安信证券
ESSENCE SECURITIES

2018 年 02 月 07 日

金融工程主题报告

证券研究报告

机器学习与量化投资：综述与反思，扬帆正当时

杨勇 分析师
SAC 执业证书编号：S1450518010002
yangyong1@essence.com.cn

周袤 分析师
SAC 执业证书编号：S1450517120007
zhoumao@essence.com.cn

相关报告

平安大华沪深 300 ETF 上市	2018-01-27
黑科技应用之如何看待“新周期”之辩	2018-01-19
FOF 和资产配置周报：富国中证 10 年期国债 ETF 开始募集	2018-01-15
FOF 和资产配置周报：广发上证 10 年期国债 ETF 开始募集	2017-12-23
FOF 和资产配置周报：广发中证 10 年期国开债 LOF 上市	2017-12-17

■**导读：**

机器学习和人工智能在量化投资的应用有很长的历史

机器学习在九十年代初的热潮中已经被大量运用于量化投资中。尽管受限于当时的计算能力和算法，但是由于在算法交易和 CTA 等领域中机器学习提供了一些更好的解决方案，机器学习在这些领域的应用一直延续到今天。

机器学习在量化投资中应用的九个思考

本报告是系列报告的第一篇，简略介绍了机器学习运用到二级市场投资过程中的一些常见问题。这些问题覆盖了策略研发常见错误，策略归因，策略失效判断，机器学习平台的建立，交易系统和机器学习平台的对接以及机器学习对冲基金的团队架构。后续系列报告将会详细围绕这些问题展开。

适当使用下的机器学习策略

本报告中使用了两个策略作为例子。其中股指短线策略的夏普 3.55，年化收益 80.36%。商品长周期策略夏普 1.06，年化收益 8.61%。

■**风险提示：**

机器学习量化策略的结果是对历史经验的总结，存在失效的可能。

内容目录

图表目录

1. 机器学习简介

在产生机器学习之前，计算机和统计两个学科完全是独立发展的。一方面，统计学科经常是试图用复杂的算法解读农业或者工业试验中的规律，而且通常来说，这些试验数据集都很小。

另一方面，计算机学科长期专注于数据的大规模存储，加工和一些非常简单运算。随着信息时代的到来，大量的数据在各行各业被产生，统计问题在复杂度和数据规模上变得越来越大，在此之后，统计学家渐渐很难用传统的方式解决问题。为了应对在数据存储，组织和搜索上的问题，一门叫数据挖掘的学科产生了。

几乎与此同时，在上世纪七十年代到八十年代，从计算机领域产生了两个影响非常深远的算法，支持向量机和神经网络。这两个算法由于在某些领域中的预测效果极好，远远超过了传统的线性模型，迅速吸引了统计学家的注意，并最终导致了统计学科的一场革命，机器学习由此产生。

机器学习产生后，继续与其他学科结合，并最终产生了现在的“数据科学”。

1.1. 机器学习的基本流程

机器学习问题可以大致分为有监督和无监督的。在有监督学习中。我们的目标是基于一系列输入，去预测一个“结果”。而在无监督学习中并没有这样一个“结果”，无监督学习的目标是解释和描述这些输入之间的关系。机器学习的框架大致可以分为下面几步：

1.1.1. 决定数据源，数据采集

在这一步，我们会收集所有有关的数据。假设我们是做互联网销售，我们可能会关注用户在各个商品上的点击率，搜索词条，以及浏览内容。如果我们是制造企业，为了预测可能要替换哪些零件，我们可能更加会关心机器运行状态和活动的日志。对于对冲基金来说，可以利用的数据源更加广泛，从市场行情数据，到公司财务数据，到宏观指标，乃至各地天气，都可以作为潜在的数据源。

1.1.2. 数据预处理

1.1.2.1. 数据清洗

现实世界的数据会有缺失值，处理缺失值经常需要对具体领域的了解。比如，在人口普查中，一些人会不愿意写他的薪水。这可能是由于这些人薪水要么太高，要么太低。所以简单的用平均值或者中位数替换并不适合。可行的方案是用一些别的去预测薪水，比如这个人的住址，这个人孩子上的学校等等。

1.1.2.2. 数据转换

数据转换是将原始数据转换成模型所需要的数据。在量化多因子模型中这一步往往被称之为建立因子库，而在机器学习中则往往被称为构造特征。所有的数据都可分为分类数据和数值型数据。举例而言分类数据可以为眼睛颜色，婚姻状况或者性别等等。而数值型数据是由实际的数字构成，比如身高体重。以多因子模型举例，收益率序列往往是数值型的，而行业的哑变量就是分类数据。

1.1.2.3. 将数据分为训练集，验证集和测试集

将所有的数据分为训练集，验证集和测试集（有时验证集不是必须的）。这里有两点需要注

意。一，对于量化投资中常见的时间序列数据，在划分训练集，验证集和测试集的时候必须考虑时间的先后顺序。这是因为在真正部署系统的时候，站在当下我们只有训练集和验证集，而测试集是我们未来才能有的。所以在模型建立阶段，也需要按照时间顺序来划分，假设有 0<T1<T2<T3，那么训练集的为 0~T1 的数据集，验证集为 T1~T2 的数据集，测试集是 T2~T3 的数据集。二，如果不是时间序列数据，那么在划分训练集，验证集和测试集的时候，必须是随机的。实际操作中，往往首先随机打乱行的顺序，然后取数据集的前 n%为训练集，n%-m%为验证集，剩下的为测试集，有 n<m。

1.1.3. 基于训练集和验证集建立模型

这一部分是最需要经验的，根据数据本身的特性选择最适合的模型。

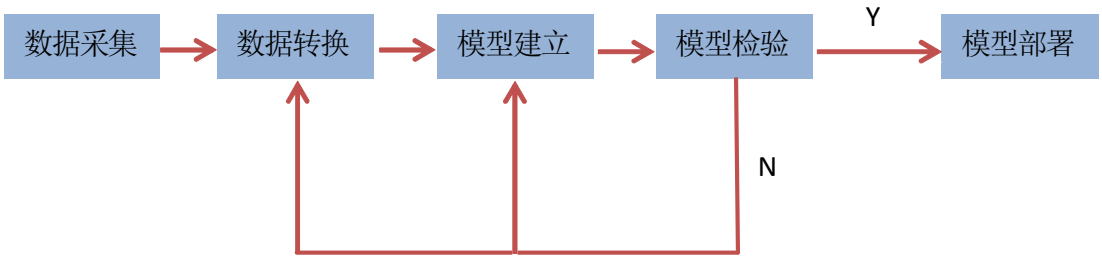
1.1.4. 在测试集上检验模型效果，如果效果不好，回到第（2,3）步，否则去第（5）步

一旦模型被建立完成，它将在测试集上被测试。首先，我们去除测试集里面的因变量。然后使用训练好的模型基于特征（因子）产生预测值。然后我们将预测值和实际值进行比较。如果预测值和实际值相近，那么说明我们的模型是非常合适的。衡量预测值和实际值是否相近的有很多方法，具体会在后文具体说明。

1.1.5. 部署至实际系统

这是最重要的一步。如果模型的速度和准确性是可以接受的，那么就应该被部署到实际的系统的当中。通常来说，数据越多，模型的预测效果越好。但是在金融投资当中不完全如此，因为太久远的数据或许不容易表现出最近的市场状态。用多久的数据经常需要人的主观经验。此外，数据越多，对计算力的要求越高……。

