

2018 年 02 月 23 日

机器学习与量化投资：避不开的那些事（1）

■从高频到低频

机器学习在高频量化策略上应用更加容易

■从线性到非线性

机器学习下的非线性比线性更能榨取数据的价值,但也更容易过度拟合,因此需要合理使用

■从单次分析到推进分析

推进分析更加符合实盘状态下盘后更新模型的实际情况

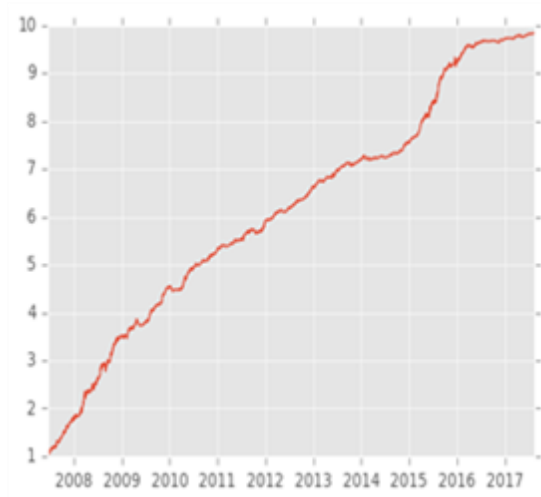
■从分类到回归

回归经常能优于简单的分成两类

■预测值相关

好的预测值不一定带来好的交易信号

■策略回测结果



回测结果：夏普 3.55，年化收益 80.36%

■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。

金融工程主题报告

证券研究报告

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

FOF 和资产配置周报：从量化驱动型策略角度说明美股调整	2018-02-10
机器学习与量化投资：综述与反思，扬帆正当时	2018-02-07
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

内容目录

1. 标准神经网络回归大盘择时策略	4
1.1. 设想和目标	4
1.2. 理论、方法及数据源	4
1.3. 交易成本与策略执行	4
1.4. 算法和模型	4
1.5. 结论	4
1.6. 策略因子归因	5
1.7. 风险点及未来的改进方向	6
2. 从低频到高频	6
2.1. 算法和模型	6
2.2. 结论	7
2.3. 高频背后的一些逻辑	8
2.3.1. 数据	8
2.3.1.1. 运算速度	8
2.3.1.2. 交易成本	8
2.3.2. 日内消息面	9
2.3.3. 行为金融	9
3. 从线性到非线性	9
3.1. 算法和模型	9
3.2. 结论	9
3.3. 非线性背后的一些逻辑和讨论	10
3.3.1. 金融市场大概率是非线性的	10
3.3.2. Bias-Variance Trade off	11
3.3.3. 人的理解方式经常是线性的	12
4. 从单次分析到推进分析	12
4.1. 算法和模型	12
4.2. 结论	13
4.3. 单次分析和推进分析的逻辑讨论	14
5. 从分类到回归	15
5.1. 算法和模型	15
5.2. 结论	15
5.3. 分类与回归的逻辑讨论	16
6. 预测值相关	17
6.1. 算法和模型	17
6.2. 结论	17
6.3. 预测值相关的逻辑	18

图表目录

图 1: 标准神经网络回归-净值	5
图 2: 标准神经网络回归-收益	5
图 3: 标准神经网络回归-回撤	5
图 4: 标准神经网络回归-分年度夏普	5
图 5: 策略因子重要程度	6

图 6: 神经网络日线策略-净值.....	7
图 7: 神经网络日线策略-收益.....	7
图 8: 神经网络日线策略-回撤.....	7
图 9: 神经网络日线策略-分年度夏普.....	7
图 10: 神经网络日线策略-净值.....	8
图 11: 标准神经网络-净值.....	8
图 12: 线性核函数支持向量机回归-净值.....	9
图 13: 线性核函数支持向量机回归-收益.....	9
图 14: 线性核函数支持向量机回归-回撤.....	10
图 15: 线性核函数支持向量机回归-分年度夏普.....	10
图 16: 线性核函数支持向量机回归-净值.....	10
图 17: 标准神经网络-净值.....	10
图 18: Dropout 算法.....	12
图 19: 单次分析标准神经网络回归-净值.....	13
图 20: 单次分析标准神经网络回归-收益.....	13
图 21: 单次分析标准神经网络回归-回撤.....	14
图 22: 单次分析标准神经网络回归-分年度夏普.....	14
图 23: 单次分析标准神经网络回归-净值.....	14
图 24: 标准神经网络-净值.....	14
图 25: 神经网络分类-净值.....	15
图 26: 神经网络分类-收益.....	15
图 27: 神经网络分类-回撤.....	16
图 28: 神经网络分类-分年度夏普.....	16
图 29: 神经网络分类-净值.....	16
图 30: 标准神经网络-净值.....	16
图 31: 神经网络回归值, 以 0 为界限-净值.....	17
图 32: 神经网络回归值, 以 0 为界限-收益.....	17
图 33: 神经网络回归值, 以 0 为界限-回撤.....	17
图 34: 神经网络回归值, 以 0 为界限-分年度夏普.....	17
图 35: 神经网络回归值, 以 0 为界限-净值.....	18
图 36: 标准神经网络-净值.....	18
表 1: 日线-分钟线策略比较.....	8
表 2: 线性核函数支持向量机回归与标准神经网络回归的比较.....	10
表 3: 单次分析示例.....	12
表 4: 推进分析示例.....	13
表 5: 单次分析标准神经网络回归与标准神经网络比较.....	14
表 6: Rolling 推进分析方法.....	15
表 7: 神经网络分类与标准神经网络回归比较.....	16
表 8: 神经网络回归值, 以 0 为界限与标准神经网络比较.....	18

1. 标准神经网络回归大盘择时策略

1.1. 设想和目标

运用机器学习对过去的模式进行识别，并预测未来。也即，用当前实时数据与过去所有数据进行模式匹配，若过去模式显示会大概率上涨下跌，则相应做多做空，否则不做操作。原本该模型是为日内策略设计的，也就是收盘平仓，但由于目前平仓手续费昂贵，所以改为第二天开盘平。

1.2. 理论、方法及数据源

和所有量化策略相似，研究假设过去发生的事情未来会重复发生（也即挑战市场弱有效的假设）。另一重要假设是指数现货和指数期货之间相关性很高，接近 1。这个假设是合理，因为在流动性充足的市场，如果现货和期货之间的任何偏差都可以造成套利机会。故可以用现货做期货。方法为传统的深度学习方法。数据源来自天软、万得，主要是中证 500 指数，沪深 300 指数，以及对应的期货主力合约。

1.3. 交易成本与策略执行

在此策略的历史数据回测中，成交成本假设为日内单边千分之一，隔日单边万分之 3。也即在成交中假设 1.5 个指数点的冲击成本。这样的假设充分包含了目前股指期货低流动性的现实。

