金融工程研究

文献启示录（第 2 期）：人工智能+投资，从认识到应用

——文献推荐系列

金融工程研究 2019 年 02 月 25 日

报告摘要：

- **认识：“人工智能+投资”到底怎么看？**

  本文以较为客观、全面的视角对 AI 投资的“有所能”和“有所不能”进行了具体介绍。作者认为“数据”在人类与 AI 的较量中起着决定性作用，AI 在历史数据丰富的战术任务上能力更强，而人类则在战略制定方面更具优势。

- **应用：“机器学习+资产定价”之方法论**

  本文主要介绍了机器学习（ML）如何应用于资产定价问题并进行了一系列实验。区别于其他研究对 ML 模型的“生搬硬套”，作者在模型选择、构建、训练等方面应用有更细节、合理的把握，实现两者的“有机结合”并取得突出的效果。

- **应用：“机器学习+资产定价”之因子降维**

  本文提出用 AutoEncoder（AE）对资产定价的隐因子进行建模。经过各种模型之间的对比实验，作者发现 AE 相比 PCA 等线性模型能对资产价格有更好的解释能力和预测能力。

- **风险提示：**

  本报告内容来源于国内外相关文献，不构成投资建议。

民生证券研究院

分析师：徐玉宁

执业证号： S0100516080001

电话： 010-85127831

邮箱： xuyuning@mszq.com

研究助理：王西之

执业证号： S0100118070034

电话： 021-60876715

邮箱： wangxizhi@mszq.com

相关研究

1、《文献启示录（第 1 期）》2019.02.17

2、《人工智能系列一：机器学习量化投资实战指南》2017.10.23

目录

一、认识：“人工智能+投资”到底怎么看？

推荐文献：How to Beat the Machines Before they Beat You？

文献来源：Vineer Bhansali (2018). LongTail Alpha.

（一）推荐理由

人工智能(AI)现在早已人尽皆知，尤其人工智能与投资的结合更是大家关注的焦点。面对“AI 投资”概念的兴起，许多人士的看法大相径庭，或是过度吹捧抑或是嗤之以鼻。本文则以更为客观、全面的视角对 AI 投资的“有所能”和“有所不能”进行了介绍。

（二）核心内容

1. 为何会有“AI 投资”的热潮？

近年来，“AI 投资”热潮的兴起主要源于四个方面：1、数据层面：金融相关数据来源更广泛、体量更大、质量更高、可得性更强；2、硬件层面：良好的 AI 生态环境，大量资本投入到 AI 基础设施建设，使得运算能力、处理速度得到大幅提升；3、开发层面：开发门槛明显下降，Python、Matlab 等高级语言的学习成本更低、开发效率更高；4、研究层面：基于上述三点支持，AI 相关算法研究得以更快发展，其它领域的突破性进展促使 AI 人才不断涌入金融这片蓝海，对传统研究方法提出了挑战。

2. 过程与结果：AI 注重“结果准确性”；人类注重“逻辑合理性”

人类倾向于解释现象，用可理解的逻辑链条一步步推导出结果；AI 则不关注拟合过程中的路径选择，专注于尽可能接近优化目标。作者认为，如果 AI 相比人类更能提升结果的准确率，与其坚守所谓的“真理”、苦寻可解释的逻辑，不如尝试缺乏理论支撑但表现更好的 AI。

3. 事件驱动：“常见事件”AI 占优；“极端特殊事件”人类占优

面对常规事件，AI 在实时信息的接收、分析、决策的环节中，其时效性、准确性、一致性（不受主观情绪、偏见的影响）和并行能力均要高于人类；而当 AI 面对罕见事件（eg. 股灾、结构性变化等），人类决策会更占优势。主要原因在于，**AI 的投资决策十分依赖于历史数据，将数据中学习到的模式（Pattern）应用到不断重复的投资问题。**

4. 人类该在何处施展所长？

尽管 AI 能力出众，但在某些场景和条件下人类仍能发挥更大优势：1、**数据量少**：模型结果难以保证有统计学意义，同时作者认为人类的“贝叶斯大脑”能做出更合理判断；**2、低信噪比**：当数据的噪音相对信号占比更大时，模型的估计和预测容易出现较大偏误，典型情况如非确定性行情（Volatile Market）；**3、战略制定**：虽然在短期战术执行上机器要明显强于人类（eg.高频交易），但在长远规划的思考和意外事件的应对上人类会更出色（eg.长期投资）；**4、预期内市场环境变化**：历史表现出色但无法快速适应环境的 AI 会缺乏竞争力，相反人类更擅长适应市场环境和经济周期变化。例如 2018 年 2 月美股闪崩带来 VIX 指数暴涨，同时引发 XIV（做空 VIX）暴跌清盘，此前配置高股票仓位的反