图 1：机器学习基本流程图



资料来源：安信证券研究中心

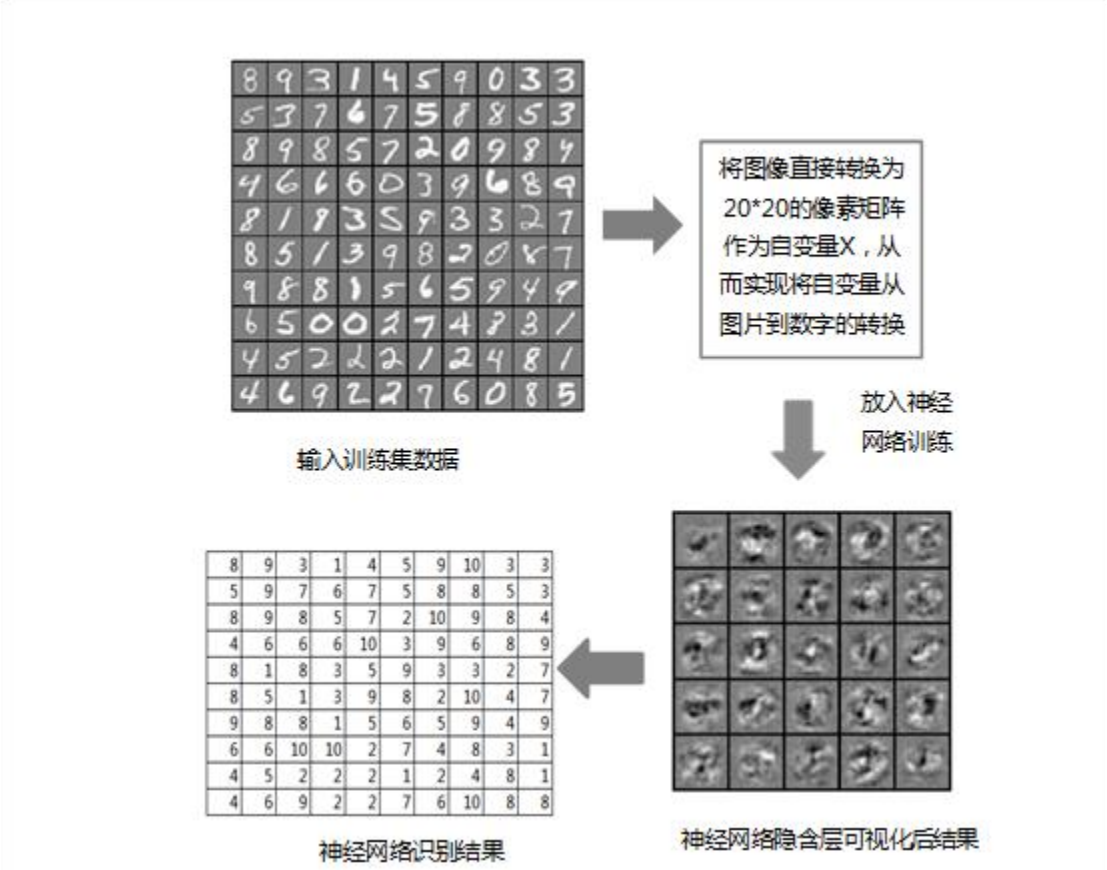
1.2. 机器学习应用场景

机器学习广泛应用于图像识别，自动驾驶，动作识别，语音识别，自然语言处理，翻译，反欺诈，推荐系统和计算生物学。

以图像识别著名的“MNIST 数据集”为例，将含有阿拉伯数字的图片转换为像素参数矩阵输入神经网络，经过多个隐含层后，神经网络将对图片中的数字进行识别，并输出具体数字结果。

可将下图中图片中的数字与神经网络识别结果进行对比

图 2：神经网络识别数字

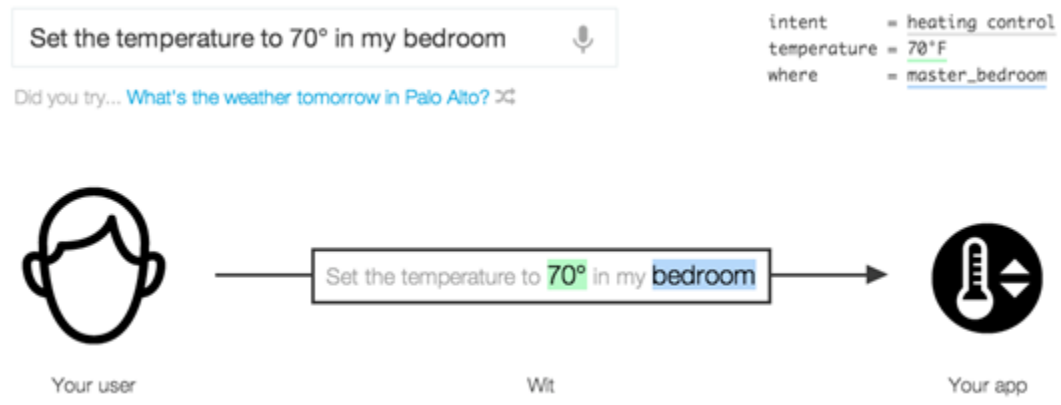


资料来源：Yann LeCun, Courant Institute, NYU

最新的 committee of 35 conv. net, 1-20-P-40-P-150-10 [elastic distortions]识别数字的正确率已经高达 99.77%

自然语言处理是当今的另一个热门方向，也是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。以智能家居为例，只需通过说话就能有效指挥你家里的电器了。如下图，“将我房间里的气温设置到 70 度（华氏）”

图 3：自然语言处理示意图



资料来源： wit.ai

1.3. 机器学习在量化投资中的应用

1.3.1. 历史

与很多人的想法相左，机器学习在量化投资中已经用了 25 年有余。事实上，人工智能曾经风靡一时。高盛早在上世纪 90 年代就组建了第一支人工智能量化投资小组。而在 1993 出版的 Neil A.Gershenfeld 和 Andreas S. Weigend 的《The Future of Time Series》的一书引言中，他们更是乐观的估计机器学习将彻底颠覆传统的时间序列分析。尽管如此，后来的结果表明，人工智能并没有成为那个金融领域的变革者。而且，不仅仅是金融，以神经网络为代表的人工智能在图像和语音识别领域也曾遭遇了重大挫折。

但这一切并不代表人工智能的最终失败。近几年来，以卷积神经网络和长短期记忆网络为代表的深度学习在科技行业产生了巨大的成功。而在量化交易中的也产生了越来越多的全部依赖人工智能的自动化交易基金，例如 Two Sigma。

