

2018 年 04 月 16 日

机器学习与量化投资：避不开的那些事（4）

■机器学习波动率预测

大多数量化策略的盈利与波动率高度相关。预知波动率对于分配每个策略的仓位至关重要。使用机器学习进行波动率预测较传统方法的预测效果有所提升。

■机器学习策略判断失效的方法

判断机器学习策略失效有独特的方法，可以在击穿最大回撤前提前下线策略。

■机器学习在量化投资中应用的杂谈

我们在这一章节中致力于打通实盘的各个环节，以及展开对机器学习对冲基金运营方式的探讨。

■风险提示：

波动率预测和策略失效判断是基于历史数据的，过去不代表未来；杂谈都是基于当前技术的探讨，技术进步可能会杂谈内容不再有效。

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

安信金工黑科技原理揭秘之一：周期分析理论	2018-04-15
大市与行业研判：风险尚可控，结构先均衡后或再偏中小创	2018-04-15
FOF 和资产配置周报：个人税收递延养老保险试点通知发布，养老目标基金开始申报	2018-04-15
FOF 和资产配置周报：广发中证京津冀 ETF 联接开始募集	2018-04-08
FOF 和资产配置周报：中融量化精选 FOF 下周开始募集，4 月维持超配债券	2018-04-02

内容目录

1. 波动率预测	3
1.1. 历史波动率概述	3
1.2. 文献综述	4
1.3. 策略简介	4
1.3.1. 策略摘要	4
1.3.2. 策略细节	4
1.3.3. 策略表现	5
1.3.4. 与简单移动平均的比较	6
2. 如何判断策略失效	6
2.1. 机器学习对数据的依赖更强	6
2.2. 判断机器学习策略失效的一些想法	7
3. 杂谈	8
3.1. 计算落地相关：我们需要什么级别的计算力？	8
3.2. 交易系统相关	10
3.3. 机器学习与主观交易	11
3.3.1. 机理不同	11
3.3.2. 数据量	11
3.3.3. 举一反三	11
3.3.4. 非结构化数据	11
3.4. 机器学习在量化投资的机遇与挑战	11
3.4.1. 数据	11
3.4.2. 算法和计算力	12
3.4.3. 用户/投资者	12
3.4.4. 机器学习对冲基金的架构	13

图表目录

图 1：中国波指	3
图 2：中证 500 按日统计波动率	3
图 3：沪深 300 按日统计波动率	4
图 4：日内波动率预测曲线	5
图 5：日内波动率实际曲线	5
图 6：日内波动率实际与预测差值	6
图 7：20 日移动平均内波动率实际与预测差值	6
图 8：滚动 IC 值	8
图 9：Hadoop 里的服务器角色	8
图 10：Apache Spark 简介	9
图 11：AWS 服务器价格	10
图 12：阿里云服务器价格	10
图 13：机器学习对冲基金分工	14
图 14：Two Sigma 是一家以机器学习见长的对冲基金	15
图 15：Two Sigma 非常强调开源	15

