

2018 年 03 月 09 日

机器学习与量化投资：避不开的那些事（2）

■从 IC、IR 到另类线性归因

基于 IC、IR 的单因子分析是传统多因子分析的基石。但是 IC、IR 分析出却不能考虑到多因子模型中因子与因子之间的相互影响。因此我们以之前报告介绍的标准神经网络回归为例，用另类线性归因对因子进行了分析。

■从线性归因到非线性归因

所有线性归因都是基于因子单调性（线性）的强假设。但是在机器学习的非线性世界中，这个强假设不复存在。非线性的机器学习算法需要非线性的归因方式。

■从相关性到因果性

所有的传统归因方式都是基于相关性的而非因果性。因果分析也是机器学习未来的一个重点。我们以 TMLE 为例介绍机器学习下的因果性分析。

■风险提示：

机器学习量化策略的归因是基于历史数据的归因，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：

MSCIA 股相关指数基金积极上报，3 月增配债券

养老基金指引发布点评：专

钱专用的养老金投资新时代

FOF 和资产配置周报：人保

稳进配置混合型 FOF 获受理

机器学习与量化投资：避不开的那些事（1）

FOF 和资产配置周报：从量

化驱动型策略角度说明美股调整

内容目录

1. 机器学习归因的意义.....	3
2. 特征工程与特征重要性.....	3
2.1. 特征工程.....	3
2.2. 特征重要性.....	3
3. 传统线性归因.....	3
3.1. 逐步回归.....	3
3.2. Ridge, Lasso, Elastic Net.....	4
4. 随机森林系列.....	7
4.1. 随机森林.....	7
4.2. Burota.....	9
5. 遗传算法.....	10
6. TMLE.....	11

图表目录

图 1: 特征系数与惩罚系数关系.....	5
图 2: Lasso Regression 的系数.....	5
图 3: Ridge Regression 的系数.....	6
图 4: ElasticNet 的系数.....	6
图 5: 变量间线性关系.....	7
图 6: 变量间非线性关系.....	7
图 7: 随机森林计算因子重要性-打乱 X 前.....	8
图 8: 随机森林计算因子重要性-打乱 X 后.....	8
图 9: 随机森林计算因子（特征）重要性.....	9
图 10: 标准神经网络回归大盘择时策略的因子（特征）重要性排名.....	10
图 11: 遗传算法变异示意图.....	11
图 12: 气温、冰淇淋和啤酒.....	12
图 13: 大盘择时策略的因子重要性归因.....	12

未找到图形项目表。

1. 机器学习归因的意义

对于传统模型，例如 logit 或者决策树而言，输入（自变量）和输出（因变量）的关系是非常明显的。你可以非常清楚的明白为什么一些样本被错误划分了，例如，比如输入因子中某个因子太小了。同样的，对于决策树，同样可以根据决策树每个分叉的逻辑（例如因子 $A >$ 某个常数）向下推演，得出错误划分的原因。但是对于其他大多数的模型，由于它们的高维和非线性，要直观的理解是非常困难的。

尽管如此，让机器学习一个非常有前景的科技让人觉得处于黑箱的状态是非常不明智的。不透明性增加了误用的概率。亚马逊的算法，决定了大多数人今天在读什么书；NSA 的算法决定了谁是潜在的恐怖分子；气候变化模型决定了二氧化碳排放量的安全范围。人不能干预和控制人所不明白的事情，这是什么要单独将机器学习归因的作为一篇报告的原因。

2. 特征工程与特征重要性

机器学习的特征在量化投资当中也被称为因子。

2.1. 特征工程

特征工程是用某些领域内的知识来构造特征的过程。

如果世界上有无穷的数据，和一个 universal function approximator（一个可以表达任何事情的模型），那么就没有特征工程存在的必要。这正是目前在图像识别领域发生的事情，卷积神经网络直接学习每个像素点，然后对图像内容进行识别，而不借助任何人手制的特征。

Coming up features is difficult, time consuming, requires expert knowledge. “Applied machine learning” is basically feature engineering.

Andrew Ng, Machine Learning and AI via Brain simulations

正如吴恩达所述，应用机器学习主要是特征工程。而金融领域的特征获取往往有两种方式。一种是从主观看盘经验来或者从经济学或者金融学的论文来；另一种是纯数据挖掘。后者经常被人诟病容易过度拟合而导致亏损。但是事实上事情可能并没有这么可怕。举例而言 WorldQuant 号称有四百万的 Alpha 因子，结合 Alpha101 来看，很多因子非常可能是数据挖掘出来的，但是从公开业绩来看，它的投资表现还是尚且可以令人满意的。

