知乎 首发于 川流不息

为什么机器学习在投资领域不好使

石川
量化交易 话题的优秀回答者

已关注

黑猫Q形态等 96 人赞同了该文章

> *The essence of data snooping is that focusing on interesting events is quite different from trying to figure out which events are interesting.*
>
> *译：关注有趣的事件与弄清楚哪些事件是有趣的是两码事，这就是数据迁就的本质。*

1 题记

最近，一条新闻引爆了投资圈：世界上最大的投资管理公司贝莱德（BlackRock）宣布将使用机器（确切的说是人工智能 artificial intelligence 或机器学习算法 machine learning algorithm）来取代一些基金经理进行选股。近年来，随着其在人脸识别，信用反欺诈乃至国际象棋和围棋领域的应用和杰出表现，人工智能被越来越多的人所熟悉。很多人开始看好在不久的将来机器学习算法在二级市场投资上将会比人取得更加优异的成绩。而贝莱德的这一宣布无疑将人工智能又一次推上了风口浪尖。这其中最根本的观点是：

机器学习通过可以使用复杂的各种非线性算法（比如神经网络、决策树、遗传算法）来从大量的历史交易数据中挖掘出人类无法看到的投资模式。根据这些模式来选股就可以取得丰厚收益。

进行对照实验（scientific control 或 controlled experiments）。这意味着虽然存在大量的金融交易数据，但是无法通过设计实验来控制自变量的变化、通过重复性试验来检验提出的假设（比如说机器学习发现的某种选股模式）。**如此的数据分析得到的大多是看似显著但实际上是欺骗式的模式（尤其对样本外数据），这个现象称作数据迁就（data snooping）。**

数据迁就（data snooping）：从数据中挖掘子虚乌有的模式（finding patterns in the data that do not exist）。

数据迁就问题存在于所有的非实验性研究中，而当我们把复杂的机器学习算法用于选股时，这种问题尤甚。这是因为复杂的非线性算法中包含大量的参数，通过这些参数的配合总能发现一些人类无法理解的、可以获得超额收益的选股模式。如果不能正确地理解并从业务上解释这些模式，数据迁就将使复杂的机器学习算法成为从历史数据中发现**无效巧合**的高效工具，正如本文开头的引用所说的那样。

2 使用伪素数选股

来看一个和股票八竿子打不着的选股算法。传统的基金经理恐怕绞尽脑汁也想不出这么个模式，但是机器学习算法可以轻易地（但是错误地）找出它。这个算法利用了素数（质数）的一个性质，它来自费马小定理的一个变种：除了 2 之外，任何一个素数 x 满足“2 的 x-1 次方被它自身除的余数为 1”。

举个例子，13 是一个素数，2 的 13-1（即 12）次方等于 4096。用它除以 13 得到 315，余数为 1。可以证明，所有 2 以外的素数都满足这个性质。但是满足这个性质的数不一定都是素数，它们被称为伪素数（又称为卡迈克尔数）。一万以内的伪素数有七个：561，1105，1729，2465，2821，6601，以及 8911。我们利用这些伪素数来对美股进行选股：选择股票编号中包含上述伪素数的股票进行投资。按照这个规则，Ametek 公司（一个制造企业，股票编号 03**1105**10）脱颖而出。更令人称奇的是，它在过去 40 年取得了 95 倍的累计收益，远超道琼斯工业或标普 500 指数。

知乎 首发于 川流不息

毫无疑问，这是一支非凡的股票，而我们的伪素数策略取得了巨大的成功。然而， 先别急着激动。我们需要好好审视一下：伪素数和选股到底有什么关系？答案是没有关系。那么这个策略是否真正找到了有效的选股模式？答案也是否定的。

有些人会马上跳出来说**“只要管用就行，为什么有用不重要！”。这种认知是非常危险的。对于选股这种非实验性问题，由于无法通过对照实验来检验假设，那么至少从业务上明白机器学习的算法为什么有效就显得格外重要。**因此，“只要管用就行”是非常不负责任的态度。

这个例子代表了很多机器学习算法的问题：我们总可以使用复杂的非线性算法（比如神经网络）、通过过度优化参数发现回测中无敌的选股模式。在这个过程中，我们已然落入了数据迁就的陷阱。

3 认知偏差加剧数据迁就

在以下这些条件下很容易发生数据迁就问题，很显然它们都存在于二级市场投资中：

1. 存在大量的数据。
2. 很多人都在使用同样的数据进行分析。
3. 缺乏业务理论或者无法控制变量。

这其中前三条是市场的客观条件，而最后一条则植根于人们的认知错误。人类认知中总是倾向于追寻不同寻常的事件。只有当一些“不同寻常”的巧合发生时，我们才往往能关注到。瑞士心理学家荣格将人们对巧合的过度关注称为**共时性（synchronicity）**。

