北京大学

硕士研究生学位论文

题目：以大数据之史鉴

人工智能之未来——案例

研究和实证分析

姓　　名：	魏　成
学　　号：	1401214425
院　　系：	国家发展研究院
专　　业：	西方经济学
研究方向：	宏观经济学
导师姓名：	胡大源　教授

二〇一七　年　五月

版权声明

摘要

回顾大数据从掀起热潮到潮水退去短短几年的历史，我们可以看到大众总是对新的技术前沿抱有不切实际的幻想。大数据带来并不是一场对传统的统计分析方法的替代革命，而是建立在非结构化数据分析上的补充。

同样地，目前站在风口的人工智能建立在问答系统和深度学习这两方面算法与应用进步的基础上，与人们过高的预期相异的是目前的人工智能仍然处于比较低级的阶段。本文梳理了 IBM Watson 和 Google AlphaGo 背后的技术细节，可以看到目前的技术进步仍然离想象中无所不能的人工智能相差甚远，在人工智能的应用场景我们离不开关注边际成本和其应用带来的边际收益。

之后本文利用 2016 年“中国经济生活大调查”数据建立预测“幸福感”的模型，并利用 Logistic 回归的预测结果为基准，比较多层人工神经网络模型的成本与收益。在耗费大量的计算成本之后，模型性能提升并不显著。再一次印证任何人工智能场景的利用都离不开成本的考量和替代收益的分析。

最后，虽然在图像识别和语音识别等领域识别出稳定特征的技术进步明显，但其技术应用上的局限仍不小，我们需要冷静看待这一次人工智能的热潮。

关键词：大数据，人工智能，深度学习

A Lesson from History of Big Data and the Future of Artificial Intelligence: Case Study and Empirical Analysis

Cheng Wei (Western Economics)
Directed by Professor Dayuan Hu

ABSTRACT

Looking back the few year history of the surge and ebb of the Big Data tide, we can see that the public always has an unrealistic fantasy about the new technology frontier. Large data is not an alternative revolution to traditional statistical analysis methods, but rather a supplement to unstructured data analysis.

Similarly, the prevalence of artificial intelligence is based on the establishment of the Q & A system and the deep learning in both the algorithm and application of the progress. Contrary to general expectation, the current artificial intelligence is still in a relatively low stage. This articles study the technical details behind IBM Watson and Google AlphaGo, and it can be seen that current technological advances are still far from imaginative artificial intelligence. We need pay attention to the trade-off between of the marginal benefit and the marginal cost when we think the scenario the application of artificial intelligence.

Then we use the 2016 "China Economic Life Survey" data to establish the model to predict "happiness". With the benchmark of Logistic Regression, we can compare the cost and benefit of multi-layer artificial neural network model. After a lot of computing costs, the model performance is not significant. Once again confirmed that the use of any artificial intelligence scene are inseparable from the consideration of cost and benefit.

In a word, although the technical progress in identifying areas such as image recognition and speech recognition is obvious, but we need to consider calmly the limitation of artificial intelligence in the mania of the boom of AI.

KEY WORDS: Big Data, Artificial Intelligence, Deep Learning

目录

第一章　引言

1.1　研究背景

随着网络时代的来临，互联网、物联网和各种传感器等产生了大量的数据。同时，计算能力持续提高、同时存储成本不断降低，使得大量的数据处理和应用在硬件上具备条件。在这些大量数据中，除了数字和符号等结构化数据之外，更多的是全文文本、图像、声音、影视和网络媒体等非结构化数据。根据 Enterprise Strategy Group 估计，2015 年全球归档非结构化数据 226716PB(1PB=10^6GB)，而结构化数据只有 32188PB，前者占比高达 87.6%。而可以对海量数据进行分布式处理的 Hadoop 技术的成熟和数据挖掘与机器学习算法的进展使得大量的数据处理和应用在软件上具备条件。

大数据的应用，使得发展停滞不前的人工智能重新兴起。以 IBM Watson 为代表的专家系统开始在医学界大放异彩；以 Google AlphaGo 为代表的深度学习在围棋上战胜之前认为不可战胜的人类专家；以新能源和自动驾驶技术为核心的 Tesla 开始重新定义汽车，其市值一度超过美国最大的汽车公司通用汽车；在金融市场上，量化交易的交易量远远超过人类；在法律领域，从事大数据分析的日本 UBIC 公司利用人工智能技术使得人们在法律诉讼时摆脱了繁杂的证据阅览业务（松尾丰，2016）。人工智能已经从各个领域进入了人们的生活，也极大地提高了生产效率。

1.2　研究目的与意义

2011 年，麦肯锡公司发布大数据的专题报告，认为大数据是创新、竞争和生产力的下一个前沿，对大数据的发展寄予厚望。回过头来看，与人们的过高期待不同的是大数据最大的进步意义来自数据处理算法的进步。而现在，人工智能同样地吸引了全世界的注意力，世界各国对人工智能的投资正在不断增加，以人工智能为卖点的初创公司如雨后春笋。与此同时，担忧的声音也不绝于耳，比尔·盖茨认为应该向代替人类工作的机器人征税，斯蒂芬·霍金甚至警告人工智能的全位发展可能导致人类灭亡。

因此，本文试图冷静地分析大数据的发展历史，来揭示人们预期过高的大数据是什么、取得了什么进展、这些进展怎么应用。我们以史为鉴，按照同样的思路分析人工智能是什么、目前人工智能取得了什么进展、这些进展怎么应用。我们试图从最基本的边际成本收益的角度去冷静地看待时下火热的人工智能。在冷静客观地分析完人工智能的“来龙”之后，我们希望能够得出人工智能的“去脉”。

1.3 本文研究思路

本文第一章简要介绍了大数据和人工智能时代特征，并简要交代了本文的选题目的和意义。

第二章研究了大数据的发展历史、大数据的本质进步。并以谷歌流感模型(Google Flu Trend, GFT)为案例来说明在大数据的进步意义之下，我们仍然要谨慎面对其缺陷。

第三章研究了人工智能到目前阶段的发展历程，也就是人工智能的“来龙”，我们既会从宏观视角来解释人工智能是什么，并以目前人工智能的两大方向专家系统的代表 Watson 和深度学习的代表 AlphaGo 为案例从微观视角来分析其具体的技术细节和进展。

在第四章，我们通过具体应用神经网络的算法做了一个幸福感预测的实证研究，并与传统的计量方法做比较，从边际成本和收益角度来分析其优劣。我们认为，在数据量并不是很大的时候，可以看到神经网络的算法与传统方法并没有明显优势，但却有着巨大的运算成本。在数据量大到足以应用深度神经网络的时候，我们仍然需要考虑其边际成本与收益。

在第五章，我们阐述本文得到的主要结论与意义。

第二章　大数据的发展与应用

2.1　大数据的兴起

随着最早提出大数据(Big Data)时代到来的是麦肯锡：“数据，已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素，人们对于海量数据的挖掘和应用，预示着新一波生产率增长和消费者盈余的到来”。波士顿咨询认为大数据必须要满足数量大、维度高、种类多和价值高等四个特征。

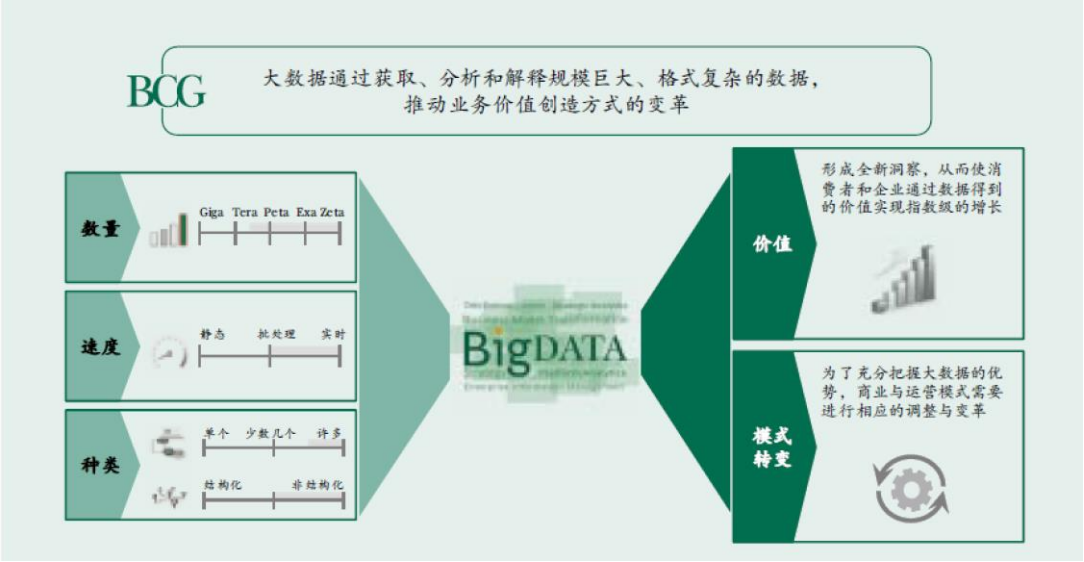

图 2.1 波士顿咨询对大数据的定义

大数据从诞生那天起，就被人们寄以厚望。在 2011 年麦肯锡的报告中，认为大数据在医疗、公共部门、零售、制造业和个人位置数据等五个领域将有着很大的潜在应用价值。其中在医疗方面，大数据通过药品研究数据、临床数据、成本数据和病人行为和情绪数据等数据，来减少医药成本并提高医疗系统的效率。在临床方面：医生可以在研究中选出更加有效的方式、建立起临床支撑系统、提高医疗数据的透明度、实现远程病人监视和病人病历的高级分析方法；在定价方面：可以避免错误诊断的自动化系统、实现基于表现和健康经济学的定价方案；在研究和开发方面：更好的分配研究资源的提前预测模型、临床测试计划的统计工具和算法、更好地分析临床测试数据、精准个人医疗和分析疾病特征。麦肯锡预期大数据将在医疗方面可以每年减少支出 2200 至 3300 亿美元。在公共部门方面，大数据可以用于提高公共部门的效率。大数据可以为公共部门创造极大的透明性，极大地为税务部门提供便利性，为政府部门复杂的决策提供极大的数据支持。在零售方面，大数据从市场调查、商品生产、运营、供应链等方面提高行业的效率。在制造业方面，数据一直不断地用于提高产品质量和效

率。利用大数据提高研发能力、管理供应链，利用数字工厂更好地管理物料、利用物联网的数据实时监控生产过程。在个人位置数据方面，用大数据可以更好地在电信、零售和媒体行业取得很好的应用。这些领域将在十年之内为服务提供商创造超过 1000 万潜在收入、为消费者和终端用户创造大约 7000 亿美元潜在收入（Mckinsey，2011）。

而美国很早就开始研究和使用大数据，美国政府也大力支持大数据的技术研究和应用，并开放政府数据。2012 年 3 月，奥巴马政府宣布投资 2 亿美元启动“大数据研究和发展计划”，旨在提高从大数据中获取信息的能力，加速在科学和工程领域的进步。2013 年 11 月，奥巴马政府推出“数据·感知·行动”计划，进一步细化了利用大数据改造国家治理、促进前沿创新、提振经济增长的路径。2014 年 5 月，美国总统办公室提交《大数据：把握机会，维护价值》政策报告，强调政府部门和私人部门紧密合作，利用大数据最大限度地促进增长和利益，减少风险。

除了美国政府以外，其他主要国家也通过开放数据，支持大数据产业的发展，通过大数据分析更好地提供公共服务和基础设施。在这些政府的战略目标里，大数据是一种宝贵资源，通过大数据分析可以提高生产效率和生产力水平；同时可以帮助更好地进行政府管理，提高政府服务质量。

同样的，中国政府也不遗余力地推动大数据在国内的发展，寄希望通过大数据产业发展，建立起各个政府部门之间的互联共享信息平台，运用大数据提升监管水平，在健康医疗、公共事业、服务贸易等领域应用大数据提高效率。2015 年 8 月，国务院通过《关于促进大数据发展的行动纲要》，该文件主要目的是推动政府信息系统和公共数据互联共享，顺应潮流引导支持大数据产业发展和强化信息安全保障并完善产业标准体系。

2.2 大数据热潮的消退

在 2014 年和 2015 年两年间，不但政府部门对大数据产业的发展寄予厚望，企业部门对大数据的火热也不相上下。2014 年，在咨询机构 Gartner 的新兴技术曲线上，大数据几乎处于狂热的顶点。美国做大数据的上市公司例如 Teradata 和 Splunk 的市值也达到了历史高点。但在 2015 之后，大数据消失在了 Gartner 的新兴技术曲线上。