未找到图形项目表。

1. 波动率预测

1.1. 历史波动率概述

从 2016 年中旬左右开始，全球的波动率就一直处于历史的低位。中国波指更是非常稳定的一路下探，从最高的 63.79 一直到了最低的 7.95。

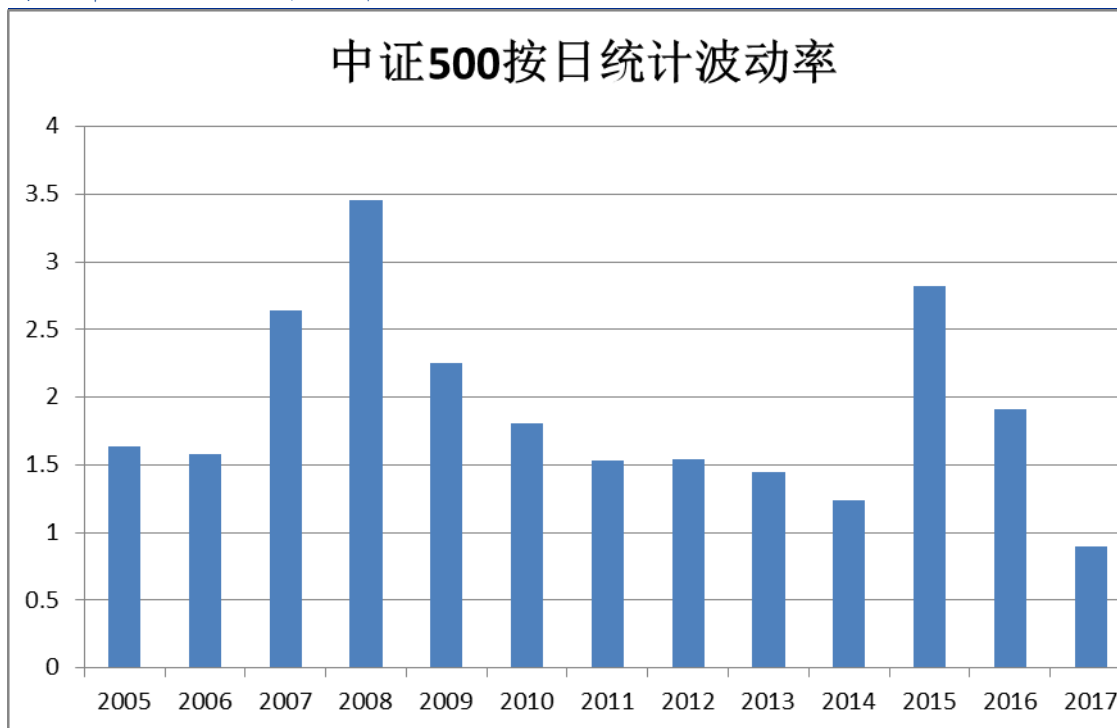
图 1：中国波指



资料来源：Wind，安信证券研究中心

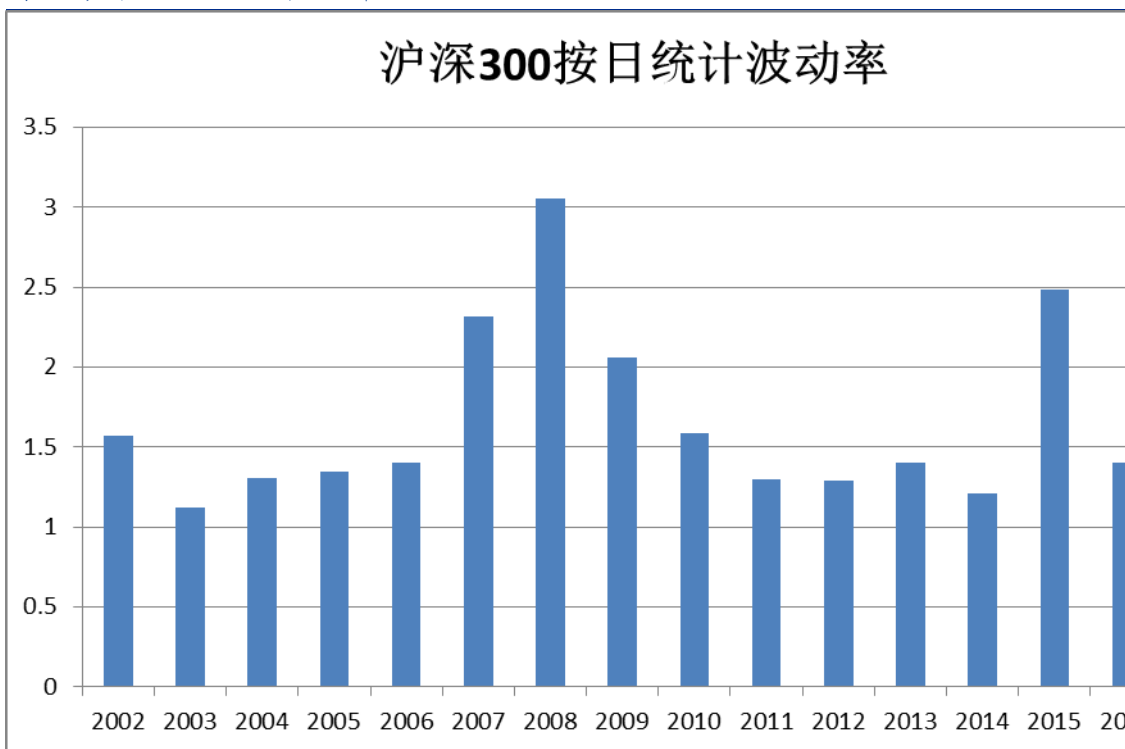
如此长时间的低波动率确实在中国市场上不常见。下面我们统计了一下中证 500 和沪深 300 指数分年度的波动率分布情况，可以看到 2017 年最近的波动率确实是处在历史最低位的。

图 2：中证 500 按日统计波动率



资料来源：Wind，安信证券研究中心

图 3：沪深 300 按日统计波动率



资料来源：Wind，安信证券研究中心

1.2. 文献综述

波动率预测的方式有下列几种；

(1) 移动平均

使用过去一段时间窗口的已实现波动率作为下一期波动率的预测值。

(2) 时间序列

使用 ARCH/GARCH 等等模型,通过时间序列的自回归,来预测下一期波动率。传统上认为,波动率具有聚集效应,也即是,高波动率伴随着高波动率,低波动率伴随着低波动率。由于时间序列模型是回看最近的过去一段时间的波动率,所以时间序列分析是最符合波动率聚集假设的。

(3) 隐含波动率

在中国市场,可以基于 50ETF 期权依赖期权定价模型来反推隐含波动率。

1.3. 策略简介

1.3.1. 策略摘要

本篇报告我们提出用监督式学习对日内波动率进行预测。结果表明,使用监督式学习的预测效果超过简单移动平均的效果。我们对日内波动率的预测为定义为每天 240 根分钟线收盘价的标准差。

1.3.2. 策略细节

策略标的：中证 500 指数

预测生成：该策略的预测目标就是下一交易日的日内波动率。预测日内波动率有很大的意义，

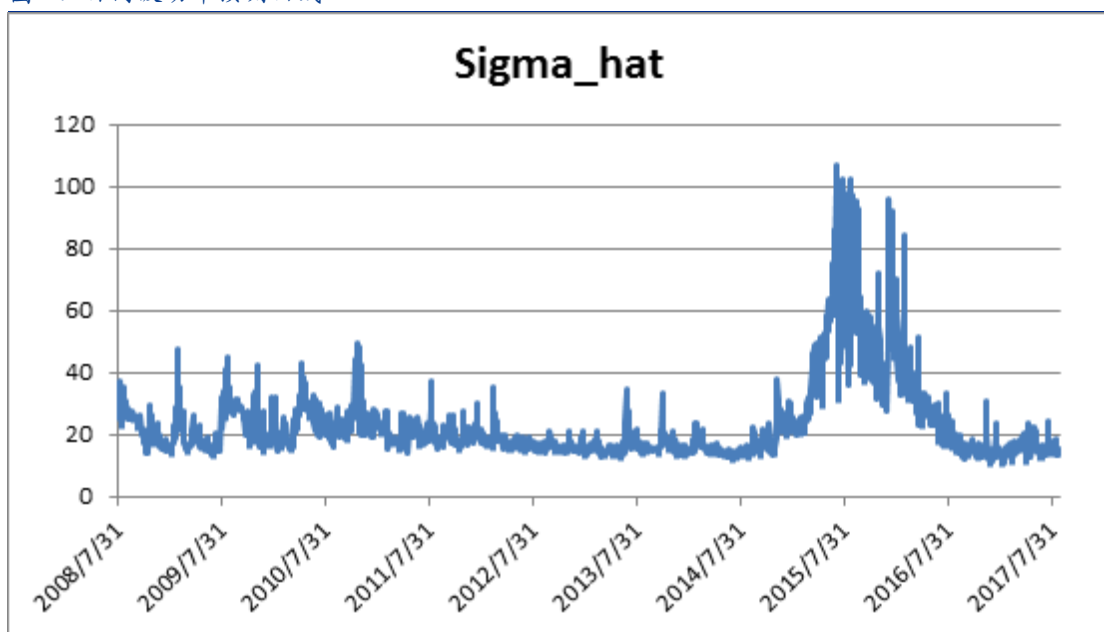
除了期权交易之外，大多数的中高频 CTA 策略的收益也与日内波动率紧密相关。

考虑到波动率的聚集效应，我们使用过去一段时间的时间序列的日内波动率作为输入变量。输入的变量例如，过去 30 天的日内波动率的移动平均值，过去 n 日的日内波动率最大值，过去 n 日日内波动率的最小值等等。

