，他们分别是感知器、决策树，支持向量机和卷积神经网络。通过这 4 个具有代表性的算法，理清机器学习的基本思想。

人工智能在金融领域应用：7 月 20 日国务院正式印发了《新一代人工智能发展规划》，**明确指出到 2030 年之前我国 AI 核心产业规模或超 1 万亿元。这是人工智能首次上升到国家战略高度**，我们认为国内人工智能在金融里的应用还处于探索阶段，未来具有很大的发展潜力。目前人工智能应用场景还是集中在和大数据分析、与互联网连接紧密的领域，本文列举大数据基金、人工智能预测和智能投顾等例子来进行阐述。

相关报告

《金融工程：HMM 指数择时研究之实战篇》 2016-09-23

《金融工程：HMM 指数择时研究之理论篇》 2017-01-19

证券分析师：高建
执业证书编号：S0550511020011

研究助理：肖承志
执业证书编号：S0550116080014
021-2036 1264 xiaocz@nesc.cn

研究助理：孙凯歌
执业证书编号：S0550117100006
13070102332 sunkg@nesc.cn

目 录

1. 人工智能概念

什么是人工智能？按照李开复博士的解释，**人工智能指的是获取某一领域的海量信息，并利用这些信息对具体案例做出判断，达成特定目标的一种技术**。举个例子，比如说互联网贷款，计算机通过海量的贷款记录训练模型，对能否对某一用户进行贷款这一事件做出判断，想达到的目标就是放贷人利益最大化。

早期人工智能的研究是基于一个很基本的假设，就是人的思维活动可以用机械的方式进行表达。在上世纪五十年代，有人断言只需要经过一代人努力，就可以创造出与人类同等智力水平的计算机机器，然而直到 20 世纪 70 年代，人工智能技术才开始扩展到各个研究领域，包括数学定理证明、机器翻译、机器人技术等。近年来，随着深度学习的发展和计算机硬件技术的提高，人工智能又进入新的发展高潮。耳熟能详的例子就是 AlphaGo 战胜李世石，之后又有 Master 以 60 连胜的记录横扫中日韩的围棋高手。2017 年 5 月份 AlphaGo 又 3 比 0 战胜了柯洁。

2017 年 7 月 20 日国务院正式印发了《新一代人工智能发展规划》，明确指出到 2030 年之前我国人工智能核心产业规模或超 1 万亿元。这是人工智能首次上升到国家战略高度。**可以想象，人类在经历了 PC 时代，网络时代，智能手机时代之后，必将进入人工智能时代。**

2. 人工智能发展历史

我们认为人工智能发展至今，可以分为 5 个阶段，分别是人工智能的形成阶段，发展的黄金阶段，第一次发展低谷及复苏、第二次发展低谷，以及目前由深度学习引领的第五个发展阶段——现代人工智能。

2.1. 形成阶段（1956-1961）

人工智能（Artificial Intelligence，AI）的形成阶段，是从 1956 年到 1961 年。人工智能概念的正式确立是在 1956 年达特茅斯学院的一次学术会议上，参会人员希望将人工智能作为一门独立科学，确立其任务和发展路径。与会者们宣称，人工智能的特征都可以被精准逻辑运算描述，精准描述后就可以用机器来模拟和实现。会议相关的摩尔(Trenchard More)、麦卡锡（John McCarthy）、明斯基（Marvin Minsky）、塞弗里奇（Oliver Selfridge）、所罗门诺夫（Solomonoff）作为 AI 领域的开创者，日后数十年间成为了 AI 领域研究的领军人物，此次会议第一次正式使用人工智能这一术语，其中参会的两人塞弗里奇(Oliver Selfridge)和纽厄尔 (Allen Newell)，塞弗里奇发表了一篇模式识别的文章，而纽厄尔探讨计算机模拟人类下棋，他们分别代表两派观点。讨论会的主持人，神经网络鼻祖之一的皮茨(Pitts)最后总结时说：“(一派人)企图模拟神经系统，而纽厄尔则企图模拟心智(Mind)……但殊途同归。”**皮茨预示了人工智能随后几十年按照符号计算和神经网络的两个路径发展。**

2.2. 黄金时代（1961-1973）

在达特茅斯会议之后的数十年间，人工智能迎来了高速发展，计算机解决了一些数学证明以及学习使用英语等问题，AI 的快速发展使得研究人员乐观情绪高涨，认为

具备人类思考能力的机器在不久的将来就会出现，与此同时，国防机构也对 AI 充满浓厚兴趣，对这一领域投入大量资金,希望获得军事上的领先。在这一时期研究成果呈井喷态势涌现。

1956 年 IBM 小组设计了一个具有自学习、自组织、自适应能力的西洋跳棋程序，这个程序可以像一个优秀棋手那样，向前看几步来下棋，它还能学习棋谱，在分析大约 175000 幅不同棋局后，可进行棋局走步预测，准确度达 48%。这是机器模拟人类学习过程卓有成就的探索。1959 年这个程序曾战胜设计者本人，1962 年还击败了美国一个州的跳棋大师。

1957 年纽厄尔和赫伯特·西蒙等人的心理学小组编制出一个称为逻辑理论机 LT(The Logic Theory Machine) 的数学定理证明程序，这是世界上第一个人工智能程序，有能力证明罗素和怀特海《数学原理》第二章中的 38 个定理。1958 年在 MIT 小组的麦卡锡（Mccarthy）建立的行动计划咨询系统以及 1960 年明斯基（Minsky）的论文“走向人工智能的步骤”，对人工智能的发展都起了积极的推动作用。1959 年麦卡锡发明的函数式处理语言 LISP，成为人工智能程序设计的主要语言，长期垄断人工智能领域的应用开发，至今仍被广泛采用。1961 年，第一台工业机器人开始在新泽西州通用汽车工厂的生产线上工作。介于以上卓越的成果和人工智能整个领域的快速发展，1965 年，赫伯特·西蒙（Herbert Simon）预测 20 年内计算机将能够取代人类智力。同年诞生了历史上第一个专家系统，费根鲍姆(Edward Feigenbaum)、布鲁斯.布坎南（Bruce G.Buchanan）、莱德伯格（Joshua Lederberg）和卡尔·杰拉西（Carl Djerassi）在斯坦福大学研究的 DENDRAL 系统，使有机化学的决策过程和问题解决自动化。而后，机器人也开始出现，日本早稻田大学在 1970 年造出第一个人形状机器人 WABOT-1。这些早期成果，充分表明人工智能作为一门新兴学科正在茁壮成长。1950-1973 年人工智能发展早期和黄金阶段主要研究工作成果如表 **1** 所示。