2.2. 特征重要性

在构造出特征之后，我们需要了解这个特征究竟对我们的预测有没有用，这就需要了解特征重要性。

特征重要性的另一作用是可以进行特征选择，例如选出前五重要性的特征作为模型输入，剩下的可以舍弃。

3. 传统线性归因

传统量化投资是基于线性的世界，在这个世界中，衡量因子的重要度是 IC、IR 等等指标。除了 IC，IR 之外，还有一些值得介绍的传统线性归因的方法。

3.1. 逐步回归

逐步回归的基本想法是，将变量逐个引入，引入变量的条件是偏回归平方和经检验是显著的，同时每引入一个新变量后，对已选入的变量要进行逐个检验，将不显著变量剔除，这样保证最后所得的变量子集中的所有变量都是显著的。这样经若干步以后便得“最优”变量子集。

3.2. Ridge, Lasso, Elastic Net

在线性回归中，损失函数定义为：

$$L = \sum_i (y_i - \hat{y}_i)^2$$

也即RSS。

线性回归的目标在于找到一组系数(w_1, w_2, \dots, w_d)使得RSS最小，但使用RSS作为损失函数可能会导致过拟合，尤其当训练集不够或者特征数量过多时（一个典型的例子是多重共线性），表现为即使实际解释力弱的特征，由于过拟合，它的系数值也较大。为了解决这个问题，在损失函数中对系数加入惩罚项：

$$L = \text{RSS} + \lambda \sum_i (w_i)^2$$

以上式最小为目标来寻找系数的方式就叫做 Ridge Regression。其中 λ 为调节参数，其大小标志着对系数的惩罚力度。 λ 越大，系数就越小。但问题是，系数只能趋近于0，当特征个数很多的时候，对那些本来系数就非常小的特征没什么影响，不能减小模型复杂度。

于是，将损失函数修改为：

$$L = \text{RSS} + \lambda \sum_i |w_i|$$

以上式最小为目标来寻找系数的方式就叫做 Lasso Regression。损失函数在收敛的过程中会使一些系数变为0。变为0的权重对结果影响较小，即对应的特征相对不重要。因此 Lasso Regression 可以筛选特征，有效减小线性模型的复杂度。

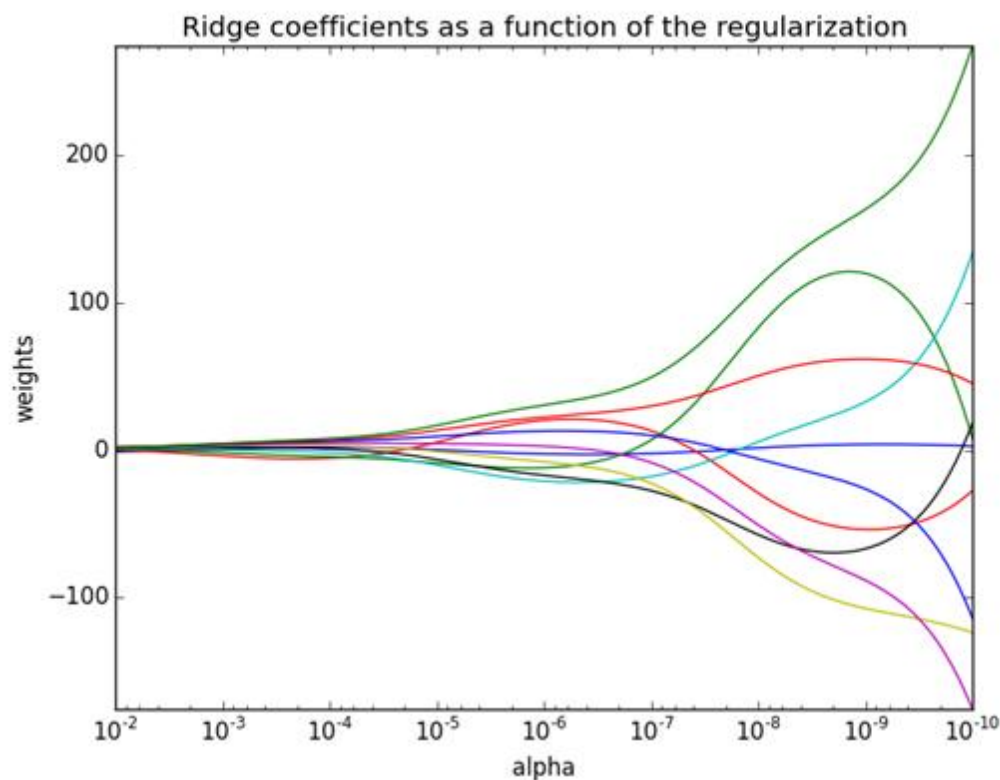
Elastic Net 实际上式 Ridge, Lasso 的综合，其损失函数表示为：

$$L = \text{RSS} + \lambda_1 \sum_i |w_i| + \lambda_2 \sum_i (w_i)^2$$

其中 L1 正则项（Lasso）产生稀疏的系数向量，减小模型复杂度。L2 正则项（Ridge）减小过拟合，消除一定的 L1 稀疏性，以产生 group effect，稳定 L1 正则项的路径。