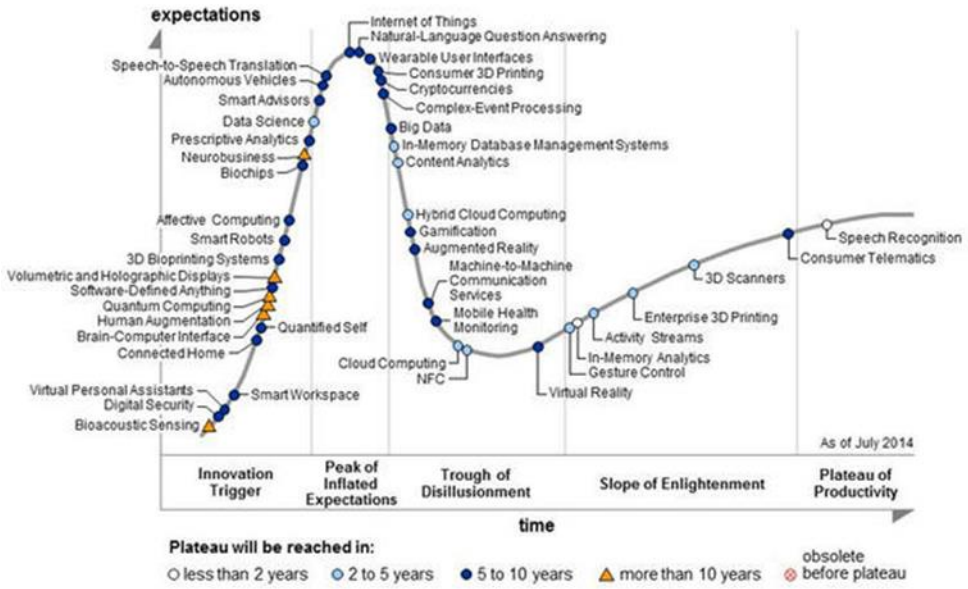

图 2. 2 Gartner 2014 新兴技术曲线

同时我们观察谷歌趋势，从 2015 年之后热度就停滞不前。而以灵活、敏捷的大数据管理能力为企业带来一场 IT 管理革命的 Slpunk 公司股价也从 2014 年 3 月份的最高的 106 美元一度下跌至 30 美元。毫无疑问，不管是学术界还是资本市场都从当时的狂热中冷静下来。

图 2. 3 谷歌趋势中的大数据热度

我们在反思这一场热潮的时候，需要回归大数据的本质。正是因为存储成本的下降和计算能力的提高使得我们在硬件上具备了处理大数据的硬件基础。而互联网和物联网时代产生了大量的图片、音频、视频、文本、XML 和 HTML 等等大量非结构化数据。相对于结构化数据来说，这些非结构化数据信息含量密度小且分析精准度较低导致其信息挖掘的价值较低；而可用的分析方法少又导致其处理成本较高。但随着 Hadoop 技

术的成熟以及机器学习算法的进步使得对非结构化数据的处理利用的边际成本低于其边际收益，因此大数据开始被广泛利用。可以看到，能够直接被计算程序处理的结构化数据分为属性数据和数值数据两种，而一般的非结构化数据通过类别分析和关联关系等数据挖掘的方法整理成属性数据以被程序使用。大数据的应用让人们开始憧憬其带来的变革，伴随着媒体的炒作，这一场热潮自然而然就产生了。而这个热潮从兴起到消退的生动写照，就是谷歌流感模型。

2.3 GFT 背后的大数据模型与传统的统计模型对比

谷歌流感模型是谷歌利用搜索引擎的大数据构建的一个通过搜索引擎查询数据来检测流感趋势的一个系统。对于这个每年导致千万人次的呼吸道疾病和 25 万至 50 万人死亡的季节性流感问题，谷歌希望能通过这个模型提前检测到流感活动，并帮助政府部门更快地介入和控制。通过监控世界各地数以百万计的用户每天提交的在线网络搜索查询健康的数据，用此模型跟踪人群中可能类似于流感疾病的发生情况，因为某些查询的相对频率与患者呈现流感症状时医生访问的百分比高度相关，通过这种相关性，谷歌可以估计美国每个地区每周流感的当前水平。

在分析此模型之前，我们有必要了解传统的流感监测系统是如何运行的。传统的监测系统，类似于美国疾病控制和预防中心(U.S. Centers for Disease Control and Prevention, CDC)和欧洲流感监测计划(European influenza Surveillance Scheme, EISS)主要依赖的是病毒学和临床数据，包括疑似流感(Influenza-like Illness, ILI)医生访问等，这些数据由 CDC 进行周度发布。在这些传统的监测方法之外，有一些监测系统通过电话治疗热线的呼叫量或者非处方药的销售量等数据来间接监测流感的爆发程度。

谷歌（Ginsberg& Mohebbi，2009）通过网络搜索日志中的数百亿次个人搜索记录和美国的 ILI 数据建立起用于流感监测的模型。通过 2003 年至 2008 年提交的网络搜索数据的汇总，研究人员计算了美国常见的 5000 万条搜索查询的周度查询次数的时间序列数据，接着用一个模型来估计医生随机访问疑似流感疾病地区的概率，这相当于 ILI 相关的医生访问的百分比。模型只有一个解释变量，也就是提交自 ILI 相关地区的随机搜索数据的比例。

$$logit(P) = \beta_0 + \beta_1 \times logit(Q) + \varepsilon$$

P 是 ILI 医生访问，Q 是 ILI 相关查询比例，β_0是截距项，β_1是相关系数，ε是误差项。$logit(P)$是 P/(1-P)的自然对数。

数据来源是 CDC 的美国流感健康网络的历史数据，CDC 为全美 9 个监测区域报告每周 ILI 相关的门诊访问的平均百分比。在不需要流感的先验知识的情况下，谷歌设计了一种选择 ILI 相关搜索查询的自动化方法。研究人员测试了如果只使用单个查询作为

解释变量 Q 的情况下，模型如何有效地拟合每个区域中的 ILI 访问数据。将数据库中 5000 万个候选查询以这种方式测试，识别出哪些搜索查询可以最准确的拟合每个区域的 ILI 访问比，并赋予拟合程度高的搜索查询以更高的权重分数。通过自动化的查询选择程序之后，生成了一系列高分搜索序列。再通过这些高分搜索序列进行组合，选择哪些序列添加到相关搜索的解释变量 Q 中，并用样本外的数据进行验证。最后，通过 ILI 相关的搜索比例作为解释变量，谷歌的模型与 CDC 报告的 ILI 比例有着很高的拟合程度，达 0.90。

在 2007-2008 年度的流感季节的时候，谷歌通过这个模型预测 ILI，并与 CDC 的数据做比较，结果表明，谷歌的模型通常能够领先 CDC 的流感监控网络 1 到 2 周。在 2009 年春季甲流爆发时，GFT 因为成功提前预测到疫情，为决策者提供了及时的疫情信息，开辟了大数据基础上疫情防治决策支持新途径。谷歌专门为 GFT 建立网站宣传，口号就是“Track Influenza Faster Than the CDC”，并为世界各地研究者提供其模型。

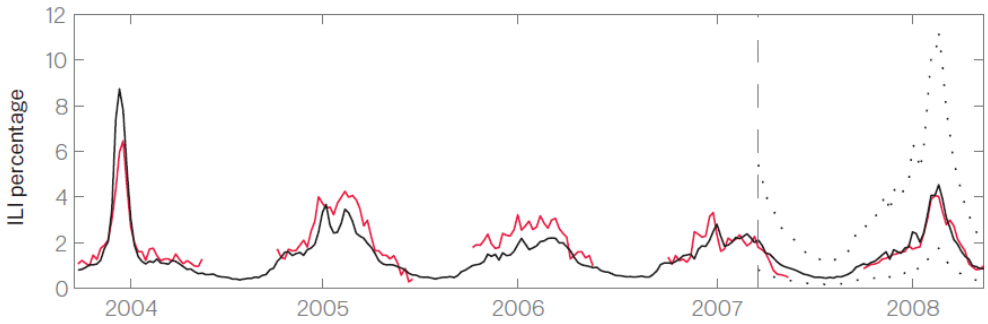

图 2. 4　模型预测(黑色)和 CDC 报告的 ILI 比例(红色)对比

但好景不长，从 2011 年 8 月至 2013 年的 9 月的 108 周中，GFT 对于流感的预测有 100 周是偏高的（Declan，2013）。GFT 为什么会犯这种错误呢？这个问题的关键一方面在于搜索或社交媒体上的数据是否可以作为自变量去做预测。之前，人们认为大数据将是传统数据收集和分析的替代而不是补充方式。对于 GFT 来说，用 5000 万条搜索词条去拟合 1152 个数据点存在着过度拟合的可能性。通过不存在结构性相关的数据来拟合流感的潜在可能性，无法准确地预测未来的可能性的概率非常大。比如说搜索词条“high school basketball”与 CDC 数据非常相关。事实上，不但在预测上大部分时间偏高，GFT 也没有预测到一些大的流行感冒。这些错误不是随机的，预测错误之间期间相关性很高，同时预测错误也有着明显的季节性。这说明 GFT 忽视大量可以通过传统的计量模型获取的信息。而有研究人员证明 GFT 的预测效果甚至不如用简单的滞后 3 周的 CDC 数据模型（Lazer & Kennedy, 2014）。

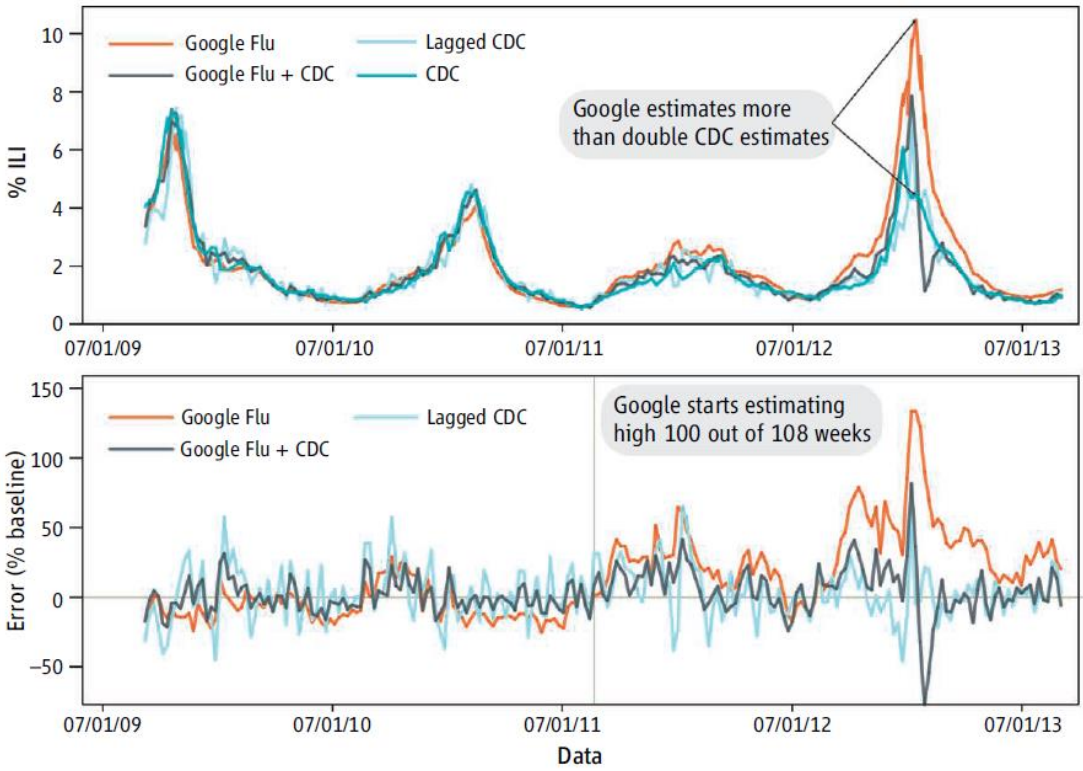

图 2. 5 GFT 和 Lagged CDC 在预测性能上的对比

数据来源：D Lazer，R Kennedy，G King，A Vespignani(2014)

另外一方面，所有的实证研究的基础是度量问题。这些变量是否有理论基础、度量方式是否在样本中保持稳定且可比较、度量误差是否是系统性等等都是需要重点关注的。但是因为谷歌本身的商业服务使得其搜索算法经常做出一些改变，而本身 GFT 的数据来源就是搜索词条。同时，因为大众传媒的炒作使得搜索词条有着自我加强的趋势。以上原因都使得 GFT 的数据的可靠性值得商榷。因此，GFT 出现的严重偏差提醒我们新方法对于传统方法来说，只能成为补充，并不能成为替代。

