中信建投证券
CHINA SECURITIES

CHINA SECURITIES RESEARCH

证券研究报告·金融工程深度报告

# 大数据、机器学习、深度学习在投资领域应用的方法论概述：　—大数据研究之五

金融工程研究

## 重要观点

**大数据特征概述**

大数据是高容量、高速度和多样性的信息资产：体积上，通过各种记录、交易信息、表格、文件存储等汇集的数据体量庞大；速度上，数据发送或接收的速度快，可以以批处理方式传输或接收，可以实时或接近实时地发送；种类上，数据通常以多种格式接收，有结构化的、半结构化的或非结构化的。

**自然语言处理将投资管理提升到新一高度**

以传统因子，比如技术面因子，基本面因子等挖掘增益信息已非常困难，而大数据，比如新闻媒体等信息，则将为投资提供更多的增益信息。利用自然语言处理技术则可从文本信息中挖掘出有效信息，提升投资管理能力。

**机器学习在投资领域的使用已逐渐成熟**

机器学习可分为监督学习（如回归和分类）及非监督学习（如因子分析和聚类）。监督学习试图找到一种规则或一个方程来预测变量。非监督则学习试图揭示数据的结构。机器学习在选股、择时等领域应用十分广泛。

**深度学习将是金融创新的新引擎**

深度学习是一种通过深度神经网络模型学习海量数据规律的方法。以循环神经网络、长短期记忆网络、卷积神经网络、受限玻尔兹曼机及深度信念网络等为代表的深度学习方法越来越成为投资和研究的热点。

**总结与展望**

本文所讨论与研究的问题是投资领域的热点，正在随着它的深入展开而受到越来越多的关注。正如文中所呈现的，大数据、机器学习、深度学习的应用提升了投资的效率，对于优化投资组合及加强投资盈利性也有促进作用。目前关于深度学习等技术在促进投资策略的优化科学上的应用仍处于探索阶段，值得深入观察和跟踪挖掘。


**丁鲁明**

dingluming@csc.com.cn

021-68821623

执业证书编号：S1440515020001

**研究助理：喻银尤**

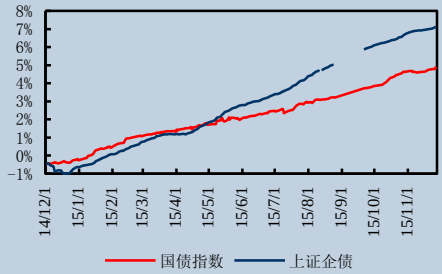
yuyinyou@csc.com.cn

021-68821600-808



发布日期：　2017 年 10 月 18 日


**市场表现**




**相关研究报告**

| | |
|---|---|
| 17.08.16 | 大数据研究之四：基于新闻热度的周期、成长、消费风格轮动配置 |
| 17.03.02 | 大数据研究之指标构建：机器学习之贝叶斯文本分类算法的实现 |
| 17.02.27 | 大数据周报：新闻热度未见突破，大盘继续谨慎；新闻情绪选股空头组合相对沪深 300 指数超额收益为-0.32% |
| 17.02.21 | 大数据周报：大盘维持谨慎，新闻情绪选股多头组合跑赢沪深 300 达 0.9% |
| 17.02.13 | 大数据周报：大盘相对谨慎，重点关注沪深 300 新闻情绪选股多空组合 |
| 16.10.12 | 大数据研究之择时：基于新闻热度的多空策略 |

目录

## 图形目录

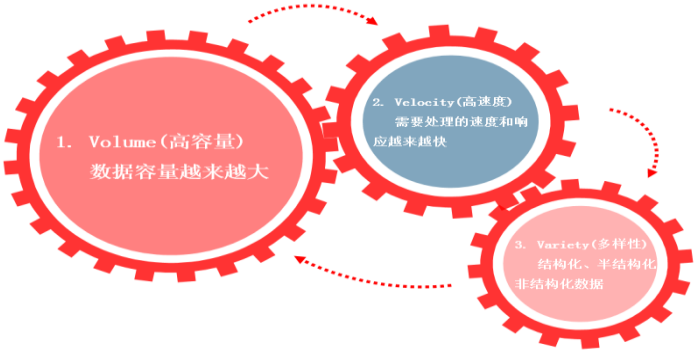


## 表格目录

# 一、 大数据与量化投资

大数据是高容量、高速度和高多样性的信息资产。“大”字鲜明体现出了大数据的三个突出特点：体积上，通过各种记录、交易信息、表格、文件存储等汇集的数据体量庞大；速度上，数据发送或接收的速度快，可以以批处理方式传输或接收，可以实时或接近实时地发送；种类上，数据通常以多种格式接收，有结构化的（例如 SQL 表或 CSV 文要件）、半结构化的（例如 JSON 或 HTML）或非结构化的（例如博客文章或视频消息）。大数据需要与之相配的新处理形式，对增强决策、发现洞见和过程优化有着积极作用。

**图 1：大数据的特点**




*数据来源：中信建投证券研究发展部*

另类数据（alternative data）指的是原始的或非结构化的数据，与公司档案、历史市场价格或投资者介绍等不同。

大数据和另类数据根据数据来源可分为三类：个体产生的数据（社交媒体、新闻及评论、互联网搜索和个人数据等），商业产生的数据（交易数据、公司数据、政府机构数据等）和由传感器产生的数据（卫星影像数据地理定位数据等）。

**图 2：大数据/另类数据的分类（数据来源视角）**


*数据来源：中信建投证券研究发展部*

此外，从金融与投资的视角，可将大数据按资产类别、投资角度、alpha 值等标准分类，如下图所示。

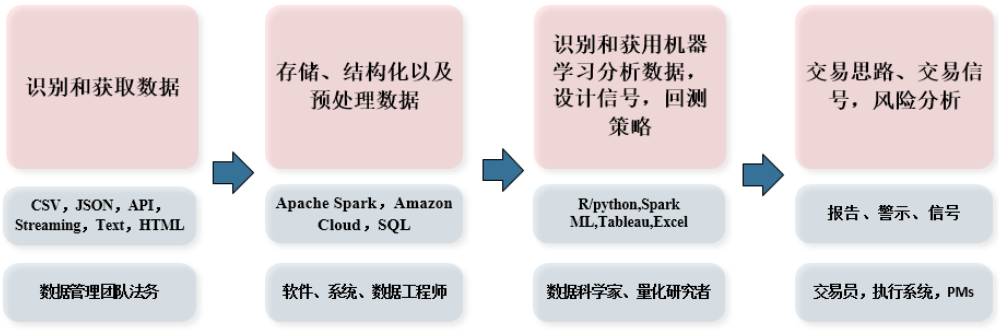
**图 3：大数据/另类数据的分类（投资视角）**

| 资产类别 | 投资类型 | 阿尔法值(扣除成本) | 可知性 | 数据处理阶段 | 数据质量 | 技术面 |
|---|---|---|---|---|---|---|
| 股票 | 宏观 | 独立可行 | 免费公开 | 原始数据 | 历史数据 | 频率 |
| 商品 | 特定部门 | 在资产组合中可行 | 十分清楚 | 半成品数据 | 极值 | 延迟 |
| 信用 | 特定股票 | 不可行 | 不知道 | 处理好完成的数据 | 缺失值 | 格式 |
| 利率 | 风险指标 | 容量 | 因保密不可知 | 交易信号 | 方法透明性 | API稳健性 |
| 外汇 | 量化信号 | 正交性 | 限制销售交易 | 研究报告和警示 | 支持结构 | 冲突和法律风险 |

*数据来源：中信建投证券研究发展部*

在采用大数据策略时，首先需要识别和获取数据。接下来要评估数据的质量，对数据进行预处理（如检测异常值、缺失值等）。此时定量研究人员通常会借助机器学习、测试策略和可视化技术等。随后，要基于数据集设计并测试交易策略。

**图 4：投资领域中大数据的工作流程**




*数据来源：中信建投证券研究发展部*

接下来将从数据的识别和获取出发，介绍不同来源的大数据的特点及其应用案例。

## 1.1 个体产生的数据

这种类型的数据源于个人的在线活动，可进一步分为社交媒体数据（推特、领英、博客等）、专业网站数据（新闻媒体、产品评论等）及网络搜索和自愿的个人资料（谷歌搜索、电子邮件收据等）。