1.3.2. 人工智能复兴的原因

有三个原因直接导致了人工智能在量化投资的再次兴起。

一）计算机比 25 年提高了一千倍，这使得更加复杂的神经网络成为可能。

二）在这 25 年来，算法有了巨大的进步，特别是反向传播的优化算法的完善，使得机器能更好和更快的完成任务。

三）数据量变得更加巨大，一方面数据量的增加提升了机器学习的效果，另一方面也使得人工处理数据变得成本高昂。

1.3.3. 关于本系列研报

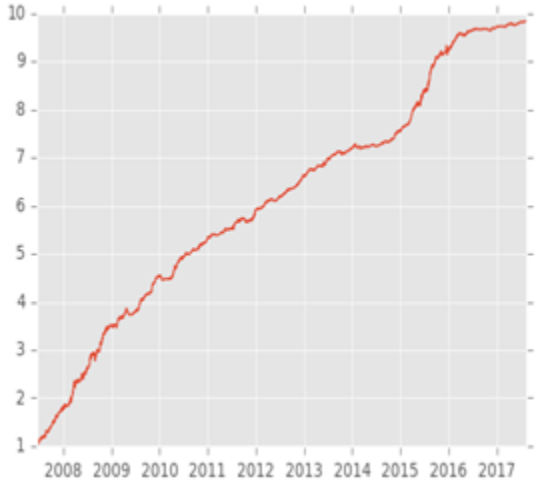
这篇和以后相关的系列报告将引入很多新奇的想法，并且分享机器学习运用到二级市场投资过程中的一些常见错误。尽管这些报告会不可避免的牵涉到一些数学细节，但是我们将更着重介绍应用机器学习的方式，而不是机器学习理论和算法本身。因此，我们希望本系列报告会不只是针对机器学习领域的量化工作者，更是针对其他领域对此感兴趣的研究人员。

2. 机器学习九个思考

为便于论述，先简要列举出这篇报告中两个经常使用的例子。

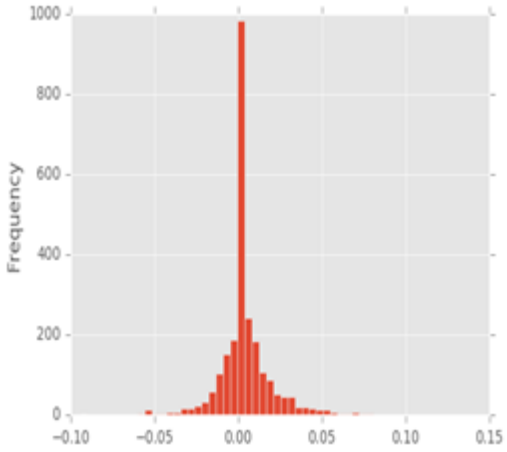
下面两个策略，一个是以半小时 K 线为基础的股指策略（为了区分其他策略，下称为标准神经网络回归），一个是以日 K 线为基础的商品策略，按周调仓（为了区分其他策略，下称为标准长周期商品策略），回测结果分别如下图

图 4：标准神经网络回归-净值



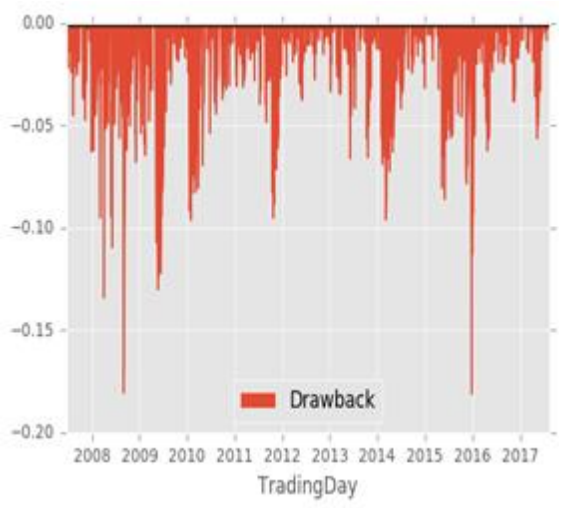
资料来源：Wind，安信证券研究中心

图 5：标准神经网络回归-收益分布



资料来源：Wind，安信证券研究中心

图 6：标准神经网络回归-回撤



资料来源：Wind，安信证券研究中心

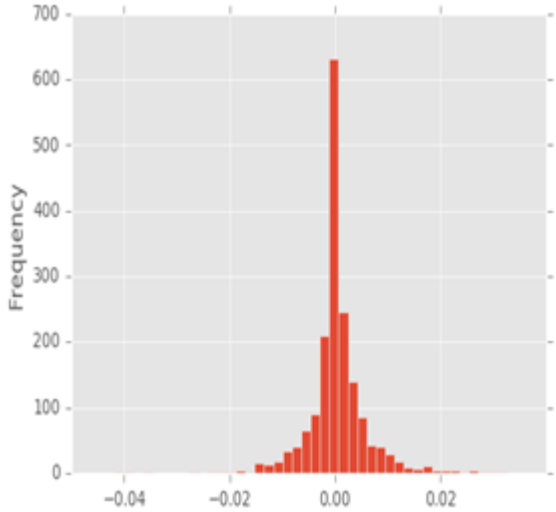
图 7：标准神经网络回归-分年度夏普

资料来源：Wind，安信证券研究中心

图 8：标准长周期商品期货策略-净值

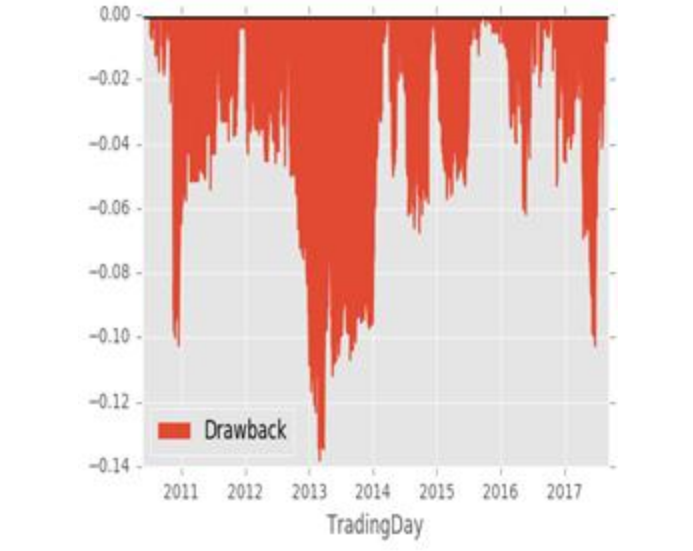
资料来源：Wind，安信证券研究中心

图 9：标准长周期商品期货策略-收益分布



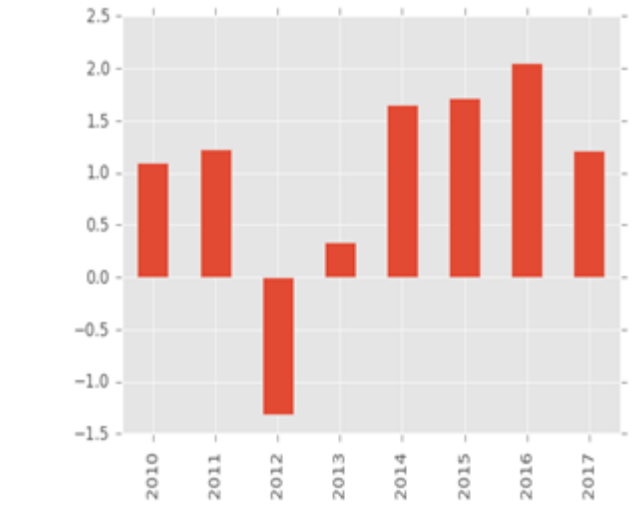
资料来源：Wind，安信证券研究中心

图 10：标准长周期商品期货策略-回撤



资料来源：Wind，安信证券研究中心

图 11：标准长周期商品期货策略-分年度夏普



资料来源：Wind，安信证券研究中心

2.1. 机器学习从线性到非线性

（一）金融市场大概率不是线性的，Abhyankar, Copeland and Wong (1997)，Ammermann and Patterson (2003)等人各自在不同的市场中发现了这一点。机器学习有效的放宽了一些假设，例如：模型不再要求收益率服从正态分布；模型也不再要求因子和下期收益存在线性关系。但是机器学习在从数据中榨取更多信息的同时，更容易出现过度拟合。机器学习有一系列防止过度拟合的方法，具体我们以后的系列报告中讨论。

