The Review of Machine Learning in Asset Pricing

Shihao Wu[1,a,*]

[1]*International School, Beijing University of Posts and Telecommunications, No.10 Xitucheng Road, Haidian District, Beijing, China*
a.1030784123@qq.com
**corresponding author*

Keywords: machine learning, asset pricing, neural network, deep learning

Abstract: With the rapid development of computer technologies, machine learning methods have been applied in addressing many practical issues in finance and shown big potential in future. In this paper, I mainly introduce from two aspects: (1) The theoretical basis of classical machine learning methodologies; (2) Application and characteristics of machine learning methods in finance. Those demonstrate the great potential of machine learning in finance, which also give us some directions in deep study. And the premise that machine learning methods can bring us considerable benefits is to constantly explore and study the reasonable combination of economic theories and modern computer technology. Hope this paper can give some inspiration to the research on the fintech industry.

1. Introduction

With the development of technology, more and more investors and financial researchers combine the traditional financial theoretical structure with machine learning technology, hoping to better predict the complex changes of financial market. We can see that from Sharpe and Lintner et al. (1964) proposed Capital Asset Pricing Model (CAPM), to Fama and French multi-factor Model. The traditional factor quantitative investment model is constantly improved and developed.

However, asset pricing models are still unsatisfactory in some specific situations, and the they may even have the opposite contrast effect under the rapidly changing external conditions[1]. Machine learning methodologies show great potential in asset pricing, which is primarily reflected in the following four aspects: (1) When non-linear variable factors need to be considered, the flexibility of machine learning methodologies enables the model to reduce the restrictions of function form. (2) Machine learning methodologies can better tackle this issue: "The Multidimensional Challenge"(Cochrane, 2011), in other words a large group of cross-sectional return predictors which need to be considered appear. (3) Portfolio sequencing and linear regression are not fit to address a tons of predictive variables. However, using machine learning methods, potent return predictability

[1] See, for example , Fama and French(2015) in a failure to explain low stock yield of some businesses. And for example, by comparing the conclusions of Zhao et al.(2016) and Li et al.(2017), the five-factor model has obvious differences in the performance effect of Chinese stock market in different periods, and the model is obviously limited. It is reasonable to think that this kind of problems are inevitable in traditional regression analysis methods.

is harvested from an existing set of variables(Gu et al., 2018). (4) Machine learning is able to erect a general framework for using big data sets and portfolios(Heaton et al., 2016), which likes a prototype of a time-varying function with adjustable performance. But we have not said machine learning methodologies are impeccable. We don't yet know if machine learning can help us build a perfect model (but it could be the best), this paper aims to help those who want to study in this direction understand its development status and prospects, promoting the practice of machine learning methodologies in the finance.

2. Methodologies of Machine Learning

From the 1950s to now, there is no uniform definition of machine learning, but there are different ways to classify it based on different conditions. And of all the methodologies, neural network models showed the greatest potential in finance. To introduce it in part, we start from feedforward network (FFN)[2] by Figure 1.

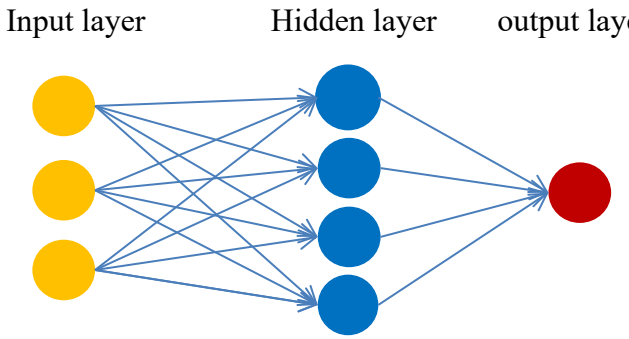

Figure 1: FFN with Single Hidden Layer.

The combinations between different layers are not single map, which is able to analysis more complex interactions between more factors. Literally, any function can be represented by a multi-layer network[3]. Additionally, backward algorithm can be used to modify the weight between layers from output layer to first hidden layer repeatedly (e.g. Back-Propagation neural network) according to errors. What is also interesting is that the combination of different networks can allow the models have better performance of prediction in finance than single[4].