1.3.3. 策略表现

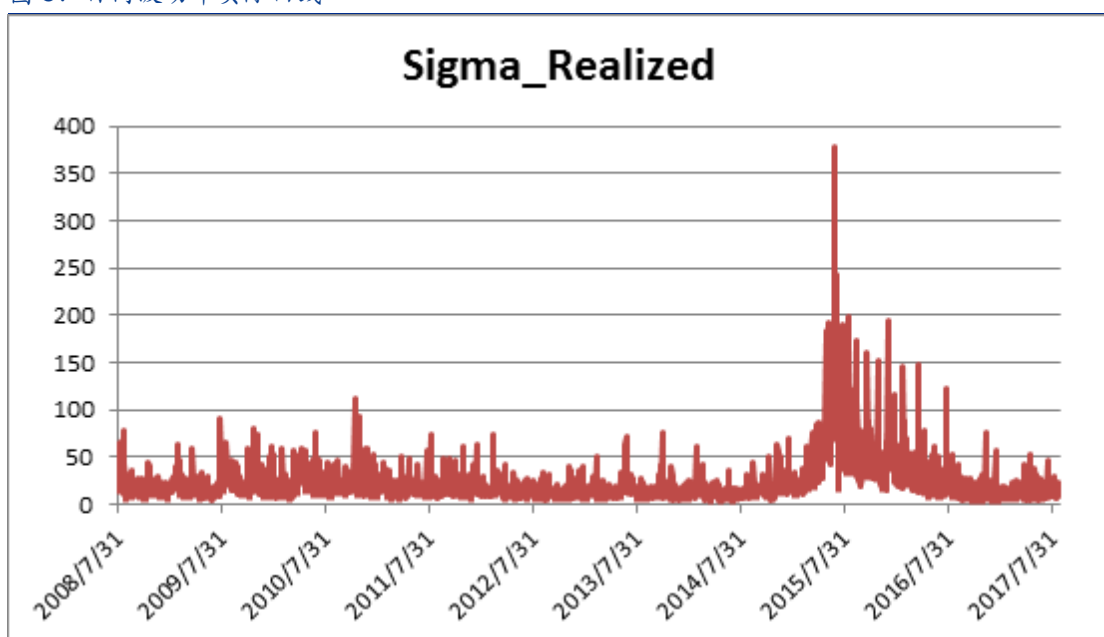
策略预测的 R 方为 37.13%，显示出策略监督式学习对未来一日日内波动率的预测是有一定效果的。下图是日内波动率的预测值和实际值之间的比较