（二）人对非线性的直观理解是非常有限的，这对传统量化基金经理是非常大的挑战。在传统多因子体系框架当中，量化分析可以拆分为单因子，然后用 IC，IR 进行单因子分析，也可以去根据因子值去进行单调性分析。但是在非线性的世界中，IC，IR 已经不能完整描述因子，而且由于非线性，因子也不必须要服从单调性的假设。为了使传统量化基金经理理解非线性，我们将力图在机器学习领域构建一套机器学习因子分析框架，具体我们也将会在以后的系列研报中讨论。

以下是简单的比较，左侧策略使用支持向量机回归，采用线性核函数。而右侧是我们标准的神经网络回归策略。可以看到从夏普，最大回撤，年化收益，日胜率和盈亏比五个方面，线性策略都逊于标准神经网络回归策略。

表 1：线性策略与标准神经网络回归策略对比

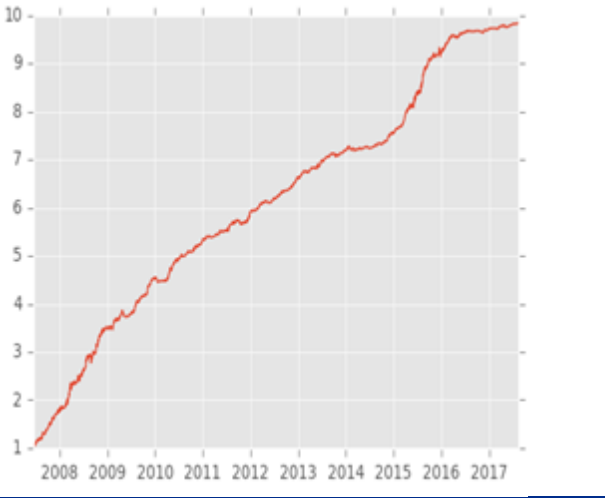
参数/策略名	SVR 线性核	标准神经网络回归
夏普	0.95	3.55
最大回撤	29.71%	17.05%
年化收益	17.67%	80.36%
日胜率	49.64%	62.69%
盈亏比	1.23	1.31

资料来源：Wind，安信证券研究中心

图 12：SVR 线性核-净值

资料来源：Wind，安信证券研究中心

图 13：标准神经网络回归-净值



资料来源：Wind，安信证券研究中心

2.2. 预测周期：从低频到高频

初学者最开始研究量化投资的时候，通常都是从股票日线开始的。但大部分既有的研究和实战结果表明，日 K 线尤其是更为低频的周 K 线等恰恰是机器学习模型不太容易把握或占优的时间尺度，理由如下。

（一）数据量高频更大，通常来说对机器学习模型，数据量越大越好。

（二）高频下人的交易行为受到的干扰因素较小，基本面不容易发生明显变化，规律更加稳定，机器更加容易学习

但是是不是越高频越好呢？答案也是否定的，现实世界中有交易成本，越高频，交易成本越高。更不用提在超高频当中需要用到大量 IT 基础设施投入导致的巨大的成本，例如 FPGA。下表是在日线上的中证 500 择时与用到日内信号的择时比较。可以看到从夏普，最大回撤，年化收益，日胜率和盈亏比五个方面，较为高频的策略（标准神经网络回归）都优于低频的策略（日线神经网络）。

表 2：日线神经网络与标准神经网络对比

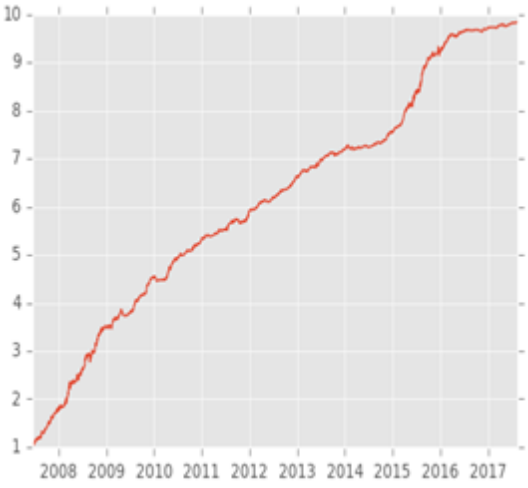
参数/策略名	日线神经网络	标准神经网络回归
夏普	0.68	3.55
最大回撤	36.92%	17.05%
年化收益	19.02%	80.36%
日胜率	53.21%	62.69%
盈亏比	0.99	1.31

资料来源：Wind，安信证券研究中心

图 14：日线神经网络-净值

资料来源：Wind，安信证券研究中心

图 15：标准神经网络-净值

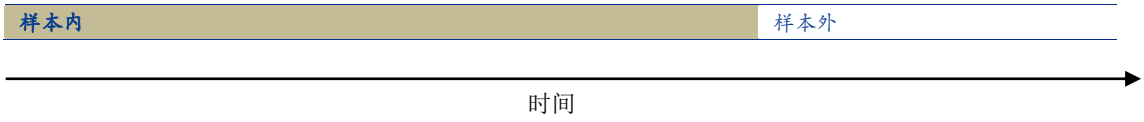


资料来源：Wind，安信证券研究中心

2.3. 从单次分析到推进分析

样本外检验是量化研究中避免过度拟合的常用方法，下面是单次分析的常见方法：

表 3：单次分析示意图

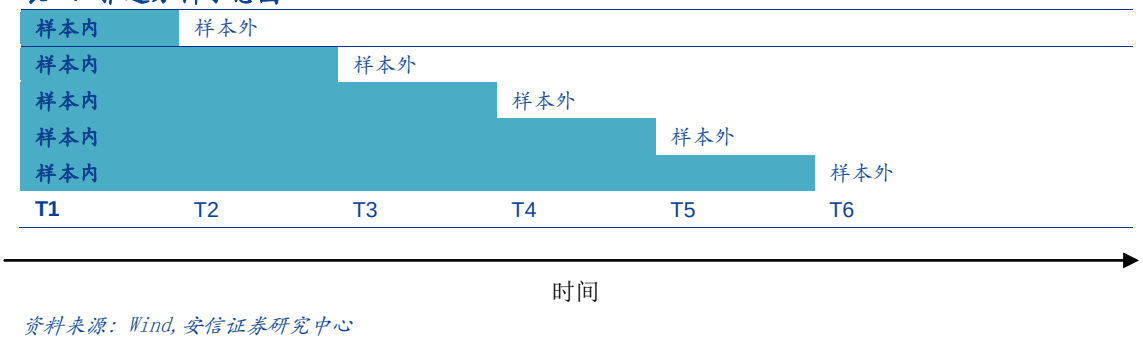


资料来源：Wind，安信证券研究中心

上图给了一个单次分析的实例。实际上单次分析就是把整个样本分为互不重叠的两个部分。白色的是样本内，灰色的是样本外。首先用样本内的数据训练机器学习模型，然后用这个建立好的机器学习模型直接放入样本外数据进行检验，如果在样本外的数据依然说明该模型效果很好，那么在一定程度上说明该模型可以处理实际的问题。