针对个体产生的数据，情感分析是常用的数据处理方法。以社交媒体情感分析为例，通常对数据与信息的处理要经过以下几个步骤：

•提取实体：算法首先要提取说话人的身份，如地址、组织、品牌。

•提取主题和类别：算法要确定正在讨论的主题，例如“联邦远足”、“中东危机”、“iphone7”等。

•识别意图和情绪：使用自然语言处理（Natural Language Processing, NLP）对文章分配情感评分。

•分析相关性和影响：根据其与交易资产的关系量化其影响程度。

①iSentium 是一个有效的情感搜索引擎（sentiment research engine），能提供基于 Twitter 消息的实时情感时间序列，为投资者提供方法来判断新闻文章、推文等对市场的潜在影响。对 iSentium Daily Directional Indicators（DDI）的回测表明社交媒体情绪分析可以被用来预测短期市场走势。J.P.Morgan 基于 DDI 指数构建了 JPUSISEN 指数，来进行标准普尔 500 指数盘中多头或空头持仓。

在构建 DDI 指数时，首先选出标准普尔 500 指数最具代表性的 100 只股票，通过推特量和实际波动率（realized volatility）测量值进行过滤。使用 NLP 算法给推特分配情感得分。通过加总推特得分，在上午 8:30 到下午 4:30 之间每分钟会有特定的情感等级。一天的情感通过以指数加权移动平均处理过去十天的数据产生。最后，通过过去两天情绪评分的线性回归预测标准普尔 500 指数的回报，β 通过卡尔曼滤波处理。

②Ravenpack 分析非结构化数据集来为专业投资认识提供结构化的 granular indicators。非结构化数据集包括高级通讯社、受监管的新闻提供者、新闻稿和 19000 多份网络出版物。从 RavenPack 提供的信息流（news feed）可计算出每日情感分数（Daily Sentiment Score）。

Ravenpack 每个事件提供了 50 个数据字段。J.P.Morgan 分析了每个资产自 2005 起的数据。首先，将某日的所有独特的事件根据货币、商品或国家名称等赋予某“ENTITY_NAME”。设定截止时间为下午 4 点，以反映纽约市场收盘价。RavenPack 提供了一个 0 到 100 间的整数，称为“RELEVANCE”。更高的 RELEVANCE 值表明前述 ENTITY 的提及对于新闻更重要。因此，RELEVANCE 发挥了过滤的作用，其值一旦小于 75，相关新闻将会去掉。此外，RavenPack 提供了在 -1.00 和 1.00 之间表示给定 ENTITY 新闻情感的“EVENT_SENTIMENT_SCORE(ESS)”。ESS 的平均值即为当天的情感值。

## 1.2 商业产生的数据

该类型数据包括由公共机构提供的数据（如政府）、商业交易数据（包括电子商务、信用卡消费、交易数据等）和其他私人机构的数据（如特定行业的供应链数据）。

BBVA US(Banco Bilbao Vizcaya Argentaria)利用情感分析分析了联邦公开市场委员会（Federal Open Market Committee，FOMC）的声明。由于不同的词语可以表达不同的情感，因此可以通过统计词语可以来估计每种情感的水平。例如，计算正面词语（“gains”、“rebound”等）的总占比，可对正面情绪进行估计。

BBVA US 测量了三种主要的情绪：积极的、消极的和不确定的。关于积极情绪的时间趋势，呈现出三大模式。首先，在经济危机之前，有几次 FOMC 声明中没有任何积极的言论。这清楚地显示出决策者对经济的担忧。第二，在经济危机期间，积极情绪显著上升，在危机结束的 2009 年 6 月达到顶峰。虽然这看起来有违直觉，但这一趋势实际上表明，美联储试图用越来越多的积极词汇来提振市场信心。第三，在经济走出衰退后，积极情绪水平基本保持稳定。尽管经济复苏一直低于预期，但 BBVA US 的分析显示，美联储试图通过保持积极情绪稳定来提振信心。

关于消极情绪的趋势，2008 年 12 月，当第一次量化宽松开始时，它达到了顶峰。在这段时间里，负面词汇的大量使用表明了美联储使用非传统手段拯救经济的决心。在经济复苏中，消极情绪逐渐消退，反映了经济状况的改善。与积极情绪类似，经济衰退后消极情绪没有出现明显的趋势。美联储显然试图在积极和消极情绪之间实现微妙的平衡，以避免对金融市场产生不良影响。

## 1.3 由传感器产生的数据

由传感器产生的数据可进一步分为卫星数据、地理定位数据、和其他的传感器产生的数据。

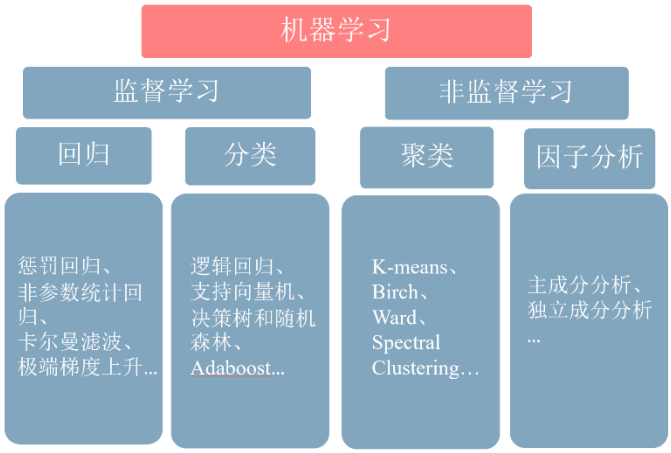
Advan Research 通过跟踪智能手机来估计实体店的客流量。数据经过客户同意后通过安装在手机上的带有地理位置码的应用程序收集。这些应用程序使用 WiFi、蓝牙和蜂窝信号等跟踪位置，以提高准确性。

Advan 的数据可用于估计公开上市交易公司的收入。该公司业务覆盖美国人口的 30%，追踪的设备约为每天 2500 万、每月 60mm，每天分析的数据点超过 3B，每天计算 1033050 处位置的交通流量，其中 498749 处被人工审核。数据可以被映射到包括零售、大卖场、超市、宾馆、医院、餐厅、电影院、游乐场、和快餐公司等 80 个不同的标准普尔 500（共有 381 股票）股票上。例如，2016 年 12 月 2 日的 Advan 研究发现 Lululemon 的 QTD 流量上升了 7%。这促使他们预测 LuLulemon 的销量将打破华尔街共识。12 月 7 日，当公司宣布销售业绩时，这一说法得到了证实。这一销售上的惊喜使该股在一天内从 58 美元涨到 67 美元。

# 二、 投资领域常用的机器学习方法

机器学习是一门多领域交叉学科，主要涉及领域包括计算机科学、统计学及数学等。机器学习的目的是使计算机能够从他们在某些任务中的经验中学习，从而随着经验的增长不断提高机器的性能。在金融学领域，机器学习尝试发现变量之间关系，在经过给定数据的训练后，能够在新数据的基础上预测结果。在机器学习工作时，共有训练数据、验证数据和测试数据三组数据，使用训练数据构建模型，使用验证数据验证模型，使用测试数据检查验证后的模型的表现。之后，完成训练的模型可以根据新数据进行预测。机器学习可分为监督学习（如回归和分类）及非监督学习（因子分析和流型识别）。监督学习试图找到一种规则或一个方程来预测变量，常运用高级回归方法来评估预测能力更高、并且对系统变化最稳定的模型。非监督则学习试图揭示数据的结构。

**图 5：机器学习方法分类**




*数据来源：中信建投证券研究发展部*

**表 1：机器学习解决的问题及使用的方法**

| 问题 | 数据分析技巧 |
|---|---|
| 给定输入集，预测资产价格方向 | 支持向量分类器，Logisitic 回归 |
| 一个资产的大幅波动是怎么影响其他资产 | 脉冲响应函数、格兰杰因果 |
| 是否一个资产的价格偏离了与其相关的资产 | 一对多分类器 |
| 那些资产价格会共同变动 | AP 聚类算法、流行嵌入 |
| 哪些因素驱动资产价格 | 主成分分析 |
| 当前的市场机制是怎样的 | 隐马尔科夫模型、soft-max 分类 |
| 事件的概率有多大 | 决策树、随机森林 |
| 市场压力的最共有的信号是什么 | K-means 聚类分析 |
| 寻找噪音数据中的信号 | SVM、Low-pass 过滤器 |
| 基于大数据样本预测波动率 | SVM |
| 一篇文献/文字的情绪基调是怎样的 | 词包、词库 |
| 一篇文献/文字的主体是什么 | 词频分析 |
| 资产价格是否会回调或大幅变动 | 成分分析 |
| 最优执行速度应该是什么 | 部分可观测的马尔科夫过程 |