表 1: 1950-1973 年人工智能主要研究成果

时间	主要事件
1950 年	阿兰·图灵（Alan Turing）发表《计算机械与智力》一文。系统性提出了甄别机器是否具有人类智能的“图灵测试”方法
1951 年	马文·明斯基（Marvin Minsky）和迪恩·爱德蒙（Dean Edmunds）用 3000 个真空管来模拟 40 个神经元规模的网络建立了人类历史上第一个人工神经网络。
1952 年	阿瑟·萨缪尔（Arthur Samuel）开发第一个计算机跳棋程序和第一个具有学习能力的计算机程序
1955 年	“人工智能”(artificial intelligence)一词在一份由约翰·麦卡锡(达特茅斯学院)、马文·明斯基（哈佛大学）、纳撒尼尔·罗彻斯特（IBM）和克劳德·香农（贝尔电话实验室）联合递交的关于召开国际人工智能会议的提案中被首次提出。
1955 年	赫伯特·西蒙（Herbert Simon）和艾伦·纽厄尔（Allen Newell）开发出世界上第一个人工智能程序“逻辑理论家”，证明了罗素和怀特海《数学原理》第二章中的 38 个定理。
1956 年	达特矛斯会议召开， 人工智能概念的正式确立。
1957 年	弗兰克·罗森布拉特（Frank Rosenblatt）开发出基于两层计算机网络能够进行模式识别的神经网络系统“Perceptron”。
1958 年	约翰·麦卡锡（John McCarthy）开发编程语言 Lisp，成为人工智能研究中最流

	行的编程语言。
1959 年	约翰·麦卡锡（John McCarthy）发表《Programs with Common Sense》，提出“Advice Taker”概念，这个假想程序可以被看成是第一个完整的人工智能系统。
1961 年	第一台工业机器人 Unimate 开始在新泽西州通用汽车工厂的生产线上工作。
1964 年	丹尼尔·鲍勃罗（Daniel Bobrow）在完成了他的麻省理工博士论文同时开发了自然语言理解程序“STUDENT”。
1965 年	斯坦福大学研究出历史上第一个专家系统 DENDRAL 系统，能够使有机化学的决策过程和问题解决自动化。
1969 年	阿瑟·布莱森（Arthur Bryson）和何毓琦（Yu-Chi Ho）描述了反向传播作为一种多阶段动态系统优化方法，可用于多层人工神经网络。
1970 年	日本早稻田大学造出第一个人形状机器人 WABOT-1
1972 年	斯坦福大学开发出名为“MYCIN”的专家系统。能够利用人工智能识别感染细菌，并推荐抗生素。

数据来源：东北证券，互联网

2.3. 第一次发展低谷及复苏（1973-1987）

第一次低谷是发生在 1974 年到 1980 年。第一次低谷的到来有必然的原因，其实在黄金十年期间人工智能的理论基础和技术发展并没有获得实质性突破，人工智能技术经历了从 1956 年开始的将近 20 年的高速发展之后，终于遇到了自己的瓶颈，于 1974 年迎来了第一次低谷期。第一次的低谷期当中，有两个比较重要的事件。第一个是，由于早期研究者对人工智能发展前景过于乐观的态度，美国高等研究计划署对麻省理工、卡内基梅隆大学这些高校的人工智能项目投入了大笔资金，但到后期逐渐发现无法实现之前的研发目标，严重打击投资者和研究者的情绪，也使研发经费被削减。第二个方面，学术界发现早期设计的逻辑器、感知器都只能做简单而且专业面很窄的任务，一旦遇到复杂环境就无法应对。人工智能领域先驱明斯基在《感知器》一书中指出，人工智能在数学基础上存在漏洞，神经网络不存在有效的学习方法，这种的悲观论调和政府支持资金的缩减最终使得人工智能的发展进入低谷。

我们认为，**人工智能在第三个阶段陷入低谷主要还是因为整个学科理论基础和技术实现都存在很大短板，遭遇危机是必然的**，可以分三个方面来解释。第一，随着程序计算复杂性上升，计算机性能满足不了学者提出的研究需求。第二，没有大容量数据库支持，研究面临数据缺失的困境，无法找到足够数据量支撑机器学习算法的训练。第三，就像明斯基所说，作为人工智能基础的数学理论还不够完善。上述三个原因都从另一方面反映了人工智能的发展，不仅是本身学科的发展，也需要依赖于其他领域的同步发展，比如计算机硬件、基础数学和数据科学。

然而这一次的低谷仅仅持续了不到七年时间便迎来了又一个七年的复苏。**推动此次复苏的标志性事件有两个，一个是 80 年代初的专家系统，另外一个是第五代计算机的研究热潮。**

1980 年卡耐基梅隆大学为 DEC 公司制造了一个专家系统，这个系统每年为公司节省 4000 万美元的开销，取得了巨大成功。此后很多公司和高校纷纷效仿，很大程度上为人工智能的发展争取了大量经费。专家系统的成功也重燃了整个社会对人工智能发展的信心。另一个事件，1981 年日本“新一代计算机技术研究所”提出研发具