共时性：指“有意义的巧合”，用于解释因果律无法解释的现象，如梦境成真，想到某人某人便出现等（“说曹操、曹操到”）。荣格认为，这些表面上无因果关系的事件之间有着非因果性、有意义的联系，这些联系常取决于人的主观经验。当两者同时发生时，便称为“共时性”现象。

通俗的说，当在时间和空间上毫无联系的两件事同时发生时，人们便会认为有一种超自然的神秘力量把它们联系在一起，并认为这种巧合具备某种意义。

比如在上面的例子中，股票标码含有伪素数和股票获得了巨大的超额收益就是一个纯粹的巧合，这样的巧合被机器学习算法发现并呈现给使用者。如果使用者不试图去理解这两者到底是否真的有关系，便会由于共时性而将这种错误的巧合赋予某种意义，即机器学习发现了一个牛逼哄哄的选股模式。

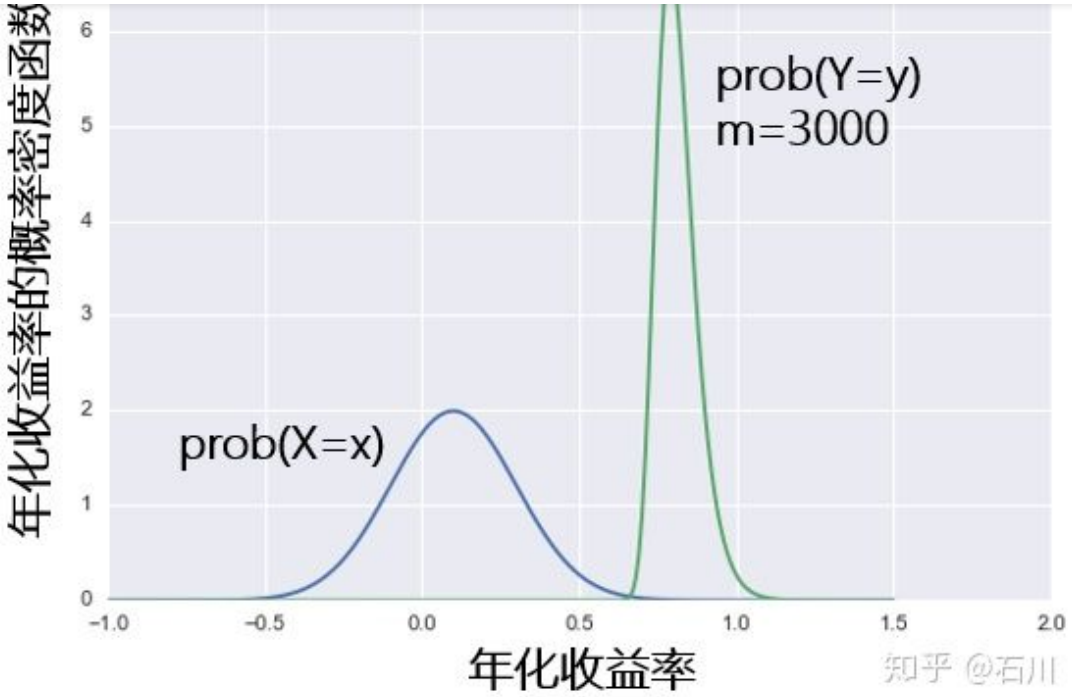
4 运气还是实力

前面说了这么多，目的当然不是为了否定人工智能和机器学习在二级市场的应用前景。但我想说，对于人工智能发现的任何模式，它有效的前提是我们能够明白无误的理解它的含义。不能以此为基础便无法分辨出好的结果到底是来自运气还是实力。