To handle the curse of dimensionality is one of the biggest advantages of machine learning in finance, and there are three prime arguments of machine learning methodologies(Mohri et al., 2018): Computational, Visualization and Feature extraction. In simple terms, they are respectively data preprocessing compression, low-dimensional mapping of data visualization and sparse extraction of data. Some specific methodologies are as follows： LASSO, which is suitable for linear and nonlinear case. It is to select variables of sample data based on the penalty method. By compressing originally tiny coefficients to zero, so that the variables corresponding to these coefficients are treated as insignificant variables which are directly discarded. And Principal component analysis (PCA) aims to obtain a new set of comprehensive variables which are called main components by integrating the original single variables. It can retain most of the information of the original variable in those main components to achieve dimensionality reduction. In addition, Kernel methods etc are also widely used which will not to be introduced in detail here.

[2] FFN is one of the simplest network. More details, see Goodfellow et al. (2016).
[3] More specific, any function can be approximated by a three-layer network within any accuracy. However, it is not easy to build an efficient algorithm.
[4] See, for example, Chen et al. (2019) constructed a three-network model for estimating asset pricing.

Machine learning provides an opportunity with us to link the advantages of human thinking with those of machine algorithms and apply them to particular fields. Combining the machine learning methodologies and traditional models in finance got an excellent theoretical performance, but it is not so easy to solve financial issues in real world equally well.

3. Application in Finance

The issues of prediction and estimation in finance have a great significant with concrete and theoretical benefits. As many practices have shown, information related to finance is also scattered in other fields' data, so different data sources make models hard to be valid out-of-sample. To deal with this, we often start from two main aspects. First is the factor explosion, Green et al. (2013) extended the list of factors to astounding around 330, which still in growth. It will be a huge workload if we want to select relative predicts to build an effective collection. The traditional regression methods are not well in tackling the high-dimension in cross-sectional, while most of data has a time dimension. Secondary, the explanatory ability, many impeccable theoretical models are failure in practical operations because the significance of the data and the complexly latent interactions of non-linear factors are not clearly illustrated by available economic theories in finance (Heaton et al., 2016).

As a consequence, how can we tame the "Factor Zoo"[5]? How can we structure an adaptive model both in theory and practice? The application of machine learning may help us approach these.

3.1. "Surviving" in Factor Explosion

How to select the usefulness we needed from a tremendous set of candidate factors, or more specific, when the predictive factors' counts are tinily different with the observant count or predictive factors are highly relevant, traditional methods will lose predictive abilities(Gu et al., 2018), but predictors usually are near relations and highly correlated. So how to choose predictor variables with independent information about the cross-sectional predictive returns and how to judge the levels of significance about tradable predictors and they contribution when we consider non-tradable predictors together. Moritz and Zimmermann(2016) used portfolio sorts of tree-based under conditions to estimate the variables through two layers by the same sort-able way and the variables and the values of sorts are typically need to be calculated primitively. But this model need to control the redundant variables in advance of the samples. Feng et al.(2019) proposed an approach of regularized two-pass cross-sectional regression. They built a model about two-stage of LASSO which estimate the stochastic discount factor (SDF)[6] both of tradable and non-tradable factors after control of the latter. To exclude the tradable predictors with little contribution to assets pricing from our set of selections. But it can conclude the fake redundant factors (also few are true) with low SDF loadings, concerning they will generate exotic influences to prediction. Admittedly, it is not easy to accurately demonstrate the cross-sectional factors of stock returns predicted by models which under diverse methods and diverse controls, because methodologies vary across the most of essays. But it is vital for finance that we take advance together as a big scholarly community (Subrahmanyam, 2010).

With the completion of factors selection comes the problems of dimension control. The methodology of Feng et al.(2019) is also valid even in high-dimension issues which myriad predictors need to be carded. LASSO (built initially for linear processing effect models) works for reducing the dimension of candidate factors. To worth noting, Giglio & Xiu(2017) used PCA to control for omissive predictors of cross-sectional regressions by utilizing the high dimension of available

[5] It quoted the paper of Feng et al. (2019). They proposed a new selection-method to deal with the issues in selecting factors, which in other words, taming The Factor Zoo.

[6] To test the SDF loadings of tradable and non-tradable factor equal to zero or not(see, more specific, Cochrane (2009))

estimating assets yet it ill-suited dynamic loading and vast dimensions, and based on that, Gu et al.(2019) proposed instrumented Principal Component Analysis (IPCA), it ameliorated the empirical deficiency in selection factors and broke the limitation of static loading, while Gu et al.(2019) combine the formulations of IPCA with neural network to auto-encoder models, a covariates method to reduce the dimension. It can compress the predictive returns into a lower-dimensional collection to avoid restriction from linearity, which is a forced dimension reduction methodology. Belloni et al.(2014) argued that it is fundamental significant to reduce dimensions when we analysis abundant confounding information, and the time-dimension is integral for predictors because the timeliness of prediction. But our estimation may be invalidly in the conditions which only have time-invariant variables in portfolios because of losing efficiency(Ang, 2009).