而推进分析的样本内外常常变化：

表 4：推进分析示意图



资料来源：Wind，安信证券研究中心

上图是一种推进分析的方法。推进分析有个最为明显的特点，就是样本外的交易长度仅为一个交易周期。同样的，首先用样本内的数据训练机器学习模型，然后用这个建立好的机器学习模型直接放入样本外数据进行检验。在 T1 时刻，用 0~T1 的数据训练模型，然后在 T1~T2

的数据去检验模型；在 T2 时刻，用 0~T2 的数据训练模型，然后在 T2~T3 的数据去检验模型；在 T3 时刻，用 0~T3 的数据训练模型，然后在 T3~T4 的数据去检验模型，以此类推。最后将所有灰色框内的检验结果汇总，就是推进分析下总的样本外结果。

推进分析除了上图所示，还有滚动 n 期等等方法。是不是推进分析总是优于单次分析，什么时候该用推进分析，什么时候该用单次分析？我们会在以后的报告中逐步讲解。

2.4. 预测目标：从收益到其他

想象一个场景，假设投资者有基础资金一百万，他可以选择两只股票。股票 A 可能亏 5 万，但是可能赚 10 万；股票 B 可能亏 10 万，但是可能赚 30 万。那么投资者应该如何选？一般人给出的答案通常是看该投资者的风险偏好。但是做量化的人都会说，投资者选 B 股票，因为 B 股票收益回撤比高。投资者完全可以买两份的 A 股票，这样 A 股票就变成了可能亏 10 万，但是可能赚 20 万，吸引力就不如 B 了。

在上述场景中，量化投资者并不关心收益，只关心收益回撤比。所以在建立模型的时候，机器学习同样可以将预测目标变成收益回撤比。但是并不是所有时候都适合拿收益回撤比做预测目标，我们将在系列研报中详细解释。

表 5：预测目标（一周累计收益/标准差）与（一周累计收益）对比

参数/策略名	预测值（一周累计收益/标准差）	预测值（一周累计收益）
夏普	1.06	1.03
最大回撤	13.02%	12.99%
年化收益	8.61%	9.24%
日胜率	54.16%	53.45%
盈亏比	1.06	1.09

资料来源：Wind, 安信证券研究中心

图 16：预测目标（一周累计收益/标准差）-净值

资料来源：Wind, 安信证券研究中心

图 17：预测目标（一周累计收益）-净值

资料来源：Wind, 安信证券研究中心

上图是当预测值为（一周累计收益/标准差）和（一周收益）的结果。结果表明，当预测值为（一周累计收益/标准差）时，收益变小，夏普变大。

2.5. 从分类到回归

监督式机器学习按照预测目标是否连续可以分为分类问题和回归问题。分类问题的因变量是离散的，而回归问题的因变量是连续的。笔者认为，分类和回归没有绝对的好和坏，各自有

各自背后的逻辑。后续的报告将结合具体实例分析，什么时候用分类，什么时候用回归。

以下是简单的比较，左侧是神经网络分类策略。分类标签是简单的分为 1（看涨）和-1（看跌）。若预测值为 1，便看多买入，预测值为-1，则看空卖出。而右侧是我们标准的回归策略。可以看到从夏普，最大回撤，年化收益，日胜率和盈亏比五个方面，分类都逊于标准回归策略。

表 6：神经网络分类与回归对比

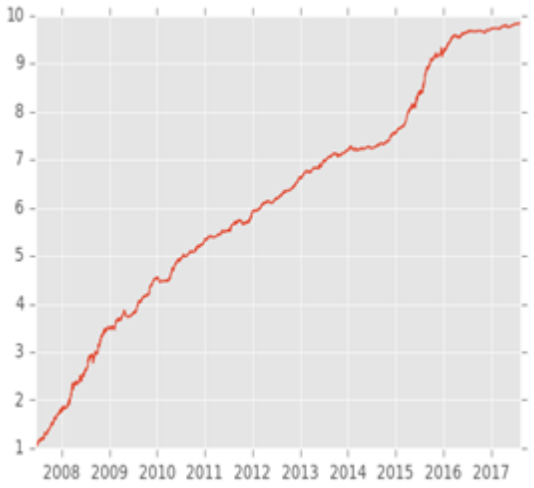
参数/策略名	神经网络分类	标准神经网络回归
夏普	1.66	3.55
最大回撤	25.30%	17.05%
年化收益	30.91%	80.36%
日胜率	49.72%	62.69%
盈亏比	1.39	1.31

资料来源：Wind，安信证券研究中心

图 18：神经网络分类-净值

资料来源：Wind，安信证券研究中心

图 19：神经网络回归-净值



资料来源：Wind，安信证券研究中心

2.6. 预测值相关

无论是分类还是回归，我们都会从机器学习中得到预测值。对预测值的有效运用是实盘成功的重要帮手。例如，假设用的是回归，那么是简单的当回归值大于 0 就做多，小于 0 就做空吗？假设用的是分类，分成两类，是分到看涨类的就看多买入，分到看跌类的就看空抛出吗？在这一系列中，我们将具体讨论。

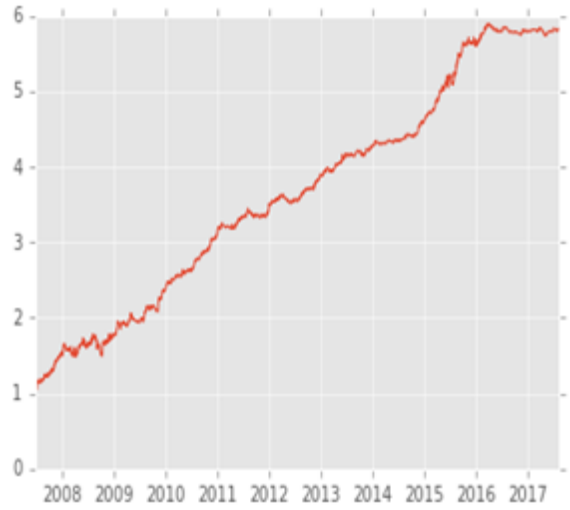
以下是简单的比较。左侧策略也是神经网络回归策略。回归值以 0 为界限，大于 0 就做多，小于 0 就做空。而右侧是标准神经网络回归策略。可以看到从夏普，最大回撤，年化收益，日胜率和盈亏比五个方面，简单以 0 为界限的策略都逊于标准回归策略。

表 7：以 0 为界与标准神经网络对比

参数/策略名	神经网络回归值，以 0 为界限	标准神经网络回归
夏普	2.17	3.55
最大回撤	26.05%	17.05%
年化收益	43.92%	80.36%
日胜率	46.68%	62.69%
盈亏比	1.75	1.31