有人工智能的第五代计算机，总投资预算达到 8.5 亿美元，并且组织富士通、夏普等著名企业配合，很多其他国家也启动类似计划，投入大量资金进入人工智能领域，用于开发第五代计算机，当时称为“人工智能计算机”。1973-1987 年人工智能领域的主要研究成果如表 2 所示。

表 2：1973-1987 年人工智能主要研究成果

时间	主要事件
1973 年	詹姆斯 莱特希尔（James Lighthill）给英国科学研究委员会报告中对人工智能持悲观态度，政府大幅度削减对 AI 研究的资金支持。
1980 年	日本早稻田大学研制出 Wabot-2 机器人，能够与人沟通、阅读乐谱并演奏电子琴。
1980 年	卡耐基梅隆大学为 DEC 公司制造了一个专家系统 RI，根据用户需求为计算机自动选择组件。
1981 年	日本“新一代计算机技术研究所”（ICOT）提出研发具有人工智能的第五代计算机，并获得日本通产省 8.5 亿美元的经费支持。英美等国家也投入巨资研发第五代计算机。
1982 年	约翰 霍普菲尔德（John Hopfield）在 1982 年发明一种能够模拟人类记忆的循环神经网络—Hopfield 神经网络
1984 年	罗杰 单克（Roger Schank）和马文 明斯基（Marvin Minsky）在年度 AAAI 会议上警告“AI 之冬”即将到来。
1986 年	恩斯特 迪克曼斯（Ernst Dickmanns）指导建造第一辆无人驾驶奔驰汽车
1986 年	鲁梅尔哈特（Rumelhart）和麦克利兰（McClelland）为首的科学家提出了 BP(back propagation)神经网络，证明了明斯基关于多层网络不存在有效学习方法的论断是错误的。

数据来源：东北证券，互联网

总的来看，**我们认为这阶段人工智能的复苏，主要得益于政府对专家系统和第五代计算机研发的巨额资金支持，人工智能在基础理论和技术创新上并没有取得特别大的进步**，这也为接下来人工智能热潮的衰退埋下隐患。

2.4. 第二次发展低谷（1987-1993）

七年的短暂复苏之后，1987 年人工智能的发展陷入第二次低谷。主要原因有两个，一是个人计算机的出现冲击了专家系统，二是“人工智能计算机”研发的失败。

当时苹果、IBM 开发的第一代个人计算机开始走向社会，价格低廉，迅速挤占了专家系统的市场，导致专家系统的需求急剧下滑。第二个是，被寄予厚望的第五代计算机“人工智能计算机”，在人机交互的关键技术没能实现突破。导致政府对于人工智能支持经费又进一步削减。在削减了大量经费之后，人工智能研究在此期间一度进入停滞期。

我们认为，第四阶段，也就是人工智能的第二次发展低谷，表面上看是专家系统失去市场和第五代计算机的研发失败，但是实际上结合人工智能发展的轨迹和当时社会环境来看，我们认为原因实质上不仅仅这样，其实是有两个更深层的原因。首先，**人工智能学科研究过于单一**，尤其前期符号计算垄断了整个学科，没有其他研究方

向来分担风险；其次，**人工智能的研究资金大部分是来自于政府机构，没有在社会上形成健全的产业链**，一旦政府研究热点发生转移，就会出现经费不足的情况。

表 3：1987-1993 年人工智能主要研究成果

时间	主要事件
1988 年	罗洛·卡彭特（Rollo Carpenter）开发了聊天机器人 Jabberwacky，能够模仿人进行幽默的聊天。这是人工智能与人类交互的最早尝试
1989 年	燕乐存（Yann LeCun）和贝尔实验室的其他研究人员成功将反向传播算法应用在多层神经网络，实现手写邮编的识别。
1991 年	德国学者 Sepp Hochreiter 第一次清晰提出梯度消失问题并阐明原因，解释了从输出层反向传播时，每经过一层，梯度衰减速度极快，学习速度变得极慢，神经网络很容易停滞于局部最优。同时，算法训练时间过长会出现过度拟合(overfit)，把噪音当成有效信号

数据来源：东北证券，互联网

2.5. 现代人工智能（1993-至今）

从 1993 年开始，数学工具不断完善，计算机性能由于摩尔定理的作用得到了大幅提高，很多学术界想法得以实现。与此同时，人工智能的任务也开始明确和简化，就是要做具有实用性的人工智能，这使人工智能重新走向繁荣。在这段时间人工智能发展也是不断自我革新，先后经历了知识管理、统计学习、机器学习，到现在的深度学习等四个不同的阶段。这一时期人工智能领域的创新性成果层出不穷，理论和应用层面均取得很大进展，包括大型图像数据库 ImageNet 的建立，以及谷歌和高校合作推出的多层神经网络。具体研究成果如表 4 所示。

表 4：1993 年-至今人工智能主要研究成果

时间	主要事件
1993 年	弗农·温格（Vernor Vinge）发表了《The Coming Technological Singularity》。认为三十年之内人类就会拥有打造超人类智能的技术。不久之后人类时代将迎来终结。
1997 年	赛普·霍克赖特（Sepp Hochreiter）和于尔根·施密德胡伯（Jürgen Schmidhuber）提出长短期记忆人工神经网络（LSTM）概念。这一概念指导下的递归神经网络在今日手写识别和语音识别中得到应用。
1997 年	IBM 研发的“深蓝”（Deep Blue）成为第一个击败人类象棋冠军的电脑程序。
1998 年	燕乐存（Yann LeCun）和约书亚·本吉奥（Yoshua Bengio）发表了关于神经网络应用于手写识别和优化反向传播的论文。
2000 年	MIT 的西蒂亚·布雷泽尔（Cynthia Breazeal）打造了 Kismet，一款可以识别和模拟人类情绪的机器人。
2001 年	斯皮尔伯格的电影《人工智能》上映。电影讲述了一个儿童机器人企图融入人类世界的故事，引发社会对人类与人工智能关系的关注
2007 年	李飞飞（Fei Fei Li ）和普林斯顿大学的同事开始建立 ImageNet。这是一个大型注释图像数据库，旨在帮助视觉对象识别软件进行研究。
2009 年	谷歌开始秘密研发无人驾驶汽车。2014 年，谷歌汽车在内华达州通过自动驾驶汽车测试
2009 年	斯坦福大学的 Rajat Raina 和吴恩达(Andrew Ng)合作发表论文《用 GPU 大规