However, to my best knowledge, the size of dimension is not the key for judging the model's prediction effect. Ossola et al.(2015) structure an estimator by econometrics, which showed well fitness and asymptotic normality when adding more sequence time-dimensions and cross-sectional sizes. Kozak et al.(2019) used Bayesian estimation and SDF in high-dimension of sparse model to achieve the robust performance, but it need enough redundancies of factors in cross-sectional and sparsity of return predictors is elusive in general. In all, the control of dimension should be analyzed flexibly according to various conditions, there is not a certain or super formula in finance for all the predictive information processing. But the appearance of machine learning might allow us closer to that for coming future.

3.2. More Explanatory Power

Appropriate selection of factors just one stage of constructing an effective model. When the predictors are input into models, they may not necessarily work well as our analysis, putting it another way, they can not have potent prediction out-of-sample always. Except for overfitting will invalidate the data, Subrahmanyam(2010) noted that the potentially relevant interaction between factors, while most of time neither predictive profitability nor expected investment can be observe directly. The undetermined proxies usually be used in tackling this condition, it is likely a gigantic simplification if we formalize all potentially relevant interactions which in practical prediction, yet many extra characteristics may have relevance of predictive profitability and planned portfolio(Kozak et al., 2019). Traditional empirical methodologies may more flexible in consideration about mutual-relationships beyond panel data, but even more complexly potential interactions will appear since increasingly large statistical set have been considered, it is likely an arbitrary determination without any basis of theories or formulas[7]. While machine learning can give an interpretable analysis of potential interaction, neural networks can imitate us to “think” and make decisions if you structure an effective “nerves” for them, its abilities about construction of function are far more than traditional statistical methodologies. As a consequence, for the impact of nonlinear interaction in data, whether or not the nonlinearity of these data can be clearly observed, machine learning can usually help traditional financial models have considerable improvement in practical prediction. It is well suited to this kind of issue as a large and complex calculating network, yet because of this, few such networks can be effectively constructed and have predictive capabilities in real-world circumstances.

A good case for this is that Moritz et al.(2016) proposed a quadruple sorting strategy by machine leaning which can considerate relatively interaction between various horizons they highlighted the abilities of models in out-of-sample estimation. Especially the non-linear interactions usually be considered in practical application of leading theoretical models because there is no justification or

[7] Kelly, B. et al. (2019) noted that”The first pre-specifies factors as sorted portfolios based on previously established knowledge about the empirical behavior of average returns.... But this is likely to be a partial understanding at best, and at worst is exactly the object of empirical interest.”

theory argued data is fully linear(Gu et al., 2019), while traditional statistical methodologies of estimations reached their limitations in practices with nonlinearities of data combinations(Refenes et al., 1994). I do not mean the linear factor model can not play a role in asset pricing, there many methodologies are better effective than non-linear under the conditional linear factors as I noted before above. Yet the nontrivial and non-linear interactions are can not be avoided in most of real issues, we should analyse from theoretical formulas and also practical conditions before choosing a methodology, whatever it is traditional or machine learning.

"Keeping pace with The Times" is another main advantage of combining machine learning methodologies with traditional models. Neural network can keep learning and modifying by itself, an example is that back propagation algorithm allows networks along the direction to lower the error constantly adjust the network connection weights and threshold of process, which through a given training mode(Supervised Learning). It is like a flexible framework which can be used in various issues and time-varying models we need to address. Yet the timeliness of statistical data is often forgotten to be considered in traditional methodologies which is also not easily to tackle this. We usually set a reserved variable for new-adding factors but it can not give us a reliable explanation based on past or forecast. Machine learning methodologies performance in the timeliness is satisfactory due to the excellent learning skills. Although its vast and complex computational processes are often difficult to comprehend totally, they are not divorced from real evidences or theories.

IPCA model of Gu et al.(2018) is suitable in dynamic factor which is given abilities to incorporate previous empirical information, but it might rely e on factors specified by people priori, which an underlying empirical error. Besides, the model of Moritz et al.(2016)can modify its weight about predictors according to practical testing which allow the predictors have more flexibility, but if there any error in estimation of sorting about selections, the models may be non-valid out-of-sample. Powerful adaptability makes machine learning require a prior setup which obtained after accurate calculation with less or even no hypothesis. Yao et al.(2000)used nerual network with back propagation algorithm to make the forecast of option price. After a appropriate period of time-to-maturity[8], the modification which according to expected value and actual output value of the weights can minimize the error of predictive estimation. But they also found neural network potent ability in predicting option price except for in low risk or low return, which means neural network is not non-limited under any conditions. Hence the basics of using machine learning methodologies in finance successfully are not only state-of-the-art technologies but also subject particular domain knowledge(Chen et al., 2019).