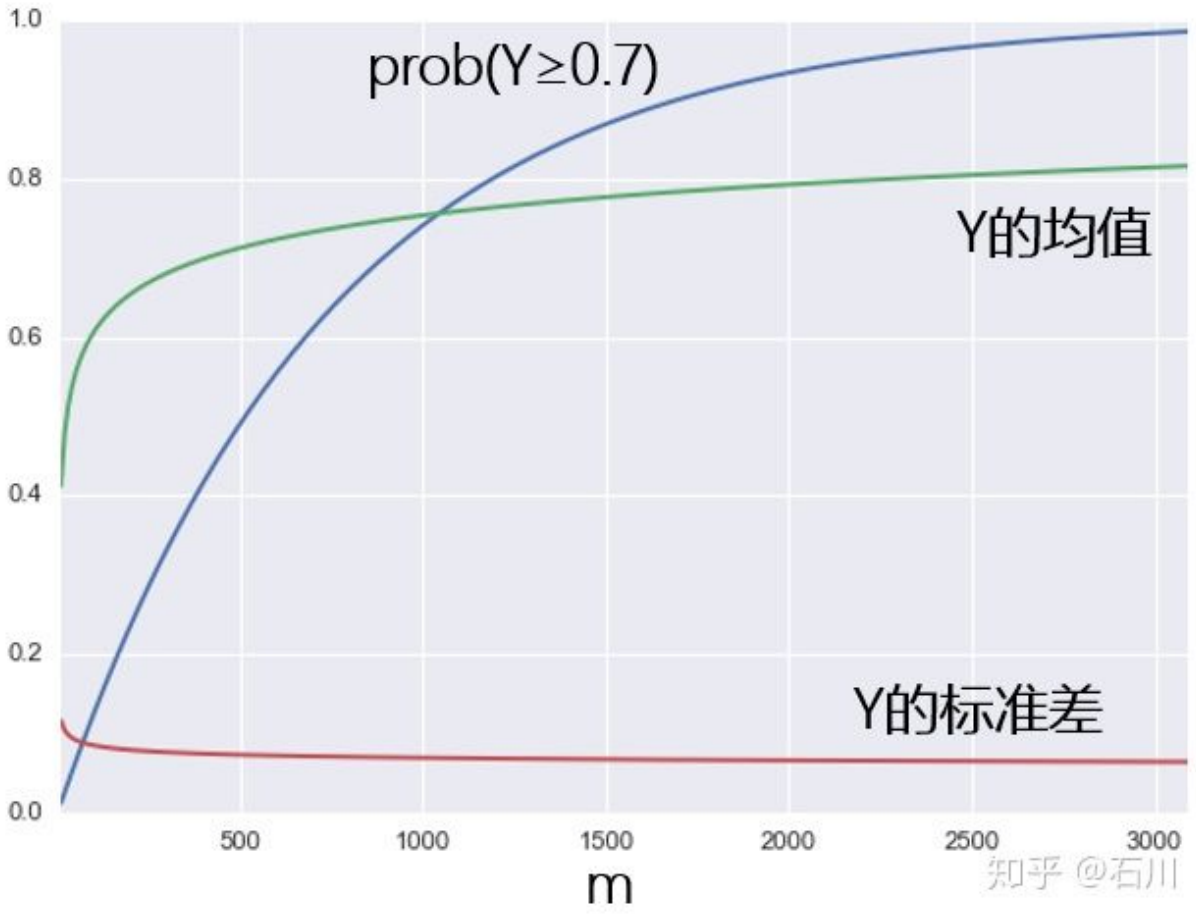
之前，我写过一篇文章《出色不如走运？》。文中使用**顺序统计量（order statistic）**解释了这样一个道理：

在众多股票中，最好的那支总会有非常优秀的收益率；在众多的策略中，最厉害的那一个总会带来令人称奇的回报率。然而，通过计算独立样本的极值（顺序统计量）分布可知，这种结果实属必然。

我们回顾一下那篇文章中的例子。假设一个股票投资策略的年化收益率 X 符合均值为 10%，标准差为 20% 的正态分布。假设市场中有 m 个不同的策略，则它们中最好的那个的收益率 Y 是 X 的函数，$Y = \max(X_1, X_2, \ldots, X_m)$。下图是当 m = 3000 时，最好的那个的收益率分布和单一策略收益率分布的比较：最优策略的收益率分布在横坐标上向右移动且变的更窄。



下图为 prob(Y≥0.7) 随策略个数 m 变化的结果。同时也给出了 Y 的均值和标准差随 m 的变化。随着 m 的增大，我们越来越确定总会有一些策略脱颖而出，年化收益率超过 70%。这种判断也同样可以被 Y 的均值和方差来证明：随着策略个数的增大，最优策略的年化收益率的均值在增加，且标准差在减小。



我们必须从业务层面弄清楚它是如何工作的。

> *As with any black box, if you don't know why it works, you won't realize when it's stopped working. Even a broken watch is right twice a day.*
>
> *译：机器学习算法犹如一个黑匣子，如果你不知道它为什么好使，你就不会知道它何时回失效。就连一块停摆的手表每天也能正确两次。*

5 人工智能前路漫漫

其实，人们使用算法来选股并不是什么新鲜事。风险多因子模型就可以算是一个算法选股的策略。当然，它之所以有效是因为它使用的因子，比如成长因子、规模因子、动量因子等，都有着清晰的业务基础。近几年，很多人使用机器学习的复杂算法，比如支持向量机，来改进多因子选股。这些非线性算法构建了很多非线性的因子。比如，如果算法告诉我们**“雄安概念板块，且对数市值 ÷ 三个月动量的 e 次方大于 π”**是一个好的模式，那我们就得好好琢磨琢磨了。