在传统的计量经济学研究里，为了获取感兴趣的自变量的标准差，我们的模型需要利用固定效应模型来考虑异质性或者其他因素。并且通过大量的检验方式来验证模型的鲁棒性。然而，应用大数据的机器学习的方法更加关注的是样本外的预测效果，而不是关注标准误。同时，传统的模型更加关注因果关系，而这是大数据模型所欠缺的。虽然大数据为我们理解人类在社交层面上的有着丰富的空间和时间动态性的交流互动，为我们发现变量之间的非线性关系等等都提供了更大的可能性。但是数据量大并不意味着我们可以忽略数据度量的基本问题以及数据的真实性、可靠性和内在关联。因此，大数据的研究的深入发展不代表传统的统计模型就可以被放弃。

回过头来看，大数据并不是想象中那样改变了人类认识和思考世界的方式，我们在接受数据的混杂性的同时，也没有放弃对精确性的追求；我们在关注相关关系的时候，

因果关系的分析仍然占据举足轻重的位置。大数据带来的更多地是数据处理算法的进步，因此在过高的预期之后，人群终于明白大数据带来的进步背后也有着不可忽视的缺陷。同样，在面对着人工智能的热潮的时候，我们要吸取大数据的经验教训，去思考人工智能是什么，目前取得了什么进展，这些进展可以应用在什么方面？

第三章　深入分析人工智能和深度学习

3.1　人工智能的发展

对于计算机专家来说，让计算机拥有类似于人类的智能是一直以来的梦想。尽管人工智能(Artificial Intelligence)这个专业术语在 1956 年才被发明出来，专家们的努力在 1940 年代就已经开始了。在阿兰·图灵（1950）中，文章提出一个问题，计算机能进行思考吗？并给出图灵测试的定义。人们开始思考计算机可以像小孩子一样通过经验进行学习的可能性。

在接下来几十年，AI 领域的研究进展几上几下，研究方向也分为三派：(1)符号学派：以数理逻辑的形式来研究人工智能，在启发式算法的基础上发展了专家系统。(2)连接学派：以模仿人脑中的神经网络来研究人工智能，从感知器到多层神经网络再到深度神经网络，目前的突破主要集中在此方面。(3)行为学派：用控制论把神经系统的工作原理和控制理论联系起来。

事实证明，在研究的过程中会遇到很多超过预期的难题并在当时的技术水平下有很多难以逾越的难题。直到 1990 年代，人工智能专家开始缩小研究范围，专注于人工智能的某一领域比如说图像识别和疾病诊断等等之后，AI 的研究进程才重新开始加速。1997 年，人工智能发展迎来里程碑，IBM 的深蓝(Deep Blue)战胜国际象棋冠军 Garry Kasparov。

从 2010 年左右开始的这一轮的人工智能的热潮主要有三个因素，大数据的发展，为机器学习算法的研究提供了充足的数据，同时这些进展又依靠计算性能更加出色的计算机。

2016 年，Google 的 DeepMind 团队利用深度神经网络构建的 AlphaGo 击败了人类围棋选手李世石，震惊世人。之前，人们认为围棋是人类智慧的最后一个堡垒，机器无法很快就攻克。AlphaGo 的成功，让人们开始意识到人工智能在不声不响中侵入了原本属于人类的领域。有人在担心人工智能会带来社会伦理问题。

对于“机器取代大部分人的工作，最后仅仅让社会最上层的人获利”的担忧一直就有。从英国的工业革命开始，这类的担忧就不绝于耳。大卫·李嘉图在 1821 年就提出应该关心机器对社会不同阶层的影响，尤其是对于劳动阶层。托马斯·卡莱尔在 1839 年写道机器恶魔将打翻整个劳动阶层。今天，这个问题看起来更加紧迫了。我们已经看到，人工智能的发展在起起伏伏之后，最近又取得了非常大突破，机器已经具备从大量的数据中获取并运用知识的能力，如果机器具有了人类的意识，再凭借其远超人类的计算能力和记忆能力，机器将在各方面碾压人类。

不管人们对于人工智能持肯定还是否定态度，人工智能的崛起是不可阻挡的历史潮流，人工智能的应用也越来越广泛。无人驾驶技术逐渐成熟、语音语义识别的精度不断提高、图像识别技术发展迅速、自然语言处理上的难题也不断被攻克。目前人工智能的两大发展方向是认知系统和深度学习。接下来，我们将以 IBM Watson 和 Google AlphaGo 分别讲解认知系统和深度学习。

3.2　认知系统与 IBM Watson

沃森(Watson)是一种认知计算系统，具有强大的理解能力、逻辑思考能力和学习能力。Watson 本质上是 IBM 制造的电脑问答(Q&A)系统，Watson 是一个集高级自然语言处理、信息检索、知识表示、自动推理、机器学习等开放式问答技术的应用，基于为假设认知和大规模的证据搜集、分析、评价而开发的深度问答技术。它的思路并非深度学习，而是更接近心智社会(Society of Mind)，是人工智能认知计算的范畴(Cognitive Computing)。

Watson 的工作流程如下（IBM，2012）：

(1) 当一个问题提交给 Watson 之后，它从语法上分析这个问题并提炼出问题的主要特征。

(2) 通过检索语料库中有潜力拥有有价值回应的段落来生成一组假设。

(3) 通过使用各种推理算法对问题和潜在答案的语言进行深入的比较。这一步时最具有挑战性的，数百种推理算法从不同角度进行比较。例如，一些算法看术语和同义词的匹配程度，一些算法看时间和空间上的特征，而另外一些看上下文信息的相关来源。

(4) 每一个推断算法生成一个或者多个得分，表明可以在多大程度上基于该算法覆盖的领域通过问题推导出潜在答案。

(5) 接着所有的得分将会通过统计模型进行加权。统计模型在 Watson 的训练阶段会评价算法在相似段落之间建立推断联系的性能，告知 Watson 在多大的置信水平上有证据认为候选答案是正确的。

(6) 对于每一个候选答案重复进行这个过程，直到出现候选答案的回馈比其他都更强。

对于 Watson 来说，最重要的莫过于知识语料库(Knowledge Corpus)，其包含了各种各样的非结构化知识，例如教科书、指南、手册、常见问题解答、福利计划和新闻等等。Watson 会将语料库的内容进行消化和整合成更加容易处理的形式，并会删除过时、不相干或者不可信的内容。因此，理解自然语言和处理非结构化数据是构建其知识语料库的基础。在这个基础上，Watson 通过假设生成将语料片段连接和整合，通过基于

证据为基础的学习能力，从大数据中提取关键信息，像人类一样拥有认知能力。

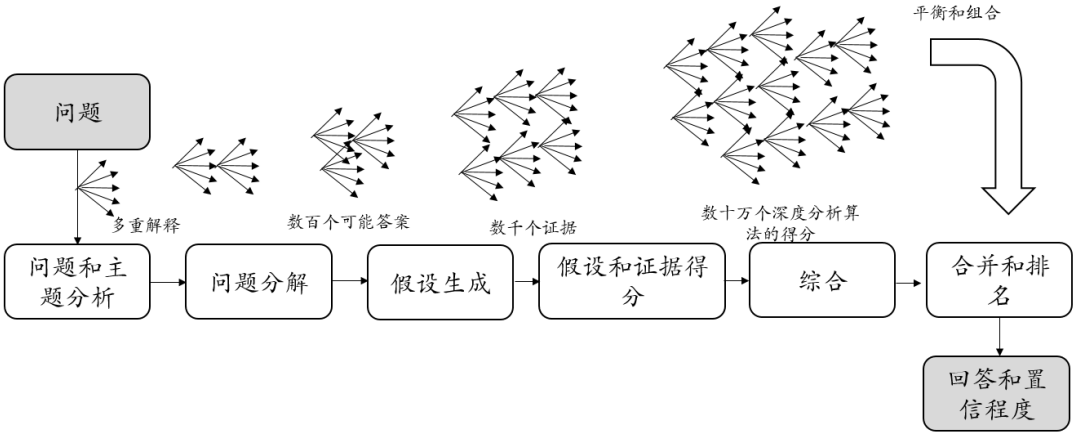

图 3.1 Watson 工作流程

从 Watson 来看，认知系统凭借其出色的理解、推理和学习能力，创造了巨大的应用价值。类似于人类，认知系统可以收集、记忆和回忆信息，就像人类的记忆一样。认知系统同样拥有交流和行动的基本能力，这些能力由以下基本行为构造：

(1) 假设生成和假设检验的能力。

(2) 对语言的分割和生成推断的能力。

(3) 提取和评估有用信息的能力。

如果将认知系统分解成关键要素，则下图中深色表示目前的认知系统所拥有的能力，浅色表示未来的认知系统所要具备的能力。

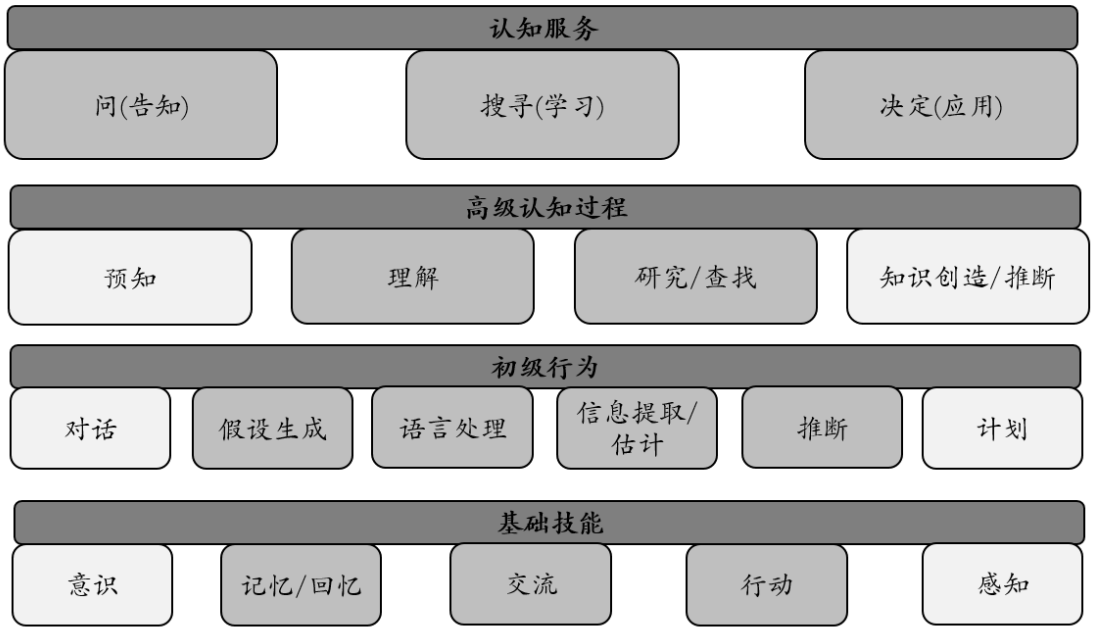

图 3.2 认知系统关键要素

未来的认知系统将与之前的计算机有着本质区别。为了研制出具有真正的思考型机器，未来的感知系统将从多种视觉、听觉、嗅觉、触觉和味觉等等感应技术中整合信息。这些感应任务需要处理大量的数据。人脑可以消耗极低的能量去高速运转 220 亿个神经元，但目前的计算机却无法做到。为了更加高效地处理数据，人们在硬件上也在不断进行研发，更加高效低能耗的计算机芯片是必不可少的条件。在 Watson 进行危险边缘的比赛时，耗能高达 85000 瓦，并且需要大量的空调进行冷却，而目前的 Watson 只有三个比萨盒子叠放的大小。

2014 年，IBM 研发了真北(True North)神经元芯片，该芯片抛弃传统计算机的内核设计不同，而是以类似于人脑的神经突触内核架构。TruneNorth 拥有 4096 个核心，每个核心约有 120 万个晶体管，其中大部分晶体管用来做数据存储和与其他核心沟通，只有少数晶体管执行数据处理和调度的任务。每个核心与其他核心的通讯方式通过模拟电路信号模拟人脑神经元与突触之间的化学信号。

认知芯片和 Watson 是互补性的技术。可以将两者看成认知系统的左脑和右脑。Watson 是左脑，主要负责语言和分析；认知芯片是右脑，负责感知。未来，IBM 的科学家希望 Watson 和 TrueNorth 能够整合成拥有意识和感知的认知系统。