*数据来源：中信建投证券研究发展部*

## 2.1 监督学习

在监督学习中，算法在已知数据及其对应的输出的情况下，通过已有的训练样本去训练得到一个最优模型，再利用这一模型对新数据进行预测。监督学习可进一步分为回归方法和分类方法。回归试图根据一些输入变量来预测输出变量。分类方法试图将输出变量分组或分类。例如，模型的输出可能是一个二进制操作，如“买入”或“卖出”。换言之，回归是对连续变量的预测，分类则是对离散 变量的预测。

具体而言，回归方法包括的模型有：

•惩罚回归：Lasso、岭回归（Ridge）、弹性网络（Elastic Net）。

•非参数统计回归：局部加权回归（Loess）、k 最邻近规则（K-Nearest Neighbor）。

•卡尔曼滤波。

•极端梯度上升。

分类方法包括的模型有：

•逻辑回归。

•支持向量机。

•决策树和随机森林。

•Adaboost。

2.1.1 回归

①惩罚回归

虽然简单回归模型可以被看作机器学习的一种方法，但是线性回归具有天然的缺陷。线性回归在处理离群值、大量变量、相互之间有相关关系的变量以及展现出非线性行为的变量时会出现问题。因此，在利用大数据时，由于模型中包括了大量的变量，并且不确定这些变量之间是否有相关关系，线性回归模型可能会提供非理性的交易策略。所以，为了对普通线性模型的问题进行修正，在存在大量潜在相关变量的情况下构建出产生更加稳健的输出结果的模型，需要采用惩罚回归模型。

以普通线性回归为例：$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \varepsilon$

在通过普通最小二乘法的方式来构建模型时，容易产生谬误回归或过大的 β 系数。惩罚回归的几种方式就是通过增加一个惩罚项来防止以上问题。

Lasso: Minimize Historical Sum of $(y - (\beta_0 + \sum_{i=1}^{n} \beta_i x_i))^2 + \alpha \sum_{i=1}^{n} |\beta_i|$

Ridge: Minimize Historical Sum of $(y - (\beta_0 + \sum_{i=1}^{n} \beta_i x_i))^2 + \alpha \sum_{i=1}^{n} \beta_i^2$

Elastic Net: Minimize Historical Sum of $(y - (\beta_0 + \sum_{i=1}^{n} \beta_i x_i))^2 + \alpha_1 \sum_{i=1}^{n} |\beta_i| + \alpha_2 \sum_{i=1}^{n} |\beta_i|^2$

Lasso 回归、岭回归、弹性网络分别增加了系数绝对值的和、系数绝对值的平方和和以及前两者的组合作为惩罚项。

可使用 Lasso 方法预测跨资产动量模型（cross-asset momentum mo）中 4 种资产一天的回报：标准普尔 500 指数、7-10 年的国债指数、美元指数和黄金。模型的输入变量选取以上四种资产过去一个月、三个月、六个月、还有十二个月的收益，在进行回归之前所有的输入变量都将会被标准化。模型使用的数据集是滚动的 500 个交易日的数据，通过移动窗口的方法进行预测，模型三个月更新一次。如果一种资产第二天的收益为正则做多，为负则做空。

如果对每一种资产都分别进行模型预测，可以发现，lasso 预测的年化收益和夏普指数都更高。

②非参数统计回归

在参数回归中，模型有一组参数比如 β 的线性组合来描述。而对于非参数回归，模型没有预先确定的形式，而是根据从数据得出的信息来构造。

K 最近邻(kNN，k-Nearest Neighbor)分类算法是非参数统计回归的一种主要方法。K 最近邻指的是 k 个最近的邻居，即每个样本都可由它最接近的 k 个邻居来代表。kNN 算法的核心思想是如果一个样本在特征空间中的 k 个最相邻的样本中的大多数属于某一个类别，那么该样本也属于这个类别，也相应具有该类别样本的特性。

kNN 算法的步骤是：首先，初始化距离为最大值，并计算未知样本和每个训练样本的距离 dist，得到目前 K 个最临近样本中的最大距离 maxdist。如果 dist 小于 maxdist，那么将该训练样本作为 K-最近邻样本。之后重复上述步骤直到算完未知样本和所有训练样本的距离。根据 K-最近邻样本中每个类标号出现的次数，频率最大的类标号即为未知样本的类标号。

③卡尔曼滤波

卡尔曼滤波是一个最优化自回归数据处理算法，利用线性系统状态方程，通过系统输入输出观测数据，对系统状态进行最优估计。卡尔曼滤波在存在不确定性的情况下，通过一系列观测值来估计和预测一个变化的系统的参数。简言之，该算法可分两步完成。第一步需得到对当前状态的估计和随之而来的估计误差，下一步中利用已有估计值和新的观测来进行预测。

该算法研究的动力系统可以用状态空间模型来描述。首先，系统现在的状态也即 t 时刻的状态是无法观察的，但根据此前 t-1 时刻的状态，能够部分推知 t 时刻状态。由于外界的随机影响，这一推测存在不确定性，如下带有高斯噪声的线性表达式所示。

$$X_t = AX_{t-1} + W_t, \quad W_t \sim N(0, Q)$$

接下来，尽管无法直接观测到 t 时刻的状态，但可以通过测量得到与 t 时刻状态相关的测量值。显然，这些测量值也受到高斯噪声的影响。

$$Z_t = HX_t + V_t, \quad V_t \sim N(0, R)$$

卡尔曼滤波将结合以上信息对系统变量进行符合高斯分布的最优化预测 $N(X_{t|t}, P_{t|t})$，其均值和协方差为：。

$$x_{t|t}=x_{t|t-1}+K(z_t-Hx_{t|t-1})$$
$$P_{t|t}=P_{t|t-1}-KHP_{t|t-1}$$

式中 K 表示卡尔曼增益，可由下式得到：$K=P_{t|t-1}H^T(HP_{t|t-1}H^T+R)^{-1}$。其中

$$\begin{pmatrix}X\\Z\end{pmatrix}\sim N\left(\begin{pmatrix}\mu_X\\\mu_Z\end{pmatrix},\begin{pmatrix}\Sigma_{11}&\Sigma_{12}\\\Sigma_{21}&\Sigma_{22}\end{pmatrix}\right)$$
$$X|(Z=z)\sim N(\hat{\mu},\hat{\Sigma})$$

$$\hat{\mu}=\mu_X+\Sigma_{12}\Sigma_{22}^{-1}(Z-\mu_Z)$$
$$\hat{\Sigma}=\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

卡尔曼滤波是在线性回归模型的基础上进行的拓展，使 β 系数能够随着时间变化，常用在统计交易以及波动性预测中。在金融中卡尔曼滤波可以用来推测趋势，为金融信号降噪、推测无法观测到的经济活动以及描绘资产和市场间的动态关系。

卡尔曼滤波可以应用在配对交易中。选择一组交易型开放式指数基金（ETFs），分别为 iShares MSCI 澳大利亚交易型开放式指数基金（Australia ETF）和加拿大交易型开放式指数基金（Canada ETF）。

首先确定 β 系数，也就是协整因子。

$$y_t=\beta_t x_t+v_t,\ v_t\sim N(0,V_v)$$

假定 β 相对时间独立、不是常数。残差$v_t$不变。出于简化考虑，假设 β 的变化是服从随机行走的：

$$\beta_t=\beta_{t-1}+w_t,\ w_t\sim N(0,V_w)$$

从而构建起状态方程和测量值方程：

$$\beta_t=I_t\beta_{t-1}+w_t,\ w_t\sim N(0,V_t)$$，其中$I_t=1, V_t=V_w$。

$$z_t=H_t x_t+v_t,\ v_t\sim N(0,R_t)$$，其中$z_t=y_t, H_t=x_t$and$R_t=V_v$。

由卡尔曼滤波估计 β 的最佳值。卡尔曼增益可以表示为$K_t=\frac{x_t}{S_t x_t+\gamma}$，其中$\gamma=\frac{P_{t-1}}{V_v}$为信号杂音比（SNR）。