图 4：日内波动率预测曲线



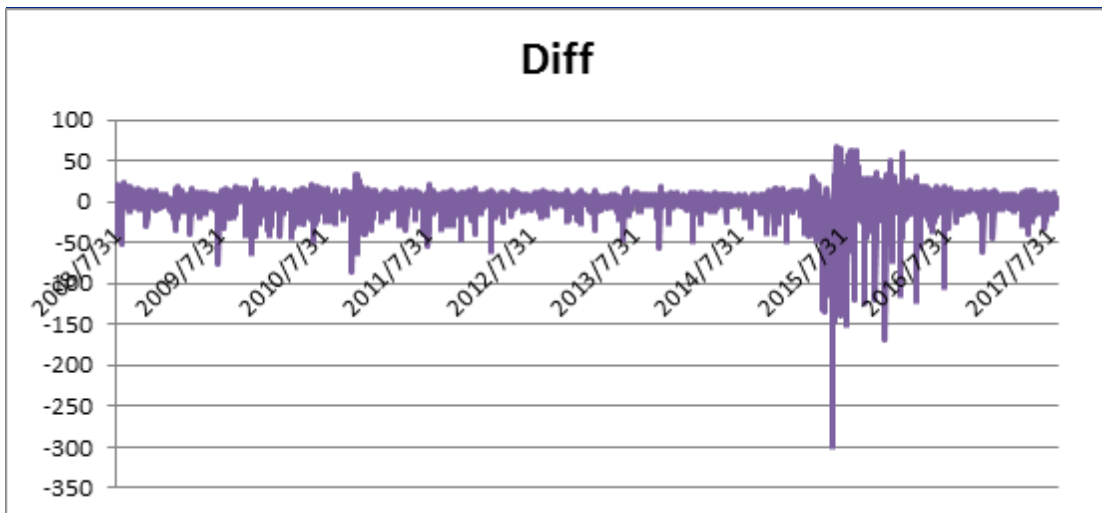
资料来源：Wind，安信证券研究中心

图 5：日内波动率实际曲线



资料来源：Wind，安信证券研究中心

图 6：日内波动率实际与预测差值



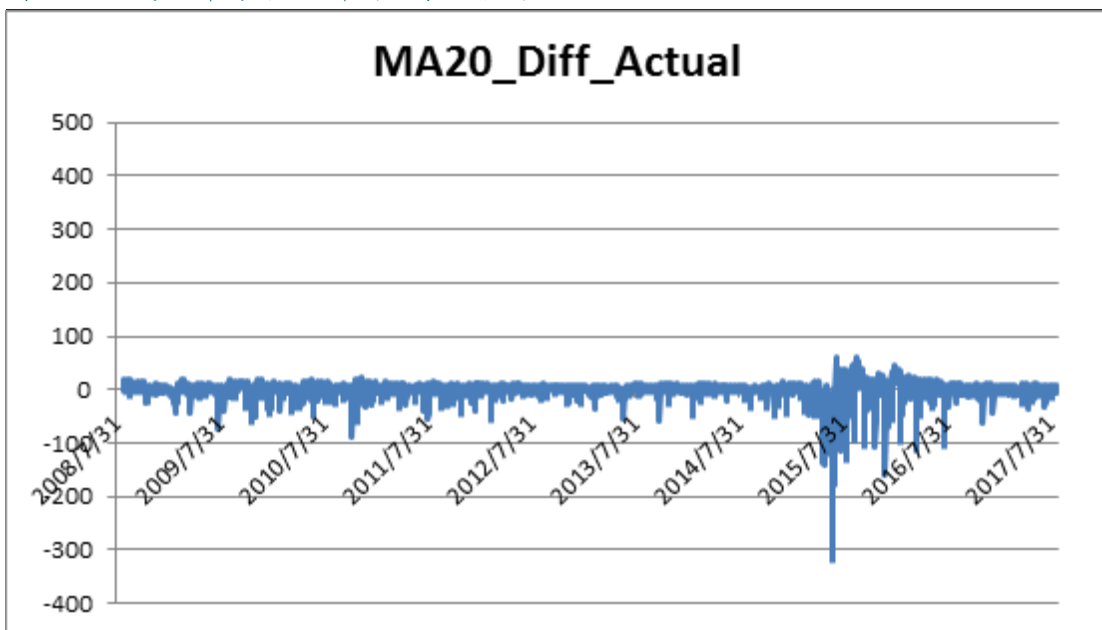
资料来源：Wind，安信证券研究中心

可以看到，除了在股灾中之外，历史上其他时候预测都比较准确。

1.3.4. 与简单移动平均的比较

简单移动平均 MA20，也即是过去 20 天的移动平均的日内波动率作为下期的预测值，R 方是 34.42%，比机器学习的预测要少 3% 左右。

图 7：20 日移动平均内波动率实际与预测差值



资料来源：Wind，安信证券研究中心

2. 如何判断策略失效

在以上几篇报告中间，我们具体讨论了一些策略，以及在研究一些机器学习策略时候常见的错误，但是我们迄今没有讨论如何判断机器学习策略的失效。在本文中我们并不将具体给出策略失效的解决方案，但是将就这个问题给出一些讨论。

2.1. 机器学习对数据的依赖更强

机器学习相对于传统策略对数据依赖性更强，所以对未来市场结构的变化会更为敏感。

量化的本质是依赖于“已有之事后必再有,已行之事后必再行,日光之下并无新事”，而对于未来市场结构的变化通常是无能为力的。例如如果一个 Alpha 策略的样本是从 2009 年到 2016 年,在不做市值中性的情况下,那么回测下来大概率小市值因子给的权重会很大。换句话说,因为历史样本中几乎都是小市值相对于大市值占优(除了 2014 年末这段短暂的时间内大市值占优),所以模型对于 2017 年的大市值占优是没有准备的,这直接导致了大部分 Alpha 策略在 2017 年大规模回撤。

尽管所有量化策略都对未来市场结构的变化无能为力,传统的量化策略在面对未来市场结构的变化还是会有一些天生的优势。因为传统的量化策略往往是从假设出发的,这些假设可能是从自己经验而来,也可能是从学术文献上来的,数据只是用来验证这些假设是否正确的。所以当传统量化策略回撤的时候,策略研究人员大致能明白是自己的哪一条假设出错了。

而在机器学习的世界中,事情却不是这个样子的,因为机器学习完全是数据驱动的 (data driven)。事先我们对问题并没有任何假设,我们不知道自变量和因变量之间有没有关系,如果有,我们也并不知道它们之间的关系是否是线性的,我们也并不一定能假设自变量与因变量的统计分布。因此,总而言之,我们最后所得出的所有结论都是机器学习基于数据给我们的。在机器学习的世界中,人的作用被缩小,而数据的作用被放大。所以,一旦新的数据与原来的数据发生系统性的不同,整个模型以及模型给出的结论都会出现系统性的偏差。因此,机器学习模型相对于传统的量化模型更容易受到市场变化的不利影响。

2.2. 判断机器学习策略失效的一些想法

下面提出一类方法。简而言之就是,基于过去一段足够长的时间 n , 计算预测值和实际值的 IC, 并判断这个 IC 是否显著。具体举例,如果模型的预测目标是 T 日到 $T+1$ 日的收益,检测天数是过去 100 天,那么我们显然就可以得到这 100 个 T 日到 $T+1$ 日的收益的预测值和实际值,然后根据 $\text{Cor}(y, \hat{y})$ 计算出 IC, y 代表实际值, \hat{y} 代表预测值,两者都是长度为 100 的向量。