3.3 深度学习与 AlphaGo

3.3.1 神经网络与深度学习

在介绍神经网络之前，有必要先介绍一下机器学习。机器学习就是通过算法，使得计算机能够从大量历史数据中获取规律，从而对新的样本做智能识别或对未来做预测。按照学习分类，主要包括两类。一类是监督学习(Supervised Learning)，通过训练数据得到模型，然后用标签和模型的预测数据之间的误差不断调整和完善模型，直到可以获得比较好的准确率，主要方法有神经网络、SVM 等等。另外一类是非监督学习(Unsupervised Learning)，与监督学习最大的区别在于此类学习没有标签可以告诉模型哪些数据是正确的，哪些是错误的。因此，非监督学习的模型主要用在关联规则的学习及聚类中。

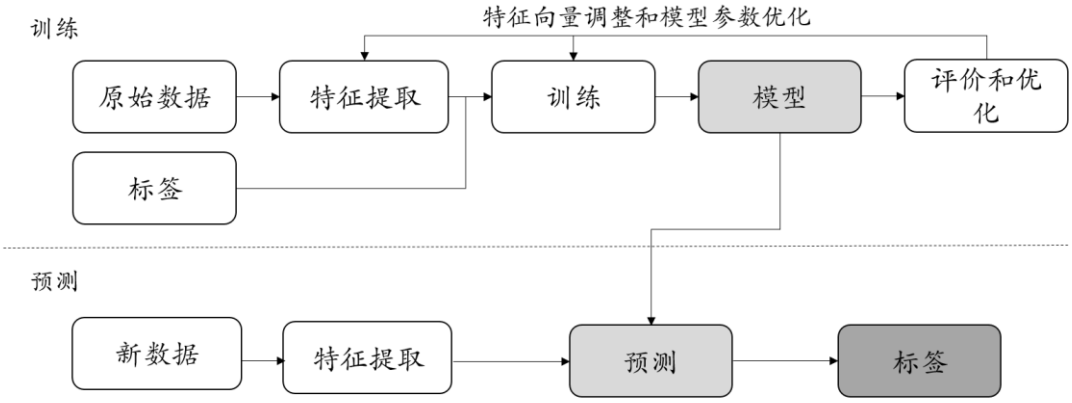

图 3.3 监督学习的流程

神经网络是机器学习中目前进展迅速的一种方法，也是深度学习的基础。神经网络建立在现代科学的基础上，旨在利用抽象数学模型来模拟人脑结构和功能。其理论基础 MP 模型（W. McCulloch &W.A Pitts, 1943）。在多年发展之后，人工神经网络的研究取得了重大进展，多种模型被提出：感知机模型（Rosenblatt, 1958），该模型可以用来线性分类；Hopfield 网络（Hopfield, J. J. & Tank, 1985），这种循环神经网络可用于联想记忆；Boltzman 机（Ackley, Hinton& Sejnowski, 1984），一种基于统计力学的随机神经网络；反向传播网络（Rumelhart & Hinton,1986），这是应用最广泛的神经网络模型，深度学习发展的基石理论。

神经网络的基本单元是神经元模型，该模型由一组用权值对各输入信号进行加权求和并加上阈值后利用一个非线性激活函数将输出映射到一定范围。不同的权重代表对不同输入信号员的敏感程度。当输入信号超过一定阈值之后才将信号输出，这也是激活函数的来源。

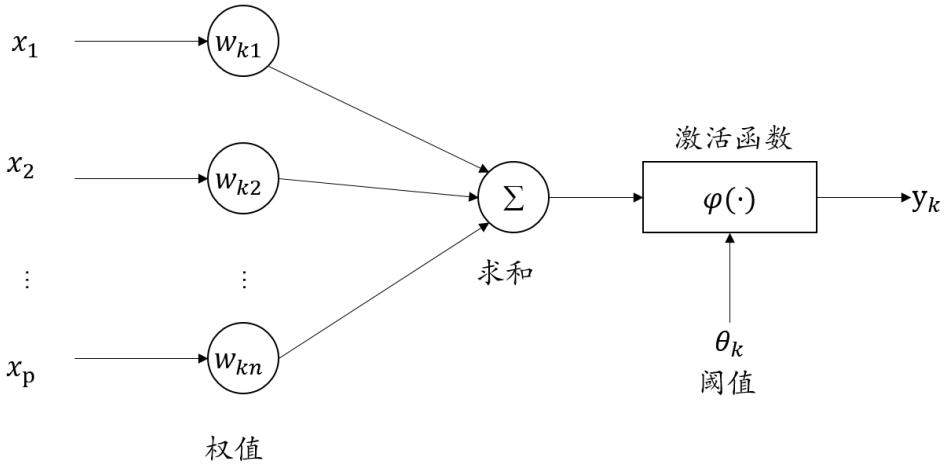

图 3.4 神经元的基本结构

上图为神经元的结构，其中$\vec{x}$为输入信号，$\vec{w}$为权重，φ为激活函数。将阈值当成输入信号输入固定为 1，则模型可以表示为下列公式：

$$y_k = \varphi(\sum_{j=0}^{p} w_{kj} * x_j)$$

目前常用的激活函数包括 Sigmoid 函数、tanh 函数和 Relu 函数。

1958 年，Rosenblatt 提出由一个输入层和输出层的神经网络，称之为感知机，虽然感知机只能做简单的线性分类任务，但因为这是首个可以学习的人工神经网络，所以当时引起了轰动。1969 年，Minsky 用数学证明了了感知机的弱点，例如无法完成异或这样的分类任务。虽然计算层增加到两层之后可以解决类似的问题但当时缺乏有效的训练算法，因此接下来十年人们放弃了神经网络的研究。

1986 年，反向传播(Back Propagation, BP)算法的提出解决了双层神经网络训练过程中的复杂计算量问题，该算法的核心思想是将误差方向传播给各层的所有单元。但因为从顶层往下传播时，误差校正值越来越小，如果网络层次比较深的话，校正信号越来越弱且容易陷入局部最优值点处。同时，该算法需要数据带有标签，否则无法进行误差校正，但现实生活中的大部分数据是没有标签的。不适用于深层次的神经网络的训练且无法训练不带标签的数据这两个缺陷制约了 BP 神经网络在实际中的应用。最后，训练神经网络仍然需要耗费大量的时间且在优化的过程中容易落入局部最优解。因此，在 1990 年代后期，神经网络和反向传播算法被机器学习、图像识别和语音识别等方向的研究者所弃用。

但在 2006 年，加拿大高级研究所(Canadian Institute for Advanced Research, CIFAR)重新开始研究前馈神经网络。研究者用非监督学习来构建特征发现层，这一可以重构底层的特征输入，在缺乏标签数据的手写数字识别上取得了很好的效果。这也是深度

学习的核心思想，将低层特征映射到更加抽象的高层特征。这就是深度神经网络，通常有多层隐藏节点，并强调特征学习的重要性，即将样本在原空间的特征映射到新的特征空间。与传统的神经网络用人工规则构造特征的方式不同的是，深度学习利用大数据来自动学习特征，因此可以从多个维度去捕捉数据之间丰富的内在信息。第一个真正成功的多层神经网络算法是卷积神经网络(ConvNet, CNN)，这种算法在图像识别领域取得了很大的成功。而在自然语言处理中，递归神经网络(Recurrent Neural Network, RNN)取得了很大的成功，RNN 可以很好地处理词向量表达、语法检查等等问题，很好地应用在文本生成、语音识别和机器翻译等领域有了较大的进展。由于篇幅有限，本文只对卷积神经网络的结构做简单的介绍。

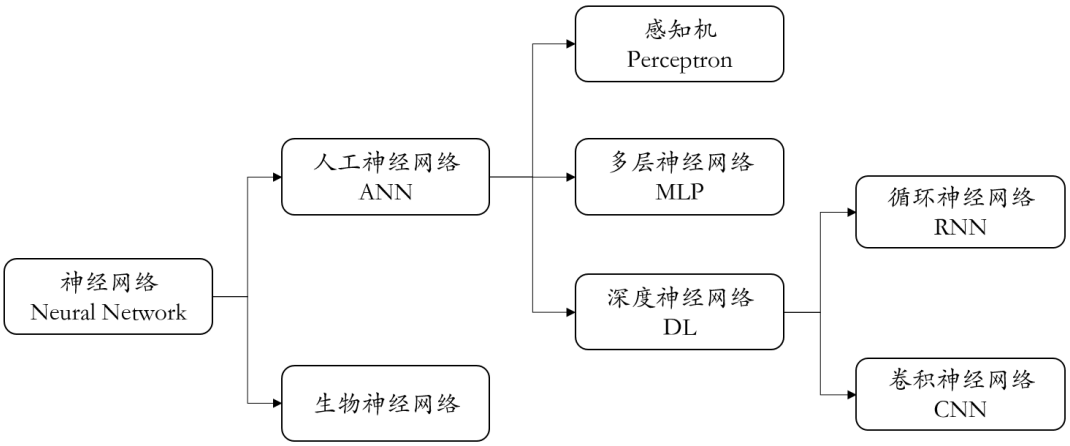

图 3.5 神经网络的类别

卷积神经网络主要由卷积层(Convolutional Layer)和子采样层(Subsample Layer)，或称为池化层(Pooling Layer)组成。卷积层为特征提取层，卷积层的每一个神经元对前一层的局部特征进行提取，其思想是利用图像的空间联系在局部较为紧密来实现局部感知；不同的神经元负责提取不同的特征，且学习到的特征可以在全局使用，称之为参数共享。这两种设计思路极大地减少了神经网络的参数。池化层为特征映射层，将相似的特征合并起来，每个特征映射层上所有的神经元的权值相同。卷积神经网络的这些构造，布局更加接近实际的生物神经网络，且因为用局部感知和参数共享降低了网络的复杂性，使之更加容易运算（Lecun& Bengio，2015）。

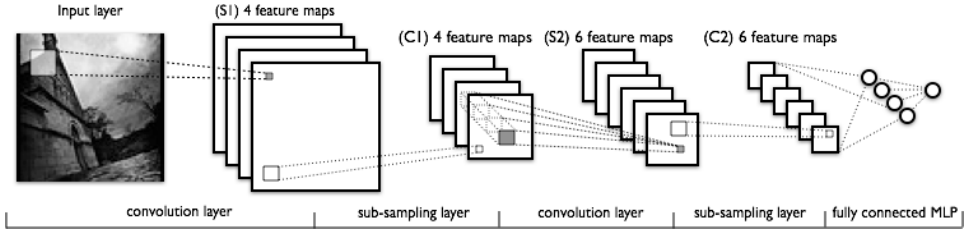

图 3.6 卷积神经网络示意图

目前的卷积神经网络架构有着 10-20 层，上百万个权值参数和几十亿个连接，硬件、软件和并行计算算法的进步，使训练时间压缩到了几个小时。这也使得卷积神经网络可以广泛应用在各种图像识别、语音处理等软件上。因为其设计特点，使得卷积神经网络可以很容易在芯片或者现场可编程门阵列(Field-Programmable Gate Arrays, FPGA)上实现。因此，类似于 NVIDIA，Mobileye, Intel, Qualcomm 和三星正在开发卷积神经网络芯片，以在手机、相机、机器人和智能驾驶等等上实现实时图像识别。

3.3.2 AlphaGo 实现技术

2016 年 3 月，Google 旗下 DeepMind 团队开发的 AlphaGo 击败了韩国围棋选手李世石，震惊全球。对于棋类博弈的游戏来说，这不是机器第一次击败人类顶尖选手，IBM 的深蓝在 1997 年就在国际象棋领域打败了世界排名第一的加里·卡斯帕罗夫。但为什么 AlphaGo 的成功会让社会这么震惊并开始关注人工智能的发展速度呢？

对于任何完全信息博弈的游戏来说，都有一个最优的评估函数，通过搜索步数的搜索树，可以算出每一步可能的b^d结果，此处b代表游戏的广度，即每一步有多少种可能性；d 代表游戏的深度，即游戏的步数。例如，对于国际象棋来说，b 约等于 35，d 约等于 80；对于围棋来说，b 约等于 250，d 约等于 150。不管是围棋还是国际象棋，这种复杂程度不可能通过穷尽所有可能性的搜索方式实现目标（Silver& Huang, 2016）。

一种方式就减少搜索的深度，在状态 s 下，将该状态之后的可能结果用评估函数来替代，搜索到 s 就截断搜索，深蓝用的就是这种方法。深蓝利用博弈树去搜索之后的六步，利用团队对国际象棋知识的理解构造评估函数(Evaluation Function)来评估节点的质量，利用极大极小算法将最坏的情况减到最低的原则选择下一步动作。此时，决定 AI 的实力一方面取决于计算能力，更强的计算能力意味着可以搜索更深更广；另外一方面取决于评估函数的质量，更高的质量意味着对棋局的判断更加准确。深蓝的评估函数在象棋大师的调整下，可以更好地判断棋局，并依靠强大的计算能力评估大量可能的局势，使得战胜了人类棋手。但这种方法并不适用于围棋，因为围棋的复杂性使得合适的评估函数的构造困难，因此，围棋被认为是人类智慧的堡垒难以被人工智能攻陷。