1.4. 算法和模型

该算法共有 7 个模型，分别对应 10:00,10:30,11:00,13:00,13:30,14:00,14:30 的决策时间点。每个模型的本质是相似的，唯一的不同只在于越向后的模型，所能拥有的供机器决策数据越多。例如在 10:30 做决策会比 10:00 做决策多出半小时的数据。每个模型本身都是监督式学习。用价量指标来预测收益。若基于机器学习的预测值触及多头开仓阈值，则做多；若基于机器学习的预测值触及空头开仓阈值，则做空。反之维持原来仓位。

1.5. 结论

在日内单边千分之一，隔日单边万分之 3 的成交假设下，策略表现如下：

夏普：3.55

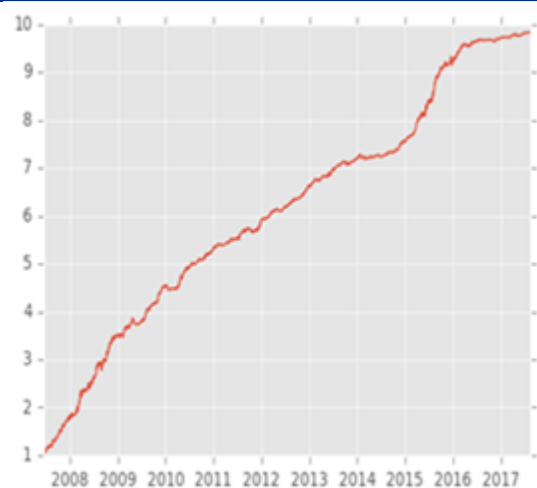
最大回撤：17.05%

胜率：62.69

盈亏比：1.31

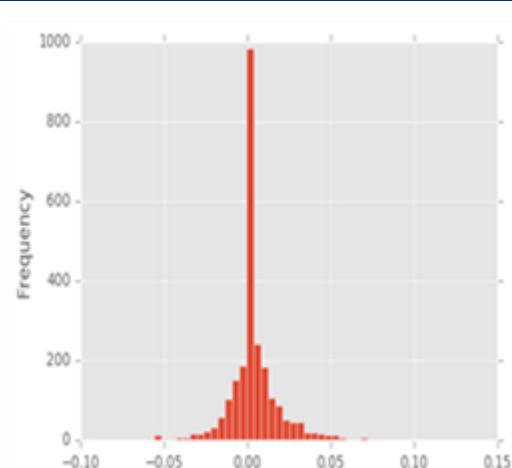
年化：80.36%

图 1：标准神经网络回归-净值



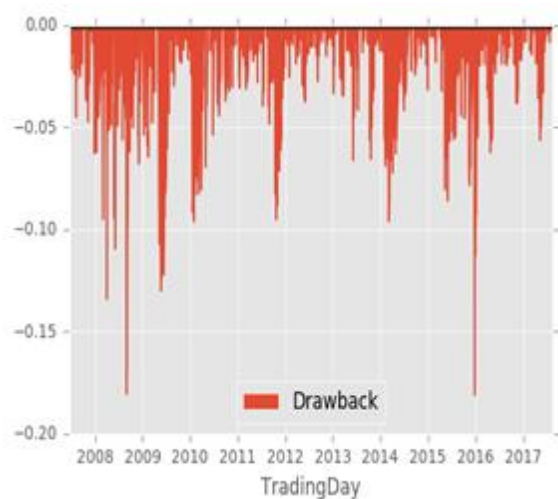
资料来源：Wind, 安信证券研究中心

图 2：标准神经网络回归-收益



资料来源：Wind, 安信证券研究中心

图 3：标准神经网络回归-回撤



资料来源：Wind, 安信证券研究中心

图 4：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

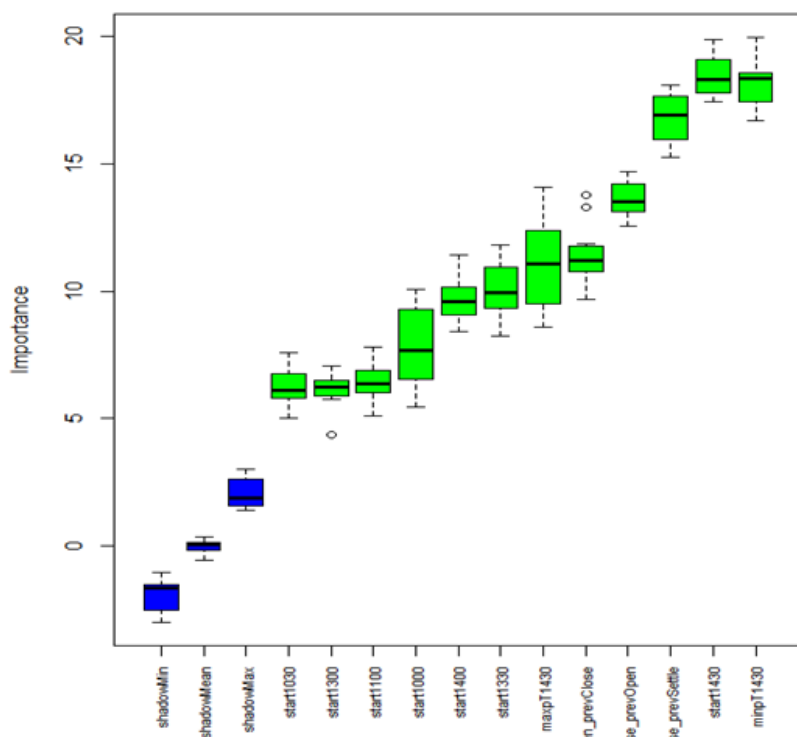
1.6. 策略因子归因

用前述的策略因子归因方法，可以看出一些非常重要的特点。

可以从图中看到，开盘的前一个小时行情数据产生的因子和收盘最后一个小时行情数据产生的因子是非常重要的，如 14:30 到 15:00 的收益。这与人的主观经验是一致的，开盘前一小时交投最活跃，基本能反应当天的市场情绪和主导全天的走势。而收盘最后一小时由于经常是对第二天情绪的猜测，所以从它的走势经常能推断出第二天市场的方向。

另外，盘中两个小时交投最不活跃，随机性也越大。单独用盘中两个小时作为因子去预测未来收益相对来说效果会差一些。

图 5：策略因子重要程度



资料来源：Wind, 安信证券研究中心

1.7. 风险点及未来的改进方向

风险点主要有

- (1) 期货和现货突然性的暂时偏离（在当前负基差的情况下和低成交量下，尤其可能发生）
- (2) 市场结构发生了深刻的变化（投资者类型，投资者风险偏好等等），导致过去的的数据不再能预测未来。例如去年股灾期间国家队的大规模救市。
- (3) 市场流动性不足，导致成交需要付出巨大成本或者无法成交。以中证 500 为例，本模型单笔收益大约在千分之三左右，如果买卖价差长期超过 5 个指数点，将对策略的盈利能力造成毁灭性的打击。目前股指期货受限以来买卖价差大约为 1-2 个指数点。
- (4) 没有合适的报撤单逻辑。如果出现单边市场，简单的用限价单的报撤会导致以最不利的价格成交。