所以从以上介绍可以看出，Ridge, Lasso, Elastic Net 前面的正则化的系数的绝对值大小直接代表了该特征的重要性。下图代表了随着惩罚系数的增加，特征前面的系数也随之缩小。

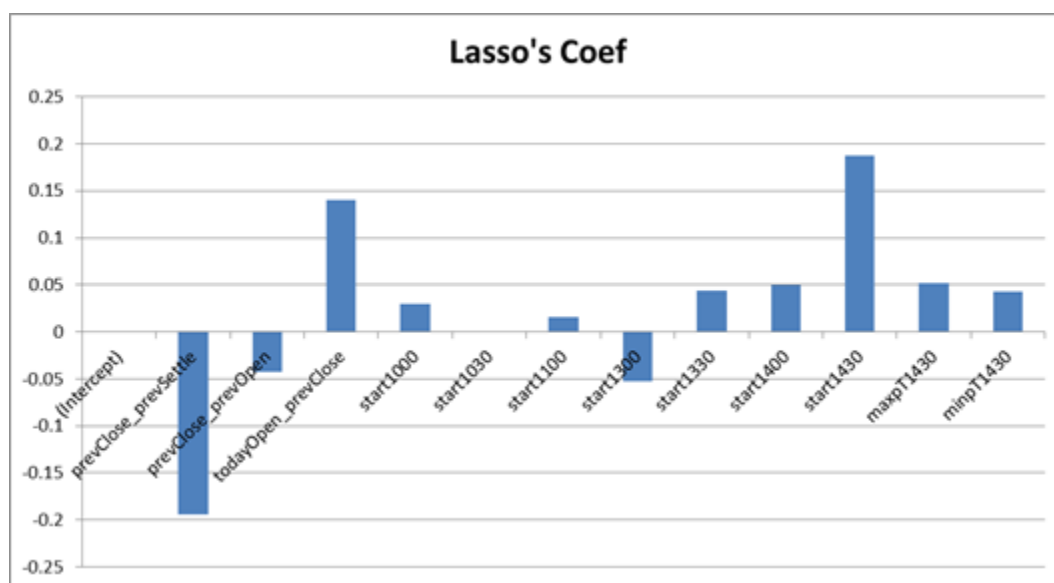
图 1：特征系数与惩罚系数关系



资料来源：Introduction to statistical learning

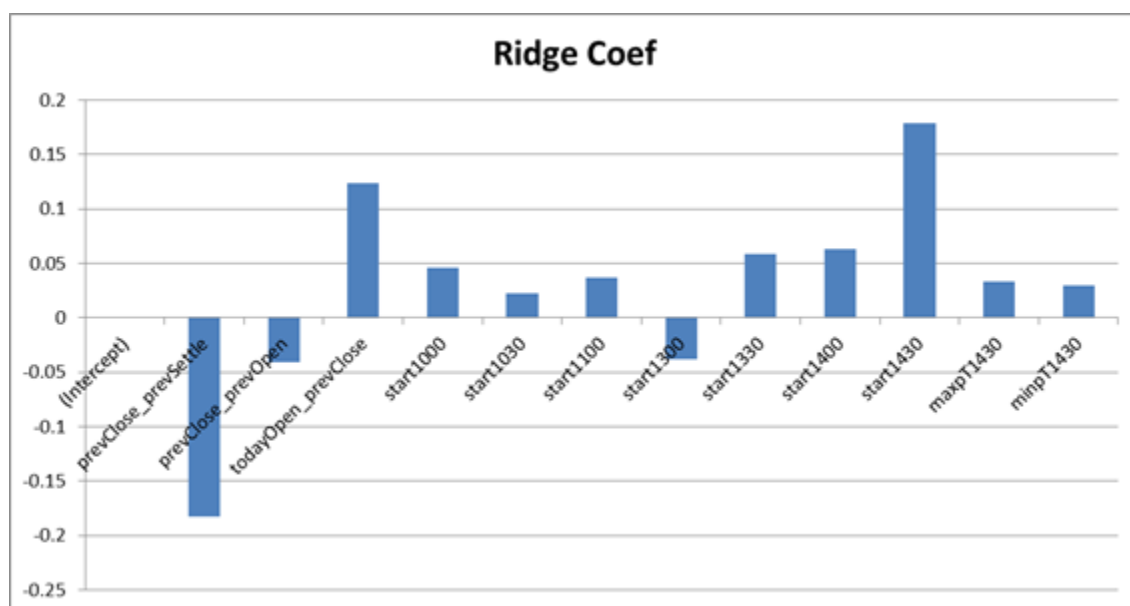
下图是之前上一篇标准神经网络回归策略的因子（特征）重要性排名，绝对值越大越重要，正负代表方向。