	模无监督深度学习》，认为运行在 GPU 上的神经网络比 CPU 快数倍。
2010 年	瑞士学者 Dan Ciresan 和合作者发表论文《Deep big simple neural nets excel on handwritten digit recognition》，在 GPU 上实现反向传播计算方法，速度比传统 CPU 快了 40 倍。
2012 年	2012 年斯坦福大学研究生黎越国领衔，和他的导师吴恩达，以及众多谷歌的科学家联合发表论文《用大规模无监督学习建造高层次特征》。黎越国的文章中使用了九层神经网络，网络的参数数量高达 10 亿，是 Ciresan 2010 年论文中的模型的 100 倍，是 2009 年 Raina 论文模型的 10 倍。而人的脑皮层接近一百五十万亿神经突触，是黎越国模型参数数量的十万倍。

数据来源：东北证券，互联网

今天，我们重新回顾人工智能的发展历史，不难发现，人工智能发展的动力主要来自于两个方面，**一个是学科的内部动力，另一个是社会目标驱动的外部动力，而社会对人工智能的期望能够带来比内部更为强大的发展驱动**。大家可能会问，这次发展热潮会不会像历史一样出现第三次消退呢？我们认为不会，主要有两个原因。

第一个是，人工智能不再局限在学院派里，工业界正试图打造人工智能完整的产业链，学科本身具备了自身造血功能。当前人工智能的发展正从过去的学术牵引转化为工业牵引。例如 IBM 围绕沃森打造人工智能生态系统；谷歌斥资 4 亿美元收购 DeepMind，DeepMind 的代表成果就是阿尔法狗；苹果和百度布局无人驾驶等。这些科技巨头投入大量的人力财力，很大程度上保证了人工智能能够持续发展下去。这和传统的人工智能要靠政府经费支持相比，有了很大的改观，也就是说，现在不靠政府补贴，人工智能在工业界领域也能自谋生路。

第二原因是，目前人类社会积累的数据量呈指数级增长，怎么对这些数据进行处理分析，找出有用的信息，都将依赖于人工智能提供的新技术方法。人工智能将不单单是简单的机器智能测试，更多是研发与人类社会相融合的智能系统，通过网络将人、计算机和其他外部事物连接起来，构建类似于智慧城市的复杂生态系统。

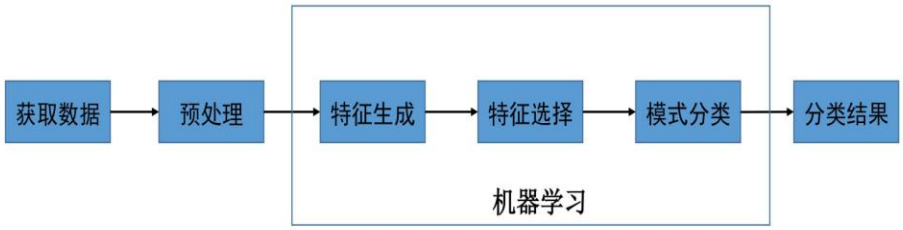
3. 人工智能算法

1950 年初期，人工智能追求研发能够像人类一样具有智力的机器，研究界把这个称之为“强人工智能”，后续出现了专家系统，在特定领域运用人工智能技术，给人工智能发展注入新的活力，然而又带来了难以移植，成本昂贵等问题。1980 年之后机器学习成为 AI 研究的主流，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的性能。2000 年左右，计算机科学家在神经网络研究基础上加入多层感知器构建深度学习模型，成功解决了图像识别、语音识别以及自然语言处理等领域的众多问题。近年来，在 IBM 等科技巨头推动下认知计算（Cognitive computing）蓬勃发展，通过学习理解语言、图像、视频等非结构化数据，更好地从海量复杂数据中获得知识，做出更为精准的决策。

机器学习是人工智能领域研究的核心问题之一，理论成果已经应用到人工智能的各个领域，机器学习算法通过模式识别系统根据事物特征将其划分到不同类别，通过对识别算法的选择和优化，使其具有更强的分类能力。**机器学习模式识别流程如图**

1 所示，包括获取数据、数据预处理、特征生成、特征选择、模式分类、和最后生成分类结果等步骤。

图 1: 机器学习模式识别流程



数据来源：东北证券

这部分我们主要介绍人工智能领域比较著名的 4 个算法，他们分别是感知器、决策树，支持向量机和深度学习中比较著名的算法—卷积神经网络。

3.1. 感知器

美国计算机科学院罗森布拉特（F.Roseblatt）于 1957 年提出感知器，是神经网络第一个里程碑算法。所谓感知器，是一种用于二分类的线性分类模型，其输入为样本的特征向量，计算这些输入的线性组合，如果输出结果大于某个阀值就输出 1，否则输出-1。作为一个线性分类器，感知器有能力解决线性分类问题，也可用于基于模式分类的学习控制中。

单个神经元的输入输出关系：

$$y_j = f(s_j)$$

f 为激活函数，其中

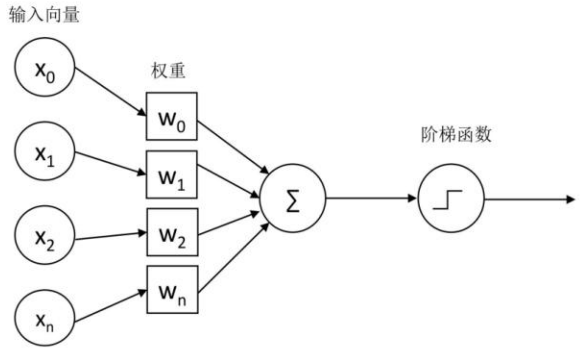
$$s_j = \sum_{i=1}^{n} \omega_i x_i - \theta = \sum_{i=1}^{n} \omega_i x_i$$

