Machine learning methods in finance

Irina Karachun[1,*], *Lyubov* Vinnichek[2] and *Andrey* Tuskov[3,4]

[1]Belarusian State University, Digital Economy Department, 220030 Minsk, Belarus
[2]Saint-Petersburg State Agrarian University, 196601 Saint-Petersburg, Russia
[3]Penza State University, Digital Economy Department, 440026 Penza, Russia
[4]K.G. Razumovsky Moscow State University of technologies and management (the First Cossack University), Moscow, Russian Federation

Abstract. This article focuses on supervised learning and reinforcement learning. These areas overlap most with econometrics, predictive modelling, and optimal control in finance. We choose to focus on how to cast machine learning into various financial modelling and decision frameworks. This work introduces the industry context for machine learning in finance, discussing the critical events that have shaped the finance industry's need for machine learning and the unique barriers to adoption. The finance industry has adopted machine learning to varying degrees of sophistication. Some key examples demonstrate the nature of machine learning and how it is used in practice. In particular, we begin to address many finance practitioner's concerns that neural networks are a "black-box" by showing how they are related to existing well-established techniques such as linear regression, logistic regression, and autoregressive time series models. Neural networks can be shown to reduce to other well-known statistical techniques and are adaptable to time series data.

1 Introduction

The development of modern information technologies entails an unprecedented growth in the volume of computing resources and large data sets. Thanks to this, machine learning methods available through open-source toolkits are gaining popularity among analysts and developers. Deep learning models have proven extremely successful in a wide range of applications, including image processing, learning gamification, neuroscience, energy conservation, and medical diagnostics.

In finance, machine learning combines several mathematical disciplines: financial econometrics, statistical computing, probabilistic and dynamic programming, and even pattern recognition. The main consumers of this technology have become finance areas focused on algorithms, such as algorithmic trading. A key challenge to understanding machine learning is the lack of well-established theories and concepts that are necessary for financial time series analysis. There are many misconceptions and limited understanding of the possibilities of this field. Effective machine learning methods remain poorly understood and often mathematically unsound.

[*] Corresponding author: karachun@bsu.by

The increasing amount of machine-readable activity data in the financial system, combined with the constant increase in computing power and storage capacity, has important implications for every aspect of financial modelling. Due to the financial crisis of 2008, the supervisory authorities of many countries began to regulate, based on data analysis, and implement stress testing programs [1].

2 Modern datasets

An increasingly important role for asset managers, traders, and decision-makers play so-called alternative data. It goes beyond the usual pricing of securities, company fundamentals, or macroeconomic indicators. The main source of such data today is social networks and messengers. Investment firms hire natural language processing (NLP) machine learning experts to work with financial news, unstructured documents, SEC 10K reports, etc. Major data providers, such as Bloomberg, Thomson Reuters, and RavenPack, provide processed data on market participants' news sentiment, adapted for systematic trading models.

It should be noted that the new alternative datasets have several properties: many of these datasets are unstructured, contain non-numeric and/or non-categorical data, such as news articles, voice recordings, or satellite images; in most cases, they are multi-dimensional (credit card transactions), and the number of variables may significantly exceed the number of observations; they may implicitly contain information about agent networks.

Methods of classical econometrics do not work with such datasets, because they are often based on linear algebra and give an error if the number of variables exceeds the number of observations. Covariance matrices cannot reflect the topological relationships that characterize networks. Machine learning methods offer numerical power and functional flexibility to identify complex patterns in a multi-dimensional space. Recent advances in machine learning make it applicable for evaluating the scientific theories validity; determining information variables (features) for explanatory and/or predictive purposes, causal inference, and visualization of large multi-dimensional, and complex data sets. [2]

3 Asset pricing modelling

Today asset pricing modelling is developing in an empirical direction. Extensive sets of company characteristics and a lot of factors are used to describe and understand differences in the asset expected return and to model the dynamics of the investment risk premium [3]. In essence, measuring the risk premium is a forecasting problem, since the risk premium is a conditional expectation of future realized excess profits. Methodologies that can reliably correlate excess returns with trading anomalies are highly valued. Machine learning provides a non-linear empirical approach to modelling real returns based on the characteristics of companies. For example, Dixon and Polson [4] consider the formulation of asset pricing models for measuring risk premiums, using 3,290 assets from the Russell 1000 index from December 1989 to January 2018. They introduced neural networks into canonical asset pricing systems. The 49-factor model generates information coefficients 1.5 times better than the usual factor model based on the least squares method.