如果 SNR 小，那么测量值噪声大、所含的信息不足，所以给予此前的$\hat{\beta}_{t|t-1}$更大权重。反之，则给观察值更大

权重。将变量代入$\hat{x}_t = \hat{x}_{t-1} + K(z_t - H\hat{x}_{t-1})$，可得

$$\hat{\beta}_t = \hat{\beta}_{t-1} + \frac{\Sigma_{t-1}S_{t,CA}}{S_{t,CA}\Sigma_{t-1}S_{t,CA} + \gamma^2}\left(S_{t,AU} - \hat{\beta}_{t-1}S_{t,CA}\right)$$

$$= \hat{\beta}_{t-1}\left(1 - \frac{S_{t,CA}\Sigma_{t-1}S_{t,CA}}{S_{t,CA}\Sigma_{t-1}S_{t,CA} + \gamma^2}\right) + \frac{\Sigma_{t-1}S_{t,CA}S_{t,AU}}{S_{t,CA}\Sigma_{t-1}S_{t,CA} + \gamma^2}$$

因此，如果信号杂音比很大，主要选择近期的观测值来估计β：$\hat{\beta}_t \approx \frac{S_{t,AU}}{S_{t,CA}}$。反之，则主要采用的是此前的信息：$\hat{\beta}_t \approx \hat{\beta}_{t-1}$。通过对比可以看出相比于线性回归方法，卡尔曼滤波对于价格的变动的反应更加灵敏。

配对交易信号主要取决于残差$v_t$，它应该在 mean zero 附近波动。在每个交易日结束时，使用 ETF 的收盘价来更新对$\beta_t$的估计，然后计算残差：

$$v_t = S_{t,AU} - \beta_t S_{t,CA}$$

还要记录残差的不确定性（uncertainty），用它来判断残差的大小是否足够大，从而触发策略：

如果$v_t \geq k\delta_t$，买入$\beta_t$单位的 MSCI Canada，卖出一单位 MSCI Australia；

如果$v_t \leq k\delta_t$，卖出$\beta_t$单位的 MSCI Canada，买入一单位 MSCI Australia。

④极端梯度上升

boosting 分类器属于集成学习模型，其基本思想是通过不断迭代使预测预测能力较弱的回归树算法具有更强大的预测能力，在迭代过程中采取梯度下降的最优化算法来不断接近目标值。极端梯度上升是梯度上升方法的优化版本。

Shubharthi Dey 和 Yash Kumar 等运用极端梯度上升来预测苹果公司（Apple Inc.）和雅虎公司（Yahoo! Inc.）的股价走势。研究者选择的两支股票信息包括了 closing price/opening price/High/Low and Volume，并用指数平滑法处理了数据。研究选用的股市技术指标包括了相对强弱指数（RSI）、随机指标（stochastic oscillator）、威廉指标（Williams %R）、指数平滑移动平均线（MACD）、价格变化率（price rate of change）和能量潮指标(On Balance Volume)。极端梯度上升算法如下：

**表 2：极端梯度上升算法**

D：训练数据

用常数初始化模型 $F_0(x) = arg\min_{\gamma}\sum_{i=1}^{n} L(y_i,\gamma)$

for do m=0 to M

计算伪残差

用基础学习器拟合伪残差

$T_i$ = 新决策树()

特征变量$_1$ = 随机特征选择函数($D_i$)

$T_i$, 训练函数($D_1$, 特征变量$_1$)

计算$\gamma_m$

更新模型

输出$F_m(x)$

*数据来源：中信建投证券研究发展部*

在计算混淆矩阵（confusion matrix）后，可得预测的准确性结果。据研究者，相比于人工神经网络(ANN, artificial neural network)81%的准确率及随机森林 78.81%的准确率，极端梯度上升的表现更好。

2.1.2 分类

①逻辑回归

逻辑回归与普通线性回归的模型形式基本相同，区别在于因变量。逻辑回归的因变量是二元变量，通常用于衡量某时间的可能性或者进行分类。

Carol Anne Hargreaves, Prateek Dixit, 和 Ankit Solanki 用逻辑回归来选股。他们选择了澳大利亚股市中健康（healthcare）和金融（finance）两个板块的股票来研究。首先运用随机森林算法（Random Forest Importance Algorithm）来选择出每个板块中的自变量。

**表 3：逻辑回归模型的输入变量**

| 健康部门 | 金融部门 |
|---|---|
| ROA | 增长率 |
| ROE | 每股净利润 |
| 每股净利润 | 账面价值 |
| 每股收入 | ROE |
| 季度收入增长率 | 每股收入 |
| 增长率 | 分析师观点 |
| EV/EBITDA | EV/R |
| 市净率 | 市销率 |

*数据来源：中信建投证券研究发展部*

在逻辑回归模型中，输出为概率得分，范围从 0 到 1。如果预测的概率得分超过 0.5，那么这些股票被认为有上升的趋势，并在未来获利。因此，根据该模型的输出，概率得分最高的六只股票被选中。

对于研究使用的交易策略，相对强弱指标（RSI）作为评价有利买卖时间的技术指标。该指标通常用于评估股票是超买或超卖。在研究中，每支股票投资 10000 澳元，每个模型的投资总和为 60000 澳元。只有当 RSI 指数在 45 以下时才买入股票。当 RSI 超过 65 时，仔细监控行情，因为这是卖出股票的信号。如果 RSI 保持上升，即使其值超过 65，也继续持有股票。而一旦 RSI 值下降，股票就被卖出。

在二十天的交易期结束时，由结果可知，逻辑回归结果比部门指数（sector index）和 AOI（All Ordinaries Index）的盈利能力都好。

②支持向量机

支持向量机是通过支持向量运算的分类器，其基本策略是寻找使到每边数据点间隔最大的超平面，也因此是最大间隔超平面。

支持向量机可用在线性分类和非线性分类中。对于线性分类，训练数据中的每个数据都有 n 个属性和一个二类类别标志，因此可以找到一个 n-1 维的超平面将数据分为两类。最佳超平面的约束条件是这个超平面到每边最近数据点的距离是最大的。对于非线性分类，可通过将线性空间经过映射变成高维度以及使用核函数解决。

③决策树与随机森林

决策树代表的是对象属性与对象之间的一种映射关系。树中的每个节点表示某个对象，而每个分叉路径则代表可能的属性值。每个叶节点则对应根据从根节点到该叶节点所经历的路径所表示的对象的值。决策树是在已知各种情况发生概率的基础上，通过构成决策树来求取净现值的期望值大于等于零的概率，评价项目风险，判断其可行性的决策分析方法，是直观运用概率分析的一种图解法。

随机森林算法是对决策树的改进，是利用多棵树对样本进行训练并预测的一种分类器。该算法用随机的方式建立起一棵棵决策树，然后由这些决策树组成一个森林，其中每棵决策树之间没有关联，当有一个新的样本输入时，就让每棵树独立的做出判断，按照多数原则决定该样本的分类结果。随机森林是在 bootstrap 和 bagging

算法基础上发展而来的。Boostrap 算法的思想是在样本数量不大且分布情况未知时，通过从原始样本中随机抽取多个样本来估计真实分布。其基本步骤为从原始数据集中有放回地抽取一定数量的样本并计算统计量，再重复前述步骤 n 次，从而得到 n 个统计量，计算 n 个统计量的样本方法以得到总体的统计量的方差。Bagging 算法则现在样本中用 Bootstrap 抽样并对样本建立分类器，之后重复前述步骤 n 次，建立 n 个分类器，在这些分类器上测试数据，最后投票选出一类。