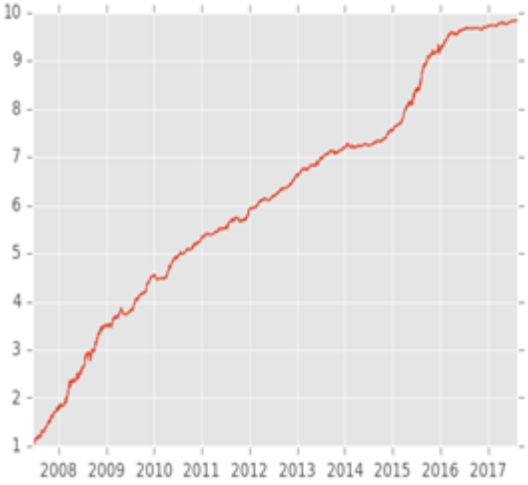
资料来源：Wind，安信证券研究中心

图 20：以 0 为界神经网络-净值



资料来源：Wind，安信证券研究中心

图 21：标准神经网络-净值



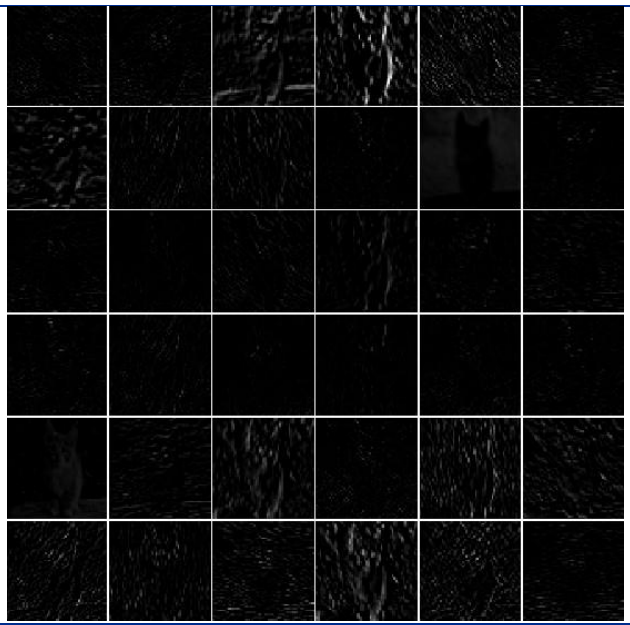
资料来源：Wind，安信证券研究中心

2.7. 如何让机器学习不那么黑箱（非线性的预测体系需要非线性的归因方式）

机器学习普遍被人诟病的部分就是可解释性差。2016 年的时候，欧盟更是发布了 General Data Protection Regulation，要求所有模型都要有可解释性，否则不得用于商业用途，在机器学习领域引起了轩然大波。

尽管机器学习模型的可解释性暂时还无法比简单的线性模型表现更加好，但是这并不意味着机器学习是一个完完全全的黑箱。比如在图像识别领域，就有一些常见的方法。

图 22：AlexNet 的卷积神经网络



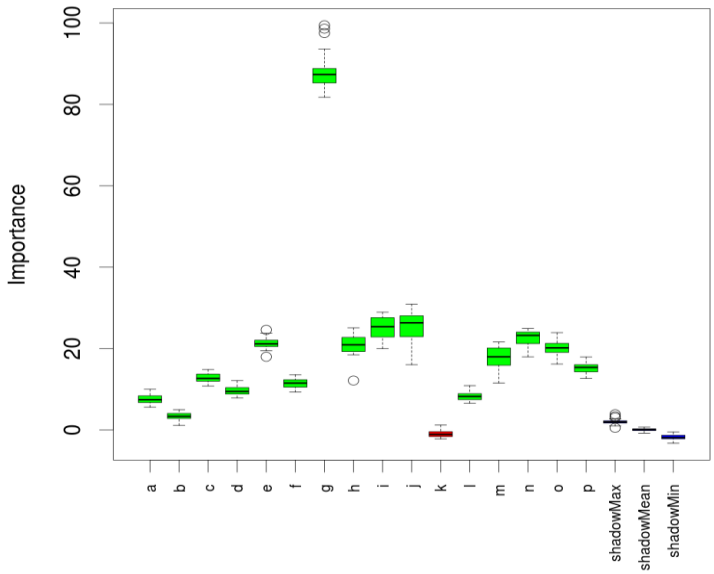
资料来源：斯坦福大学

上图的模型是一种被称为 AlexNet 的卷积神经网络。该 AlexNet 已经训练完毕，它的目标是

从 25000 张动物的图像集合识别出猫。上图是 AlexNet 是第一卷积层，越白的地方，激发值越高。可以清晰的看到，在不少图上，通过观察神经元上的激发值，还是能看到猫的全貌的，例如第一行第四列。

同样的，在机器学习领域也可以进行因子的分析，看看哪些因子贡献最大。例如对传统的多因子模型，我们可以用机器学习的归因方法做出如下图。每个因子在非线性空间下的重要性以如下盒状图的方式展示出来：

图 23：机器学习的因子分析方法



资料来源：Wind，安信证券研究中心

由于是在非线性空间下面计算的，重要性并不代表正负方向，它代表的是对下期收益的预测能力。因子的重要度越大，在 Y 轴上的位置就越高。可以看出因子 g 在 y 轴处于非常高的位置，所以非常重要。而相对应的，因子 k 在非常低的位置，甚至是红色标出，说明非常不重要（红色是因为机器认为重要程度甚至没有统计显著性！）。因子 j 的上下四分位距离差距很大，说明因子 j 非常不稳定（重要性的方差非常大）shadowMax 是指假设随机建立一个因子，这个因子重要性的最大值是多少，shadowMean 是指假设随机建立一个因子，这个因子重要性的平均值是多少，shadowMin 是指假设随机建立一个因子，这个因子重要性的最小值是多少，所有小于 shadowMax 的因子，都在统计意义上不能称之为有效的因子。

在机器学习领域，有很多种这样分析因子有效性的方法，我们将这些方法统一称为特征工程。我们将在后续系列报告中统一介绍。

2.8. 如何判断策略失效

任何策略都有失效的时候，机器学习策略也一样。无论是机器学习还是量化投资，都基于一个非常重要的假设，那就是历史能够重演。传统策略经常使用的失效标准是当最大回撤击穿历史最大回撤，或者是历史最大回撤的 1.5 倍。而机器学习策略还有一些独特的方法。通常来说，高频策略由于交易次数频繁，容易达到统计显著性，失效相对容易发现。而低频策略的判断更需要对策略本身逻辑的深刻的理解。我们将从机器学习和量化本身的原理出发，结合策略本身，在后续的系列报告中探讨这个问题。

2.9. 机器学习杂谈

杂谈 1 到杂谈 4 是我们将在最后一篇报告中讨论的。

杂谈 1：计算落地相关：我们需要什么级别的计算力？

Apache Spark 和 Hadoop 是两个主流的大数据框架。大多数的科技公司都会或多或少用到这些框架，如今一些对冲基金需要考虑使用 Apache Spark 或者 Hadoop 吗？我们将从需求和成本的角度加以讨论。

杂谈 2：交易系统相关

考虑到绝大多数机器学习系统也都是基于 Linux 的，这一杂谈将举例说明如何将机器学习系统与基于 Windows 的交易系统结合在一起。

杂谈 3：机器学习与主观交易

最近有关机器学习将要取代人类主观交易员的言论非常多，这可能是一个误区。人类做出的决定的过程相对于机器学习做出的决定的过程是有非常巨大的区别。一个非常直观的理解是，人类决策模糊但稳定性高，机器学习决策准确却脆弱。我们将从机理、数据、推理联想等等方面说明两者的区别。