The emergence of the fintech industry is also due to the growth of data volumes and the introduction of machine learning methods in technological business models, digital innovations in the financial sector. Of course, the central place in fintech is now occupied by cryptocurrencies and blockchain technologies, new digital consulting and trading systems, mobile payment systems, peer-to-peer lending, and crowdfunding. A critical aspect of product design and risk management required for consumer-focused business models is predicting agent behaviour. All market participants are provided with a well-defined variation, but they have unknown economic needs and limitations, and often do not behave economically rationally, which makes it difficult to calculate based on classical models. The emergence of robo-advisors, which provide financial advice or portfolio management services with minimal human intervention, has simplified many processes and expanded the range of financial capabilities of the individual. The majority of such technical solutions provide services for managing investment portfolios, but there are also managers of personal finances and pension savings (Betterment, Wealthfront, WiseBanyan, FutureAdvisor, Blooom, Motif Investing, Personal Capital). The level of their complexity and the use of machine learning systems is constantly growing.

PwC's Global Economic Crime and Fraud Survey 2020 notes that 47% of more than 5,000 respondents have been victims of economic crime in the past 2 years, and on average, companies have experienced 6 incidents during this period. According to this study, the main types of economic crimes are customer fraud 35%, cybercrime 34%, misappropriation of assets 31%, bribery and corruption 30%. The detection of economic crimes is one of the most successful applications of machine learning in the financial services industry. The constant growth of e-commerce (including due to the pandemic) is leading to new types of financial fraud and market manipulation. For example, exchanges are now exploring the possibility of using deep learning to counter spoofing. The application of machine learning algorithms to published corporate data allows us to identify patterns that indicate securities fraud, manipulations with the disclosure of certain items of financial statements that reflect the integrity of companies.

Blockchain technology, a distributed public ledger that records transactions, provides secure peer-to-peer communication between timestamped information blocks and transaction data. The first decentralized digital currency was Bitcoin, which uses the blockchain for distributed open storage of transaction information to reduce the shortcomings of the financial industry. The new data representation allows for a new form of financial econometrics with a focus on topological network structures, rather than just the covariance of historical price time series. A new area of research is the role of users, organizations and their interaction in the formation and dynamics of the risk of cryptocurrency investments, financial predictive analytics.

Every year, finance relies more and more on computational methods. At the same time, the growth of machine-readable data for monitoring, recording, and sharing information about financial system activities has important implications for the modelling approach. The success of artificial intelligence and computer learning algorithms depends on several factors that go beyond computer hardware and software. Machines can model complex, multi-dimensional data generation processes, deploy millions of model configurations, perform robust calculations, and adjust models to meet new information. At the same time, you can simultaneously consider several competing models to choose the optimal one in a particular market situation. At the same time, the introduction of machine learning gradually changes the behaviour of market participants, allowing them to reason, experiment and form their views based on data, and empirically manage decision-making processes.

4 Machine learning techniques in finance

Machine learning with a teacher is usually an algorithmic form of statistical evaluation of a model, in which the process of generating data is assumed to be unknown. Model selection and output are automated with a focus on processing large amounts of data. This can be seen as a highly efficient data compression technique to provide predictors in complex conditions where the relationships between input and output variables are nonlinear and the input space is multi-dimensional. Machine learning balances data filtering to make accurate and reliable decisions. This is fundamentally different from the maximum likelihood estimation method used in standard statistical models, which assume that the data was created by the model and usually have difficulty fitting, especially with multi-dimensional datasets. Because modern data sets are very complex, whether they are application lists or multi-dimensional financial time series, it is often impossible to make a conclusion based on a known data generation process. Even if an economic interpretation of the data generation process can be given, its exact form cannot be known at every point in time.

The machine learning paradigm for data analysis is very different from the traditional structure of statistical modelling and testing. Traditional fit metrics such as R^2, t-values, p-values, and statistical significance are being replaced by out-of-sample prediction and understanding the trade-off between bias and variation. Machine learning focuses on finding structure in large data sets, and the main tools for predictor selection are regularization and dropout.

Table 1. Statistical inference Maximum likelihood estimation and machine learning with a teacher.