向波动率策略（eg.风险平价）措手不及。然而，早前就有市场人士提出避险情绪上升和高波动行情的预期。

（三）总结

不难发现，“数据”在这场人类与 AI 的较量中起着决定性作用，**再强的 AI 脱离了数据支持也会变得不堪一击**，显然 AI 难以完全替代人类；而面对 AI 在多个领域的惊人表现（state-of-the-art），人类也无法摒弃 AI。未来的投资领域，不该是人类与 AI 对立的局面而是**强强联合的时代，AI 在战术执行层面的时效性和准确性，加上人类在战略制定层面的创造力**，无疑会取得更多突破。

二、应用：“机器学习+资产定价”之方法论

推荐文献：Empirical Asset Pricing via Machine Learning

文献来源：Gu, S., B. T. Kelly, and D. Xiu (2018). NBER

（一）推荐理由

本文主要介绍了机器学习（ML）如何应用于资产定价问题并进行了一系列实验。区别于其他研究对 ML 模型的“生搬硬套”，作者在模型选择、构建、训练等方面的应用有更细节合理的把握，实现两者的“有机结合”并取得突出的效果。

（二）核心内容

1. ML+资产定价的优劣

优势：**1、丰富的“模型库”**：对于预测问题，ML 领域发展了大量经典模型，诸如 SVM、Decision Tree 等；**2、模型拟合能力强**：ML 摆脱线性形式的约束，从变量的高维信息和交互信息中发现复杂规律；同时也有 Penalty、Dropout 等 Trick 来防止模型复杂程度过高带来的过拟合问题；**3、有效处理高维数据**：降维方法能有效缩减高相关、冗余的变量。

劣势：**ML 无法为我们解释经济运行的逻辑和机制。**

2. 细节决定成败

在模型选择、构建、训练等细节处理方面，作者进行较为深入的探讨：**1、Robust 的目标函数**：例如引入加权使模型倾向于更重要的样本、采用 Huber Loss 降低对 Outlier 样本的惩罚力度以取得更稳健预测；**2、惩罚项防止过拟合**：加入惩罚项防止模型对样本内噪音变量的过度拟合，从而在样本外有更好预测，例如包含 Elastic Net（ENet）的 Lasso 等；**3、降维减少冗余变量**：使用 PCR 等降维方法减少高相关变量；**4、增强变量间交互作用**：采用 Gradient Boosting Regression Trees（GBRT）、Random Forests（RF）等模型，树结构能达到对多个变量多次最优分层的效果，Boosting 能使多个弱预测结合为强预测；**5、模型融合**：针对复杂程度更高、拟合能力更强的 Neural Network（NN），设定不同 Random Seed 独立训练多个模型，再用 Averaging 的方法组合模型的预测结果以减小预测的 Variance；**6、滚动训练**：用近期数据或历史累计数据做滚动训练。

3. ML 预测效果明显优于传统模型

作者主要用历史面板数据训练不同模型，并比较模型对未来收益预测能力强弱。数据来源是 CRSP 上 1957 年 3 月至 2016 年 12 月美国市场个股和宏观经济的月度数据，涉及 900 多个变量包括：94 个基础公司特征、74 个行业哑变量、8 个宏观经济变量、公司层面变量和宏观层面变量之间的两两交互项。基础设定包括：1、初始目标函数均为最小化模型预测收益与实际收益之间的 MSE；2、以样本外的 R 方作为模型预测效果的评判标准；3、用 Impurities 减少情况（树结构模型）或 Marginal Effect 平均大小（回归和 NN）来刻画变量重要程度。

从多组对比实验中发现：

1、**Robust 目标函数相比一般 MSE 能提升预测表现；**

2、**对数量庞大的变量集采取降维能在一定程度上改善模型预测；**

3、**NN、GBRT、RF 的预测效果明显更优，但模型层数不宜太深；**

4、**模型大多对大市值股票预测效果更优（小市值股票“信噪比”较低）；**

5、**技术类变量（如趋势、流动性等）对模型预测的贡献程度更高；**

6、**基于 NN 预测结果构建的投资组合在样本外取得最高的夏普比率。**

（三）总结

随着研究不断深入和因子库不断扩充，因子与因子之间、因子与资产价格之间的关系变得愈发错综复杂，而拥有强大拟合能力的 ML 模型或许是解决难题的答案。**更为重要的是：真相往往藏于细节之中，模型应用的细节与合理性会直接决定最终结果的好坏。**

三、应用：“机器学习+资产定价”之因子降维

推荐文献：Autoencoder Asset Pricing Models

文献来源：Gu, S., B. T. Kelly, and D. Xiu (2019). SSRN

（一）推荐理由

本文提出用 AutoEncoder（AE）对资产定价的隐因子进行建模。经过各种模型之间的对比实验，作者发现 AE 相比 PCA 等线性模型能对资产价格有更好的解释能力和预测能力。