图 2：Lasso Regression 的系数



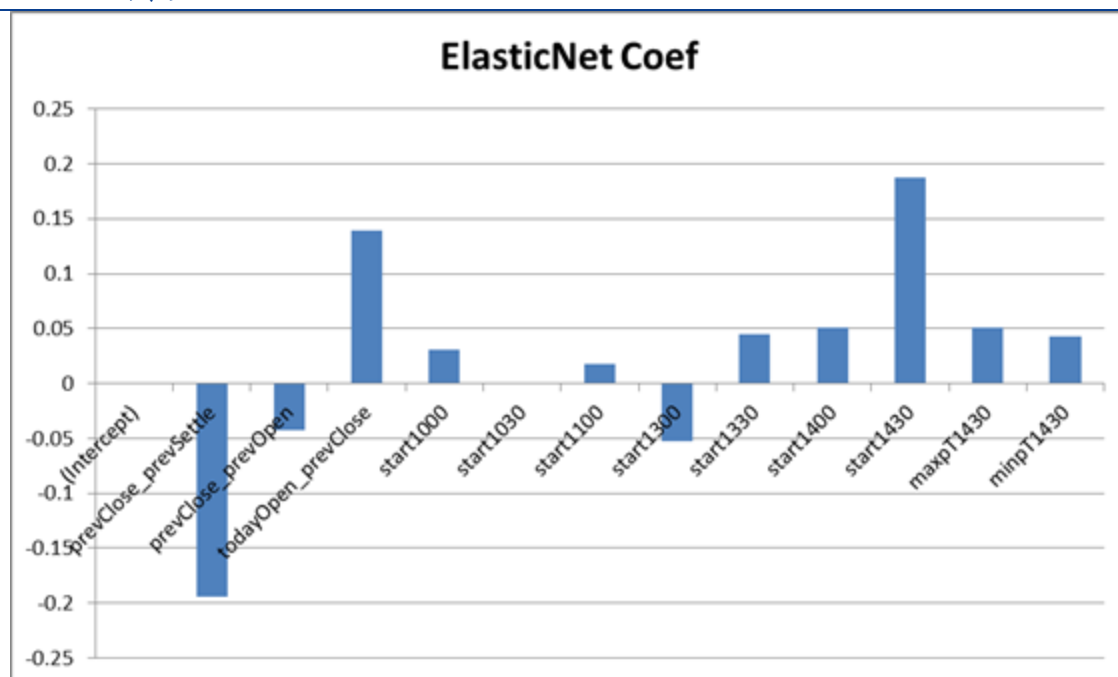
资料来源：Wind, 安信证券研究中心

图 3: Ridge Regression 的系数



资料来源: Wind, 安信证券研究中心

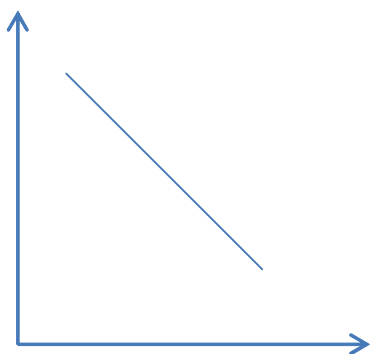
图 4: ElasticNet 的系数



资料来源: Wind, 安信证券研究中心

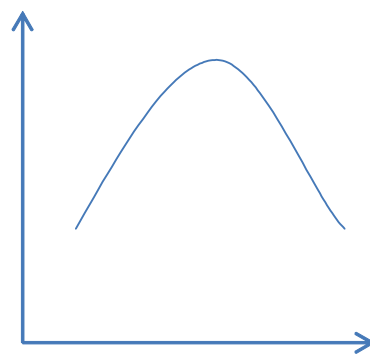
传统量化投资是基于线性的世界。到了非线性的世界中，线性归因显然不能满足要求。

图 5：变量间线性关系



资料来源：安信证券研究中心

图 6：变量间非线性关系



资料来源：安信证券研究中心

例如，在上图左中，变量之间的关系是线性的，而在右图中，线性归因显然是不能反映出真实的变量之间的相关关系。

为了在非线性的世界中衡量因子的重要性，一系列不同的算法被开发出来了。

4. 随机森林系列

4.1. 随机森林

随机森林属于集成学习，可以视为是 bagging 算法在决策树上的运用。

机器学习中决策树主要用于分类和回归，树中的每一个节点表示某一特征的判断条件，其分支表示符合节点条件的对象。叶子节点表示对象所属的预测结果。

随机森林则由许多决策树构成，每棵决策树都由随机的部分样本的部分特征进行训练，它只接受了部分的训练数据，因此每棵决策树都是一个弱学习器。然后，通过 bagging 所有的弱学习器——决策树，比如投票（分类问题）或者取均值（回归问题），得到一个强学习器——随机森林。

由于每一棵树的输入样本不是全部的样本，每一棵树的特征不是全部特征，基于此基础上进行集成，预测结果相对不容易出现过拟合。并且由于训练的样本是随机、独立地进行选取，对各棵树的训练可以并行进行，训练速度相对快。

用随机森林计算因子重要性的方法有很多种，下面介绍其中一种

1: 对于随机森林中的决策树 i , 使用相应的 OOB(Out of Bag 袋外数据)数据来计算它的袋外数据误差, 记为 err_{OOB1_i} .