2. 从低频到高频

为了比较高频分钟线与日线策略的区别，我们也设计了一个日线策略。这个策略是基于传统技术指标做特征，例如昨日收盘价相对于过去几日的均线的位置，以及高开低开情况等等。去预测未来的一日的收益。

2.1. 算法和模型

该算法每日决策一次，每个模型本身都是监督式学习。用价量指标来预测收益。若基于机器学习的预测值触及多头开仓阈值，则做多；若基于机器学习的预测值触及空头开仓阈值，则做空。反之维持原来仓位。

2.2. 结论

在日内单边千分之一，隔日单边万分之3的成交假设下，策略表现如下：

夏普：0.68

最大回撤：36.92%

胜率：53.21

盈亏比：0.99

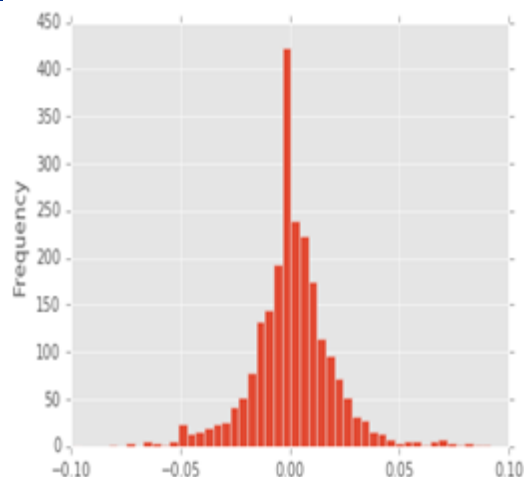
年化：19.02%

图 6：神经网络日线策略-净值



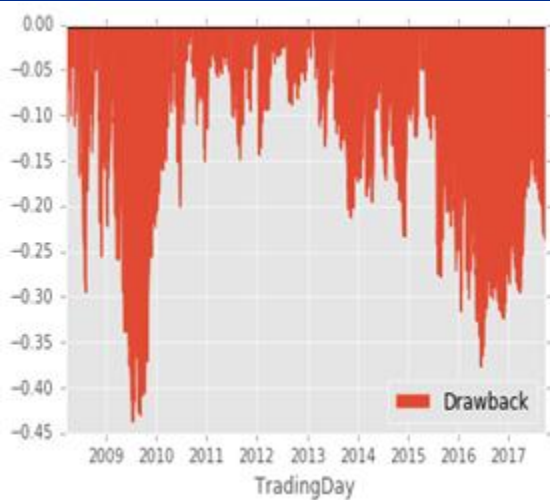
资料来源：Wind, 安信证券研究中心

图 7：神经网络日线策略-收益



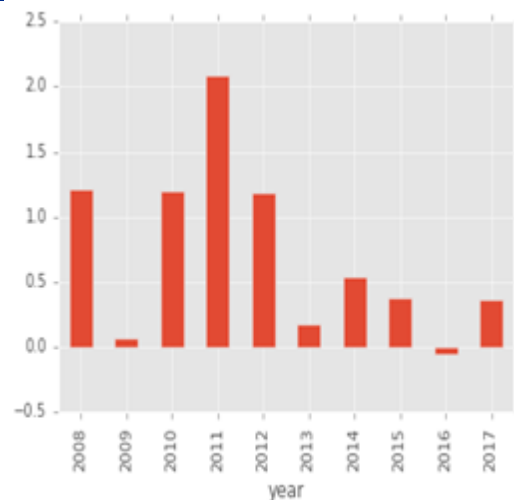
资料来源：Wind, 安信证券研究中心

图 8：神经网络日线策略-回撤



资料来源：Wind, 安信证券研究中心

图 9：神经网络日线策略-分年度夏普



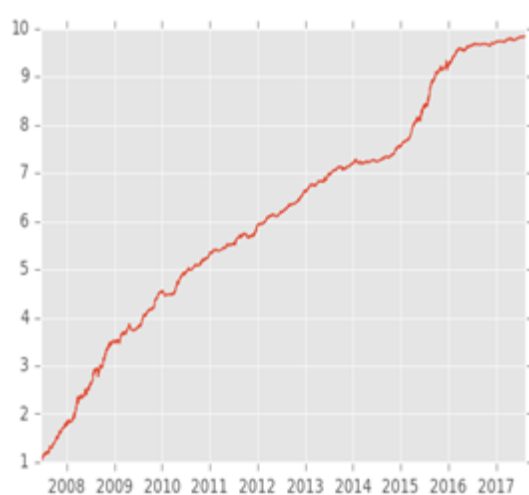
资料来源：Wind, 安信证券研究中心

图 10: 神经网络日线策略-净值



资料来源: Wind, 安信证券研究中心

图 11: 标准神经网络-净值



资料来源: Wind, 安信证券研究中心

表 1: 日线-分钟线策略比较

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

2.3. 高频背后的一些逻辑

2.3.1. 数据

通常来说,对机器学习模型,数据量越大越好。在假设能反映出目前市场的前提下,尽可能多的增加训练集的长度,对机器学习的模型算法收敛和模型稳定性是大有裨益的。假设是日线,以每年 250 个交易日为例,那么 2010 年到 2017 年就是大约 1700 个数据点。但如果是分钟线,同样是每年 250 个交易日,每个交易日 240 根分钟线,那么一共就有 $1700 \times 240 = 408,000$ 的数据点。显然后者就比前者多了好几个数量级。

但是并不是数据量的增加可以无限的,数据量的增加会收到其他客观条件所约束,如运算速度和交易成本。

2.3.1.1. 运算速度

举例期货的例子来说,交易所每 500 毫秒推送一个 tick,所以理论上,2010 年到 2017 年就可以有 49,000,000 个数据点。如果假设交易策略是简单的每 500 毫秒预测一次,那么数据点的增加在实盘中就并没有什么用处。因为在 CPU 下,神经网络的计算用时不太可能在 500 毫秒之内。所以在一个决策时间点内没有算完,就已经进入了下一个决策时间点,实盘中根本交易不到。

2.3.1.2. 交易成本

同样上一个期货的例子。如果是非做市类策略,那么算上冲击成本后的交易成本通常双边至少要达到千分之一。在 500 毫秒乃至更长一些的时间尺度,由于时间尺度偏短,波动很难非常大,所以这是一个非常难覆盖交易成本。