感知器结构如图 **2** 所示，具体包括:

- ✓ 输入向量（input),即为用来训练感知器的原始数据
- ✓ 阶梯函数(step function),可以通过生物上的神经元阈值来理解，当输入向量和权重相乘之后，如果结果大于阈值（比如 0），则神经元激活(返回 1)，反之则神经元未激活(返回 0)
- ✓ 权重(weight),感知器通过数据训练，学习到的权向量通过将它和输入向量点乘，把乘积带入阶梯函数后我们可以得到期待的结果

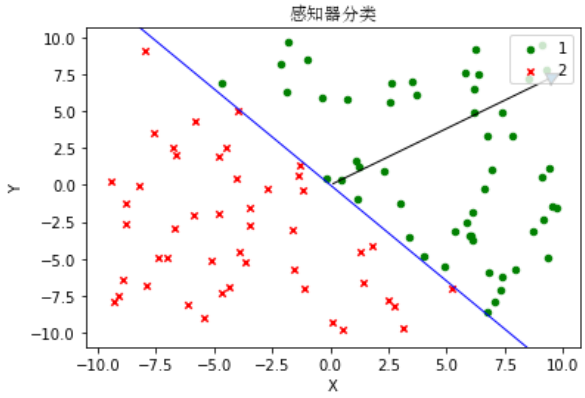
由于感知器自身结构的限制，使其应用被限制在一定的范围内。所以在采用感

图 2：感知器结构



数据来源：东北证券

图 3：感知器分类



数据来源：东北证券

感知器解决具体问题时器有以下局限性：

- ✓ 由于感知器的激活函数采用的是阀值函数，输出矢量只能取 0 或 1，所以只能用它来解决简单的分类问题；
- ✓ 感知器是一种线性分类器，在迭代过程中，如果训练数据不是线性可分的，可能导致训练最终无法收敛，最终得不到一个稳定的权重向量。

感知器只能解决简单的线性分类问题，应用面很窄，但是在人工智能发展中起到了很大推动作用，由于是第一个神经网络算法，吸引了大量学者对神经网络开展研究，同时感知器也为后期更复杂算法如深度学习奠定基础。

3.2. 聚类算法

从机器学习的角度，聚类算法是一种“无监督学习”，训练样本的标记信息是未知的，根据数据的相似性和距离来划分，聚类的数目和结构没有事先给定。聚类的目的是寻找数据簇中潜在的分组结构和关联关系，通过聚类使得同一个簇内的数据对象的相似性尽可能大，同时不在同一个簇中的数据对象的差异性也尽可能地大。在人工智能中，聚类分析亦被称为“无先验学习”，是机器学习中的重要算法，目前被广泛应用于各种自然科学和工程领域，如心理学、生物学、医学等。

目前已经提出多种聚类算法，可分为：划分方法、层次方法、基于密度的方法、基于网格的方法和基于模型的方法。其中著名的分类算法 k-means 算法就是基于划分的聚类算法。我们以 k-means 算法为例，先对数据原型进行初始化，然后对原型进行迭代更新求解，算法描述如表 5 所示：

表 5：k-means 算法描述

输入： 样本集 $D=\{x_1,x_2,...,x_m\}$ ，聚类簇 k

过程：

1: 从 D 中随机选择 k 个不样本作为初始质心 $\{\mu_1,\mu_2,...,\mu_k\}$

2: **repeat**

3: 计算样本 x_j 与质心 μ_i 的距离：$d_{ji}=\left\|x_j-u_i\right\|_2$

4: 将样本 x_j 划入距离最近的簇 C_i

5: 计算每个聚类簇的新的质心 $u_i^{'} = \frac{1}{|C_i|}\sum_{x \in C_i} x$

6: **until** 所有聚类簇质心不发生变化或者达到最大迭代次数

数据来源：东北证券、《机器学习》

3.3. 决策树

决策树是一种简单却使用广泛的分类器，**通过训练数建立决策树对未知数据进行高效分类**。一棵决策树一般包括根结点、内部结点和叶子结点；叶子结点对应最终决策结果，每一次划分过程遍历所有划分属性找到最好分割方式。

决策树的目标是将数据按照对应的类属性进行分类，通过特征属性的选择将不同类别数据集合贴上对应的类别标签，使分类后的数据集纯度最高，而且能够通过选择合适的特征尽量使分类速度最快，减少决策树深度。

决策树生成过程一般分为三个步骤：

- ✓ 特征选择：是指从训练数据中众多的特征中选择一个特征作为当前节点的分裂标准。如何选择特征有着很多不同量化评估标准，从而衍生出不同的决策树算法。
- ✓ 决策树生成：根据选择的特征评估标准，从上至下递归地生成子节点，直到数据集不可分则停止决策树生长。
- ✓ 剪枝：决策树容易过拟合，一般来需要剪枝，缩小树结构规模、缓解过拟合。剪枝技术有预剪枝和后剪枝两种。

3.4. 支持向量机

支持向量机 SVM（Support Vector Machine）是由 Cortes 和 Vapnik 于 1995 年首先提出的，它是一种基于统计学习的机器学习方法，在小样本分类上也能获得良好统计规律。同时由于在文本分类中表现出特有的优势，成为当时机器学习领域研究的热点。SVM 的学习方法主要包括：线性可分向量机、线性支持向量机以及非线性支持向量机。