（二）核心内容

1. 特征→隐因子→收益

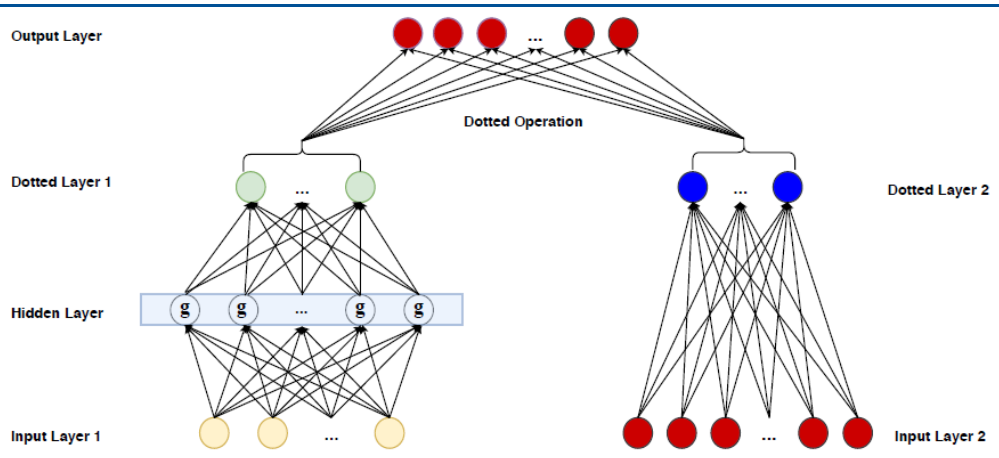
相关研究认为：有些所谓的“市场异象”只是表象，相应的 Risk Compensation 解释略显牵强并且这些特征只是公共风险因子的代理因子。另外，Fama 和 French 曾提到没有理论能定义公共风险因子的具体形式，诸如 Size、BM 等因子都是依据实证结果挑选得到的。因此，**区别于许多研究将观测到的特征数据直接用于回归估计，作者将风险因子认为是无法观测到的隐因子（Latent Factor），建立起“特征→隐因子→收益”的联系。**

2. CAE 模型

对隐因子的转换涉及机器学习的降维技术，前人多会采用 PCA 或 IPCA 等变种模型，作者则选择拥有神经网络结构的 AutoEncoder（AE）。**AE 可以看作是 PCA 的升级版（PCA 等同于单个隐藏层+线性激活函数的 AE），具备更强的非线性刻画能力。**

在 AE 基础上，作者借鉴 IPCA 的思路构建了 Conditional Autoencoder（CAE），使得 Factor loading 的估计更为合理。以单个隐藏层的 CAE 为例，其结构如图 1 所示。其中，右半部分是资产收益（Input Layer2）→隐因子（Dotted Layer2）；左半部分是滞后公司特征（Input Layer1）→Factor Loading（Dotted Layer1）；最顶层是 Factor Loading 和隐因子结合后→资产收益（Output Layer），即资产 i 在 t 时刻收益$r_{i,t} = \beta'_{i,t-1} f_t + u_{i,t}$。

图 1：CAE 模型结构图



资料来源：《Autoencoder Asset Pricing Models》，民生证券研究院

3. CAE 效果不凡

作者比较了不同模型对股票收益的解释能力和预测能力，包括“观测因子模型”（Fama-French 多因子模型）、“线性隐因子模型”（PCA 和 IPCA）和本文提出的“非线性隐因子模型”（各种结构的 CAE）。数据来源是 CRSP 上 1957 年至 2016 年美国市场月度数据，涉及 94 个特征变量。在模型细节设定方面，为防止过拟合而采用 L1-Penalization、Early Stopping 设定和 Ensemble 多个模型的方法；优化算法部分选择 ADAM 并加入 Batch Normalization。评价指标包含对同期收益解释能力的“总体 R 方”和对未来收益预测能力的“预测 R 方”。

从实证结果来看：

1、收益解释能力上，“隐因子模型”明显强于“观测因子模型”；

2、收益预测能力上，“非线性隐因子模型”强于“线性隐因子模型”且远强于“观测因子模型”。

（三）总结

因子降维是多因子模型构建中重要的一环，对因子逻辑的把握和降维方法的选择会直影响模型的效果。本文将机器学习中的降维技术与资产定价进行了深度结合，突破了传统模型的上限，也为其它投资问题的探索开阔了思路。

插图目录

分析师与研究助理简介

徐玉宁，金融工程分析师，理学学士，经济学硕士，2014年加入民生证券研究院。
王西之，金融工程研究助理，管理学学士，管理学硕士，2018年加入民生证券研究院。

分析师承诺

作者具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，保证报告所采用的数据均来自合规渠道，分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求客观、公正，结论不受任何第三方的授意、影响，特此声明。

民生证券研究院：

北京：北京市东城区建国门内大街28号民生金融中心A座17层； 100005
上海：上海市浦东新区世纪大道1239号世纪大都会1201A-C单元； 200122
深圳：广东省深圳市深南东路5016号京基一百大厦A座6701-01单元； 518001

免责声明