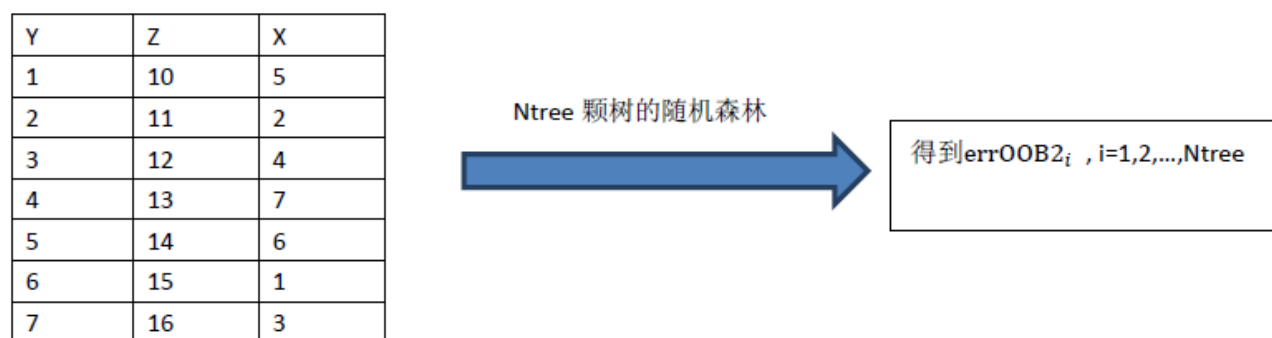
2: 随机地对袋外数据 OOB 所有样本的特征 X 加入噪声干扰(例如可以把 X 重新打乱顺序, 常见的方法是就可以随机的改变样本在特征 X 处的值), 再次计算它的袋外数据误差, 记为 err_{OOB2_i} .

图 7：随机森林计算因子重要性-打乱 X 前



资料来源：安信证券研究中心

图 8：随机森林计算因子重要性-打乱 X 后



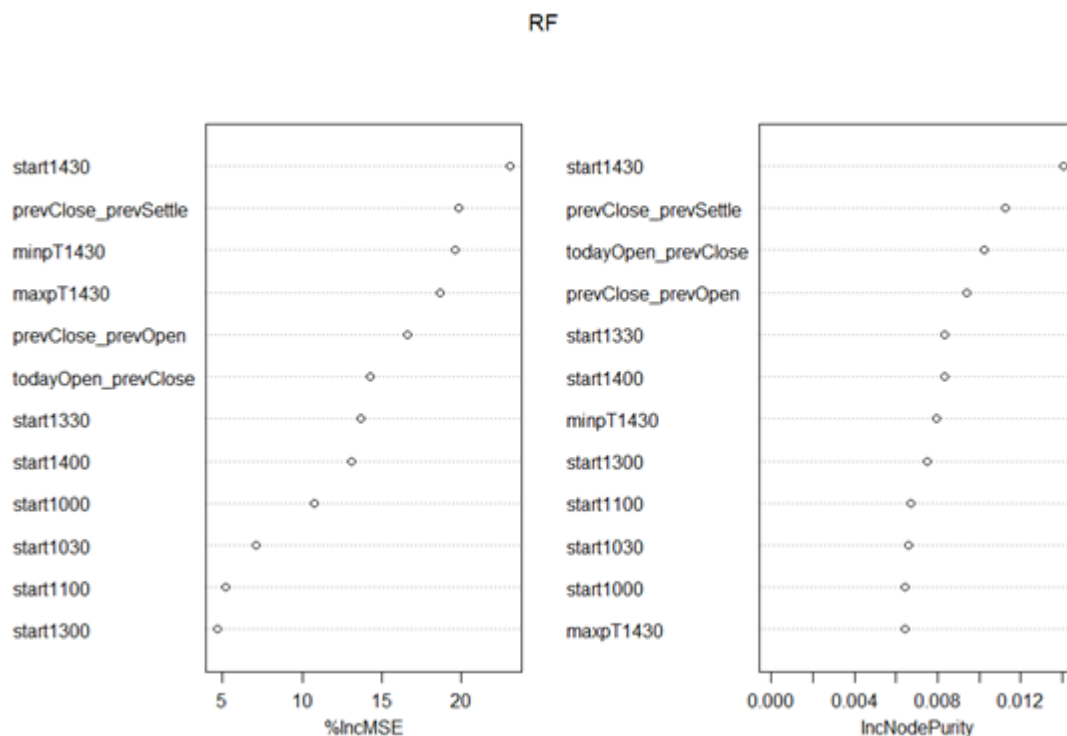
资料来源：安信证券研究中心

3：假设随机森林中有 Ntree 棵树,那么对于特征 X 的重要性为

$$\sum_i \frac{err_{OOB2_i} - err_{OOB1_i}}{Ntree}$$

之所以可以用这个表达式来作为相应特征的重要性的度量值是因为：若给某个特征随机加入噪声之后,袋外的准确率大幅度降低,则说明这个特征对于样本的分类结果影响很大,也就是说它的重要程度比较高。下图是随机森林计算因子重要性的结果图。