SVM 主要思想是，建立一个最优决策超平面，使得该平面两侧距平面最近的两类样本之间的距离最大化，从而对分类问题提供良好的泛化能力。将复杂的模式分类问题非线性投射到更高维空间变成线性可分的，因此支持向量机算法在特征空间建立分类平面，可解决非线性可分的问题,其学习策略是间隔最大化，将分类问题转化为一个凸二次规划问题的求解。SVM 采用核函数技巧将原始特征映射到更高维空间，解决原始低纬空间线性不可分的问题。

设样本集为(x_i, y_i)，$x_i \in \mathbb{R}^m$ ，$y_i \in 0,1$， $i = 1,2,\dots,n$,

在线性可分的情况下，最优超平面的构建转化成下面最优化问题：

$$min \quad \omega^T \omega$$
$$s.t. \quad y_i(\omega^T \phi(x_i) + b) - 1 \geq 0$$

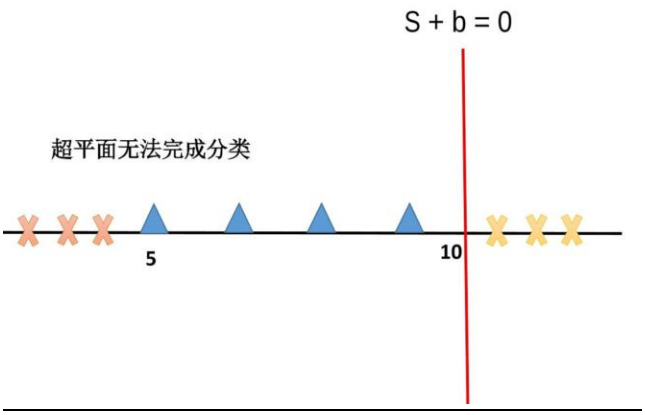
在线性不可分情况下，引入误差带宽和惩罚函数，构造并求解约束最优化问题：

$$min \quad \frac{1}{2}\omega^T \omega + C\sum_{i=1}^{i} \xi_i$$
$$s.t. \quad y_i(\omega^T \phi(x_i) + b) + \xi_i - 1 \geq 0$$
$$\xi_i \geq 0$$

其中 ω 是分类向量，b 是常数，ξ_i 是样本点 i 的松弛变量，C 是对松弛变量的惩罚系数，$\phi(x_i)$ 是将原始特征值映射到高维空间的核函数。

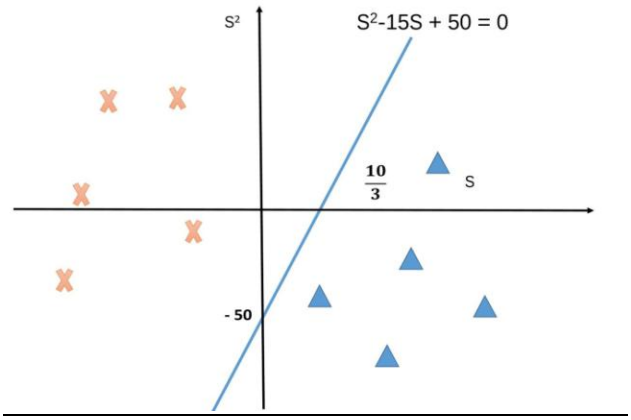
SVM 的核心问题是选取合适的核函数，将低维空间的原始特征映射到高维空间。举个例子，按照市值规模进行选股时，我们需要选取市值大于 5 亿而小于 10 亿的股票，输入市值 S 的原始向量数据，5<S<10 类别记为 1，其他情况记为-1。如果采用一维空间，如图 4，是不可能找到一个分离超平面正好能把一维空间分成两个符合要求的类别。如果将一维向量映射到二维空间 S^2 和 S，通过构建超平面 S^2-15S + 50 = 0，当 S^2-15S + 50 < 0 时，可判为属于类别 1，否则属于类别-1，如图 5 所示。

图 4：一维空间分类



数据来源：东北证券

图 5：二维空间分类



数据来源：东北证券

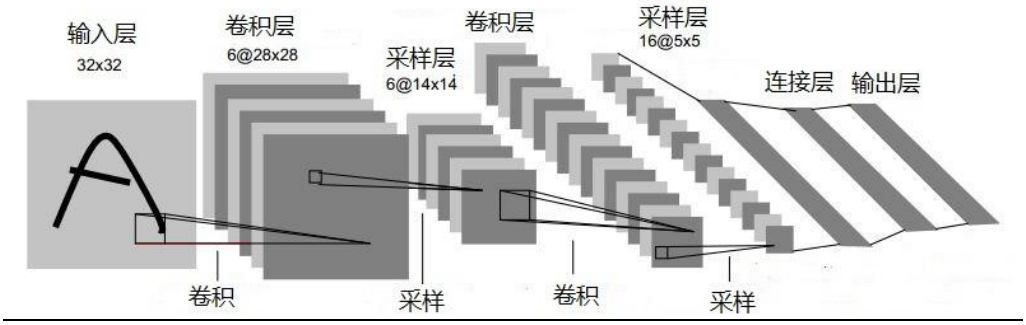
3.5. 卷积神经网络

当人工智能领域在 20 世纪 50 年代起步的时候，生物学家开始提出简单的数学理论，来解释智力和学习的能力如何产生于大脑神经元之间的信号传递。当时的核心思想一直保留到现在，**如果这些细胞之间频繁通信，神经元之间的联系将得到加强**。神经学研究表明，人类大脑在接收到外部信号时，不是直接对数据进行处理，而是通过一个多层的网络模型来获取数据的规律。这种层次结构的感知系统使视觉系统需要处理的数据量大大减少，并保留了物体有用的结构信息。由于这些信息的结构一般都很复杂，因此构造深度的机器学习算法去实现一些人类的认知活动是很有必要的。