2.3.2. 日内消息面

国内股市实行每日交易时间是四份小时。四个小时内出现基本面新消息的概率较小。而隔日的话，各种消息面容易打破股价自身的运行规律。使得预测的准确性大幅降低。

2.3.3. 行为金融

人的行为在短期内是比较固定的。比如日内短线的追涨杀跌等等，这些都是由人性所决定的。但是随着时间的拉长，特别是两个交易日之间，人会冷静下来，情绪会淡化。

3. 从线性到非线性

为了比较线性模型与非线性模型的区别，我们也设计了一个线性模型。这个策略是基于传统线性核函数的支持向量机回归，使用标准的神经网络回归策略一样的因子和预测目标。

3.1. 算法和模型

算法与模型基本和标准的神经网络回归策略一样，不同的是，神经网络被替换成了线性核函数的支持向量机回归。

3.2. 结论

在日内单边千分之一，隔日单边万分之3的成交假设下，策略表现如下：

夏普：0.95

最大回撤：29.71%

胜率：49.64

盈亏比：1.23

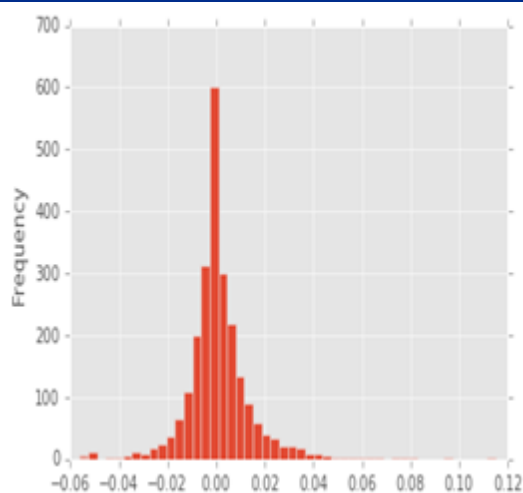
年化：17.67%

图 12：线性核函数支持向量机回归-净值



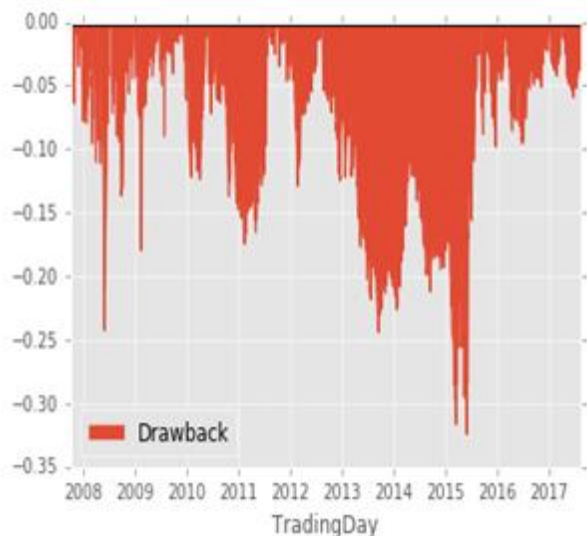
资料来源：Wind，安信证券研究中心

图 13：线性核函数支持向量机回归-收益



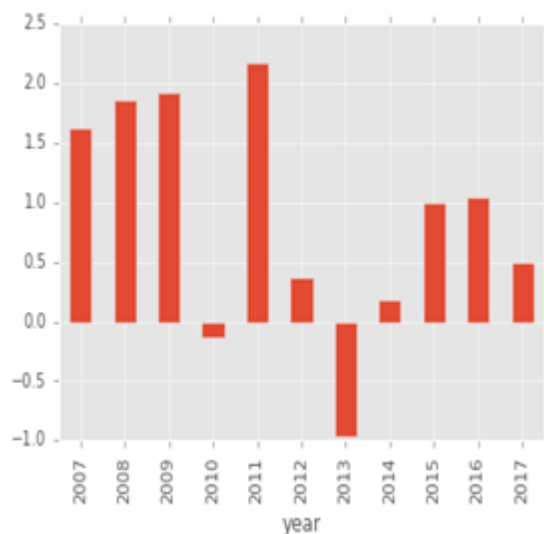
资料来源：Wind，安信证券研究中心

图 14：线性核函数支持向量机回归-回撤



资料来源：Wind, 安信证券研究中心

图 15：线性核函数支持向量机回归-分年度夏普



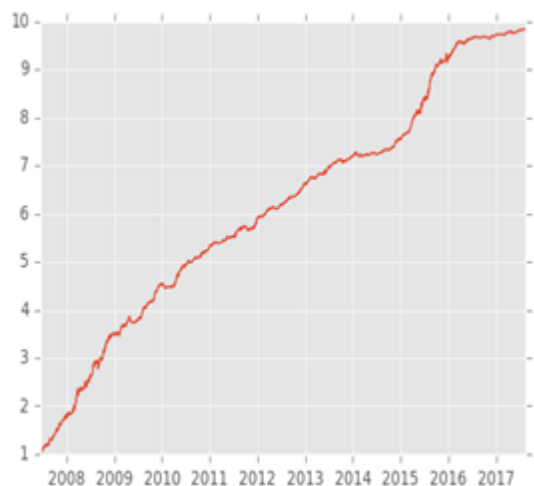
资料来源：Wind, 安信证券研究中心

图 16：线性核函数支持向量机回归-净值



资料来源：Wind, 安信证券研究中心

图 17：标准神经网络-净值



资料来源：Wind, 安信证券研究中心

表 2：线性核函数支持向量机回归与标准神经网络回归的比较

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

3.3. 非线性背后的一些逻辑和讨论

3.3.1. 金融市场大概率是非线性的

金融市场大概率是非线性的，举例而言，业内研究发现，不是高开幅度越大，当日的后续走势就越向上。如果当日高开 0~0.5% 左右，那么当日大概率是上扬的，但是如果高开的过大，当日就容易高开低走。

3.3.2. Bias-Variance Trade off

如果我们假设 $Y = f(X) + \varepsilon$ ，且有 $E(\varepsilon) = 0$ 和 $\text{Var}(\varepsilon) = \sigma_\varepsilon^2$ ，我们可以推导得到，回归 $\hat{f}(X)$ 在 $X = x_0$ 处的平方误差表达式：

$$\begin{aligned}\text{Err}(x_0) &= E[(Y - \hat{f}(x_0))^2 | X = x_0] \\ &= \sigma_\varepsilon^2 + [E\hat{f}(x_0) - f(x_0)]^2 + E[\hat{f}(x_0) - E\hat{f}(x_0)]^2 \\ &= \sigma_\varepsilon^2 + \text{Bias}^2(\hat{f}(x_0)) + \text{Var}(\hat{f}(x_0)) \\ &= \text{Irreducible Error} + \text{Bias}^2 + \text{Variance}\end{aligned}$$

上式中第一项为估计值和其真值之间的不可避免的误差，无论拟合多么精确，该项都无法被消除，除非 $\sigma_\varepsilon^2 = 0$ 。式中第二项为偏差平方项，代表拟合估计与真值之间的平均偏差。最后一项为方差项，代表估计值 $\hat{f}(x_0)$ 偏离其均值的平方和的期望。一般来说，拟合模型 \hat{f} 越复杂，偏差越小，而方差越大。也就是说， \hat{f} 越复杂，对训练集的拟合效果越好，就容易在测试集上泛化效果越差。非线性模型相对于线性模型， \hat{f} 越复杂了很多，泛化效果容易变差。