另外一种方法就是减少搜索的广度，通过在状态 s 下的行动的概率分布抽样来达到目标。蒙特卡洛搜索树(Monte Carlo tree search, MCTS)利用蒙特卡洛算法来估算搜索树每一个状态的值。随着越来越多的模拟，搜索树不断地扩大，相关值会越来越准确。AlphaGo 利用估值网络(Value Network)来估计棋局的状态，利用走棋策略网络(Policy Network)来决定下一步棋的走法，最后利用树搜索讲前两者结合在一起，模拟下一步会发生什么，并通过策略网络选择最佳的落子位置。

DeepMind 用卷积神经网络(Conventional Neural Network, CNN)训练走棋策略网络

(Policy Network)，以目前状态的棋局作为训练集，将下一步棋局形态作为标签，做监督式的监督学习。训练求最优解的过程用的随机梯度上升法来求最大人类在状态s下选择a步的可能性。

$$\Delta\sigma \propto \frac{\partial \log p_\sigma(a|s)}{\partial \sigma}$$

经过拥有 13 层的深度神经网络，将 3000 万的人类对弈的位置信息拆解为训练集，反复训练，最后达到 57%的准确度。同时研究团队也训练了快速走棋策略网络$p_\pi(a|s)$，这个策略网络通过一些简单特征做 Softmax 回归，回归权值为π，特点就是运算时间非常快，只需要 2 微秒而前者需要 3 毫秒，不过准确率只有 24.2%。接下来，DeepMind 团队用加强学习的方法对走棋策略网络进一步训练。加强学习策略网络在结构上与初始的有监督学习网络一致，参数初始值也保持一致。然后用目前状态的策略网络p_ρ和随机选取的之前迭代版本的策略网络进行对弈，通过随机选取对手可以避免过度拟合出现。最后训练估值网络来判断棋局的状态。用策略 p 在状态 s 下的预期结果。

$$v^p(s) = \mathbb{E}[z_t|s_t = s, a_{t\ldots T}\sim p]$$

用参数θ的神经网络的来近似替代对当前状态的判断,这个神经网络与走棋策略网络的结构一样,但输出为对盘面局势的判断。用状态——结果对*(s, z)*来训练神经网络，通过随机梯度下降算法求解预测值$v_\theta(s)$和真实结果 z 的最小均方误差(Mean Squared Error)来获得参数。

$$\Delta\theta \propto \frac{\partial v_\theta(s)}{\partial \theta}(z - v_\theta(s))$$

最后一步，AlphaGo 在 MCTS 中结合策略网络和估值网络。搜索树的每一条边*(s, a)*记录了落子的估值 *Q(s, a)*，访问次数 *N(s, a)*和先验概率 *P(s, a)*。接着从根部模拟对搜索树的访问，对于每一次模拟的第 t 步来说，落子位置a_t按如下公式进行选择：

$$a_t = argmax(Q(s_t, a) + u(s_t, a))$$

其中

$$u(s, a) \propto \frac{P(s, a)}{1 + N(s, a)}$$

可以看出，*u(s, a)*正比于先验概率，但随着访问次数减少以此来激励向其他地方进行搜索。当搜索在第 L 步时访问到叶节点s_L，用策略网络(SL policy network)计算出下一步的落子的先验概率 *P(s, a)*= $p_\sigma(a|s)$。同时，用两种方式来评估叶节点：第一种是估值网络$v_\theta(s_L)$，第二种是利用快速走棋策略p_π来预测终局z_L；两种方式结合起来，得到叶边值 $\mathrm{V}(s_L)$。而$Q(s, a)$的值取决于叶边值和访问次数。

$$\mathrm{V}(s_L) = (1 - \lambda)v_\theta(s_L) + \lambda z_L$$

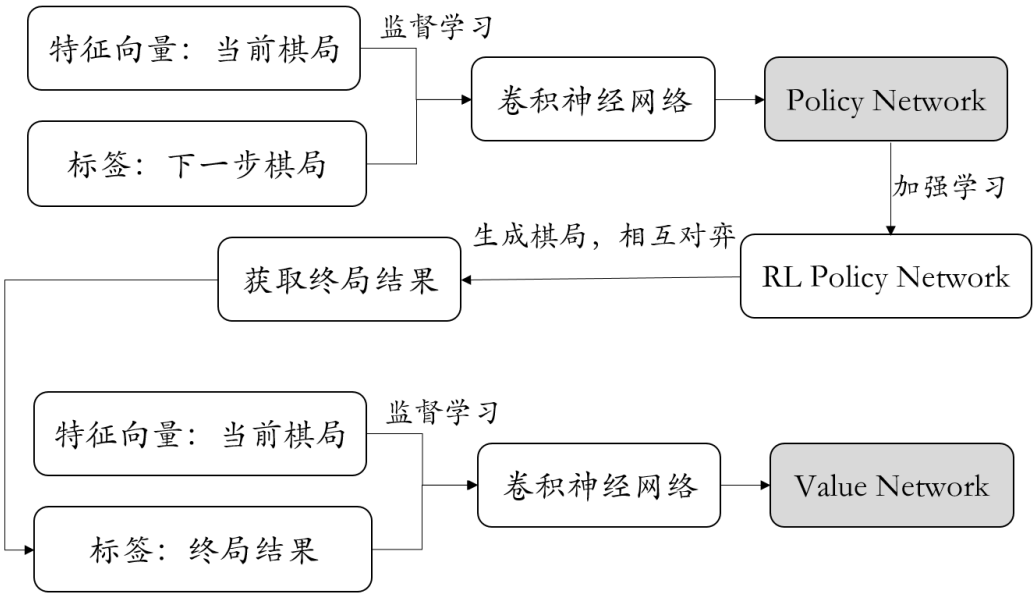

图 3.7 深度神经网络生成 Policy Network 和 Value Network 的过程

上文是 AlphaGo 实现的技术细节。总结来说，AlphaGo 学习的过程可以分为三个阶段。上图详细地描述了 AlphaGo 的训练流程。

(1) 利用 3000 万专业棋手棋谱位置来训练策略网络。用当前盘面状态做神经网络的输入，输出下一步棋在其他位置的落子概率。

(2) 利用第 t 步的策略网络与先前训练的策略网络互相对弈，用强化学习的方法来修正第 t 步的网络参数。

(3) 利用策略网络生成棋局的前 M-1 步，随机决定第 M 步的位置，利用增强策略网络完成后面的博弈过程，直到分出胜负，然后将胜负作为标签，第 M 步的盘面作为特征变量输入，深度学习一个估值网络，用于判断结果的输赢概率。这是 AlphaGo 的创新点。

总而言之，AlphaGo 本质上就是利用深度学习网络通过棋局训练得到两个网络，然后用 MCTS 高效地寻找最优的落子位置。这个过程，核心技术就是深度神经网络。

3.3.3 神经网络发展的回顾和总结

上文简单回顾了神经网络起起落落的发展历史，并仔细分析了基于卷积神经网络的 AlphaGo 的实现技术。我们可以看到，在人工智能的任何进展都是循序渐进的。从 1958 年感知机的兴起到 1969 的人工智能的第一个冬天，到 1986 年 BP 算法的提出使得神经网络第二次兴起再到遇到瓶颈之后的第二个冬天，到 2006 年深度神经网络的提出再到 2016 年 AlphaGo 利用深度神经网络技术在围棋上战胜了李世石，人工智能迎来其第三

次热潮。回顾历史，人工智能的起起落落也让我们谨慎面对目前的这一轮狂热。

人工智能发展的历史，也是计算机硬件和算法的进步史。在感知机的时代，计算能力还是晶体管，数十个数据量；而到了多层神经网络时代，我们可以方便地在性能良好的 CPU 上进行计算，用成千上万的数据来训练模型；到现在的深度学习的时代，我们使用分布式计算机集群，可以在 GPU 甚至 TPU 等专用芯片上加速训练模型，动辄使用千万至亿条数据进行训练。

总之，人工智能的发展历史与数据量变大、硬件性能提升、理论和算法的进步息息相关，我们不应该孤立地看待人工智能。在当前人工智能的热潮之中，冷静看待这一轮技术进步是非常有必要的。

3.4 计算机硬件的发展降低边际成本

不管是Google的流感预测模型，还是IBM的Watson，又或者是Google的AlphaGo，在简单的原理背后是需要大量数据训练模型。Watson 是一个通过海量的知识语料库进行假设生成和基于证据验证的问答系统，而震惊世界的 AlphaGo 只是将棋盘当成一个图，用大量选手的对弈棋局去训练它。这个过程离不开算法的进步和硬件性能提升的支持，Google 流感模型在数百台机器上运行其分布式计算系统，Watson 刚开始的时候体积是一间房，AlphaGo 用了分布式的 1202 个 CPU 和 176 个 GPU 去支撑其 40 个搜索线程。

现阶段，硬件的发展非常迅速。Google 在 2016 年发布了 TPU(Tesnor Processing Unit)，这款产品是专用芯片，可以用来加速训练神经网络，处理速度比当下的 GPU 和 CPU 快 15 到 30 倍；能效比是 GPU 和 CPU 的 30 到 80 倍（Jouppi& Norman，2017）。而 IBM 的 TrueNorth，与 CPU、GPU 或 TPU 是完全不同的芯片架构（Esser& Merolla 2016），这种脉冲神经元的芯片结构的精度较差的问题一直阻碍着该芯片的应用，但 IBM Almaden 中心的研究员在 2016 年解决了这个问题，用此芯片在低能耗的条件下完成了接近目前先进水平的图像和语音识别精度，这预示类脑计算的硬件和深度学习可以完美结合。美国空军实验室(Air Force Research Lab, AFRL)在 2017 年初报告该芯片可以在雷达生成的图像中很好地识别军用和民用车辆，而能耗不到传统计算机的 20%。在卫星、高空飞机、空军基地等能源和空间有限但又需要高级视觉识别系统的地方，这种芯片无疑具有广阔的应用前景。

目前的芯片或者往专用化的方式发展，类似于 TPU，可以快速用来训练神经网络；或者往类脑方向发展，以完全不同于传统芯片的架构去设计芯片。与此同时，通用处理器的进步也很关键。总之，人工智能的发展离不开硬件的支持，低能耗小体积高效率的芯片是未来人工智能应用更加广泛的必要条件，硬件的发展减少了人工智能方法

应用的边际成本。

3.5　人工智能的应用前景

如同大数据一样，现在人们又对人工智能寄予很高的期望。但冷静分析，目前的人工智能研究的前沿中的专家系统在上实际八十年代就已经出现。大数据处理方式的进步，使得 Watson 在前人的研究基础上更进一步，在实际中开始应用。Watson 利用其出色的理解、推理和学习能力在医疗行业迅速发展。研究表明，到 2020 年医疗数据每 73 天都会翻一番，且大部分数据是类似于图像和视频的非结构化数据。一个人在一生中会产生 100G 与健康相关的数据，医生需要每周花费 160 小时来学习新的医疗知识。而 Watson 可以每秒学习 267 万页文献，利用出色的学习推理能力辅助医生进行决策。2013 年，Memorial Sloan-Kettering(MSK)和 Watson 合作发展 Watson for Oncology，利用 Watson 分析大量的医疗文献、病历信息和临床测试来提供基于证据学习的个体性治疗。在总计 15000 个小时的训练的过程中，由医生和研究人员组成的团队为 Watson 输入了数千个病历、接近 500 本医学杂志和教材、1200 万页医学文献和 MSK 定制的研究方式，将 Watson 训练成一个博学的专家。在后续的过程中，专家持续不断地为系统输入新的数据，使系统保持足够的敏锐直觉。在治疗的过程中，Watson 首先分析患者的病历，从中提取中关键信息；接着，Watson 将患者的属性数据与训练过程中获取的临床经验、学术研究和数据项结合，获取潜在的治疗方案并为每种方案提供证据支持，帮助医生研究个性化的治疗方案。

而人工智能的另外一个研究前沿，深度学习目前最大的应用场景在于图像识别和语音处理。AlphaGo 除了下棋之后，在别的方面的应用目前乏善可陈。不过，深度学习在图像识别上的进步使得在医学上的应用成为可能。Enlitic，一家将深度学习应用在医学上的初创公司。系统利用深度学习在图像识别上的优势，以识别病人肺部的 CT 图为例，系统会识别某嫌疑位置是否是血管还是可疑的肿瘤，并将需要进一步检测的位置标记出来。在测试中，Enlitic 的系统在识别肿瘤的性能上提高 50%，并且漏诊率为 0，而人类的漏诊率为 7%。Enlitic 另外一个检测 X-rays 的系统性能也优于人类专家。