杂谈 4：机器学习在量化投资的机遇与挑战

机器学习在量化投资领域已经经历了几十个年头。随着计算能力的进一步发展，机器学习必将在量化领域发挥更大的作用。在后续的系列报告中，笔者将从数据，算法，计算力，用户和公司组织架构五个方面对机器学习在量化投资的机遇与挑战做出讨论。

3. 监督式学习简介

3.1. 线性模型

对于输入$x=(x1,x2,\ldots,xd)$和输出$f(x)$，机器学习中的线性模型就是利用训练集中（x,y）数据，找到一个线性组合来尽可能好地描述：$f(x)=w_1x_1+w_2x_2+\ldots+w_dx_d+b$，也即$y=w^Tx+b$。

寻找这个线性组合的方式有许多，我们最为熟悉的是线性回归，通过拟合得到的变量方程，可以求出需要预测的变量。

但是线性模型不仅仅是线性回归，只要能够表达为$y=w^Tx+b$形式（或$y=g^{-1}(w^Tx+b)$）的均属于线性模型范畴。从直观上讲线性模型是在高维样本空间中找到一组超平面，通过这个超平面得到 x 到 y 的映射（回归），或者由这个超平面将空间划分成不同区域，每个区域对应不同类别（分类）。因此，诸如 Logit 模型等可以化为$y=g^{-1}(w^Tx+b)$形式的均属于线性模型。

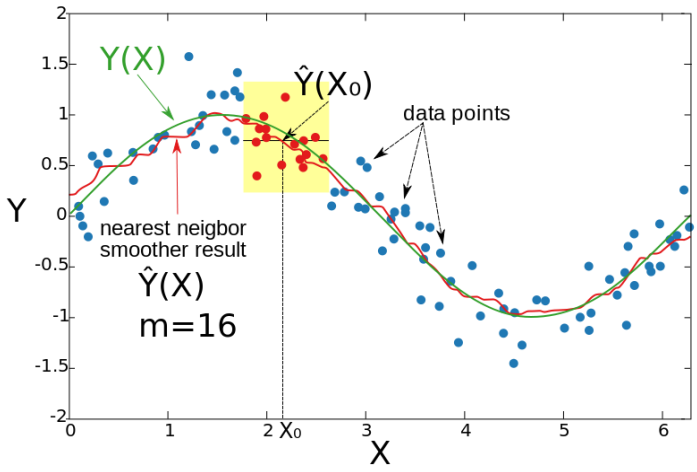
线性模型的优点在于简单直观，解释性强，通过观察向量 w 可以简便地观察各特征的权重。其缺点在于模型过于简单，往往不能解释复杂的联系。

3.2. Kernel Smoothing

在使用上述线性模型时，是站在整体上得到模型，利用它去拟合样本空间中所有的点。但有时，找不到一个对于整个样本空间放之四海而皆准的模型，这时我们需要从 local 的角度对模型进行训练。

我们最为熟悉的 local 训练方法是 KNN，选取距离最近的 K 个样本，取均值后得到预测值。这种方式会带来不连续、不光滑的情况，如下图：

图 24：KNN 算法



资料来源：Wind，安信证券研究中心

图中绿线为真值，蓝点为样本点，红线为 KNN 拟合曲线。

为使曲线更加平滑，可以采用 Kernal Smoothing 的方法，对 K 个近邻，不直接平均而是先赋予不同权重，距离越近，权重越大，距离越远，权重越小。

也即从直接平均：

$$f(x) = Ave(y_i|x_i \in N_k(x))$$

变为加权平均：

$$f(x) = \frac{\sum_{i=1}^{N} K_\lambda(x_0, x_i) y_i}{\sum_{i=1}^{N} K_\lambda(x_0, x_i)}$$

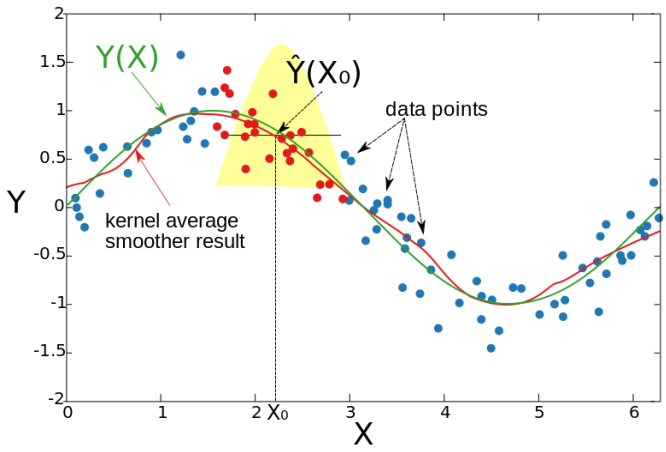
上式中的K_λ即为核函数（kernel），有多种选取方式，一个常用的核函数是高斯核函数，这里给出称为 Epanechnikov quadratic kernel 的核函数，如下：

$$K_\lambda(x_0, x_i) = D(\frac{|x - x_0|}{\lambda})$$

$$D = \begin{cases} \frac{3}{4}(1 - t^2) \text{ if } |t| \le 1 \\ 0 \quad \text{otherwise} \end{cases}$$

通过核函数加权平均后，如下图：

图 25：核函数下的 KNN 算法



资料来源：Wind，安信证券研究中心

红色的拟合曲线比原来更加平滑，此即核平滑（Kernel Smoothing）一个比较简单的例子。

3.3. 树状模型：Bagging 和 Boosting

Bagging 和 Boosting 均属于集成学习方法。其基本思路相同，都是从训练集中选出许多子集，喂给多个同类学习器，也即每个学习器得到的训练集不是完全相同的，然后将它们集合起来，得出预测结果。

选取训练集时，Bagging 的训练集是在原始集中有放回选取的，从原始集中选出的各轮训练集之间是独立的；而 Boosting 的每一轮的训练集的 X 是不变的，在分类问题中，每个样例在分类器中的权重发生变化，在回归问题中，每个样例的 Y 会有一定调整，两者都是根据上一轮的分类错误的样本进行调整。

在样例权重上，Bagging 使用均匀取样，每个样例的权重相；而 Boosting 则会根据错误率不断调整样例的权值，错误率越大则权重越大。

在集成多个训练结果时，Bagging 对于所有预测函数的权重相等；而 Boosting 的每个预测函数都有相应的权重，对于分类误差小的预测函数有更大的权重。

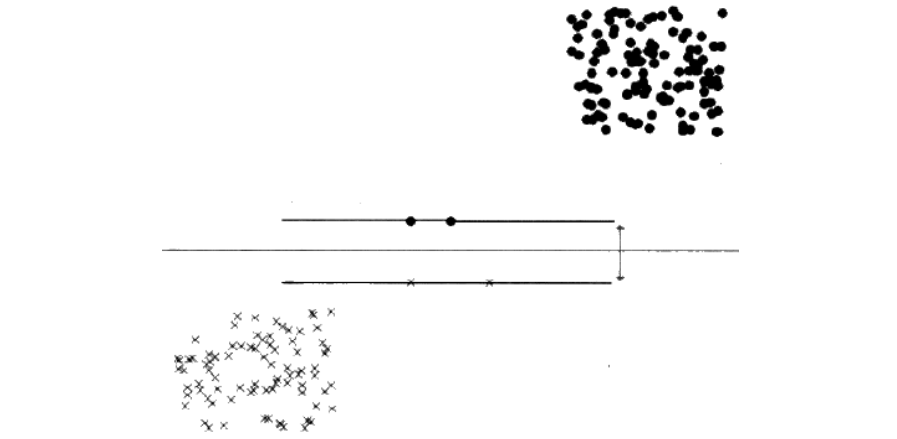
在并行计算方面，Bagging 的各个预测函数可以并行生成；而 Boosting 的各个预测函数只能顺序生成，因为后一个模型参数需要前一轮模型的结果。