这里主要介绍一个经典的深度学习算法：卷积神经网络（CNNs）。卷积神经网络

（CNN）是近年发展起来，并引起广泛重视的一种高效识别方法。受生物自然视觉认知机制启发而来。1959 年，休博尔（Hubel）等人发现，动物视觉皮层细胞负责检测光学信号。受此启发，1980 年福岛邦彦（Kunihiko Fukushima）提出了 CNN 的前身—神经认知机（neocognitron）。20 世纪 90 年代，燕乐纯（LeCun et al.）等人发表论文，设计了一种多层的人工神经网络，取名叫做 LeNet-5，可以对手写数字做分类，LeNet-5 确立了 CNN 的现代结构，在每一个采样层前加入卷积层。在图像识别领域，CNNs 已经成为一种高效的识别方法。

图 6: 卷积神经网络 LeNet-5 结构图



数据来源：东北证券，互联网

一般地， CNN 的基本结构包括两层，**其一为特征提取层**，每个神经元的输入与前一层的局部接受域相连，并提取该局部的特征。一旦该局部特征被提取后，它与其它特征间的位置关系也随之确定下来；**其二是特征映射层**，网络的每个计算层由多个特征映射组成，每个特征映射是一个平面，平面上所有神经元的权值相等。特征映射结构采用影响函数核小的 sigmoid 函数作为卷积网络的激活函数，使得特征映射具有位移不变性。此外，由于一个映射面上的神经元共享权值，因而减少了网络自由参数的个数。卷积神经网络中的每一个卷积层都紧跟着一个用来求局部平均与二次提取的计算层，这种特有的两次特征提取结构减小了特征分辨率。

4. 在金融领域的应用

据高盛统计，人工智能为美国金融业每年节约和新增收入 340-430 亿美元。但是，我们认为目前人工智能在国内金融领域的应用还处于探索阶段，应用场景还是集中在和大数据分析、与互联网连接紧密的领域，**我们分别列举大数据基金、人工智能预测和智能投顾三个例子来进行阐述**。

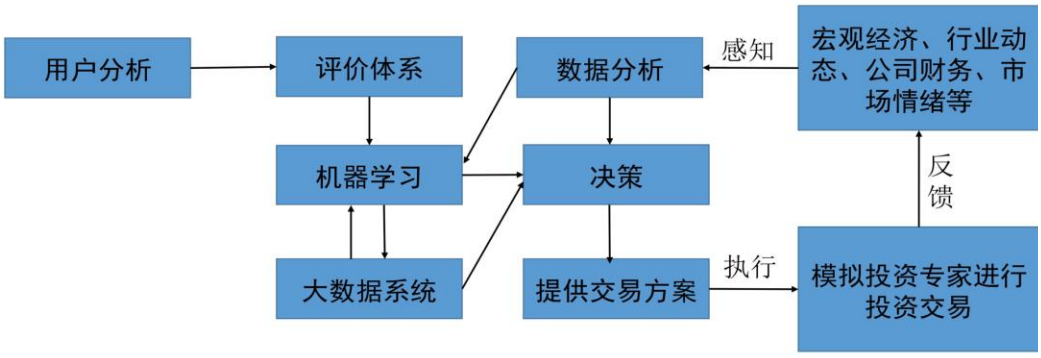
1、**大数据基金**。2012 年 5 月，世界首家基于社交媒体的对冲基金 Derwent Capital Markets 在屡次跳票后终于上线，它会即时关注 Twitter 中的公众情绪，通过大规模文本数据分析，来指导投资。其实早在上世纪 80 年代，就诞生了所谓媒体指标，通过收集电视广播和纸质媒体上关于经济和股市的看法，用算法拟合公众情绪，然后作为市场操作中的重要参考指标。事实上，大数据基金是一个相对简单的应用，**通过构建大数据因子，来反应市场或者单个股票特性对市场上的波动做出及时反应**。从目前大数据指因子的编制方法上看，可分为情绪类、行业类和专家类。其中，情绪类大数据因子适用于市场情绪高涨的牛市；行业大数据因子适用于侧重研究公司基本面的投资者；而专家类因子相当于跟随专业投资者进行调仓操作，但是会存在专家对市场误判的风险。大数据基金产品，目前国内市场上主要有淘宝大数据

100，百发 100 指数，雪球智选大数据 100 等。

2、**运用人工智能进行预测**。这属于比较前沿的领域，主要列举两个国外成功的例子。第一个例子是，瑞贝林基金，它是全球第一个纯人工智能驱动的基金，曾成功预测了 2008 年全球股市的动荡，并在 2009 年给希腊债券信用评级是 F，比惠誉调低评级还提前了一个月。第二个例子是日本三菱公司发明的智能机器算法，每月 10 号预测日本股市在 30 天后是上涨还是下跌。经过四年左右的测试，模型的正确率高达 68%。目前国内这方面的研究和实践已经有了一点基础，这几年有不少卖方研究在关注这一领域，但真正可以用于实盘交易的策略还不多见。现在有不少开放平台比如优矿、米筐、聚宽等，可以帮助投资者在线进行简单策略实现，也包括在线回测机器学习算法。

3、**智能投顾。智能投顾一大功能是为用户画像**，结合个人客户的风险偏好和理财目标、结合宏观经济、行业动态和公司财务等指标，通过大数据分析，用智能算法模拟投资专家为客户推荐合适的资管方案。智能投顾的运行如图 6 所示，一般包括用户画像、大数据分析、机器决策、交易执行等部分。我们各举一个国内和国外的例子，国内的例子是芝麻信用，用淘宝消费记录和余额宝投资数据为用户刻画出真实立体的信用画像，结合蚂蚁聚宝这些理财平台给用户提供合适的理财产品。国外的例子是，2014 年成立的日本企业艾帕卡，利用深度学习进行图像识别，通过卫星遥感图像实时分析一国经济运行状况，比如港口的吞吐量，农作物种植面积，在建工程数量等，根据这些信息辅助用户进行投资决策。