对此，机器学习有一系列的解决方案。

(1) 区分训练集，验证集和测试集，做模型评价的时候基于测试集去做评价。

(2) 如果数据集太小，不能有效区分训练集，验证集和测试集，那么可以使用 AIC，BIC，MDL，SRM 来近似。

其中 AIC 称为赤池信息准则，它建立在熵的概念上，提供了权衡估计模型复杂度和拟合数据优良性的标准，通常定义为： $AIC = 2k - 2\ln(L)$ ，其中 k 是模型参数个数， L 是似然函数。使得上式最小的模型即为最优模型，AIC 不仅要提高模型拟合度，而且对模型参数引入了惩罚项，有助于降低过拟合的可能性。

BIC 与 AIC 类似，称为贝叶斯信息准则，一般表示为 $BIC = k\ln(n) - 2\ln(L)$ ，其中 k 是模型参数个数， L 是似然函数， n 为样本个数。BIC 对复杂度的惩罚力度比 AIC 大，当样本数量过多时，可有效防止模型精度过高造成的模型复杂度过高。

MDL 同样是用在模型集中寻找最优模型，目标是使式

$$L(D, M^i) = L(M^i) + L(D|M^i)$$

达到最小，此时选出的即为最佳模型。式中， $L(M^i)$ 描述的是模型 M^i 的长度， $L(D|M^i)$ 则是在模型 M^i 假设下观测数据的长度。其重要特性就是避免过度拟合现象。

SRM 称为结构风险最小化，结构风险由经验风险 $\text{Remp}(w)$ 和置信风险 $\varphi(n/h)$ 构成。在追求经验风险最小，也即找到训练集中最优模型时，极易产生过拟合现象，置信风险则代表了多大程度上可以相信模型给出的结果。使得两者的和最小，也即结构风险最小化，可以尽可能降低真实风险，因为真实风险 $R(w)$ 满足：

$$R(w) \leq \text{Remp}(w) + \varphi(n/h)$$

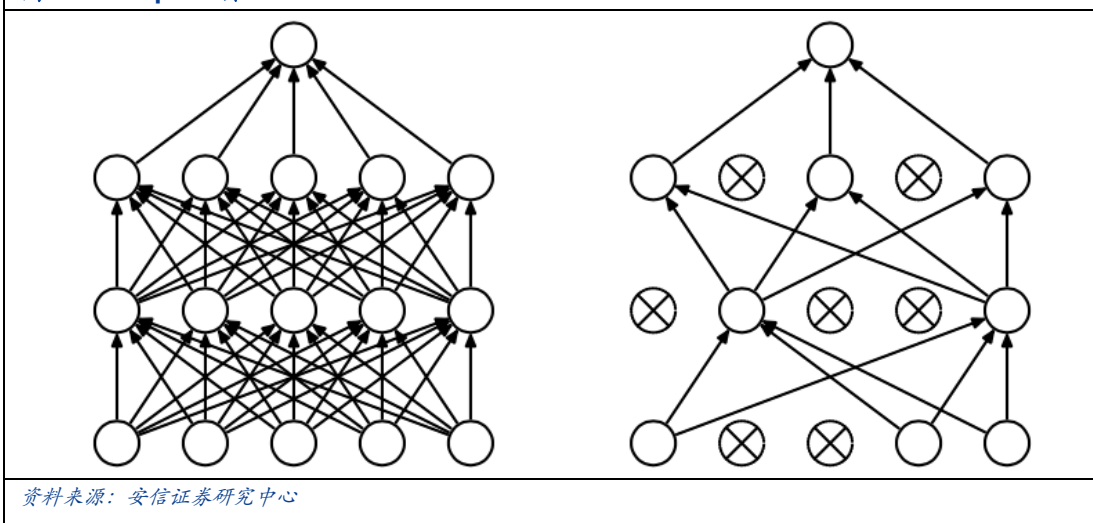
(3) 结合算法的其他方法

例如如果使用神经网络，那么可以用 Dropout 来有效平衡 Bias-Variance Trade off。

Dropout 的思想是训练整体 DNN，并平均整个集合的结果，而不是训练单个 DNN。DNNs 是以概率 p 舍弃部分神经元，其它神经元以概率 $q=1-p$ 被保留，舍去的神经元的输出都被设置

为零。

图 18: Dropout 算法



如上图所示,被舍弃的神经元用X表示出来了。所有指向它和从它发出的有向箭头都被斩断。

3.3.3. 人的理解方式经常是线性的

正向线性思维的特点是，思维从某一个点开始，沿着正向向前以线性拓展，经过一个或是几个点，最终达到思维的正确结果。举例而言，经常有人会认为，如果过去 100 年平均每年的生产力提升是一千亿，那么未来一年统计上的期望生产力提升也是一千亿。这里就犯了一个常见错误。人类生产力的提升经常是指数级别上升的，所以未来一年统计上的期望生产力提升应该不止一千亿。