最后,我们通过

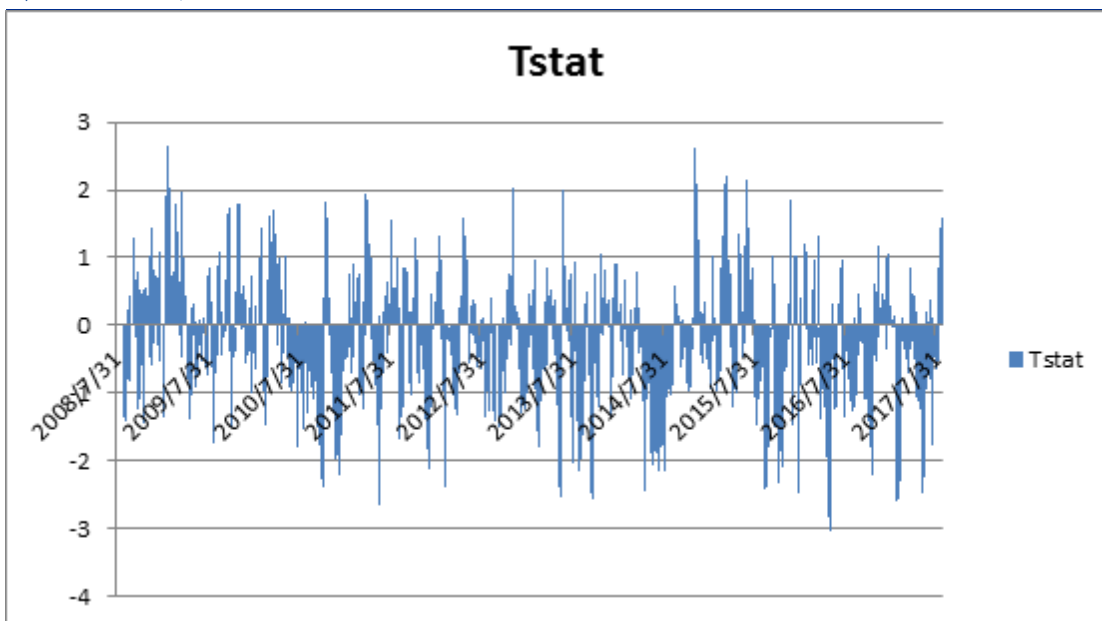
$$t = \frac{r\sqrt{n-2}}{1-r^2}$$

来计算出 t 值,并且与 t_{n-2} 进行比较,看是否显著。这里 $r = \text{Cor}(y, \hat{y}), n=100$ 。如果 t 值并不显著,那么策略可能已经失效。

对 n 的选择非常关键,对于一天几百笔的高频策略,可能一天就足够达到统计显著需要的量了。但对于低频策略,可能需要天数数目非常大。

以波动率为例子,我们可以得到如下滚动的 IC 值,这里选取 $n=20$

图 8: 滚动 IC 值



资料来源: Wind, 安信证券研究中心

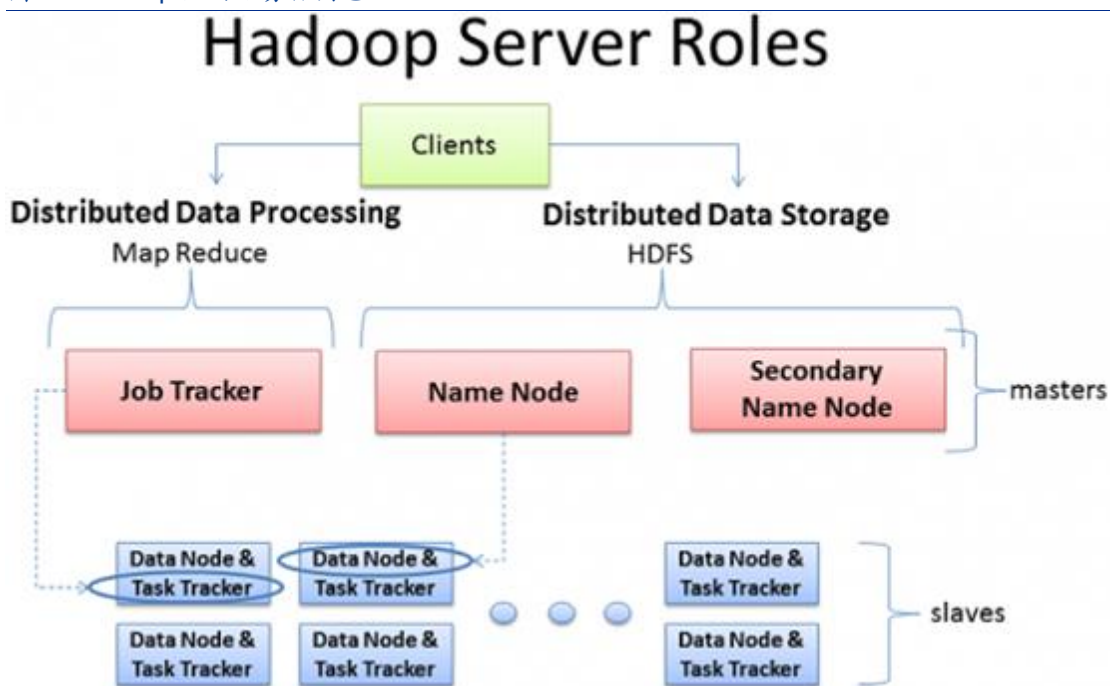
可以看到, 在 2016 年 7 月到 8 月之间, 有一次明显的失效, t 值小于 -3, 达到统计显著水平。

3. 杂谈

3.1. 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, 如今一些对冲基金也在考虑使用 Apache Spark 或者 Hadoop。

图 9: Hadoop 里的服务器角色



资料来源: Cloudera

上图是 Hadoop 里的服务器角色。Hadoop 主要的任务部署分为 3 个部分，分别是：Client 机器，主节点和从节点。主节点主要负责 Hadoop 两个关键功能模块 HDFS、Map Reduce 的监督（HDFS 是 Hadoop Distributed File System 的简称，Map Reduce 是一种算法）。当 Job Tracker 使用 Map Reduce 进行监控和调度数据的并行处理时，名称节点则负责 HDFS 监视和调度。从节点负责了机器运行的绝大部分，担当所有数据储存和指令计算的苦差。每个从节点既扮演者数据节点的角色又充当与他们主节点通信的守护进程。守护进程隶属于 Job Tracker，数据节点归属于名称节点。

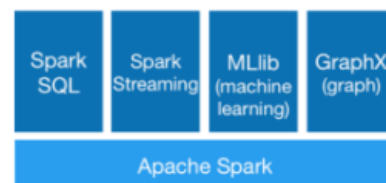
Client 机器集合了 Hadoop 上所有的集群设置，但既不包括主节点也不包括从节点。取而代之的是客户端机器的作用是把数据加载到集群中，递交给 Map Reduce 数据处理工作的描述，并在工作结束后取回或者查看结果。在小的集群中可能会面对单物理设备处理多任务，比如同时 Job Tracker 和名称节点。作为大集群的中间件，一般情况下都是用独立的服务器去处理单个任务。

图 10: Apache Spark 简介

Generality

Combine SQL, streaming, and complex analytics.

Spark powers a stack of libraries including SQL and DataFrames, MLlib for machine learning, GraphX, and Spark Streaming. You can combine these libraries seamlessly in the same application.



