Scientific Research Publishing

Journal of Mathematical Finance, 2019, 9, 691-697
https://www.scirp.org/journal/jmf
ISSN Online: 2162-2442
ISSN Print: 2162-2434

Asset Return Prediction via Machine Learning

Liangliang Zhang

NatWest Markets Securities, Ltd., Stamford, CT, USA
Email: liangliangzhang81@qq.com

How to cite this paper: Zhang, L.L. (2019) Asset Return Prediction via Machine Learning. *Journal of Mathematical Finance*, **9**, 691-697.
https://doi.org/10.4236/jmf.2019.94035

Received: October 6, 2019
Accepted: October 27, 2019
Published: October 30, 2019

Abstract

In this paper, we provide insights on the prediction of asset returns via novel machine learning methodologies. Machine learning clustering-enhanced classification and regression techniques to predict future asset return movements are proposed and compared. Numerical experiments show good applicability of the methodologies and backtesting unveils superior results in China A-shares markets.

Keywords

Clustering, Classification, Regression, Unsupervised Learning, Supervised Learning, Deep Neural Networks, Machine Learning, Asset Returns, Prediction, Investment Strategies, Universal Approximation Theorem

1. Introduction

Predicting asset returns is of central importance for empirical, theoretical and practical considerations. Essentially, prediction is the computation of the orthogonal projection of future asset returns, *i.e.*, asset price movements, often modeled as stochastic processes, onto the information structure that we observe today. Described in mathematical language, prediction involves the computation of conditional expected asset returns.

From an empirical point of view, [1] proposes to decompose asset returns with respect to the underlying risk factors. More description can be found in [2]. The approach is, in principle, the same as relative pricing of financial derivatives, where the underlying risk factors are taken as primitives, modeled econometrically and the derivatives prices, *i.e.*, the conditional expected discounted future cash flows, are expressed as a functional of them.

In this paper, we propose new methodologies to compute the conditional expected asset returns under the risk-decomposition framework of [1]. We use non-linear functions to model the dependency of expected asset returns on risk factors and the nonlinearity is recovered from the market data in a model-free manner. [3] runs a comparison of different machine learning methods on a monthly frequency in US equity market. [4] extends the analysis to fixed income markets. In [5], the authors recover the stochastic discount factor, project it onto the asset span and obtain the optimal weights considering the no-arbitrage constraints. In all the references, the authors use a brute-force supervised learning approach to predict the asset returns as a regression problem. In this paper, we factor in the clustering techniques to compute conditional expected asset returns, first proposed in a derivative pricing setting in [6] to evaluate the conditional expectation at each point of time, expressed as function of risk factors. The method utilizes non-supervised learning techniques, such as k-means clustering, to partition the factor space and in each of the sub-spaces, a simple functional form is used to approximate the non-linear relationship between the future asset returns and current values of underlying risk factors.

The contribution of this paper is three-folds. First, methodologically speaking, we combine unsupervised learning and supervised learning techniques to enhance the computation efficiency for regression and classification problems. Second, under our framework, machine learning techniques and classical function approximation methodologies jointly deliver high-performance methods. This helps because, under limited computational resources, we can use as many past data as possible to train the model and meantime enjoy a fast-computational speed. Third, in terms of prediction, we propose to measure the expected absolute returns and the signs of the future price movements separately to increase the forecasting accuracy. Moreover, new stock selection criteria are proposed. This methodology enables us to use a larger number of factors and past data than the artificial neural network approach and meanwhile achieving a faster computational speed. Empirical studies in China A-share markets reveal superior results.

The organization of this paper is as follows. Section 2 describes the proposed techniques. Section 3 discusses backtesting methodologies and the results and Section 4 concludes. All the theoretical justifications can be found in the Appendix.

2. The Prediction Methodology

In this section, we introduce the main methodologies of this paper. We show that both the currently used machine learning regression and classification problems can be embedded into our theoretical framework as special cases[1]. Then, we show that we can, instead of predicting asset returns in a brute force manner, enhance the prediction precision by separating the prediction of magnitude of the asset price changes with the directions.