图 9：随机森林计算因子（特征）重要性



资料来源：Wind，安信证券研究中心

4.2. Boruta

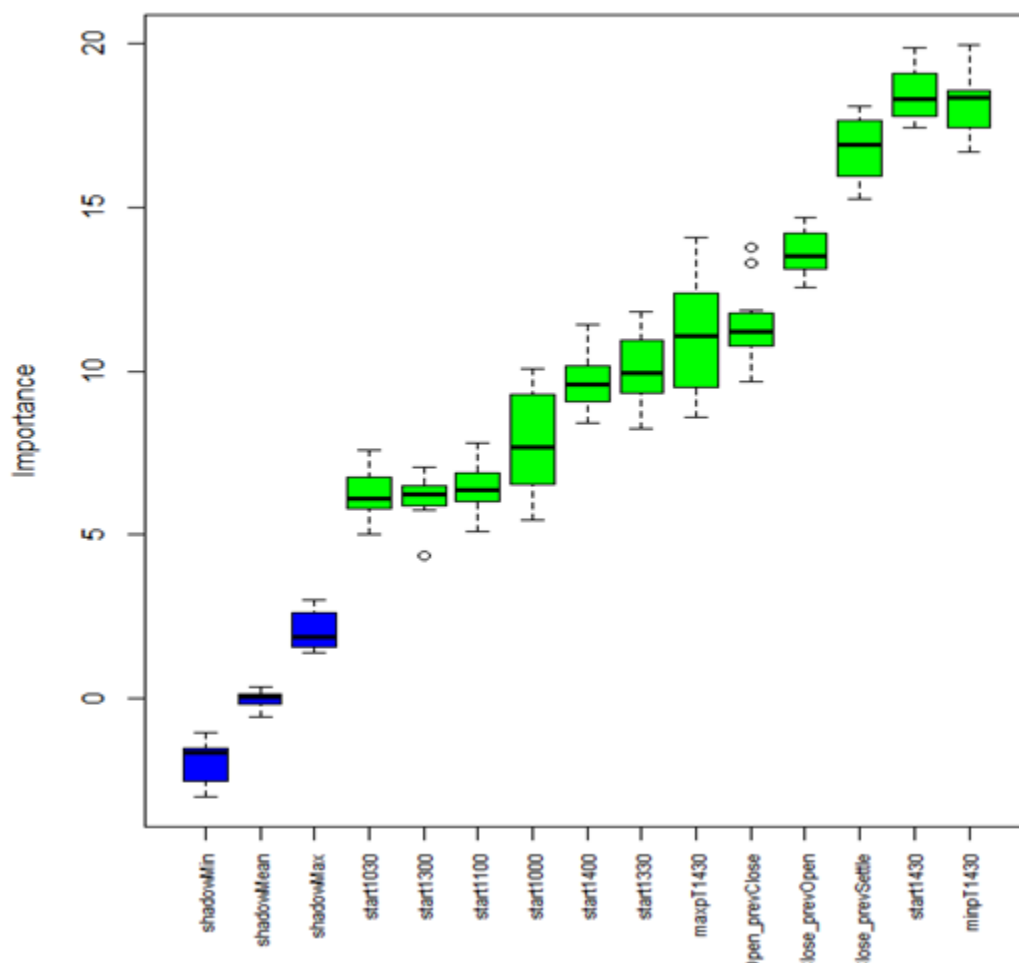
Boruta 是一种特征选择算法。精确地说，它是随机森林周围的一种延伸算法。

下面是 Boruta 算法运行的步骤：

1. 首先，它通过创建混合副本的所有特征（即阴影特征）为给定的数据集增加了随机性。阴影特征就是把许多打乱后的特征作为新的特征
2. 然后，它训练一个随机森林分类的扩展数据集，并计算特征重要性，以评估的每个特征的重要性，越高则意味着越重要。
3. 在每次迭代中，它检查一个真实特征是否比最好的阴影特征具有统计显著的更高（低）的重要性（即该特征是否比最大的阴影特征得分更高），如果是，则确认（拒绝）。它会删除它视为拒绝的特征，然后回到第 1 步。
4. 最后，当所有特征得到确认或拒绝，或算法达到随机森林运行的一个规定的限制时，算法停止。