因此，站在目前的技术水平我们去思考人工智能可以应用在何处时，首先，我们要考虑其技术局限性。目前的技术，离不开专家系统或者是深度学习。一个任务能够用这些技术来描述是其可以应用的必要条件。其次，任务被机器替代后的收益也是我们需要考虑的。例如，阿迪达斯首席执行官罗斯德(Kasper Roested)认为制造业大举回流美国是幻想因为在生产阿迪达斯的大约 120 道工序中，有部分很难自动化。制鞋业面临最大挑战是如何开发出为鞋子穿鞋带的机器人。既然这个任务既难以被目前的技术所智能化，且替代的人力成本又比较低，因此也就很难去应用。

现阶段，机器已经可以完成许多形式的例行程序性的工作，因此薪酬很高但工作却比较固定的会计等行业的被人工智能化是即将发生的。Frey & Osborne(2013)通过对 702 个详细的职位建模，得出 47%的就业将要被电脑取代。同时，工资和教育程度与职业被机器替代的可能性有着很强的负相关。未来的就业市场，中端制造业的工作被替代的可能性是最大的，而高技能工作和低技能工作的岗位在增加。而其中最有可能被替代的工作是电话销售员、会计师和审计师、零售销售员和房地产销售机构等等，而最不可能被替代的包括化学工程师、牧师、运动员训练师和牙医等等。

第四章　神经网络在实际应用中与经典回归模型比较——以幸福感研究为例

大数据算法和人工智能应用极大的改变了社会，机器替代了大量的人工工作，自动驾驶也将会把人类从枯燥驾驶中解放出来，这些技术的进步将极大的重新塑造这个社会的形态，也会给经济学研究者提供新的工具。而我们在这一章试图利用神经网络建立起幸福感预测的模型并与传统方式做比较，以分析模型的成本与收益。

4.1　相关研究简介

Giesecke (2016)论述了用深度学习的模型来研究抵押贷款的违约风险情况。使用了从 1995 年至 2014 年全美超过 35 亿条初级和次级贷款的数据，数据包含贷款者的个人特征、贷款有效期内每个月贷款状态的更新数据、地区水平的经济指标。研究者用机器学习的模型开发、估计和测试贷款的提前还款、拖欠和丧失抵押赎回权，并将地理相似性和共同风险因素（类似于地区的经济指标）带来的贷款之间的相关性考虑在模型之中。该文最核心的一部分就是利用深度神经网络来处理解释变量和贷款表现之间的非线性关系。文章选择了最大似然估计方法来拟合深度学习模型，由于数据庞大大约 35 亿条抵押贷款的月度观察数据，且每一条数据都有 300 个解释变量，因此研究者利用了 GPU 进行并行计算来处理如此大的数据集。拟合算法运行在亚马逊云服务(Amazon Web Services, AWS)的节点集群上。求解最优化参数的时候使用梯度下降算法，同时使用了正则化和交叉验证来避免过度拟合。研究证明，影响抵押贷款的拖欠或者提前还款的关键因素是当地的经济因素例如就业率、房价和当地丧失抵押品赎回权概率的滞后变量等等。深度学习模型在样本外预测效果上比类似于逻辑回归等线性模型有了非常明显的提升。

对于幸福感研究，首先值得关注的问题就是“收入的增长是否能够带来幸福的增长”。经济学家对于此问题，主要有两种理论：(1)相对效用论(Easterlin, 1974)认为幸福感既取决于收入页取决于社会可比标准；(2)绝对效用论(Veenhoven, 1991)认为收入的提高可以提高人的幸福感。胡霄俊(2010)构建了基于 4 个基本人口特征和 6 个社会需求因素的主观幸福感函数，发现幸福感影响因子的地方差异性和支持收入的相对效用的证据。同时胡霄俊(2010)提出，主观幸福感影响因素分为时间和空间两个维度。传统的幸福感研究理论有人格理论、相随理论和适应性理论。外部影响因素包括文化差异、收入因素和婚姻；内部影响因素包括人格、性别和年龄的差异。黄耀峰(2011)将个人主观幸福的影响因素分解为以下四种因素：(1)包括年龄和文化程度的基本因素；(2)包括住房、

就业、收入、医疗、教育、养老等社会问题因素；(3)工作时间和休闲时间因素；(4)收入和预期收入的变化影响因素。

4.2 数据描述和变量选择

本章以《中国经济生活大调查(2015-2016)》数据为数据集，数据来源于 CCTV 财经频道“中国经济生活大调查”项目（以下简称为“大调查”)。“大调查”是由 CCTV 联合国家统计局、中国邮政集团退出的大型社会问卷调查，是国内覆盖面最广的民间调查。此调查从 2006 年开始，但因为每年的问卷的会有调整，且受访者对象不确定，因此选取 2015-2016 年的截面数据。本数据集有效问卷 88415 份，问卷问题有关于受访者对于社会、经济看法，大部分问题集中于受访者的收入、工作满意度、目前的生活感受、影响生活幸福的因素、性别、文化程度、婚姻状况等等。在问卷中，受访者直接关于幸福的有效回答 85913 份，认为自己很幸福的人数为 13840，占比 16.1%；认为自己比较幸福的人数为 29813，占比 34.7%；认为自己一般的人数为 32822，占比 38.2%；认为自己比较不幸福的人数为 5899，占比 6.9%；认为自己很不幸福的人数占比 4.1%。

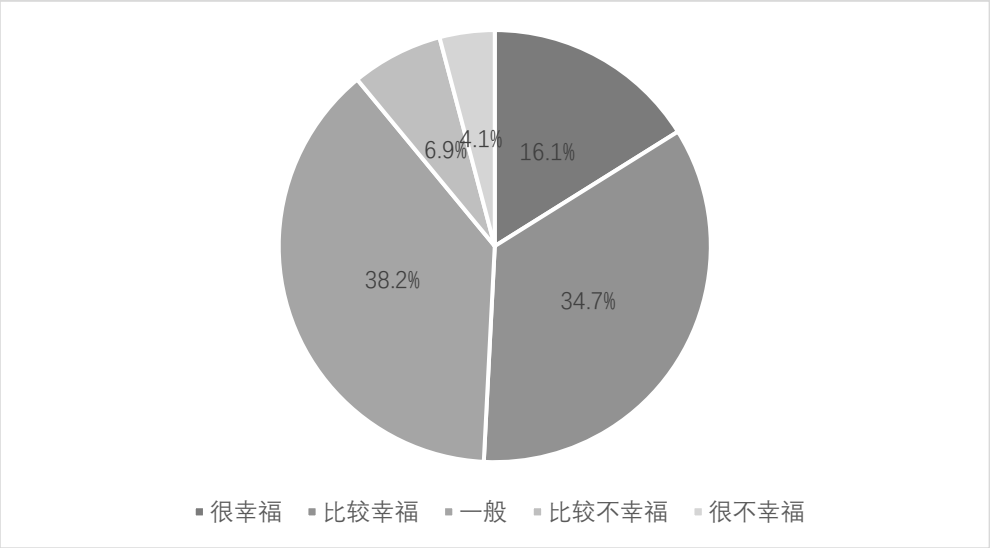

图 4.1 受访者对幸福感的主观判定

在问卷中，有问题直接调查受访者认为的影响幸福的主要因素。受访者可以在社会保障、健康状况、婚姻情感等十项中选取不多于三项作为自己认为的影响生活幸福的主要因素。从数据来看，大部分受访者认为社会保障、健康状况、婚姻情感、环境卫生和收入水平这五项因素是影响幸福感的主要因素。

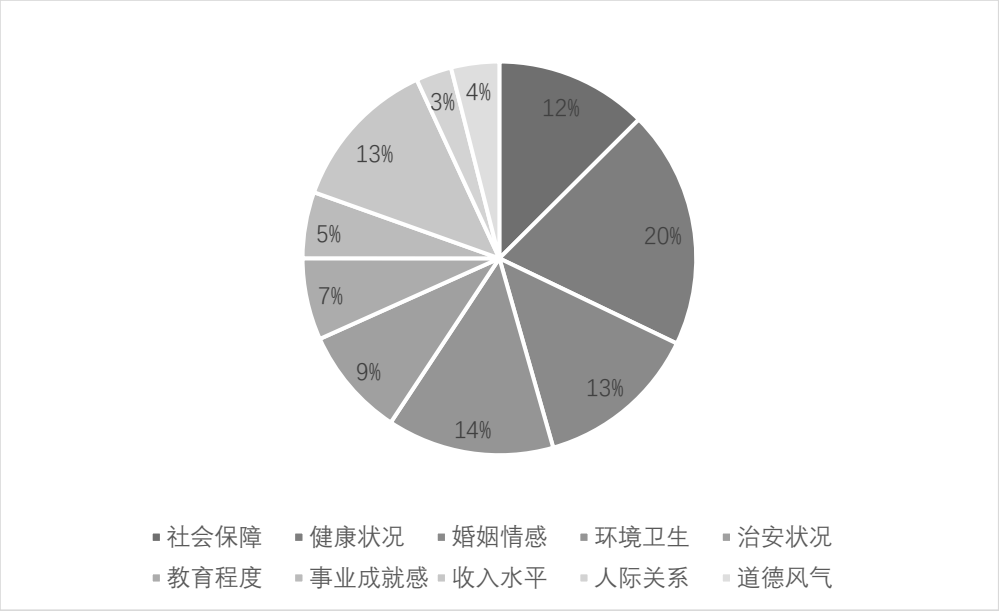

图 4.2 受访者认为影响幸福感的主要因素

根据幸福感影响因素理论和本问卷的具体选项，选择性别、年龄、文化程度、婚姻状况、常住地、职业、家庭年收入、住房情况、生态环境、工作满意度、预期收入等11个变量作为个体的特征向量，是否幸福作为标签进行建模。在删除缺失值数据之后，有效数据61018条。变量处理细节如下：

(1) 性别 Gender。采用问卷编码，“男”为0，“女”为1。

(2) 年龄 Age。采用问卷编码，“18~25岁”为0，“26~35”为1，“36~45”为2，“46~59”为3，“60岁及以上”为4。

(3) 文化程度 Edu。采用问卷编码，“小学及以下”为0，“中学及中专”为1，“大专”为2，“本科”为3，“硕士”为4，“博士”为5。

(4) 婚姻状况 Marital。采用问卷编码，“未婚无恋人”为0，未婚有恋人”为1，“已婚”为2，“离异”为3，“丧偶”为4。

(5) 常住地 Residence。采用问卷编码，“城市”为0，“农村”为1。

(6) 职业 Occupation。采用问卷编码，“行政事业单位人员”为0，“企业管理人员”为1，“城市户籍企业职工”为2，“在校学生”为3，“务农农民”为4，“进城务工人员”为5，“离退休人员”为6，“待业/失业”为7，“自由职业者”为8。

(7) 家庭年收入 Income。以城镇居民家庭年平均收入为参考，将数据重新分类，“6万以下”为0，“7~15万”为1，“15~30万”为2，“30万以上”为3。

(8) 住房情况 House。采用问卷编码，“自有房（大产权）”为0，“自有房（小产权）”为1，“农村住房”为2，“公租房”为3，“自租房”为4。

(9) 生态环境 Environment。因为问卷中关于生态环境有多个选项，涵盖空气质量是

否变好，城市绿化是否变好，自来水质是否变好和垃圾清运是否变好等四个方面和 8 个选项，受访者选择不多于三个选项作答。因此，将变好项计+1，将变坏项计-1，并将目前比较关心的空气质量赋予较高的权重 2，其他为 1，得出样本对环境的综合评价。取值范围为+4 至-4。

(10) 工作满意度 JobSatisfaction。采用问卷编码，“很满意”为 0，“比较满意”为 1，“一般”为 2，“比较不满意”为 3，“很不满意”为 4。

(11) 预期收入 Expectation。采用问卷编码。“增加 20%以上”为 0，“增加 10~20%”为 1，“增加 10%以内”为 2，“持平”为 3，“减少 10%以内”为 4，“减少 10%以上”为 5。

(12) 幸福感 Happiness。问卷中有“很幸福”，“比较幸福”，“一般”，“比较不幸福”，“很不幸福”五个选项，显然，如果将“一般”归类为“不幸福”或者“幸福”都会造成比较大的误差，但因为样本数量有限，如果删除一般后样本数量过少且“幸福”和“不幸福”的样本数量过于悬殊。因此，将“很幸福”和“比较幸福”这两类分为“幸福”，值为 0；将“一般”、“比较不幸福”和“很不幸福”归类为“不幸福”，值为 1。