资料来源: Spark

Apache Spark 是一个新兴的大数据处理的引擎，主要特点是提供了一个集群的分布式内存抽象，以支持需要工作集的应用。相比于 Hadoop，它更快，更加容易使用。

如今这些年，对冲基金如果不是处于数据安全性的考虑，并不需要独自搭建这些复杂的系统。Amazon Web Service 或者阿里云提供的公有云中自带了这些系统。租金对于对冲基金来说，也不会太贵。

下面我们将从用 Apache Spark 和 Hadoop 导致的成本和产生的效用角度来讨论在对冲基金中是否需要用 Apache Spark 和 Hadoop。

从需求出发，Apache Spark 和 Hadoop 首先要运用于大数据，所以如果一个国内对冲基金只是用到交易数据，那么数据量不会超过 10 个 T。在这种情况下，没有必要用 Apache Spark 和 Hadoop。真正有需求是使用超越市场数据的数据。例如使用社交网络数据，使用图片等等。

从供给成本出发，我们截图了 AWS 和阿里云的定价，对于大多数对冲基金而言，假设使用 m4 级别，以一小时 2 美金算，一天 48 美金，基本这个成本可以忽略不计。

图 11: AWS 服务器价格

m4.large	2	6.5	8	仅限于 EBS	\$0.1 每小时
m4.xlarge	4	13	16	仅限于 EBS	\$0.2 每小时
m4.2xlarge	8	26	32	仅限于 EBS	\$0.4 每小时
m4.4xlarge	16	53.5	64	仅限于 EBS	\$0.8 每小时
m4.10xlarge	40	124.5	160	仅限于 EBS	\$2 每小时
m4.16xlarge	64	188	256	仅限于 EBS	\$3.2 每小时

资料来源: Amazon

图 12: 阿里云服务器价格

规格族	实例ID	vCPU	内存(GB)	按量(小时)	月价(月付)	月价(年付)	月价(2年付)	月价(3年付)
通用型 (g5)	ecs.g5.large	2	8	0.988	300.0	255.00	210.00	150.00
通用型 (g5)	ecs.g5.xlarge	4	16	1.976	600.0	510.00	420.00	300.00
通用型 (g5)	ecs.g5.2xlarge	8	32	3.963	1200.0	1020.00	840.00	600.00
通用型 (g5)	ecs.g5.4xlarge	16	64	7.914	2400.0	2040.00	1680.00	1200.00
通用型 (g5)	ecs.g5.6xlarge	24	96	11.875	3600.0	3060.00	2520.00	1800.00
通用型 (g5)	ecs.g5.8xlarge	32	128	15.837	4800.0	4080.00	3360.00	2400.00
通用型 (g5)	ecs.g5.16xlarge	64	256	31.663	9600.0	8160.00	6720.00	4800.00

资料来源: 阿里巴巴

所以从成本角度考虑, 用大型互联网公司的云也并不是非常贵, 对冲基金可以考虑投资。

3.2. 交易系统相关

在笔者读书的时候, 用的深度学习系统是 Caffe (Convolutional Architecture for Fast Feature Embedding), 当时的 Caffe 只支持 Ubuntu, Red Hat, OS X。与 Caffe 一样, 当时包括 TensorFlow 在内的绝大多数机器学习系统也都是基于 Linux 的。尽管现在机器学习社区渐渐开始支持 Windows, 例如出现了 Caffe2 和 MXNET 和 TensorFlow 的 Windows 版, 但是许多算法依然只有在 Linux 中才能实现。与此相对应, 很多交易系统是仅支持 Windows (不是跨平台的), 对于如何衔接机器学习系统和交易系统, 这里提出一个有效的方案。

从一般的 CTA 策略来说, 是可以完全分成两部分的, 一个是信号产生端, 一个是交易系统端, 各自基本是可以独立的, 它们唯一的交集是信号产生端需要交易系统端的数据, 交易系统端需要信号产生端产生的交易信号。由于一般 CTA 策略对实时性要求不高, 故网络造成的延迟并不重要。所以完全可以用 Client Server 双向通信的架构解决问题。信号产生端通过通信拿数据, 交易系统端通过通信拿信号。

3.3. 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。因为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，两者不太能够直接比较。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。以下是一些详细的理由。

3.3.1. 机理不同

众所周知在深度学习神经网络当中，梯度下降是让这些深度学习具有优秀表现的最重要的算法。然而梯度下降跟大脑的工作方式很不一样，它不是来自于大脑的启发，而是来自于数学。神经科学中研究的是具有复杂空间分布的神经细胞，它们可以有穿透或者非线性的行为，它们之间的连接方式无比复杂。事实上，我们到今天对神经还没有研究清楚。所以说深度学习神经网络像动物神经细胞间的连接只是一种比喻，深度学习网络的本质是数学、计算机，而不是靠仿造大脑的结构。

3.3.2. 数据量

人类做出决定通常是通过几个例子做出的，例如，当人们在说价值投资的好的时候，往往只是由于茅台等等个别的股票，也就是依赖非常小的数据量做出的决策。但是当机器学习做出决策的时候，一定是大样本的，大多数的机器学习模型依赖大量的数据使得算法得到收敛。

3.3.3. 举一反三

机器学习能使用大数据，但是这不代表机器学习就能清楚地观察全局、能理解抽象的概念、能做长期的推理、能改变事情发生的可能性或是对事情做出预见。其实推理是很难的，在多步骤的、具有不确定性的状况下，人类能做得非常好，但是人工智能机器学习系统就需要用大量的数据反复训练，然后才能在某些步骤中模仿人类，但尽管如此，还是没办法达到人类的水平。

3.3.4. 非结构化数据

人类对于非结构化数据（图片，声音等等）的处理是非常好的，好到难以置信。人类在做出任何决定的时候，几乎都是视觉，听觉，甚至于味觉同时作用的。人类不仅仅能明白他看到了什么，听到了什么，尝到了什么，更能够把这些东西整合在一起形成一个结论进而做出决策。然而机器学习却在这方面做得很差。首先，机器学习依赖于人的标注，例如在识别猫的时候，机器需要首先得到一个受到标注的数据集，标注清楚哪些是猫，哪些不是猫。其次，这些猫的图案还必须在长宽高上面彼此一致，否则识别不出或者错误率很高。而对于人的话，如果第一次是用高分辨率的图，第二次改用低分辨率的图，照样能识别出猫。第二，假设机器能很好的用 CNN 识别图案，用 RNN 识别语音，并且假设都预测全对，但是机器却不能将 RNN 和 CNN 非常好的结合，这是一个很大的问题（很难像人一样把这些东西整合在一起形成一个结论进而做出决策）。