普特南投资（Putnam Investments）的 Eric H.Sorensen、所罗门美邦国际(Salomon Smith Barney)的 Keith L.Miller 及 Chee K.OOL 曾利用决策树模型来选股。他们根据罗素 1000 指数(Russell 1000 Index)的分类标准从中选择了“技术”股（technology）这一门类来进行研究，采集了从 1992 年到 1999 年 10 月末的数据。首先，他们计算了每只股票在每个月的总回报。之后，通过将单个股票的月回报减去股票回归的中位数来衡量股票的相对回报，从而可将股票每月的优劣表现程度归类（the level of outperformance or underperformance）。显然，某月一只股票可能在趋于下降的市场中提供很好的收益，也可能在趋于上升的市场中仅仅提供较差的收益水平。研究的目标是从选出那些高于中位数水平的股票。所以，因变量自然为二元变量，将样本均分。自变量则选择了市销率(Price-to-sales ratio,PS)、EPS 及 ROA 变化等，如下表所示。每月的每个变量都进行了五分位数处理（quintiled）。

在第一个静态树模型中，研究者假定了输入变量和股票市场表现间关系的高度稳定性。数据集中 1993 年 3 月到 1995 年 12 月的数据作为训练数据，1996 年 1 月到 1999 年 10 月数据作为测试数据。决策树最顶层的划分标准选取为 EPS MOM 也即预测修正（estimate revision），接续的划分标准为 ROA。相应的，在不考虑交易费用的情况下，long minus short 的组合一年的平均回报约为 17.95%，标准差约为 13.92%。Wilcoxon 测试显示，被划分为表现优异的股票（outperform）比表现不佳的（underperform）股票平均多 1.40%的额外回报，在 95%的置信区间上显著。

上述静态模型对股票选择有一定的预测力。研究者还建立了一个动态重复估计（dynamic re-estimation）的“变化的树”（evolving tree）模型，试图获得更好的预测效果。这一模型每个月都会重新利用从起始月份到该月的数据来重新估计决策树。从 1995 年 12 月底开始，用 1993 年 2 月到 1995 年 12 月的数据来预测一棵树，并参考预测结果来为 1996 年 1 月的股票分类。在接下来的每个月里，将新数据加入到数据集中并进行新的预测。

④隐马尔科夫模型

隐马尔可夫模型（Hidden Markov Model，HMM）用来描述一个含有隐含未知参数的马尔可夫过程。它从可观察的参数中确定该过程的隐含参数。然后利用这些参数来作进一步的分析。在简单的马尔可夫模型（如马尔可夫链），所述状态是直接可见的观察者，因此状态转移概率是唯一的参数。在隐马尔可夫模型中，状态是不直接可见的，但输出依赖于该状态下，是可见的。每个状态通过可能的输出记号有了可能的概率分布。因此，通过一个 HMM 产生标记序列提供了有关状态的一些序列的信息。隐马尔可夫模型常用于估计经济趋势以及预测指数价格。

隐马尔可夫模型的组成为：

$\lambda=(A,B,\pi)$

简记为 $\lambda=(A,B,\pi)$

$N$ 状态数目

$M$ 可能观测数目

$A$ 状态转移概率矩阵

$B$ 给定状态下，观测概率分布

$\pi$ 初始状态的概率分布

状态转移概率矩阵为：

$$A=[a_{ij}]_{N\times N}$$
$$a_{ij}=P(i_{t+1}=S_j\mid i_t=S_i),\quad 1\le i,j\le N$$
$$a_{ij}\ge 0,\ \sum_{j=1}^{N}a_{ij}=1$$

观察值概率分布矩阵为：

$$B=[b_j(k)]_{N\times M}$$
$$b_j(k)=P(O_t=v_k\mid i_t=S_j),\quad 1\le j\le N,1\le k\le M$$
$$b_j(k)\ge 0,\ \sum_{k=1}^{M}b_j(k)=1$$

初始状态概率分布为：

$$\pi=[\pi_i]_{N}$$
$$\pi_i=P(i_1=S_i),\quad 1\le i\le N$$
$$\pi_i\ge 0,\ \sum_{i=1}^{N}\pi_i=1$$

隐马尔可夫模型的几种算法为：

1、向前算法。

向前变量：$\alpha_t(i)=P(o_1,o_2,\ldots,o_t,i_t=q_i\mid\lambda),\quad 1\le t\le T$

算法：

1）初始化：$\alpha_1(i)=\pi_i b_i(o_1),\quad 1\le i\le N$

2）递归：$\alpha_{t+1}(j)=\left[\sum_{i=1}^{N}\alpha_t(i)a_{ij}\right]b_j(O_{t+1})$，$t=1,2,\ldots,T-1$；$j=1,2,\ldots,N$

3）终结：$P(O|\lambda)=\sum_{i=1}^{N}\alpha_T(i)$

2、向后算法。

向后变量：$\beta_t(i)=P(O_{t+1},O_{t+2},\ldots,O_T|q_t=S_i,\lambda)$，$t=T-1,T-2,\ldots,1$

算法：

1）初始化：$\beta_T(i)=1$，$i=1,2,\ldots,N$

2）递归：$\beta_t(i)=\sum_{j=1}^{N}a_{ij}b_j(O_{t+1})\beta_{t+1}(j)$，$t=T-1,T-2,\ldots,1$；$i=1,2,\ldots,N$

3）终结：$P(O|\lambda)=\sum_{i=1}^{N}\pi_i b_i(O_1)\beta_1(i)$

3、Viterbi 算法。

1）初始化：$\delta_1(i)=\pi_i b_i(O_1)$，$\psi_1(i)=0$，$i=1,2,\ldots,N$

2）递归：$\delta_t(j)=\left[\max_{1\le i\le N}\delta_{t-1}(i)a_{ij}\right]b_j(O_t)$，$t=2,3,\ldots,T$；$j=1,2,\ldots,N$

$\psi_t(j)=\arg\max_{1\le i\le N}\left[\delta_{t-1}(i)a_{ij}\right]b_j(O_t)$，$t=2,3,\ldots,T$；$j=1,2,\ldots,N$

3）终结：$P^*=\max_{1\le i\le N}\left[\delta_T(i)\right]$，$q_T^*=\arg\max_{1\le i\le N}\left[\delta_T(i)\right]$

4）路径回溯：$q_t^*=\psi_{t+1}(q_{t+1}^*)$，$t=T-1,T-2,\ldots,1$

4、Baum-Welch 算法。

1）初始化：随机地给 $\pi_i$，$a_{ij}$，$b_j(k)$ 赋值（满足概率条件），得到模型 $\lambda_0$，设 $l=1$。

2）EM 步骤

E 步骤：由 $\lambda_l$ 根据公式（1）和（2），计算期望值 $\xi_t(i,j)$ 和 $\gamma_t(i)$。

M 步骤：用 E 步骤所得的期望值，根据公式（3）重新估计 $\pi_i$，$a_{ij}$，$b_j(k)$，得到模型 $\lambda_{l+1}$

3）循环计算：$l=l+1$，重复 EM 步骤，直至 $\pi_i$，$a_{ij}$，$b_j(k)$ 值收敛。

公式（1）：

给定 HMM 和观察序列，在时间 t 位于状态 *i*，时间 *t*+1 位于状态 *j* 的概率：

$$\xi_t(i,j)=P(q_t=S_i,q_{t+1}=S_j|O,\lambda)$$
$$=\frac{P(q_t=S_i,q_{t+1}=S_j,O|\lambda)}{P(O|\lambda)}$$
$$=\frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{P(O|\lambda)}$$
$$=\frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}$$

公式（2）：

给定 HMM 和观察序列，在时间 t 位于状态 *i* 的概率：

$$\gamma_t(i)=\sum_{j=1}^{N}\xi_t(i,j)$$

公式（3）：

$$\pi_i=q_i$$ 在时刻 $t=1$ 处于状态 $S_i$ 的概率 $=\gamma_1(i)$

$$q_{ij}=\frac{Q\text{从状态}i\text{转移到状态}j\text{的期数}}{Q\text{从状态}i\text{转移到任何状态的期数}}=\frac{\sum_{t=1}^{N-1}\delta_t(i,j)}{\sum_{t=1}^{N-1}\lambda_t(i)}$$

$$b_j(k)=\frac{Q\text{处于状态}j\text{并观测到}v_k\text{的期数}}{Q\text{处于状态}j\text{的期数}}=\frac{\sum_{t=1}^{N}\lambda_t(j)\times\delta(O_t,v_k)}{\sum_{t=1}^{N}\lambda_t(j)}$$