正是因为人的思维方式是线性的，在理解非线性的时候会直观上比较困难。为了增加读者对非线性的直观理解，我们将在下一篇中重点阐述。

4. 从单次分析到推进分析

4.1. 算法和模型

算法与模型基本和标准的神经网络回归策略一样，不同的是，我们这次要比较第一篇报告中写了单次分析和推进分析。

下面是单次分析的常见方法

表 3: 单次分析示例

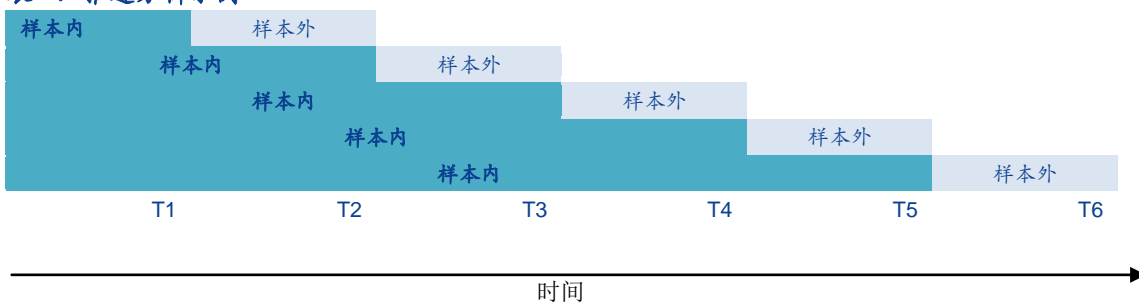
样本内	样本外
时间 →	

资料来源：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：推进分析示例



资料来源：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。

4.2. 结论

在日内单边千分之一，隔日单边万分之 3 的成交假设下，策略表现如下：

夏普：2.66

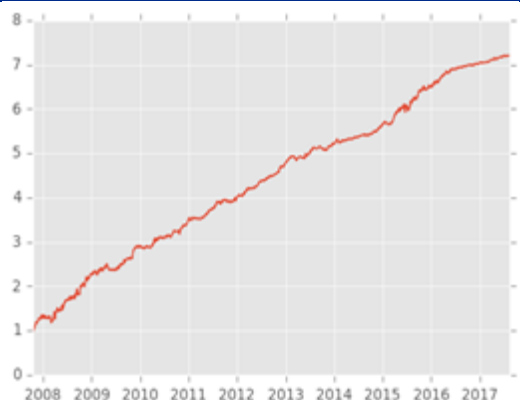
最大回撤：17.24%

胜率：57.56

盈亏比：1.22

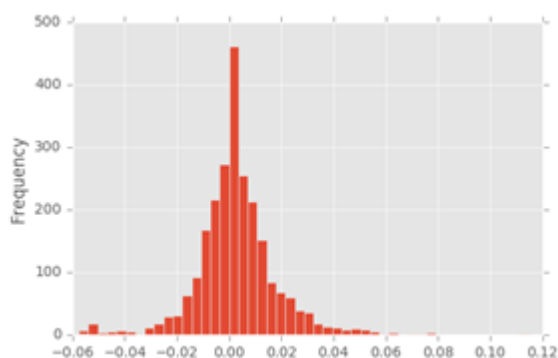
年化：56.38%

图 19：单次分析标准神经网络回归-净值



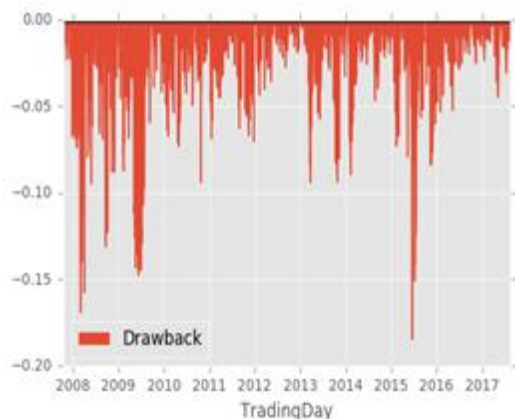
资料来源：Wind，安信证券研究中心

图 20：单次分析标准神经网络回归-收益



资料来源：Wind，安信证券研究中心

图 21：单次分析标准神经网络回归-回撤



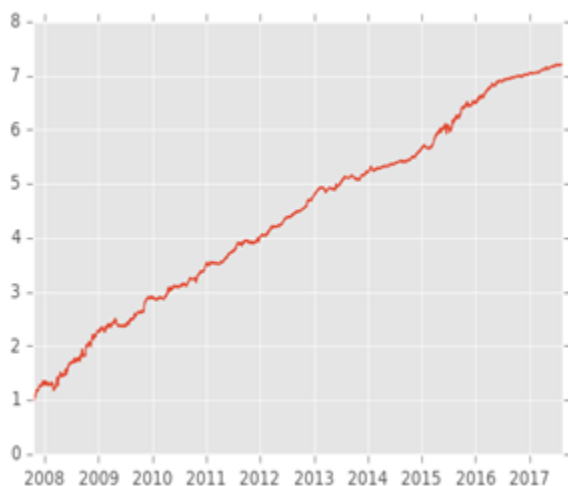
资料来源：Wind，安信证券研究中心

图 22：单次分析标准神经网络回归-分年度夏普



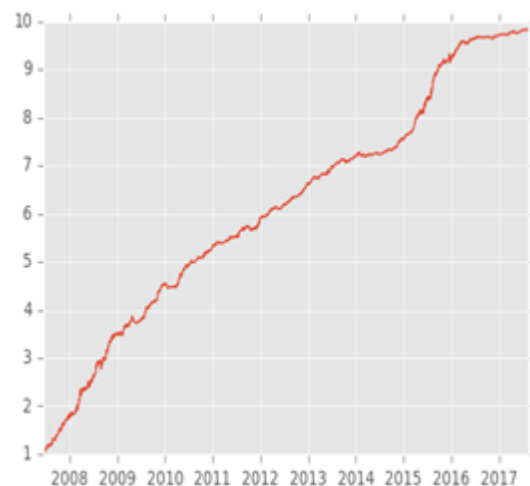
资料来源：Wind，安信证券研究中心

图 23：单次分析标准神经网络回归-净值



资料来源：Wind，安信证券研究中心

图 24：标准神经网络-净值



资料来源：Wind，安信证券研究中心

表 5：单次分析标准神经网络回归与标准神经网络比较

参数/策略名	单次分析标准神经网络回归	标准神经网络回归
夏普	2.66	3.55
最大回撤	17.24%	17.05%
年化收益	56.38%	80.36%
日胜率	57.56%	62.69%
盈亏比	1.22	1.31

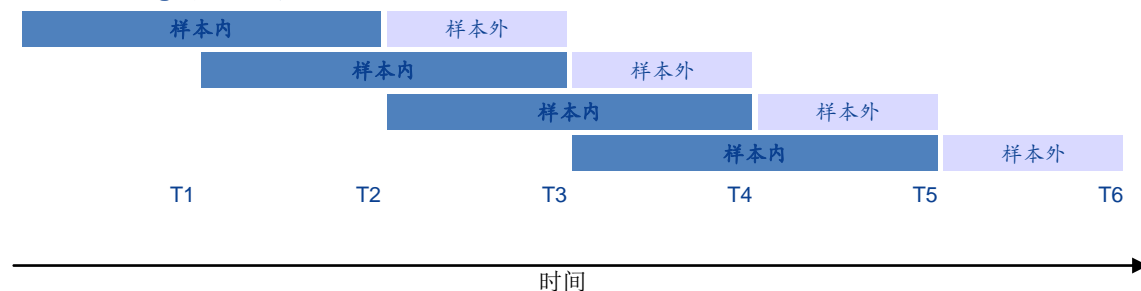
资料来源：Wind，安信证券研究中心

4.3. 单次分析和推进分析的逻辑讨论

在逻辑上，推进分析更接近实盘。因为在实盘中，经常地，模型会每日在盘后更新。所以如果是这样，在回测时候，也应该假定，在T日末，模型会被重新训练，也即是站在T日的模型和站在T+1的模型也应当是不一样的。这样做的显著好处是，不论是回测还是实盘，每个模型都能用到站在当前时点上最新的数据。坏处也很显著，在回测时候，对历史上的每一天都要建立一个模型，这样的计算量是巨大的。

另外推进分析也有不同的方法。

表 6: Rolling 推进分析方法



资料来源：安信证券研究中心

上图是另一种推进分析的方法 (Rolling)。与之前推进分析的方法不同，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，我们并不像之前一样，用 0~T3 的数据训练模型，而是用 T1~T3 的数据训练模型，然后同样的在 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。

使用全样本做推进分析和使用过去 n 期样本做推进分析之间没有优劣之分。选择时候大体上要遵循两个基本原则，一个是数据要具有对当前市场状态的代表性，另一个数据量要尽可能多。使用过去 n 期样本通常能对当前市场状态的代表性，使用全样本做推进分析。