4. Conclusions

Machine learning, as a newborn and powerful tool being applied to the field of finance, has begun to show some advantages over traditional models by degrees, but its methodologies and models are still unknown for many people, and machine learning models are often criticized as black-box prediction. The paper aims to help more people who want to learn about machine learning methods and their applications in finance understand this new field by summarizing the applications of machine learning and contributing to the development of fintech industry.

In this paper, I introduce several representative machine learning models like neural network, LASSO and PCA etc. and briefly analyze the advantages and disadvantages of practically using these models in literatures. Additionally, theoretical models constructed from existing basis of axioms likely will not have comparable abilities to deep learning models in the performance of predication

[8] More specific, see Yao et al.(2000). They argued it is not fair for neural network if we do not give enough time for network to "learn" and expect a long-time prediction.

(Heaton et al., 2016), yet it is not easy to successfully integrate machine learning methodologies with traditional financial models and stand in empirical tests. The benefits to the financial world of machine learning methodologies will undoubtedly be enormous, but we are only now taking our first steps towards success.

References

[1] *Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. Journal of financial economics, 116(1), 1-22.*

[2] *Zhao, S. M., Yan, H. L., & Zhang, k.. (2016). Is the fama-french five-factor model better than the three-factor model? -- empirical evidence from China's a-share market. Nankai Economic Studies, (2), 41-59.(in Chinese)*

[3] *Li, Z. B., Yang, G. Y., Feng, Y. C., & Jing, L.. (2018). Empirical test of fama-french five-factor model in Chinese stock market. Journal of Financial Research, 444(6), 191-206.(in Chinese)*

[4] *Cochrane, J. H. (2011). Presidential address: Discount rates. The Journal of finance, 66(4), 1047-1108.*

[5] *Gu, S., Kelly, B., & Xiu, D. (2018). Empirical asset pricing via machine learning (No. w25398). National Bureau of Economic Research.*

[6] *Heaton, J. B., Polson, N. G., & Witte, J. H. (2016). Deep learning in finance. arXiv preprint arXiv:1602.06561.*

[7] *Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.*

[8] *Chen, L., Pelger, M., & Zhu, J. (2019). Deep learning in asset pricing. Available at SSRN 3350138.*

[9] *Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). Foundations of machine learning. MIT press.*

[10] *Green, J., Hand, J. R., & Zhang, X. F. (2013). The supraview of return predictive signals. Review of Accounting Studies, 18(3), 692-730.*

[11] *Feng, G., Giglio, S., & Xiu, D. (2019). Taming the factor zoo: A test of new factors (No. w25481). National Bureau of Economic Research.*

[12] *Moritz, B., & Zimmermann, T. (2016). Tree-based conditional portfolio sorts: The relation between past and future stock returns. Available at SSRN 2740751.*

[13] *Cochrane, J. H. (2009). Asset pricing: Revised edition. Princeton university press.*

[14] *Subrahmanyam, A. (2010). The cross-section of expected stock returns: what have we learnt from the past twenty-five years of research?. European Financial Management, 16(1), 27-42.*

[15] *Giglio, S., & Xiu, D. (2017). Inference on risk premia in the presence of omitted factors (No. w23527). National Bureau of Economic Research.*

[16] *Belloni, A., Chernozhukov, V., & Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. Journal of Economic Perspectives, 28(2), 29-50.*

[17] *Ang, A., Liu, J., & Schwarz, K. (2009). Using individual stocks or portfolios in tests of factor models.*

[18] *Ossola, E., Gagilardini, P., & Scaillet, O. (2015). Time-varying risk premium in large cross-sectional equity datasets.*

[19] *Kozak, S., Nagel, S., & Santosh, S. (2019). Shrinking the cross-section. Journal of Financial Economics.*

[20] *Gu, S., Kelly, B. T., & Xiu, D. (2019). Autoencoder asset pricing models. Available at SSRN.*

[21] *Kelly, B. T., Pruitt, S., & Su, Y. (2019). Characteristics are covariances: A unified model of risk and return. Journal of Financial Economics.*

[22] *Refenes, A. N., Zapranis, A., & Francis, G. (1994). Stock performance modeling using neural networks: a comparative study with regression models. Neural networks, 7(2), 375-388.*

[23] *Yao, J., Li, Y., & Tan, C. L. (2000). Option price forecasting using neural networks. Omega, 28(4), 455-466.*