Youngstown State University 的 Nguyet Nguyen 和 Ned Davis Research Group 的 Dung Nguyen 尝试基于经济指标用隐马尔可夫模型来进行月度选股。该模型的 4 个宏观经济指标为：通货膨胀，以 12 个月 CPI 变化衡量（12-month changes (%)in CPI）；工业生产指数（INDPRO）；股票市场指数，以 1 个月内标准普尔 500 指数变化衡量（one-month changes of the S&P 500 index）、市场波动性（以芝加哥期权交易所（CBOE）波动性指数衡量）。数据集包括从 1990 年 1 月到 2014 年 12 月的所有 S&P500 的股票。研究者选择的股票收益因子来自股票估值和成长性两方面，估值方面选择了 E/P（earnings/price）、自由现金流与公司价值之比（the free cash flow/enterprise value）和销售额与公司价值之比(the sales/enterprise value)。出于数据处理的考虑，对传统指标进行了转换。股票成长性方面选择了：长期每股收益增长率（Long-term earnings per share growth, L-T EPS growth），即基于 5 年变动趋势回归线的预计的长期每股收益增长率；长期销售额增长率，计算方法与 L-T EPS 增长率相似。

研究的出发点是注意到了股市在不同的经济状态下表现不同，并且在同样的经济状态下指数有类似的表现。因此，研究者选择了受宏观经济状态影响最大的股票因子并用它们为股票排序。每个月研究者从 S&P500 的股票中根据股票排行（stock ranking）选择 50 支加入组合。每月底，首先运用宏观经济变量的月度数据计算隐马尔可夫模型的参数，将变量划分为某一状态并预测下个月变量所处的状态。完成预测后，返回最近 20 年的历史数据中寻找与之相似的情形，检查前述股票因子的表现，并据此对因子们从 1-5 进行排序，算出每个因子的权重（其排序值/1 到 5 的和）。每支股票所得分是每个因子的排序值乘权重后的总和，被限制在 1-100 内。由此选出分数最高的 50 支作为股票组合。每个月，买进组合中新增股票，卖出离开组合的股票。

2.2 非监督学习

非监督学习直接面对的是数据集中整体的资产的回报，其中没有区分哪些是解释变量，哪些是被解释变量。聚类和因子分析是两种常用的非监督学习方法。

2.2.1 聚类

聚类是根据客体属性对一系列未分类的客体进行类别的识别，把一组个体按照相似性归成若干类。

①K-means 算法

该算法首先随机地选择 k 个对象作为初始的 k 个簇的质心；然后对剩余的每个对象，根据其与各个质心的距离，将它赋给最近的簇，然后重新计算每个簇的质心；这个过程不断重复，直到准则函数收敛。通常采用的

准则函数为平方误差和准则函数，即 SSE(sum of the squared error)，其定义为：$SSE=\sum_{i=1}^{k}\sum_{p\in C_i}\|p-m_i\|^2$。SSE 是数据库中所有对象的平方误差总和，p 为数据对象，mi 是簇 Ci 的平均值。这个准则函数使生成的结果尽可能的紧凑和独立。

k-means 算法的具体步骤为：

(l) 给定大小为 n 的数据集，令 I=1，选取 k 个初始聚类中心

Zj(I)，j=1，2，3，…，k；

(2) 计算每个数据对象与聚类中心的距离 D(xi，Zj(I))，i=1，

2，3…n，j=l，2，3，…，k，如果满足

D(xi，Zk(I)) =min{D(xi，Zj(I))，i=l，2，3，…n}

则 xi∈C k；

(3) 计算 k 个新的聚类中心：

即取聚类中所有元素各自维度的算术平均数；

(4) 判断：若 Zj(I+1)≠Zj(I)，j=l，2，3，…，k，则 I=I+1，

返回(2)；否则算法结束。

其中的距离 D 可由欧几里得距离、曼哈顿距离和闵可夫斯基距离得到。

由于通过股价之间的相关关系来划定股票之间的相似性往往在经济下行时期不适用，更好的办法是用股票此前的成功和公司增长的潜力等来判定股票之间的相似性。研究者选择的相似性判断标准为两个财务指标的加权平均：收入/资产（revenues to assets）和净利润/资产（net income to assets），可表示为 $\frac{Revenues}{Assets}x+\frac{Net\ Income}{Assets}(1-x)$。通过该加权平均值之间的区别，股票被分为许多簇，从每一簇中挑选出夏普比率最高的股票即可组成一个投资组合。此外，延迟效应（a delay in effect from previous quarters of data）也被考虑进来。研究者考察了三种不同的延迟模型：one period，用最近一个季度公布的数据来为下一个季度进行投资组合；two period，用最近一个季度之前的季度的数据来进行投资组合；average，取前两种方法的平均值。

研究者的数据集为 2000 年到 2015 年的 NYSE 和 NASDAQ 的 229 支包含了每季度收入、净利润和资产数据的股票，并最终选定收入/资产的权重为 50%以及选用 one period 延迟模型。结果显示，在从 2000 年到 2007 年的前金融危机时期和 2009 年到 2015 年的后金融危机时期，算法选出了表现很好的投资组合，相比于 S&P500，该组合有更多的高峰值（high peaks）和更少的低谷值（low peaks）。金融危机期间则缺少明显的大的峰值。总体上，算法选出的投资组合比 S&P500 波动性更大但是表现也更好。

2.2.2 因子分析

因子分析是无监督学习的一个重要环节。因子分析的目的是确定数据的主要因素，或者确定数据的最佳代表。在一个多资产组合中，因子分析将识别主要因素，如动量、价值、进位、波动性、流动性等。

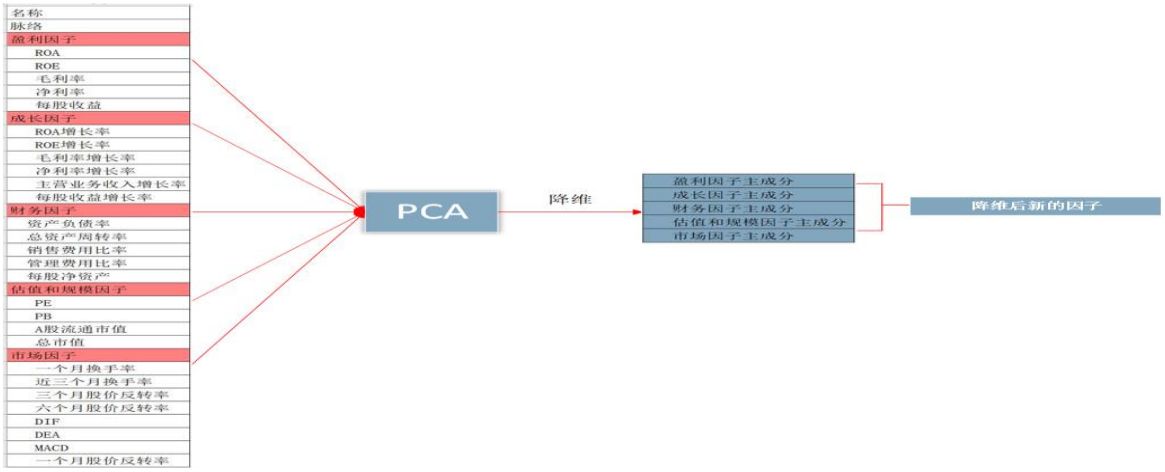
①主成分分析（PCA）

主成分分析（PCA）是降维的统计工具。给定市场数据，主成分分析可以将每个时间序列分解成不相关因素的线性组合。此外，PCA 可以通过计算由给定因素解释的方差的比例来量化每个因素对时间序列的影响。

在二级市场中，影响股票价格的因子有很多，大致上可以分为这么几类，分别是盈利因子、成长因子、财务因子、估值和规模因子、市场因子等。而每一个大类因子下又有许多因子，例如 ROA、ROE、毛利率、净利率、每股收益这些指标都可以视为盈利因子。而大量的因子数量无疑增加了因子选股的难度，而通过主成分分析，降低数据的维度不失为一种好的办法。

如下图所示，通过主成分分析，对盈利类因子、成长类因子、财务类因子、估值和规模类因子以及市场类因子进行降维，从而将 28 个因子根据他们的分类归结为五个重要因子。

**图 6：主成分分析的变量和因子**




*数据来源：中信建投证券研究发展部*

例如可以先计算五个盈利因子的协方差矩阵，再计算该矩阵的特征值和特征向量。接着选取累积贡献量较大的前几个主成分，再根据各自的贡献率构建它们的线性组合。最终得到的就是降低维度后的盈利因子。其他几个因子的降维方式也通盈利因子的降维方式一样。