5. 从分类到回归

5.1. 算法和模型

算法与模型基本和标准的神经网络回归策略一样，不同的是，预测目标不再是某段时间的收益，而是一个二分类。也即，大于 0 的时候是上涨分类，小于 0 的时候是下跌分类。

5.2. 结论

在日内单边千分之一，隔日单边万分之 3 的成交假设下，策略表现如下：

夏普：1.66

最大回撤：25.30%

胜率：49.72%

盈亏比：1.39

年化：30.91%

图 25：神经网络分类-净值

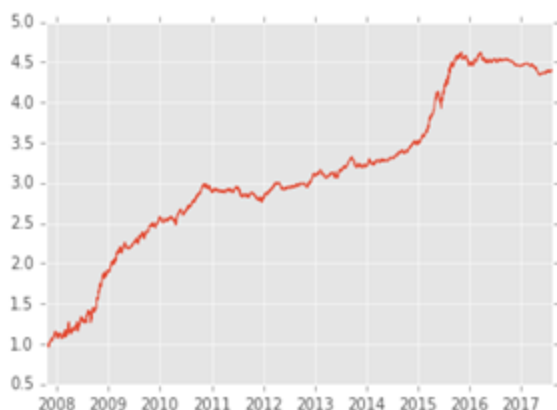
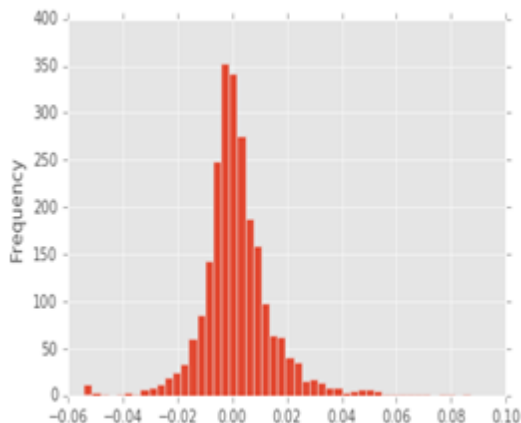
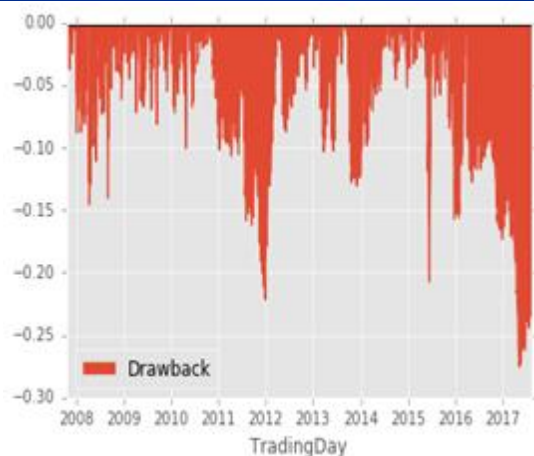


图 26：神经网络分类-收益



资料来源：Wind，安信证券研究中心

图 27：神经网络分类-回撤



资料来源：Wind，安信证券研究中心

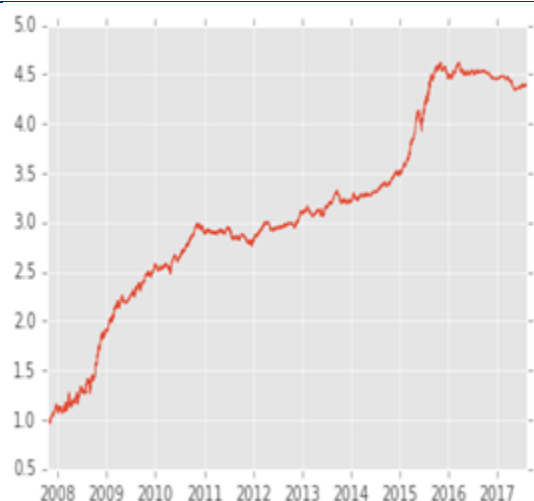
资料来源：Wind，安信证券研究中心

图 28：神经网络分类-分年度夏普



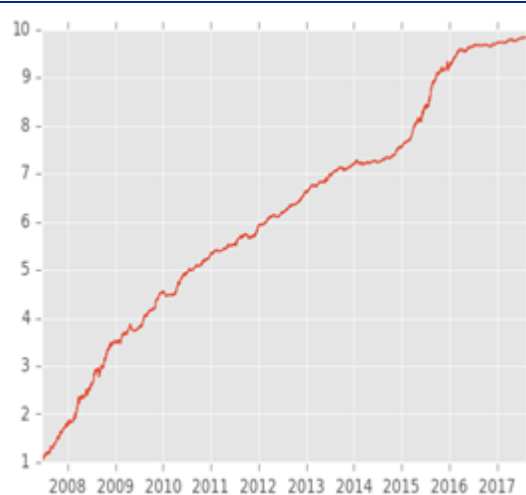
资料来源：Wind，安信证券研究中心

图 29：神经网络分类-净值



资料来源：Wind，安信证券研究中心

图 30：标准神经网络-净值



资料来源：Wind，安信证券研究中心

表 7：神经网络分类与标准神经网络回归比较

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

5.3. 分类与回归的逻辑讨论

分类的逻辑是，市场的状态是离散可分的。如果按照上例，涨 0.1%和涨 10%都会归结到上涨一栏。但是事实上，涨 0.1%和涨 10%是截然不同的，前者很可能是随机扰动，而后者一定是市场情绪的体现。但是随机扰动与市场情绪的分界点是很难确定的，涨 0.1%是随机扰动，但是涨 0.5%是不是随机扰动呢？所以分类有个天生的问题，以什么标准来划分类，如何划分类？我们也曾经尝试过划分成 5 类，7 类，但是由于划分的类过多，效果也不及二分

类。

6. 预测值相关

6.1. 算法和模型

算法与模型基本和标准的神经网络回归策略一样。在标准神经网络回归，我们只有当大于一个阈值的时候，做多。小于一个阈值的时候做空。但是在该策略中，再算出预测值后，直接预测值大于 0 就做多，小于 0 就做空。