3.4. 支持向量机

支持向量机（SVM）一般用于分类问题，解决分类问题的思路在于找到一条离所有分类尽可能远的分界线，选择这条分界线的依据是：寻找两条经过各分类相应坐标点的平行线，并使其与分界线的距离尽可能远。

在判断新的数据点的类别时，只需要看其落在分界线的哪一边即可。值得注意的是，确定分界线只需要间隔区边缘的坐标点，其余数据可以被去掉，那么这条分界线附件的坐标点称为支持向量，寻找支持向量，并利用支持向量来寻找分界线的算法便是支持向量机。

图 26：支持向量机



资料来源：Wind，安信证券研究中心

支持向量机可以作为线性分类器，也可以利用核技法 kernel trick 运用于非线性分类。

支持向量回归（SVR）利用了支持向量的思想，来解决回归问题，只是它的目标与 SVM 不同。直观来讲，SVM 试图找到一条线，忽略边缘的点，利用周围的点分割空间。而 SVR 试图找到一条线，忽略周围的点，对剩余的点进行回归。同样可以利用核技法 kernel trick 运用于非线性回归。（注：此处 kernel trick 与 1.2 kernel smoothing 并非一个概念）

3.5. 深度学习：CNN, DNN 和 LSTM

NN：

NN 即 Neural Network，是最基本感知机加上多个隐含层组成的，通过反向传播 BP 算法，由节点模拟神经元对激励的响应，来训练机器学习模型。

DNN：

神经网络的层数直接决定了它对现实的刻画能力，但是随着神经网络层数的加深，优化函数越来越容易陷入局部最优解，且面临“梯度消失”现象（在 BP 反向传播训练时，由于衰减，如果层数太多，低层几乎接受不到有效训练信号）。DNN 即深度神经网络，则是利用各种方法（比如预训练、传输函数用 maxout 代替 sigmoid 等），克服了这个问题，增加了层数，得到了更好的结果。

CNN：

DNN 成功增加了 NN 的层数，但层数的有效增加直接带来的问题在于节点数猛增，参数数量爆炸，这不仅容易过拟合，还容易陷入局部最优。于是 CNN 即卷积神经网络，通过“卷积核”作为中介，极大降低了隐含层中参数的数量。一定程度解决了上述问题。

RNN：

DNN 还存在另一个问题在于：无法对时间序列上的变化进行建模，在样本出现的时间顺序非常重要的情景下，RNN 即循环神经网络解决了这个问题。在 RNN 中，神经元的输出可以在下一个时间戳直接作用到自身，t+1 时刻网络的最终结果是该时刻输入和所有历史共同作用的结果，这就达到了对时间序列建模的目的。

LSTM：

RNN 解决了对时间序列建模的问题，但是“梯度消失”现象又出现了，对于 t 时刻来说，它

产生的梯度在时间轴上向历史传播几层之后就消失了，根本就无法影响太遥远的过去。为了解决时间上的梯度消失，机器学习领域发展出了 LSTM 长短时记忆单元，通过门的开关实现时间上记忆功能，防止梯度消失。

4. 风险提示

机器学习量化策略的结果是对历史经验的总结，存在失效的可能。

分析师声明

杨勇、周袤声明，本人具有中国证券业协会授予的证券投资咨询执业资格，勤勉尽责、诚实守信。本人对本报告的内容和观点负责，保证信息来源合法合规、研究方法专业审慎、研究观点独立公正、分析结论具有合理依据，特此声明。

本公司具备证券投资咨询业务资格的说明

安信证券股份有限公司（以下简称“本公司”）经中国证券监督管理委员会核准，取得证券投资咨询业务许可。本公司及其投资咨询人员可以为证券投资人或客户提供证券投资分析、预测或者建议等直接或间接的有偿咨询服务。发布证券研究报告，是证券投资咨询业务的一种基本形式，本公司可以对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向本公司的客户发布。

免责声明

本报告仅供安信证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因为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。

本报告基于已公开的资料或信息撰写，但本公司不保证该等信息及资料的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映本公司于本报告发布当日的判断，本报告中的证券或投资标的价格、价值及投资带来的收入可能会波动。在不同时期，本公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。本公司不保证本报告所含信息及资料保持在最新状态，本公司将随时补充、更新和修订有关信息及资料，但不保证及时公开发布。同时，本公司有权对本报告所含信息在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。任何有关本报告的摘要或节选都不代表本报告正式完整的观点，一切须以本公司向客户发布的本报告完整版本为准，如有需要，客户可以向本公司投资顾问进一步咨询。

在法律许可的情况下，本公司及所属关联机构可能会持有报告中提到的公司所发行的证券或期权并进行证券或期权交易，也可能为这些公司提供或者争取提供投资银行、财务顾问或者金融产品等相关服务，提请客户充分注意。客户不应将本报告为作出其投资决策的惟一参考因素，亦不应认为本报告可以取代客户自身的投资判断与决策。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投资建议，无论是否已经明示或暗示，本报告不能作为道义的、责任的和法律的依据或者凭证。在任何情况下，本公司亦不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。

安信证券股份有限公司对本声明条款具有惟一修改权和最终解释权。

■ 销售联系人

上海联系人	葛娇妤	021-35082701	gejy@essence.com.cn
	朱贤	021-35082852	zhuxian@essence.com.cn
	许敏	021-35082953	xumin@essence.com.cn
	章政	021-35082861	zhangzheng@essence.com.cn
	孟硕丰	021-35082788	mengsf@essence.com.cn
	李栋	021-35082821	lidong1@essence.com.cn
	侯海霞	021-35082870	houhx@essence.com.cn
	潘艳	021-35082957	panyan@essence.com.cn
	刘恭懿	021-35082961	liugy@essence.com.cn
	孟昊琳	021-35082963	menghl@essence.com.cn
北京联系人	王秋实	010-83321351	wangqs@essence.com.cn
	田星汉	010-83321362	tianxh@essence.com.cn
	李倩	010-83321355	liqian1@essence.com.cn
	周蓉	010-83321367	zhourong@essence.com.cn
	温鹏	010-83321350	wenpeng@essence.com.cn
	张莹	010-83321366	zhangying1@essence.com.cn
深圳联系人	胡珍	0755-82558073	huzhen@essence.com.cn
	范洪群	0755-82558044	fanhq@essence.com.cn
	巢莫雯	0755-82558183	chaomw@essence.com.cn
	黎欢	0755-82558045	lihuan@essence.com.cn

安信证券研究中心

深圳市

地　址： 深圳市福田区深南大道 2008 号中国凤凰大厦 1 栋 7 层

邮　编： 518026

上海市

地　址： 上海市虹口区东大名路638号国投大厦3层

邮　编： 200080

北京市

地　址： 北京市西城区阜成门北大街 2 号楼国投金融大厦 15 层

邮　编： 100034