下图是之前上一篇标准神经网络回归大盘择时策略的因子（特征）重要性排名，从左到右依次从重要到不重要。

图 10：标准神经网络回归大盘择时策略的因子（特征）重要性排名



资料来源：Wind, 安信证券研究中心

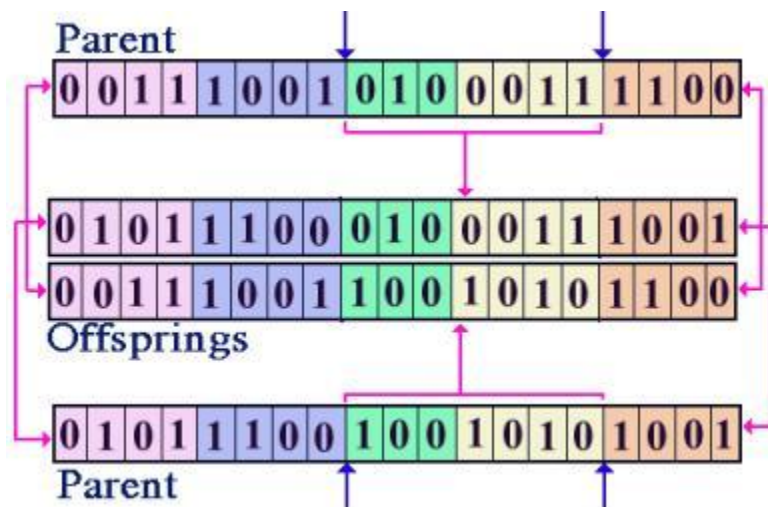
5. 遗传算法

遗传算法主要应用于优化问题，来源于种群进化的想法。首先需要确定题解的形式，一般为向量形式 (x_1, x_2, \dots, x_d) 。开始时，随机生成大量的向量，作为初始种群。然后从该种群中挑选出最优题解，形成新的种群。然后，对它们做出修改，重新挑选出最优题解，依此反复进行这一过程。

修改题解的方法有变异和交叉：变异是对一个既有题解进行微小、简单、随机的改变，比如随机修改向量中一个元素 x_i ；交叉则是选取 2 个最优题解，将它们按某种方式结合，比如 $x_1 \dots x_i$ 来自 a 向量，而 $x_{i+1} \dots x_d$ 来自 b 向量，组成新的向量 c。

变异如下图：

图 11：遗传算法变异示意图



资料来源：安信证券研究中心

新的种群是通过对上一种群中的最优解，进行随机的变异和交叉构造出来的，它的大小通常与旧种群相同。这一过程会一直重复进行，达到指定的迭代次数，或者经数代后题解没有得到改善，结束整个过程。

遗传算法的归因往往需要结合特定的算法。举例来说,如果要从 m 个特征中,选出 n 个特征,使得一个线性回归的拟合效果最好。除了用 $\binom{m}{n}$ 的遍历方法之外,就可以用遗传算法来减少运算量。

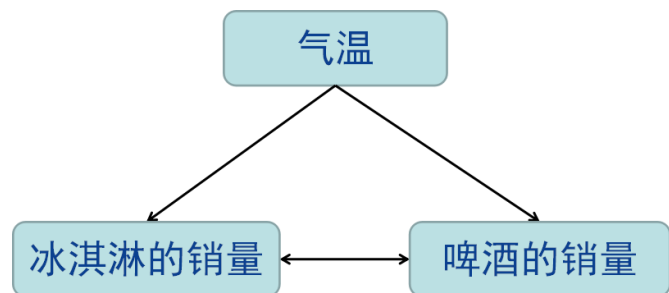
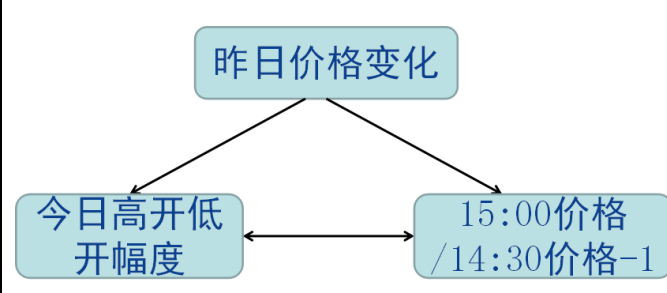
6. TMLE

传统的机器学习模型往往是考虑相关性,但是不考虑因果性。相关性单纯指出 A 和 B 是有联系的,而因果性会指出是由于 A 导致了 B 还是由于 B 导致了 A。

更复杂的因果性可以从下图看到,气温升高导致了冰淇淋的销量和啤酒的销量的增加,两者是因果性的关系。冰淇淋的销量和啤酒的销量的增加虽然有强相关,但是两者都是受气温驱动,两者没有因果联系。

同样的,在大盘择时策略中,我们也可以提出下列问题,昨日价格变化这个因子是否部分决定了今日高开低开幅度和 15:00 价格/14:30 价格-1? 今日高开低开幅度和 15:00 价格/14:30 价格-1 是不是只有相关性,没有因果性?