# 三、 投资领域常用的深度学习方法

深度学习有多种定义，如：使用多层次的非线性处理单元进行特征提取和转换的算法；基于对数据的多个特征或表示的（无监督）学习，更高层次的特征是从低级特征派生而成的；更广泛的学习数据表示的机器学习领域中的一部分，能够促进端到端的优化（end-to-end optimation）。这些定义都共同强调了深度学习使用多层次

的非线性处理单元及更高层次的特征从低级特征派生而成的特点。

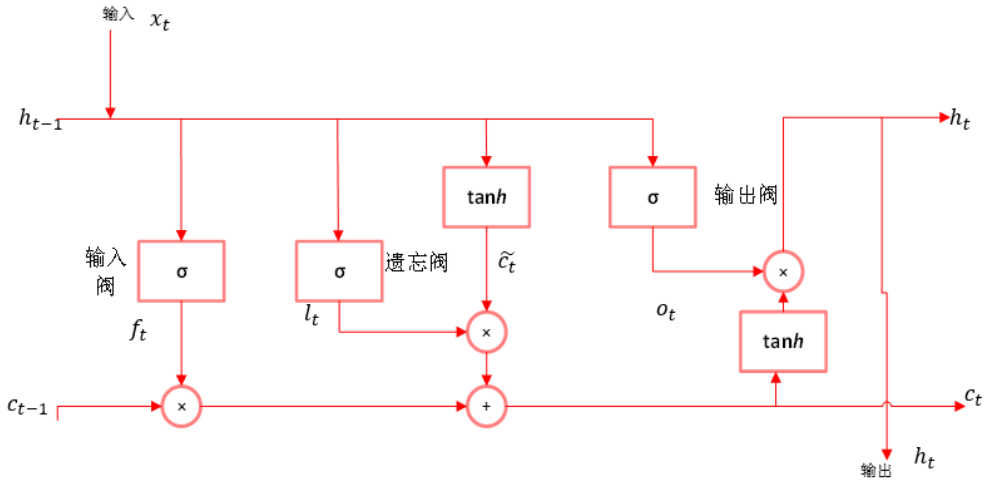
## 3.1 循环神经网络

传统的神经网络无法实现利用此前的信息推断后续事件的功能，循环神经网络（RNN, Recurrent Neural Networks）则解决了这个问题，因为它包含循环的网络，容许信息的持久化，这些循环使得信息可以从当前的步骤传递到下一步骤，但是这种信息的传达从前向后的传递，有着长期依赖的特征，比如在语言模型的预测中，如果已知信息和预测词之间位置之间的间隔很小，那么，RNN 可以轻松的使用此前的信息，但是如果场景变得更加复杂，已知信息和预测位置之间的间隔变得更大，RNN 将丧失连接已知信息到当前步骤的能力。

## 3.2 长短期记忆网络

经济学家靠 ARIMA 模型预测的时间序列模型。该模型对小数据集效果很好，可容纳时间序列的记忆效应，如持久性、均值回归、季节性等。在深入学习中，长短期记忆（Long short-term memory，LSTM）可类比于 ARIMA。LSTM 是一个循环神经网络，能记忆通过网络预先输入的信息。LSTM 对 RNN 进行了结构上的修改，来避免长期依赖问题。

**图 7：LSTM 神经元结构**




*数据来源：中信建投证券研究发展部*

这里，ht 代表神经元的输出，而 Ct 代表神经元的状态。系统的运行分为以下步骤：

A.计算需要被细胞丢弃的信息：$f_t=\sigma(W_f[h_{t-1},x_t]+b_f)$

B.计算需要添加到细胞中的信息 $i_t=\sigma(W_i[h_{t-1},x_t]+b_i)\ \tilde{c}_t=\tanh(W_c[h_{t-1},x_t]+b_c)$

C.更新细胞状态 $C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$

D.计算输出信息 $o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$ and $h_t = o_t * \tanh(C_t)$

Xiong 等人通过估计开盘价、走高、走低和收盘价来预测标准普尔 500 指数的每日波动。研究使用包含一个 LSTM 模块的单一 LSTM 隐藏层，输入每日标准普尔 500 指数的收益和波动。研究纳入了 25 个国内的谷歌趋势（domestic Google trends），涵盖行业和经济的主要领域。研究中使用每次批处理（batch）有 32 个采样器的 Adam 方法，以平均绝对百分误差（MAPE）作为目标损失函数。研究设置了最大的 LSTM 延迟以包含 10 个连续的观察。

结果表明 LSTM 方法比 GARCH、Ridge 和 LASSO 的效果都好。

## 3.3 卷积神经网络

卷积神经网络（Convolutional Neural Network，CNN）一种专门处理图像的特殊的多层神经网络，包括卷积层(alternating convolutional layer)和池层(pooling layer)。CNN 的基本结构一般包括两层，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。随着该局部特征被提取，它与其它特征间的位置关系也确定下来。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。

Ding 等人(2015) 利用从新闻头条中提取出的结构化信息来预测每天标准普尔 500 指数的走势。以 OpenIE 处理头条，以获得结构化的事件表示（执行器、动作、对象、时间）。相比于标准神经网络，神经张量网络通过成倍地结合事件参数来学习语义组合性（semantic compositionality）。

研究结合事件的短期和长期影响，使用了一个 CNN 来执行输入事件序列的语义组合。研究在卷积层顶部使用一个最大池层，使网络只保留由卷积层产生的最有效特性。长期和中期事件存于单独的卷积层。这两个层，以及用于短期事件的输入层都会将信息送到一个隐藏层中，之后送到两个输出节点。

研究从路透社和彭博新闻里提取出了 1000 万个事件。为了训练，使用随机参数替换事件参数的方式来破坏事件。在训练过程中，假设真实事件比损坏的事件得分更高。反之，则更新模型参数。

相比于股市预测词汇，结构化事件具有更好的特征。模型中使用的方法的效果比基准方法好 6%。研究对标准普尔 500 指数和 15 只股票进行了预测，结果显示可以以 65%的准确率预测标准普尔 500 指数。

## 3.4 受限玻尔兹曼机

受限玻尔兹曼机（restricted Boltzmann machine，RBM）是一种基于降维技术的神经网络。神经元在 RBM 形成两层，一层为可见单元（反映资产收益），一层为隐单元（反映潜在因素）。一层之内的神经元无论是可见的还是隐的相互之间没有连接。RBM 结构的灵感来自于统计物理学，特别地，Boltzman 分布被用在算法设计中，

使之命名为 RBM。

Takeuchi 和 Lee（2013）通过预测哪支股票比中位数有更高或更低的月度收益来提升动量效应。研究使用由堆栈 RBM 组成的自动编码器提取来自股票价格的特征，然后将之传送到前馈神经网络分类器。每个 RBM 包含由对称链连接的一层可见单位和一层隐藏单位。第一层有 33 个单位，用于输入某个时候某支股票的特征。对每个月 t 而言，特征包括 t-2 月到 t-13 月 12 个月份的特征，以及对应 t 月的近似 20 天的回报。通过计算关于所有股票每个月或每天的横切面的 z-得分，正则化每个回报特征。编码器最后一层隐藏单元的数量急剧减少，迫使降维。输出层有 2 个单元，对应于股票是否在本月收益的中位数之上或以下。最后一层的大小是 33-40-4-50-2。

在预训练期间，他们将数据集分割更小的、不重叠的小批量（mini-batches）。之后，展开 RBM 形成一个编码解码器，通过反向传播进行精调。研究考虑了所有在 NYSE、AMEX 或纳斯达克上交易的价格高于 5 美元的股票，以 1965 年至 1989 年的数据（每月 848,000 支股票样本）进行训练，以 1999 年至 2009 年的数据（每月 924,300 支股票样本）进行测试。一些训练数据用于验证层数和每层单位数。