图 7：智能投顾运行模式



数据来源：东北证券

另外，有很多人担忧智能投顾的发展是不是会取代分析师的地位，是不是会代替分析师写报告等等，那么我们对此观点也是比较谨慎的。短期来看，一些人力成本较高却又技术含量低的领域，如信贷审查员等岗位可能是首先被冲击，分析师这个行业总体来说还是一个需要靠长期研究和实践，积攒经验、具有一定护城河的行业。人工智能的发展也许会帮助分析师减轻很多的工作量，运用得当，分析师可以更高效的发挥其个人主观能动性。其实大家做研究就会发现，很多时候，我们需要的是 60 分到 85 分研究价值的报告，可是为了达到这个水平，我们往往得先付出相当一大部分精力去做一些的重复的、机械性基础性的工作，而这些工作只能到达 60 分的水平，以至于留给 60 分到 85 分这种高价值工作的时间和精力反而不多。我们应该乐观的预见，未来固然是充满挑战的，但是如果运用得当，分析师是可以结合人工智能的技术以更高的效率做到比之前更好的研究。从这个意义上看，智能投顾尚

且不能代替分析师，反而能对其起到辅助作用。

5. 总结

总的来看，人工智能从诞生伊始就成为研究热点，发展过程中又不断遭受质疑，一直在曲折中前进。

经过数十年发展，人工智能研究重点从早期的符号计算迁移到神经网络，特别是近年来，随着深度学习等算法的成熟和机器计算能力的提高，人工智能在各领域的应用不断深化，目前在金融领域，就已经相继出现了大数据基金、智能预测和智能投顾等产品，但是整体仍处于探索阶段，随着近期国务院正式印发了《新一代人工智能发展规划》，人工智能上升到国家战略高度，未来人工智能在金融领域的应用具有极大的发展潜力。

本文作为人工智能系列报告的第一篇，围绕人工智能的发展历史，人工智能主要算法以及人工智能在金融领域的应用展开综合论述，力求先为投资者勾勒出一幅人工智能全景图，后续我们会继续推出人工智能系列实战篇，结合多种机器学习算法，构造量化选股框架，为广大投资者提供成熟有效的策略。

参考文献:

1、 Mitchell T M. Machine learning[M]. New York: McGraw-Hill, 1997.

2、 Osuna E, Freund R, Girosi F. Training SVM: An application to face detection[C], 1997.

3、 Big Data and AI Strategies:Machine Learning and Alternative Data Approach to Investing. J.P.Morgan，2017.

4、 周志华. 机器学习 ：Machine learning[M]. 清华大学出版社, 2016.

分析师简介:

肖承志：金融工程研究助理，同济大学应用数学硕士，2016年加入东北证券研究所。

孙凯歌：金融工程研究助理，中科院大学金融硕士，2017年加入东北证券研究所。

重要声明

本报告由东北证券股份有限公司（以下称“本公司”）制作并仅向本公司客户发布，本公司不会因任何机构或个人接收到本报告而视其为本公司的当然客户。

本公司具有中国证监会核准的证券投资咨询业务资格。

本报告中的信息均来源于公开资料，本公司对这些信息的准确性和完整性不作任何保证。报告中的内容和意见仅反映本公司于发布本报告当日的判断，不保证所包含的内容和意见不发生变化。

本报告仅供参考，并不构成对所述证券买卖的出价或征价。在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的证券买卖建议。本公司及其雇员不承诺投资者一定获利，不与投资者分享投资收益，在任何情况下，我公司及其雇员对任何人使用本报告及其内容所引发的任何直接或间接损失概不负责。

本公司或其关联机构可能会持有本报告中涉及到的公司所发行的证券头寸并进行交易，并在法律许可的情况下不进行披露；可能为这些公司提供或争取提供投资银行业务、财务顾问等相关服务。

若本公司客户（以下称“该客户”）向第三方发送本报告，则由该客户独自为此发送行为负责。提醒通过此途径获得本报告的投资者注意，本公司不对通过此种途径获得本报告所引起的任何损失承担任何责任。

分析师声明

作者具有中国证券业协会授予的证券投资咨询执业资格，并在中国证券业协会注册登记为证券分析师。本报告遵循合规、客观、专业、审慎的制作原则，所采用数据、资料的来源合法合规，文字阐述反映了作者的真实观点，报告结论未受任何第三方的授意或影响，特此声明。

投资评级说明

股票投资评级说明	买入	未来 6 个月内，股价涨幅超越市场基准 15%以上。
	增持	未来 6 个月内，股价涨幅超越市场基准 5%至 15%之间。
	中性	未来 6 个月内，股价涨幅介于市场基准-5%至 5%之间。
	减持	在未来 6 个月内，股价涨幅落后市场基准 5%至 15%之间。
	卖出	未来 6 个月内，股价涨幅落后市场基准 15%以上。
行业投资评级说明	优于大势	未来 6 个月内，行业指数的收益超越市场平均收益。
	同步大势	未来 6 个月内，行业指数的收益与市场平均收益持平。
	落后大势	未来 6 个月内，行业指数的收益落后于市场平均收益。

东北证券股份有限公司

中国 吉林省长春市
生态大街6666号
邮编：130119
电话：4006000686
传真：(0431)85680032
网址：http://www.nesc.cn

中国 北京市西城区
锦什坊街28号
恒奥中心D座
邮编：100033
电话：(010)63210800
传真：(010)63210867

中国 上海市浦东新区
杨高南路729号
邮编：200127
电话：(021)20361009
传真：(021)20361258

中国 深圳南山区
大冲商务中心1栋2号楼24D
邮编：518000

机构销售

华北地区
销售总监 李航
电话：(010) 63210890
手机：185-1501-8255
邮箱：lihang@nesc.cn

华东地区
销售总监 袁颖
电话：(021) 20361100
手机：136-2169-3507
邮箱：yuanying@nesc.cn

华南地区
销售总监 邱晓星
电话：（0755）33975865
手机：186-6457-9712
邮箱：qiuxx@nesc.cn