因此我们需要用全新的方法去解决因果性的问题。

<p>图 12：气温、冰淇淋和啤酒</p>  <p>资料来源：安信证券研究中心</p>	<p>图 13：大盘择时策略的因子重要性归因</p>  <p>资料来源：安信证券研究中心</p>
---	--

TMLE(Target Maximum Likelihood Estimation)是一种非参数估计的方法

。它能够非常健壮的估计信心区间，以及统计显著性估计。在 TMLE 中，我们运用控制变量法，称被控制不变的变量为 W ，控制变化的变量为 A 。那么我们有：

$$L(O) = P(Y|A, W)P(A|W)P(W)$$

我们定义

$$Q(A, W) \equiv E(Y|A, W)$$

$$g(A, W) \equiv P(A|W)$$

$Q(A, W)$ 是可以直接从数据当中估计的。

TMLE 可以表示成为

$$\varphi_n^{\text{TMLE}} = \frac{1}{n} \sum_{i=1}^n Q_n^*(A = 1, W_i) - Q_n^*(A = 0, W_i)$$

$Q_n^*(A, W_i)$ 是一个分布估计，所以相减之后 φ_n^{TMLE} 也是一个分布估计。上式大致可以理解成为，在固定除了变量 A 以外的变量的情况下， A 的变化会对 Y 有多大的影响（影响用差值表示）

具体地，TMLE 可以由下列方法估计

(1) 估计 $E_0(Y|A, W) \equiv \bar{Q}_0(A, W)$ ，例如可以使用 Super Learner 建立模型

(2) 对每一个 Y ，用 (1) 所述的模型，产生两个对应的预测值。也即是，对每一个 i ，有 $\bar{Q}_n^0(A_i = 1, W_i)$

$$\bar{Q}_n^0(A_i = 0, W_i)$$

(3) 估计调整变量在基准变量下的概率

$$g_0(a|W) \equiv P_0(A = a|W)$$

(4) 为每个 Y ，计算如下

$$H_n^*(A_i|W_i) \equiv \left(\frac{I(A_i = 1)}{g_n(A_i = 1|W_i)} - \frac{I(A_i = 0)}{g_n(A_i = 0|W_i)} \right)$$

(5) 更新最初的估计 $E_0(Y|A, W)$ ，这一步通过两个小步来实现。

(5.1) 用 Y 对 $H_n^*(A|W)$ 和 $\bar{Q}_n^0(A, W)$ 做逻辑回归，得到 ϵ_n

$$Y \sim \epsilon_n H_n^*(A|W) + \frac{e^{\bar{Q}_n^0(A,W)}}{1 + e^{\bar{Q}_n^0(A,W)}}$$

所谓 ϵ_n 就是最大似然法下面对 $H_n^*(A|W)$ 系数的估计, $\frac{e^{\bar{Q}_n^0(A,W)}}{1+e^{\bar{Q}_n^0(A,W)}}$ 是 fixed effect

(5.2) 更新最开始的估计

$$\frac{e^{\bar{Q}_n^1(A,W)}}{1 + e^{\bar{Q}_n^1(A,W)}} = \frac{e^{\bar{Q}_n^0(A,W)}}{1 + e^{\bar{Q}_n^0(A,W)}} + \epsilon_n H_n^*(A|W)$$

(6) 对每个i的每种情况分别计算

$$\frac{e^{\bar{Q}_n^1(A_i=1,W_i)}}{1 + e^{\bar{Q}_n^1(A_i=1,W_i)}} = \frac{e^{\bar{Q}_n^0(A_i=1,W_i)}}{1 + e^{\bar{Q}_n^0(A_i=1,W_i)}} + \epsilon_n H_n^*(A_i = 1, W_i)$$

$$\frac{e^{\bar{Q}_n^1(A_i=0,W_i)}}{1 + e^{\bar{Q}_n^1(A_i=0,W_i)}} = \frac{e^{\bar{Q}_n^0(A_i=0,W_i)}}{1 + e^{\bar{Q}_n^0(A_i=0,W_i)}} + \epsilon_n H_n^*(A_i = 0, W_i)$$

(7) 计算 TMLE

$$\hat{\Phi}_{TMLE}(P_n) = \frac{1}{n} \sum_{i=1}^n \bar{Q}_n^1(A_i = 1, W_i) - \bar{Q}_n^1(A_i = 0, W_i)$$

在之前的标准神经网络回归大盘择时策略当中, 如果我们对“今日高低开幅度(昨日收盘价/昨日开盘价-1)”因子做 TMLE(调用 R 包: <https://github.com/chizhangucb/tmleCommunity>), 可以得到如下结果:

TMLE: -0.059

95%信心区间: [-0.0605,-0.0577]

所以可以说明这个因子是有效的。

■ 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