对于人工智能在二级市场投资的应用，一位具有丰富实战经验的量化投资前辈阐述过如下的观点，我对此十分认可：

> *我们可以相信它（人工智能）能够捕获到那些人类根本无法察觉到的细微模式。**但是，这些模式能够持续吗？这些模式会不会只是一些不会重复的随机噪声？**人工智能领域的专家向我们保证他们有许多防范措施用以过滤那些瞬间噪声。并且，这些工具确实在消费者营销和信用卡欺诈检测上效果显著。消费者行为和诈骗行为的模式显然都具有较长的持续期，这使得这些人工智能算法即使包含大量参数也能有效运行。然而，以我的经验来看，要对金融市场进行预测，这种防范措施是远远不够的，并且对历史数据噪声的过度拟合还会带来严重后果。……相对于可以获取的大量相互独立的消费者行为和信用交易数据，**我们能够获取的在统计学意义上相互独立的金融数据是非常有限的。**你可能会说，我们拥有大量分时金融数据可供使用。但实际上，**这些数据是序列相关的，并不是相互独立的。***

这位前辈对于人工智能何时有效给出了自己的见解：

1. 基于正确的计量经济学或理论基础，而不是随机发现的模式。
2. 所需的参数用到历史数据较少。
3. 只用到线性回归，并未使用复杂的非线性函数。
4. 概念上很简单。
5. 所有优化都必须在不含未来未知数据的移动窗口中实现，并且这种优化的效果必须不断地被未来未知的数据所证实。

策略的规则越多，模型的参数越多，就越有可能发生数据迁就。能经得起时间考验的往往是简单的模型。

作为全球最大的资产管理公司，贝莱德宣布使用人工智能代替基金经理无法令人忽视，且必然会一石激起千层浪。有机构预测，到 2025 年，全球金融机构将有 10% 的人工会被机器取代。这恐怕和越来越高昂的 α 不无关系。毕竟，从长期来看，绝大多数基金经理都跑不赢指数，那么要这些基金经理还有什么用呢?

引用我的合伙人高老板的话也许可以更好的理解贝莱德的这个决定:

> *超额收益越来越贵，开源不行，就想办法节流。最终投资市场的均衡状态是超额收益的边际成本恰好等于超额收益。这样成本高的投资基金终将不断被成本低的基金挤出市场。*

（全文完）

免责声明：文章内容不可视为投资意见。市场有风险，入市需谨慎。

编辑于 2019-07-02

机器学习 人工智能 量化交易

文章被以下专栏收录

推荐阅读

金融交易中
金融 的
味着传统预
和约 与
争， 尔

知乎 首发于 川流不息

felix 发表于有意思的数... 凯风 发表于国民经略 babyquant

19 条评论 ⇌ 切换为时间排序

写下你的评论...

Zhang fusu 1 年前
在投资领域不好使，但是用来发投资领域的论文再好使不过了
8

石川 (作者) 回复 Zhang fusu 1 年前
😗😂
赞

宋泽元 回复 Zhang fusu 3 个月前
道破了天机
赞

蓝海平 1 年前
嘿嘿，用来做营销挺好的。
赞

BugCreater 1 年前
数据迁就和过拟合是一个意思不
赞

石川 (作者) 回复 BugCreater 1 年前
是。
赞

曲百万 10 个月前
机器学习在投资中目前最大的应用价值是统计模型，也有人叫预测模型或者回归模型。有两种应用方式，第一个就是通过线性回归算法来解释某个factor对于return的影响因素，当然这是建立在统计学基础上的，它的假设是return正态分布，从而构建多因子模型，暴露因子alpha；第二种是做很多feature，喂给线性或非线性模型，找出R方大的feature构建预测模型，预测未来一段时间的市场涨跌。目前第一种主要应用在股票市场的指数增强或中性策略，

知乎 首发于 川流不息

Rainey 回复 曲百万 10 个月前

第一种是否算第二种的特例呢？这两个方法是不是都不能算严格意义上的机器学习，更像是传统的统计模型？

赞

曾若辰 回复 曲百万 10 个月前

这就是统计学一直在做的 好几十年了

赞

查看全部 8 条回复

Rainey 10 个月前

但机器学习厉害不就厉害在它能描述一些无法用简单形式描述的规律吗？如果它的拟合结果能被直观理解了，也就用不上这么复杂的算法了吧？所以机器学习在投资领域究竟该怎么应用呢？

赞

曾若辰 回复 Rainey 10 个月前

主要还是取决于数据的信噪比。信噪比低的数据没有什么复杂规律可言

赞 回复 踩 举报

江海 回复 Rainey 6 个月前

看我的文章

赞

Accelerator 9 个月前

机器学习：一门量化的玄学

赞