4.3 经典 Logistic 回归模型

Logistic 回归是一种经典的机器学习算法，该算法根据已知的一系列因变量估计离散数值，即通过数据拟合成一个逻辑函数对数据进行线性分类，预估一个事件出现的概率。训练的过程就是寻找最佳拟合参数。Logistic 函数定义如下：

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

其中：L为函数的最大值，x_0为曲线的中心点的值，k为 Sigmoid 曲线的陡程度度。

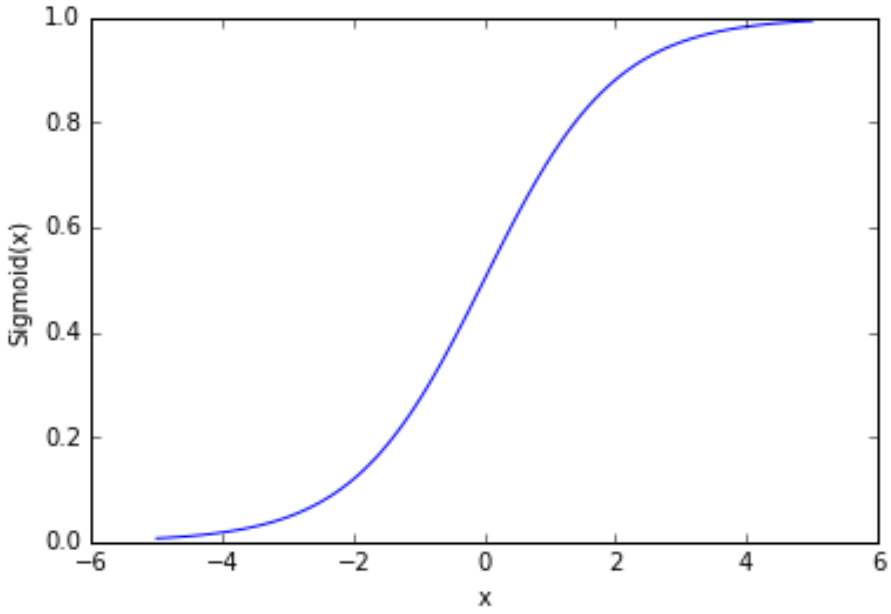

图 4.3 标准 Logistic Sigmoid 函数(k=1, x_0=0)

训练数据不够或者过度训练会导致过度拟合(Overfitting)，因为模型在训练集上的误差减少，但在验证集上的误差却会逐渐加大。因此在实际的训练过程中，将数据分成三组：训练集(Training Data)，验证集(Validation Data)和测试集(Testing Data)。为了避免过度拟合，在训练过程中一些参数（例如学习率 Learning Rate）由验证集而不是训练集来确定，训练集的作用保留为计算梯度更新权重。同时，使用权重衰减(L2 Regularization)来避免过度拟合。

通过求解正则化的训练误差来求解模型的参数：

$$E(w,b) = \frac{1}{n}\sum_{i=1}^{n} L\big(y_i, f(x_i)\big) + \alpha R(w)$$

其中，L 为代表模型预测的误差函数，R 代表正则项。我们选取 L2 正则方式，因此最后的目标函数如下：

$$\min_{w,c} \frac{1}{2} w^T w + C\sum_{i=1}^{n} \log\Big(exp\Big(-y_i(X_i^T w + c)\Big) + 1\Big)$$

利用随机梯度下降法(Stochastic Gradient Descent)算法求解最优值。权值迭代更新公式如下：

$$w \leftarrow w - \eta(\alpha\frac{\Delta R(w)}{\Delta w} + \frac{\partial L(X_i^T w + b,\ y_i)}{\partial w})$$

其中η为用来控制在参数解空间寻找最优解的步长的学习率。

在 Python 上实现 LR 模型，最后结果如下，模型的 AUC 为 0.791。在分类性能上，

当样本为幸福时，判断正确的概率为 72.0%；当样本为不幸福时，判断正确的概率为77.6%。

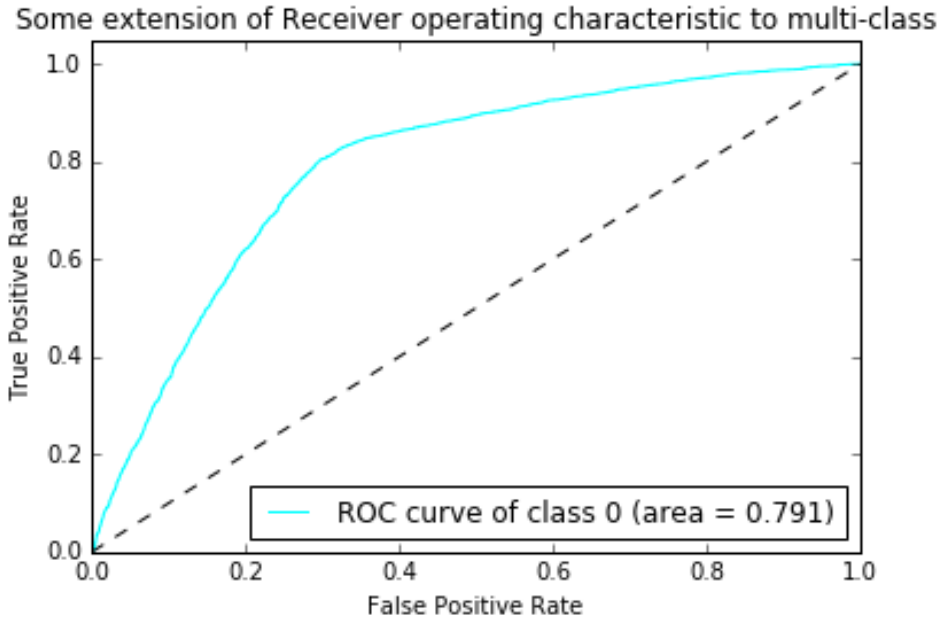

图 4. 4 LR 模型的 ROC 曲线

表 4. 1 LR 模型对测试集的分类结果

	判断为幸福	判断为不幸福
真实为幸福	72.03%	27.97%
真实为不幸福	22.41%	77.59%

4.4 神经网络模型

本文构建多层感知器神经网络(Multi-layer Perception Neural Network)模型，模型有四层结构，第一层为特征输入层，第二层和第三层分别为隐藏层一和隐藏层二，第四层为输出层。同一层之间的神经元互不连接，相邻层之间的神经元全链接，具体结构如下图所示：

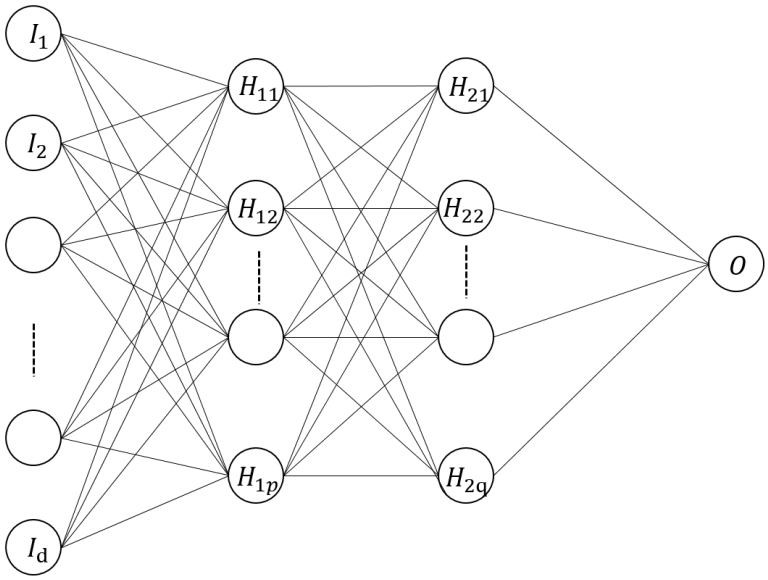

图 4.5 MLPNN 模型结构

模型的参数求解利用反向传播(Back Propagation)算法完成。即首先利用目前的参数得到输出结果，再将输出结果和正确结果进行比较，得到误差；误差再反向对神经网中的连接权重进行反馈修正，从而得到最优解。通过正向和反向的过程，在权值向量的空间执行误差函数梯度下降策略，使得误差最小化，在模型的求解过程中，使用链式规则计算梯度。模型的符号定义如下：

表 4.2 符号定义列表

变量	定义
d	输入层的神经元数量，即特征向量的维度，$d=11$
m	输出层的神经元数量，输出幸福感，m=1
$\vec{x}$	输入特征向量$(x_1, x_2, \ldots x_d)$
y	输出的真实值
$\hat{y}$	输出的预测值
p	隐藏层一的神经元数量
q	隐藏层二的神经元数量
v_{hi}	输入层的神经元h和隐藏层一的神经元i间的连接权值
w_{ij}	隐藏层一的神经元i和隐藏层二的神经元j间的连接权值
u_j	隐藏层二的神经元j和输出神经元间的连接权值
a_i	隐藏层一的第i个神经元的输入
A_i	隐藏层一的第i个神经元的输出
b_i	隐藏层二的第j个神经元的输入
B_i	隐藏层二的第j个神经元的输出
γ	输出层神经元的输入
f_1, f_2, f_3	分别为隐藏层一、隐藏层二和输出层的激活函数

首先，由数据的正向传播计算出模型预测值。

$$A_i = f_1(a_i) = f_1(\sum_{h=0}^{d} v_{hi}\, x_h)$$

$$B_j = f_2(b_j) = f_2(\sum_{i=0}^{p} w_{ij}\, A_i)$$

$$\hat{y} = f_3(\gamma) = f_3(\sum_{j=0}^{q} u_j\, B_j)$$

其次，采用均方误差定义误差函数(损失函数)L

$$L(v, w, u) = \frac{1}{2}(y - \hat{y})^2$$

接下来，通过梯度下降法迭代求解最优参数。定义学习率参数为η，则参数迭代更新公式如下：

$$\delta \leftarrow \delta - \eta \frac{\partial L}{\partial \delta}$$

调整隐藏层二的神经元j和输出神经元间的连接权值u_j，

$$\Delta u_j = -\eta \frac{\partial L}{\partial u_j} = -\eta \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \gamma} \frac{\partial \gamma}{\partial u_j} = \eta(y - \hat{y}) f_3'(\gamma) B_j$$

调整隐藏层一的神经元i和隐藏层二的神经元j间的连接权值w_{ij}，

$$\Delta w_{ij} = -\eta \frac{\partial L}{\partial w_{ij}} = -\eta \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \gamma} \frac{\partial \gamma}{\partial B_j} \frac{\partial B_j}{\partial b_j} \frac{\partial b_j}{\partial w_{ij}} = \eta(y - \hat{y}) f_3'(\gamma) u_j f_2'(b_j) A_i$$

调整输入层的神经元h和隐藏层一的神经元i间的连接权值v_{hi}，

$$\begin{aligned}
\Delta v_{hi} &= -\eta \frac{\partial L}{\partial v_{hi}} = -\eta \frac{\partial L}{\partial a_i} \frac{\partial a_i}{\partial v_{hi}} = -\eta \frac{\partial L}{\partial A_i} \frac{\partial A_i}{\partial a_i} \frac{\partial a_i}{\partial v_{hi}} \\
&= -\eta \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \gamma} \sum_{j=0}^{q} \left(\frac{\partial \gamma}{\partial B_j} \frac{\partial B_j}{\partial b_j} \frac{\partial b_j}{\partial A_i}\right) \frac{\partial A_i}{\partial a_i} \frac{\partial a_i}{\partial v_{hi}} \\
&= \eta(y - \hat{y}) f_3'(\gamma) \sum_{j=0}^{q} (u_j f_2'(b_j) w_{ij}) f_1'(a_i)\, x_h
\end{aligned}$$

选择标准形式的 Sigmoid 函数作为激活函数，因 Sigmoid 函数具有如下性质：

$$f'(x) = f(x)(1 - f(x))$$

因此，在计算中不需要求导，只需要直接带入公式，可以节省很多计算量。

在 Python 上实现多层神经网络模型，最后结果如下，AUC 值为 0.808。在分类性能上，当样本为幸福时，判断正确的概率为 71.7%；当样本为不幸福时，判断正确的概率为 80.4%。