结果显示，总体的准确率大约是 53%。考虑到预测前面十分之一股票和后面十分之一股票之间的差别，月收益约为 3.35%，年化收益约为 45.93%。

### 3.5 深度信念网络

深度信念网络是（Deep Belief Network，DBN）一种生成模型，通过训练其神经元间的权重，使整个神经网络按照最大概率来生成训练数据。DBN 由多层神经元构成，即显性神经元和隐性神经元。显性神经元用于接受输入，隐形神经元用于提取特征。最顶上的两层间的连接是无向的，组成联合内存 (associative memory)。较低的其他层之间有连接上下的有向连接。最底层代表了数据向量 (data vectors)，每一个神经元代表数据向量的一维。DBN 的组成元件是受限玻尔兹曼机。训练 DBN 的过程是一层一层地进行的。在每一层中，用数据向量来推断隐层，再把这一隐层当作高一层的数据向量。

Sharang 和 Rao（2015）使用一个在技术指标上训练的 DBN 交易美国国债期货的投资组合。他们使用一个包含 2 个堆栈 RBM 的 DBN。第一个 RBM 是 Gaussian - Bernoulli（15 个节点），第二个 RBM 是 Bernoulli（20 节点）。DBN 产生隐藏的特征，这些特征被输入 3 种不同的分类器：正则化逻辑回归、支持向量机和有 2 个隐层的神经网络。如果 5 天内投资组合上升，结果是 1，反之是-1。

研究使用对比差异算法训练 DBN。从 1985 年开始，根据开仓、走高、走低、收盘利益和成交量数据计算信号。2008 年金融危机期间的一些点被移除了。研究使用 20 个特征：在不同时间框架下计算“日常趋势”，然后进行规范化。所有的参数使用验证数据集选择。当训练神经网络分类器时，在小批量梯度下降训练中使用动量参数来在每次更新时将系数缩减一半。

使用 PCA 构建的投资组合对第一主成分没有什么影响。投资组合是人为的对工具的延展，所以实际上交易是在 ZF 和 ZN 合同之间完成的。所有的输入价格是中间价，所以买卖差价被忽略了。结果看起来有利可图，三种分类模型的准确率比随机预测高 5%-10%。

## 四、 总结和展望

随着投资者从电子设备及互联网获取信息变得愈加便捷，如何以更高的效率从大量文字和数据中挖掘有益的信息便成为焦点。大数据策略、机器学习和深度学习因其强大的信息处理和学习技能而相应地在投资领域得

到越来越广泛的应用。本文将对投资中常用的大数据策略、机器学习和深度学习方法进行了介绍。

大数据因高容量、高速度、高多样性的特点为量化投资创造出新的维度。从数据来源的角度，大数据可分为个体产生的数据、商业产生的数据和由传感器产生的数据。针对前两者，情感分析是常用的数据处理方法。 自然语言处理技术，在对实体、主题和类别进行分析，可预测所分析的数据资料与交易资产的关系并量化其影响，因而在判断宏观经济形势、形成股市交易策略等方面多有应用。

机器学习是一门多领域交叉学科，主要涉及领域包括计算机科学、统计学及数学等。机器学习的目的是使计算机能够从他们在某些任务中的经验中学习，从而随着经验的增长不断提高机器的性能。在金融学领域，机器学习尝试发现变量之间关系，在经过给定数据的训练后，能够在新数据的基础上预测结果。机器学习可分为监督学习（如回归和分类）及非监督学习（因子分析和流型识别）。监督学习试图找到一种规则或一个方程来预测变量，常运用回归方法来评估预测能力更高、并且对系统变化最稳定的模型。非监督则学习试图揭示数据的结构。正因其强大功能，机器学习在选股、择时方面广泛应用。

深度学习有多种定义，如：使用多层次的非线性处理单元进行特征提取和转换的算法；基于对数据的多个特征或表示的（无监督）学习，更高层次的特征是从低级特征派生而成的；更广泛的学习数据表示的机器学习领域中的一部分，能够促进端到端的优化（end-to-end optimation）。这些定义都共同强调了深度学习使用多层次的非线性处理单元及更高层次的特征从低级特征派生而成的特点。以循环神经网络、长短期记忆网络、卷积神经网络、受限玻尔兹曼机及深度信念网络为代表的深度学习方法在预测股市走势、选择投资组合等领域多有应用。

本文所讨论与研究的问题是投资领域的热点，正在随着它的深入展开而受到越来越多的关注。正如文中案例所呈现的，大数据、机器学习、深度学习的应用提升了投资的效率，对于优化投资组合及加强投资盈利性也有促进作用。目前关于深度学习等技术在促进投资策略的优化科学上的应用仍处于探索阶段，值得深入观察和跟踪挖掘。

## 分析师介绍

**丁鲁明：**同济大学金融数学硕士，中国准精算师，现任中信建投证券研究发展部金融工程方向负责人，首席分析师。9 年证券从业，历任海通证券研究所金融工程高级研究员、量化资产配置方向负责人；先后从事转债、选股、高频交易、行业配置、大类资产配置等领域的量化策略研究，对大类资产配置、资产择时领域研究深入，创立国内“量化基本面”投研体系。多次荣获团队荣誉：新财富最佳分析师 2009 第 4、2012 第 4、2013 第 1、2014 第 3 等；水晶球最佳分析师 2009 第 1、2013 第 1 等。

**研究助理 喻银尤：**021-68821600-808 yuyinyou@csc.com.cn
复旦大学计算机硕士，通过 CFA 三级，两年上交所相关部门工作经验，专注于大数据、多因子、人工智能等相关策略研究。

## 研究服务

**社保基金销售经理**

彭砚苹 010-85130892 pengyanping@csc.com.cn
姜东亚 010-85156405 jiangdongya@csc.com.cn

**机构销售负责人**

赵海兰 010-85130909 zhaohailan@csc.com.cn

**北京非公募组**

张博 010-85130905 zhangbo@csc.com.cn
朱燕 010-85156403 zhuyan@csc.com.cn
李祉瑶 010-85130464 lizhiyao@csc.com.cn
李静 010-85130595 lijing@csc.com.cn
赵倩 010-85159313 zhaoqian@csc.com.cn
周瑞 18611606170 zhourui@csc.com.cn
刘凯 010-86451013 liukaizgs@csc.com.cn

**北京公募组**

黄玮 010-85130318 huangwei@csc.com.cn
黄杉 010-85156350 huangshan@csc.com.cn
任师蕙 010-85159274 renshihui@csc.com.cn
王健 010-65608249 wangjianyf@csc.com.cn
罗刚 15810539988 luogang@csc.com.cn

**上海地区销售经理**

陈诗泓 021-68821600 chenshihong@csc.com.cn
邓欣 021-68821600 dengxin@csc.com.cn
黄方禅 021-68821615 huangfangchan@csc.com.cn
戴悦放 021-68821617 daiyuefang@csc.com.cn
李岚 021-68821618 lilan@csc.com.cn
肖垚 021-68821631 xiaoyao@csc.com.cn
吉佳 021-68821600 jijia@csc.com.cn
朱丽 021-68821600 zhuli@csc.com.cn
杨晶 021-68821600 yangjingzgs@csc.com.cn
谈祺阳 021-68821600 tanqiyang@csc.com.cn
翁起帆 021-68821600 wengqifan@csc.com.cn

**深广地区销售经理**

胡倩 0755-23953859 huqian@csc.com.cn
张苗苗 020-38381071 zhangmiaomiao@csc.com.cn
许舒枫 0755-23953843 xushufeng@csc.com.cn
廖成涛 0755-22663051 liaochengtao@csc.com.cn

## 评级说明

以上证指数或者深证综指的涨跌幅为基准。

买入：未来 6 个月内相对超出市场表现 15%以上；

增持：未来 6 个月内相对超出市场表现 5—15%；

中性：未来 6 个月内相对市场表现在-5—5%之间；

减持：未来 6 个月内相对弱于市场表现 5—15%；

卖出：未来 6 个月内相对弱于市场表现 15%以上。

## 重要声明



## 地址

**北京中信建投证券研究发展部**

中国北京 100010

东城区朝内大街 2 号凯恒中心 B 座 12 层

电话：(8610) 8513-0588

传真：(8610) 6518-0322

**上海中信建投证券研究发展部**

中国上海 200120

浦东新区浦东南路 528 号上海证券大厦北塔 22 楼 2201 室

电话：(8621) 6882-1612

传真：(8621) 6882-1622