Specification	**Statistical inference**	**Supervised machine learning**
Applied methods	Probabilistic	Algorithmic and probabilistic
Data	Data is generated by the model	The data creation process is assumed to be unknown
Modeling base	Information criterion	Numerical optimization
Model type	Mostly linear	Non linear
Scalability	Is limited to lower-dimensional data	Scaled for large-dimensional input data
Diagnostics	Extensive	Limited
Reliability	The tendency to over-fit	Works regardless of the sample
Objective	Causal models with high explanatory power	Prediction, often with limited explanatory power

Table 1 provides a generalized comparison of model inference based on maximum likelihood estimation and machine learning with a teacher. Although in reality, these two approaches are two opposing boundaries of a variety of data modeling methods. For example, some linear regression models, such as LASSO, ridge regression, or Elastic Net hybrids, are somewhere in the middle. They provide a combination of the explanatory power of the maximum likelihood estimation method while maintaining predictive out-of-sample efficiency for multi-dimensional datasets. Machine learning and statistical methods can be further characterized depending on whether they are parametric (OLS, polynomial regression, neural networks, and hidden Markov models) or nonparametric (kernel methods). Their comparison is presented in Table 2. It is worth noting that neural networks can be either parametric or nonparametric, depending on how they are configured.

From the point of view of the modeling paradigm, there are probabilistic and deterministic models. The first group treats the parameters as random, and the second group assumes that the parameters are defined. In probabilistic modeling, a special place is occupied by the so-called state space models. These models assume the existence of some unobservable latent process, the evolution of which sets in motion a certain observed process. The evolution of the latent process and the observed process dependence on the latent process can be represented in stochastic probabilistic terms. It puts the state space models in the domain of probabilistic modeling. A deterministic model can give probabilistic outputs. For example, logistic regression gives the probability that the answer will be positive given the input variables.

Table 2. Comparison of parametric and nonparametric models

Specification	Parametric models	Nonparametric models
Parameters	A finite set of parameters	The parameter space is infinite-dimensional
Response	Function of input variables and parameters	It has no strict restrictions
Flexibility	They can't capture complex patterns in big data	The structure is usually not defined a priori and can become more complex as the number of data increases

The main problem of machine learning, and especially deep learning, is the tendency to over-fit taking into account the number of model parameters. In frequency statistics, over-fitting is solved by introducing a penalty for the likelihood function. A common approach is to select models based on the Akaike information criteria [5], assuming that the model error is Gaussian. Machine learning methods such as the least absolute shrinkage and selection operator (LASSO) and ridge regression are more convenient for direct optimization of the loss function with a penalty. Moreover, the approach is not limited to assumptions about the distribution of modeling errors. LASSO regularization promotes sparser parametrization, while ridge regression reduces the significance of the parameters. Regularization may be the reason for the success of machine learning methods in finance. Conversely, due to its absence, neural networks lost popularity in the financial industry in the 1990s.

The natural platform for machine learning is the algorithmic trading industry. Trading decisions should be based on data, not intuition, hence it should be possible to automate this decision-making process using an algorithm, either given or learned. The advantages of algorithmic trading include recognizing complex market patterns, reducing the number of human-made errors, being able to test on historical data, and so on. Recently, as more and more information is digitized, the possibilities of algorithmic trading are increasing dramatically.

5 Conclusions

To sum up, some key elements of machine learning should be noted. Machine learning with a teacher is an algorithmic approach to statistical inference that is independent of the data generation process; evaluates a parameterized map between inputs and outputs with a functional form defined by the methodology; automates model selection using regularization and averaging methods to iterate through possible models and obtain the model with the best characteristics regardless of the sample; is well suited for big multi-dimensional data.

Due to this, machine learning is a more reliable approach than many methods of parametric financial econometrics used today. The key to implementing machine learning in finance is to be able to run machine learning alongside parametric methods, observing over time the differences and limitations of parametric modeling based on fit metrics in the sample. Statistical tests must be used to characterize the data and select the algorithm. Another advantage is the ability to easily scale up the data, but only if the data is of sufficiently high quality and adds a new source of information. At the same time, the use of machine learning requires strong skills of scientific justification. And it is not a panacea for automatic decision-making.

References

1. M. Flood, H.V. Jagadish, L. Raschid, *Big data challenges and opportunities in financial stability monitoring*. FSR, **20**, 129–142 (2016)
2. M.L. de Prado, *Beyond econometrics: A roadmap towards financial machine learning*. SSRN (2019)
3. S. Gu, B.T. Kelly, D. Xiu, *Empirical asset pricing via machine learning*. The Review of Finan. St., **33** (5), 2223–2273 (2020)
4. M.F. Dixon, N.G. Polson, *Short Communication: Deep Fundamental Factor Models*. SIAM J. Finan. Math. **11** (3), SC-26-SC-37 (2020)
5. H. Akaike, *Information Theory and an Extension of the Maximum Likelihood Principle*. In: Parzen E., Tanabe K., Kitagawa G. (eds) Selected Papers of Hirotugu Akaike. Springer Series in Statistics, 199–213 (1998)