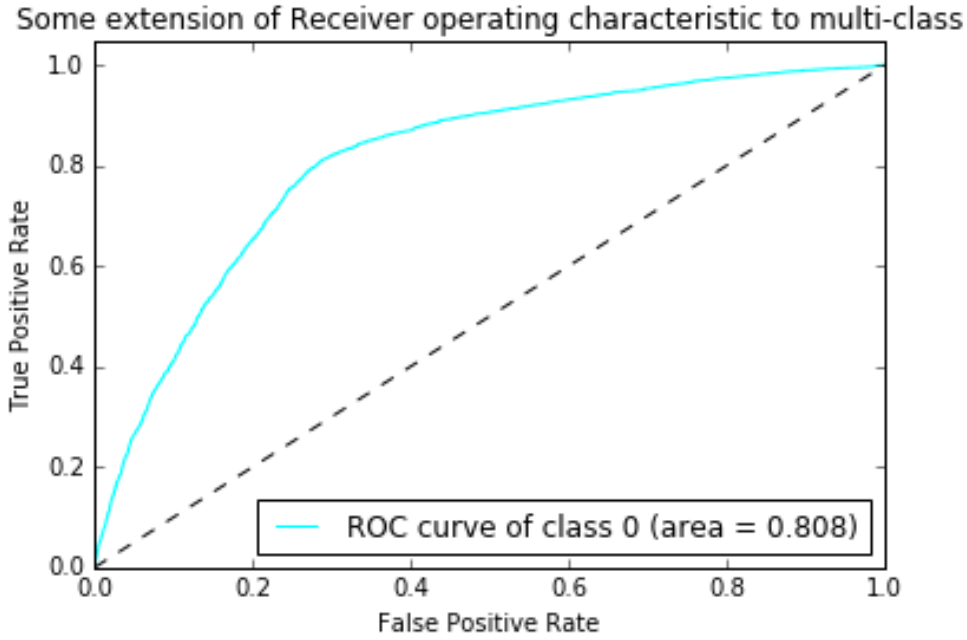

图 4. 6 MLPNN 模型的 ROC 曲线

表 4. 3 MLPNN 对测试集的分类结果

	判断为幸福	判断为不幸福
真实为幸福	71.73%	28.27%
真实为不幸福	19.64%	80.36%

4.5　模型对比和总结

从时间来看，因为需要求解的参数多且收敛时需要迭代的次数多，MLPNN 的时间是 LR 的很多倍。从模型结果来看，多层神经网络在分类的总体性能上略微好于 LR 模型，AUC 值从 0.791 提升到 0.808。LR 模型的 Type I 错误为 29.97%，Type II 错误为 22.41%。MLPNN 模型的 Type I 错误为 28.27%，Type II 错误为 19.64%。因为 MLPNN 模型可以更多地揭示数据之间的非线性关系，因此在性能上的提升是情理之中的。MLPNN 模型拥有更多的参数，从更多的维度去刻画数据特征，但是对数据量的要求更高，且需要更多的计算时间。

表 4. 4 MLPNN 对测试集的分类结果

	计算成本			性能提升		
	参数	收敛迭代次数	计算时间(分)	Type I Error	Type II Error	AUC
LR	15	100	3.8	27.97%	22.41%	0.791
MLPNN	529	300	352.9	28.27%	19.64%	0.808

在数据量更多且数据维度更高的时候，将程序运行在 GPU 上进行并行加速是必不可少的。Giesecke (2016)的结果也证明了神经网络在数据量大时优于线性回归模型 LR。但我们并不认为深度学习的方法会很快改变经济学的研究方法，首先，深度学习模型需要大量的数据来训练，Giesecke (2016)用了 35 亿条数据，AlphaGo 用了 3000 万个棋局来获取最初的走棋策略网络，任何一个深度学习的模型想要拥有比较好的效果都需要通过大量数据训练，而在经济学实际应用中，很多问题难以获得大量的数据。金融数据可能是个例外，但金融市场的变化太快，难以从数据中提取不变的特征。其次，经济学研究更关心因果推断，而深度学习模型缺乏对这方面的研究。

第五章　结论

我们可以看到，随着大数据处理技术的进步和深度学习带来人工智能技术的再一次飞跃，但人工智能并不意味着万能，目前的人工智能技术无法进行抽象思考，很多工作也仅仅是只能只能由人工来完成。目前的深度学习最大的进步来自于从图像获取特征表示及概念，可以提高图像识别的精度。同时，通过其他例如视频和语音等非结构性数据获取多形态的特征表示，比如说环境识别已经很好地应用在自动驾驶上。但上述的这些特征学习能力离高级智能还相差甚远。机器从自身的行动中去提取特征、并真正意义上理解人类的语言到从语言中获取真正的知识以像人类一样思考还是一个非常困难的过程。

从大数据的历史经验来看，我们对于新兴的技术总是给予过高的预期。人工智能也一样，谷歌的流感模型预测失败后结束项目，人们津津乐道的 AlphaGo 创新性的利用卷积神经网络来构造围棋的价值评估函数，但将这个项目用在别的方面并不会像人们想象中的那么容易，而目前能够应用并直接创造价值的 Watson 在医疗方面辅助医生进行决策。

目前阶段的人工智能在图像识别、语音识别和自然语言处理上有了长足的进步，也就是初步具备了视觉和听觉，但机器想要具备主观思考能力和意识还有漫长的路要走。至于人工智能对经济学研究的影响，我们认为并不会很大。因为深度学习首先需要大量的数据来训练模型，但大部分经济学研究的问题并没有如此多的数据；其次虽然此类问题可能对最后的实证结果得出改进，但是我们也需要考虑这个运算成本和时间成本，训练一个大的深度学习的模型往往需要分布式的计算能力和大量的时间，这也是大量社会学研究者不具备的。最后因为深度学习模型是为了提取固定的特征，但在此类研究问题上因为涉及到人的主观参与往往很难呈现出固定不变的特征，金融市场的反身性就是明显的代表。

但人工智能对工作的替代将是持续的，特别是工作内容是重复性且可以用程序描述或者实现的话。最后，人工智能将重塑整个社会形态，但目前离技术的“奇点”还有很漫长的路要走。在这个过程中，我们既要从微观上分析预期的应用在技术上的可行程度，又要冷静分析人工智能带来的边际收益与边际成本。

参考文献

大数据战略重点实验室. DT 时代[M]. 中信出版社, 2015.

胡霄俊. 主观幸福、地区差异和财富效应——基于 OLR 的中国经验实证研究[D]. 北京大学, 2010.

黄耀锋. 主观幸福感影响因素探究——基于我国居民休闲工作时间分配的实证检验[D]. 北京大学, 2011.

松尾丰. 人工智能狂潮:机器人会超越人类吗?[M]. 机械工业出版社, 2016:134-151.

吴岸城. 神经网络与深度学习[M]. 电子工业出版社, 2016:114-124.

约翰・E・凯利, 史蒂夫・哈姆. 机器智能[M]. 中信出版社, 2016.

Declan B. When Google got flu wrong.[J]. Nature, 2013, 494(7436):155-156.

Easterlin R A. Does Economic Growth Improve the Human Lot? Some Empirical Evidence[J]. Nations & Households in Economic Growth, 1974:89-125.

Einav L, Levin J. Economics in the age of big data.[J]. Science, 2014, 346(6210):1243089.

Esser S K, Merolla P A, Arthur J V, et al. Convolutional networks for fast, energy-efficient neuromorphic computing[J]. Proceedings of the National Academy of Sciences of the United States of America, 2016, 113(41):11441.

Frey C B, Osborne M A. The future of employment: How susceptible are jobs to computerisation? *[J]. Technological Forecasting & Social Change, 2017, 114.

Ginsberg J, Mohebbi M H, Patel R S, et al. Detecting influenza epidemics using search engine query data.[J]. Nature, 2009, 457(7232):1012-4.

Hinton G E, Sejnowski T J, Ackley D H. Boltzmann machines: constrained satisfaction networks that learn[J]. 1984.

Hinton G E, Mcclelland J L, Rumelhart D E. Disctributed representations, Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations."[J]. Language, 1986, 63(4).

Jouppi, Norman P, Young, Cliff, Patil, Nishant, et al. In-Datacenter Performance Analysis of a Tensor Processing Unit[J]. 2017.

Lazer D, Kennedy R, King G, et al. The Parable of Google Flu: Traps in Big Data Analysis[J]. Science, 2014, 343(6176):1203-5.

Lecun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553):436-444.

Manyika J, Chui M, Brown B, et al. Big Data: The Next Frontier For Innovation, Competition, And Productivity[J]. Analytics, 2011.

Mcculloch W S, Pitts W H. A logical calculus of ideas imminent in nervous activity[J]. 1943.

Min K L, Kusbit D, Metsky E, et al. Working with Machines:The Impact of Algorithmic and Data-Driven Management on Human Workers[C]// CHI '15 Proceedings of the, ACM Conference on Human Factors in Computing Systems. ACM, 2015:1603-1612.

Norvig P, Russell S J. Artificial intelligence :a modern approach[J]. Applied Mechanics & Materials, 2003, 263(5):2829-2833.

Rosenblatt. The perception: a probabilistic model for information storage and organization in the brain[J]. Psychological Review, 1958, 65(6):386.

Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search.[J]. Nature, 2016, 529(7587):484.

Sirignano J, Sadhwani A, Giesecke K. Deep Learning for Mortgage Risk[J]. Social Science Electronic Publishing, 2016.

Stuart G, Spruston N, Sakmann B, et al. Action potential initiation and backpropagation in neurons of the mammalian CNS[J]. Trends in Neurosciences, 1997, 20(3):125-131.

The Era of Cognitive Systems: An Inside Look at IBM Watson and How it Works, IBM, 2012.

Tom Standage. Artificial Intelligence: The return of the machinery question[J]. The Economist, June 25 2016.

TrevorHastie, RobertTibshirani, JeromeFriedman. The Elements of Statistical Learning[M]. Springer New York, 2009.

Veenhoven R. Questions on Happiness[J]. 1992.

附录 A　神经网络模型参数求解算法

输入：训练集 S={(x_i, y_i)|i=1,2,⋯,N}

参数学习率，迭代终止条件

输出：神经网络的权值参数

1. 初始化神经网络的参数v_{hi},w_{ij},u_j
2. 计算 S 中的列向量的均值$(\overline{\mu}_1, \overline{\mu}_2, \dots, \overline{\mu}_d, \overline{\mu}_y,)$和标准差$(\sigma_1, \sigma_2, \dots, \sigma_d, \sigma_y,)$
3. 定义空集S'
4. 对于 S 的每一个样本点：
5. $x' = (\frac{x_1-\overline{\mu}_1}{\sigma_1}, \frac{x_2-\overline{\mu}_2}{\sigma_2}, \dots, \frac{x_d-\overline{\mu}_d}{\sigma_d})$
6. $y' = \frac{y-\overline{\mu}_y}{\sigma_y}$
7. 将(x', y')加入集合S'
8. 当迭代条件满足时做下列循环
9. 对于S'中的每个(x', y')：
10. 计算神经网络的输出，得到$A_i, B_j, \hat{y}$
11. 计算损失函数
12. 根据公式计算各个参数的反馈更新，得到$\Delta u_j, \Delta w_{ij}, \Delta v_{hi}$
13. 更新网络的权重参数v_{hi},w_{ij},u_j
14. 如果迭代满足终止条件时：
15. 结束迭代
16. 返回神经网络的参数v_{hi},w_{ij},u_j

致谢

在此学位论文完成之际，我要向在论文完成工作中所有指导和帮助我的人们致以深深的谢意。

首先，向我敬爱的导师胡大源教授表示最衷心的感谢！这篇论文是在导师的指点下完成的。论文的选题、构思、模型的构建、写作和修改都离不开导师的谆谆教导。每当在写作过程中遇到难以处理的问题之时，导师总是会在繁忙的工作之中抽出时间不辞劳苦地为我答疑解惑。胡老师除了在学业上为我指点迷津之外，在生活中也是我学习的榜样，胡老师在治学上严谨、在工作中勤奋，无疑都是我学习的楷模。胡老师知识渊博，心系国家经世济民的高尚品德，不断激励我在以后的工作中勤奋正直。在此谨向我的导师胡大源教授致以由衷的敬意。

同时，本论文的完成，也离不开其他老师同学的帮助。各位老师的教导给我打下了扎实的理论基础，各位同学的帮助也让我能够顺利攻克遇到的难关。在此特别感谢我的大学同学魏华，在我求解神经网络模型遇到困难之时他用出色的专业能力帮我解决问题；感谢我的同窗苏熊和侯国栋，我们互相讨论问题、相互鞭策成长，不管在生活上还是工作上我们建立起了深厚的友谊。

最后，我要感谢我的家人。感谢父亲和母亲给我最大的支持和关心，父母的爱是我前行的动力。感谢两位姐姐，虽然从求学之后就聚少离多，但感觉你们的支持总是在我最需要的时候出现。

魏成

2017 年 5 月

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名：　　日期：　　年　　月　　日

学位论文使用授权说明

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
- 因某种特殊原因需要延迟发布学位论文电子版，授权学校□一年/□两年/□三年以后，在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名：　　导师签名：

日期：　　年　　月　　日