3.4. 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。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。

3.4.1. 数据

在数据科学时代，如果算法和计算力是发动机，那么数据就是石油。机器学习可以充分发挥大数据的优势。

尽管到目前为止还没有对大数据的非常明确、权威的定义，但是我们依然可以总结出大数据三大特点：全体性（即全体数据而不再抽样）、混杂性（即数据不追求精确）和相关性（即找相关性重于因果关系），大数据还具有数据量大、数据种类多样等等特点。

市场行情数据经常是 TB 级别的，但是相对于互联网行业庞大的用户数据而言，TB 级别的市场行情数据非常小。所以为了充分在这么小的数据集上实现机器学习算法，在仅仅使用市场行情的情况下，机器学习在量化投资中的运用大多数情况下都必须是高频的（就如同避开不开的那些事（2）所述，小数据上算法容易不收敛）。

另外所有的数据大体上可以分为结构化数据和非结构化数据，金融领域的市场行情数据显然是结构化数据，但是市场上还有大量的没有被利用的非结构化数据。包括大家都非常热衷的舆情数据，也包括暂时没有能力处理的大量的图片和语音数据，这些数据或许对股票市场的预测作用举足轻重。

下面给出一些非结构化数据的例子，

（1）通过客户留下的信息在外部能获取的信息：通过家庭地址获取房产及小区信息，通过微信微博等获取互联网社交信息

（2）上市公司重点产品在重点区域的经销记录（价格、销量、利益分配方式）

（3）上市公司相关新闻、在社交媒体中对该公司相关的评论信息，该公司的重要合作伙伴的信息和线索

（4）上市公司的高管个人的评价信息、高管人员的微信微博

值得注意的是，数据量的增加会冲击传统的法律法规，使得内幕交易的边界变得模糊。例如，一个在商业银行的从业者，如果能拿到消费者在零售业巨头的信用卡刷卡记录，他根据这些信用卡的刷卡记录，预判零售业巨头们的销售与盈利情况，以此来交易二级市场的股票牟取大量利润。这究竟是否可以判断为内幕交易？

3.4.2. 算法和计算力

算法和计算力本质上是现代机器学习的核心矛盾。为此，从算法上，Hilton 提出了要放弃反向传播；算力上，Google 研发出了 TPU，英伟达的黄仁勋更是祭出了 TensorRT3 这一大杀器。此外，一些面向人工智能的专用硬件架构也开始出现，比如说用 FPGA 去做专用的人工智能加速芯片和加速的基础设施，微软的数据中心就大量的运用了 FPGA 技术。但是现在传统方式依然是租用亚马逊云或者阿里云，或者自己使用 GPU。未来，随着量子计算机的发展，当量子芯片中的量子比特数量达到一定数量后，计算能力将满足人工智能对运算能力的需求，人工智能将不再依赖于大型服务器集群。在算法和计算力不断提升的背景下，机器学习有望在量化投资领域发挥更大的作用。

3.4.3. 用户/投资者

机器学习对用户端的作用很明显，它的作用可以类比阿里巴巴的推荐商品。

例如可以用来判别用户属于那种风险偏好，作出对应的风险偏好分析。具体来说，可以用来匹配用户和交易策略；可以通过分析流失客户的行为特征而采取挽留行动；分析高亏损客户的交易行为特征而有针对性的加强这些客户的投资咨询服务；可以通过分析客户群体的财经浏览信息、自选股信息的变化，分析判读客户群体的焦点；分析客户产品购买行为与客户其他数据的关系（基本资料、浏览行为等等），设计相应的产品提供给类似客户群体。

3.4.4. 机器学习对冲基金的架构

机器学习对冲基金本质上是科技公司。正如同科技公司的生产流程可以划分成前端后端一样，对冲基金也可以。

本质上机器学习对冲基金可以分为：

数据工程师

数据工程师是大型数据库的设计者，建造者和管理者。他们的目标是把数据整理好，存储成本低，查询效率高的形式。一般地，数据工程师具有良好的计算机背景。

机器学习平台工程师

机器学习平台工程师负责搭建和维护机器学习平台。市面上的开源机器学习平台大多能满足要求，但是要能部署到本地的服务器集群还是需要一定的技巧。一般地，机器学习平台工程师具有良好的计算机背景。

数据科学家/量化研究员

数据科学家或者量化研究员负责机器学习的建模。并且产生有价值的交易决定。他们会用到数据工程师搭建的数据库和机器学习平台工程师搭建的机器学习平台，他们往往需要具有统计学计算机和金融的交叉背景。