6.2. 结论

在日内单边千分之一，隔日单边万分之 3 的成交假设下，策略表现如下：

夏普：2.17

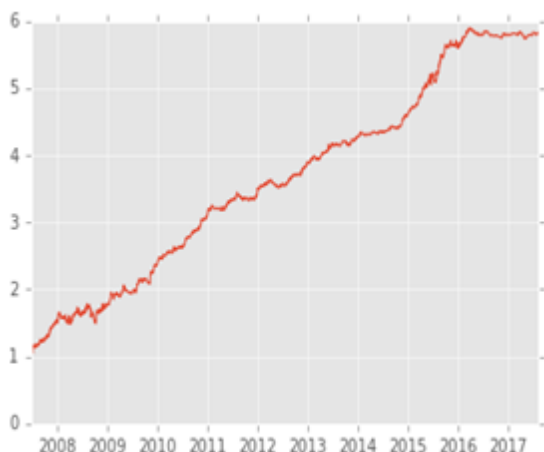
最大回撤：26.05%

胜率：46.68%

盈亏比：1.75

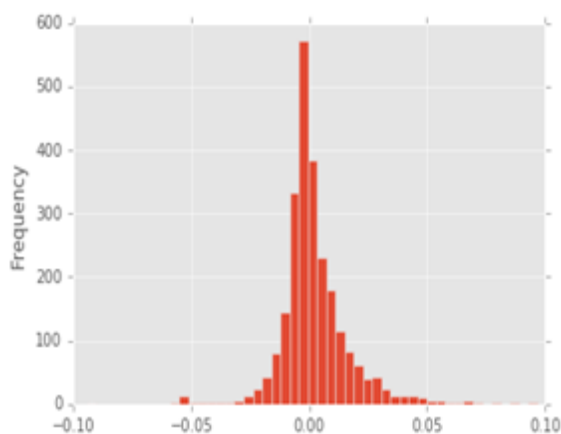
年化：43.92%

图 31：神经网络回归值，以 0 为界限-净值



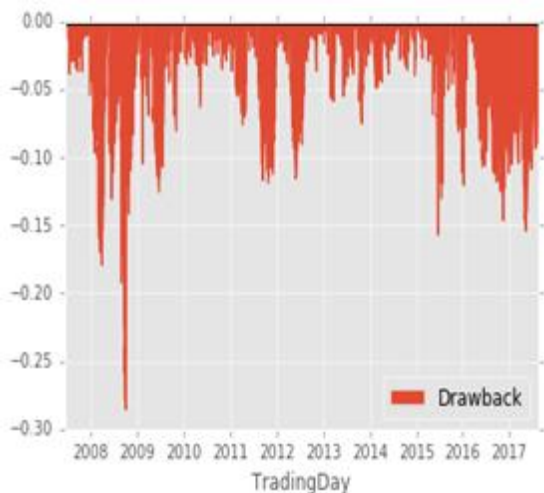
资料来源：Wind，安信证券研究中心

图 32：神经网络回归值，以 0 为界限-收益



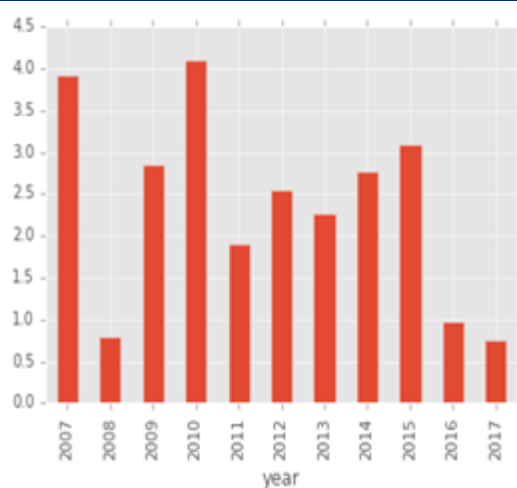
资料来源：Wind，安信证券研究中心

图 33：神经网络回归值，以 0 为界限-回撤



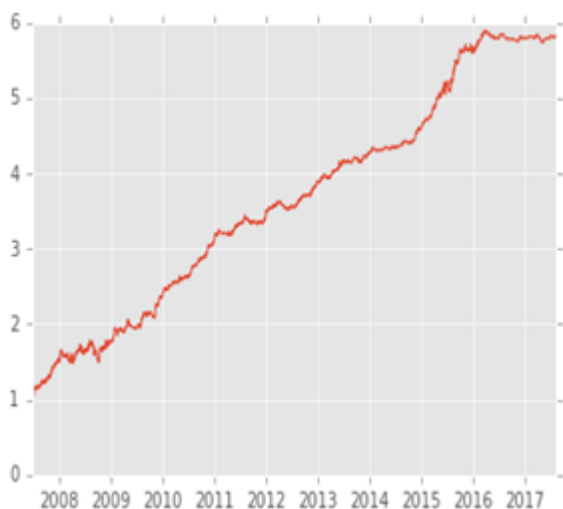
资料来源：Wind，安信证券研究中心

图 34：神经网络回归值，以 0 为界限-分年度夏普



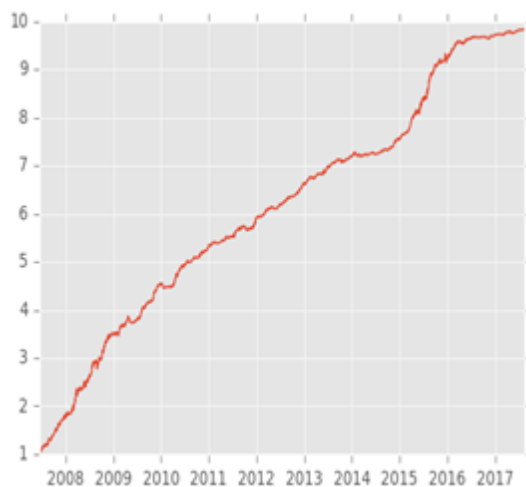
资料来源：Wind，安信证券研究中心

图 35：神经网络回归值，以 0 为界限-净值



资料来源：Wind，安信证券研究中心

图 36：标准神经网络-净值



资料来源：Wind，安信证券研究中心

表 8：神经网络回归值，以 0 为界限与标准神经网络比较

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

6.3. 预测值相关的逻辑

预测值可以从上述的回测看到，预测值如果按照简单的大于 0，小于 0 交易，效果并不是特别出色。目前业内比较公认的结论是，预测值的强度代表方向的概率。举例而言，如果一个预测值是 0.1%，一个是 1%，那么后者实际上涨的概率大于前者。因此，选择一个合适的阈值变的至关重要。可能的确定阈值的方法可以是，历史上预测值的平均数加上一个历史上预测值的标准差作为看多阈值；历史上预测值的平均数减去一个历史上预测值的标准差作为看空阈值。

一个绝对值较大的预测阈值容易漏掉一些真正的上涨机会（统计上的 Type II error），而一个绝对值比较小的预测阈值容易错误的开多仓（统计上的 Type I error）。同时，绝对值预测阈值越小，越容易达到阈值，越容易触发交易，交易频率就越高，交易成本就越高。反之交易就越不频繁，交易成本就越低。

■ 分析师声明

杨勇、周袁声明，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。

本报告版权仅为本公司所有，未经事先书面许可，任何机构和个人不得以任何形式翻版、复制、发表、转发或引用本报告的任何部分。如征得本公司同意进行引用、刊发的，需在允许的范围内使用，并注明出处为“安信证券股份有限公司研究中心”，且不得对本报告进行任何有悖原意的引用、删节和修改。

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
	胡珍	0755-82558073	huzhen@essence.com.cn
深圳联系人	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