以上三类是科技公司也需要配备的，但是对于对冲基金而言，还需要如下两类，

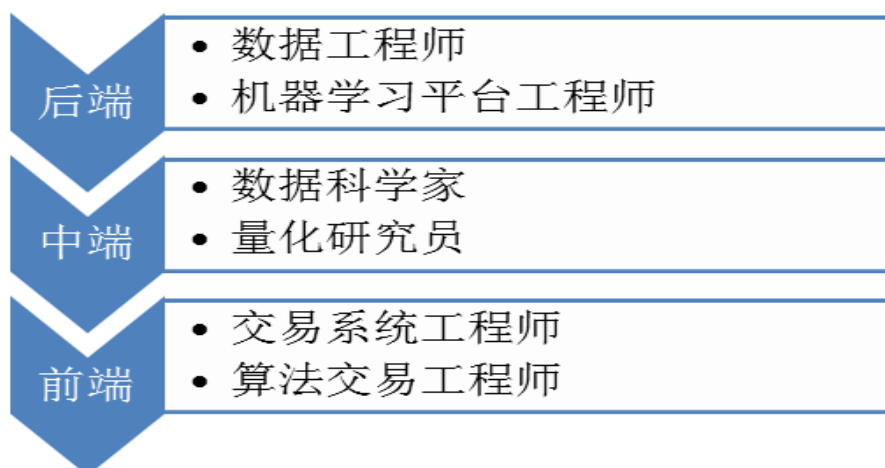
交易系统工程师

交易系统工程师负责搭建和维护交易系统，这里所说的交易系统通常包括实时风控、报单执行和报单管理系统。

算法交易工程师

算法交易工程师负责处理算法交易，减小冲击成本等等。算法交易也可以与机器学习结合。事实上，算法交易与机器学习的结合正是长期以来被认为的最为成功的一个方向。

图 13：机器学习对冲基金分工



资料来源：安信证券研究中心

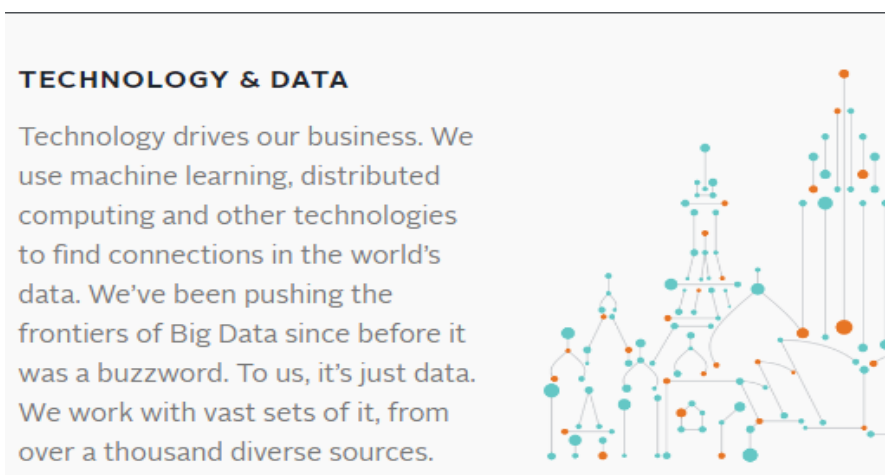
一个优秀的机器学习对冲基金还应有工程师的文化。

工程师文化是一切以解决问题为导向的工作文化。这种文化并不要求当事人本身是个工程师。这种文化以解决问题的第一线人员为核心，除部分公司发展方向的制定者外，所有其他人员均为第一线人员服务。第一线要具有极大的自由度，层级要足够扁平，权责要向下转移。一个优秀的一线问题解决者的收入高于中层管理人员是很正常的。在工程师文化里面，也有领导层级，但在具体的问题上，只有一个判断谁的意见更重要的准则，那就是谁的方案产生的结果更好。背景、资历、年龄、官位，所有的一切都不顶用，顶用的是当事人的方案好用而且好落地。一个人的地位、声誉、威望全都归结到这个人在多大程度上能够解决问题。就算当事人是常春藤毕业的，方案没有一个排名几百名学校的毕业生好用，当事人的地位、声誉和威望就比不上后者。工程师文化不讲意识形态，不看怎么说，只看怎么做。在工程里面，只有“这管用”，没有“这正确”。工程里面信奉的只有一条：“实践是判断真理的唯一标准。”吹得再天花乱坠，未经实践检验，都会被怀疑。

另外，传统上，金融公司的文化往往不够开放。一个不够开放的文化不容易吸引强调自由与开源的工程师的加入。这也解释了为什么像高盛，摩根或者城堡对冲基金这样的金融机构难以获得像 FLAG(Facebook, LinkedIn, Amazon, Google)这样优秀的技术人才。

以 Two Sigma 为例，Two Sigma 是一家以机器学习见长的对冲基金。下面是它的介绍，'Technology drives our business. We use machine learning, distributed computing and other technologies to find connections in the world's data'。

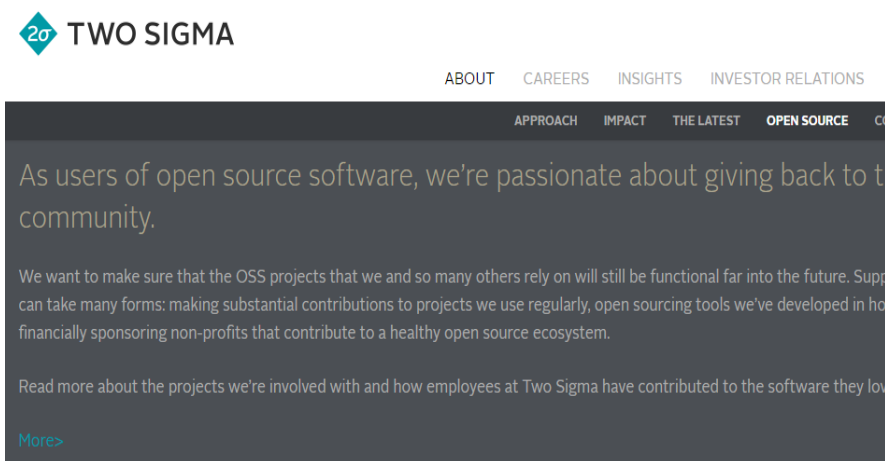
图 14: Two Sigma 是一家以机器学习见长的对冲基金



资料来源: Two Sigma

可以看到，与传统对冲基金不同，Two Sigma 非常强调开源。"As users of open source software, we are passionate about giving back to the community"

图 15: Two Sigma 非常强调开源



资料来源: Two Sigma

■ 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
	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
	林立	021-68766209	linli1@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	温鹏	010-83321350	wenpeng@essence.com.cn
	田星汉	010-83321362	tianxh@essence.com.cn
	王秋实	010-83321351	wangqs@essence.com.cn
	张莹	010-83321366	zhangying1@essence.com.cn
	李倩	010-83321355	liqian1@essence.com.cn
	高思雨	021-35082350	gaosy@essence.com.cn
	姜雪	010-59113596	jiangxue1@essence.com.cn
	周蓉	010-83321367	zhourong@essence.com.cn
	胡珍	0755-82558073	huzhen@essence.com.cn
	范洪群	0755-82558044	fanhq@essence.com.cn
深圳联系人	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