2.1. Clustering-Based Regression

Let us first assume that the conditional expected asset returns can be expressed

[1] That is, when the number of clusters is 1, our method degenerates to the brute-force machine learning regression and classification algorithms.

as continuous functions of risk factor values. The case with discontinuous functions is analogous with mollifiers. According to the definition of a continuous function, if its argument values are sufficiently close, then the dependent variable values are also very close. If we further assume first order differentiation of the target function, it can be shown that, in a small region of the domain of the continuous function, we can approximate it well with linear functions. This observation inspires us to use a clustering-based approach to enhance the classification or regression prediction for asset returns.

Suppose that the risk factors are denoted by an *r*-dimensional vector X_t. The target function is φ. We are trying to compute $\varphi(t,h,X_t)=E_t\left[R_{t+h}\right]$. Suppose that at time *t*, the state space of the risk factor X_t is D_t. In what follows, we are seeking a partition of this state space, denoted by $\left\{U_t^k\right\}_{k=1}^K$ such that in each of the subspace U_t^k, we use a linear function $\varphi_k(t,h,X_t)=a(t,h,k)+b(t,h,k)X_t$ to approximate φ. The rigorous mathematical justifications of this approach are given in the Appendix.

The steps are as following. Given m assets, whose rate of return processes are denoted by $\left\{R_t^i\right\}_{i=1}^m$, suppose that we want to consider T periods of data. Therefore, there are $m\times T$ observations in total for the *r*-dimensional factors. Partition, using MiniBatchKMeans function in Python, the $m\times T$ observations into K clusters. In each of the cluster, use a simple neural network or just a linear regression model to fit the data via equation $E_t\left[R_{t+h}^i 1_{X_t\in U_t^k}\right]=a(t,h,i,k)+b(t,h,i,k)X_t$. Then, for each new observation X_{t+h}, we first use predict function in python to decide which cluster it belongs to, then use equation $E_t\left[R_{t+2h}^i 1_{X_t\in U_t^k}\right]=a(t,h,i,k)+b(t,h,i,k)X_{t+h}$ to compute the expected return.

2.2. Clustering-Based Classification

Taking a two-category logistic regression-based classification as an example, we know that a classification problem is essentially a regression one. Therefore, we can use the clustering-based method introduced in Section Clustering-Based Regression to run the regression and conduct the classification. Multi-category classification problems are analogous.

3. Empirical Study

3.1. The Methodology

In this empirical study, we use two sets of machine learning architectures, which will be documented below.

3.1.1. Forecasting Magnitude and Directions of Asset Price Movements via Deep Learning

Previous methods try to build regression models to forecast the future asset returns. Assume that the confidence interval and point estimate are (c_l,c_u) and

c_p. If $c_u - c_l$ is large, the point estimate is useless since it is indicated that the forecasting errors might be large, and the realized values can deviate from the point estimate c_p. We hope to propose a method to narrow the confidence interval and therefore make the point estimate more reliable. The key is to separate the estimation of the magnitude and sign of the future asset price returns. Denote by R_t the asset return at time *t*. Then, we will try to take two steps. The first step is to compute $E_t\left[\left|R_{t+h}\right|\right]$, $E_t\left[R_{t+h}^2\right]$ and therefore $VAR_t\left[\left|R_{t+h}\right|\right]$. The second step is to use a two-category classification algorithm to label 0 if $R_{t+h} \le 0$ and 1 otherwise. If the probability associated with the classification of label 0 is larger than a threshold α, then we categorize that the future return will be negative. On the other hand, if the probability of label 1 is larger than α, then we categorize that the future return will be positive. After we determine the sign of the future expected returns, we can use the result from the first step, *i.e.*, the estimates of $E_t\left[\left|R_{t+h}\right|\right]$ as the magnitude of the expected returns. If R_{t+h} is estimated to be positive and $E_t\left[\left|R_{t+h}\right|\right] - q_\alpha \sqrt{VAR_t\left[\left|R_{t+h}\right|\right]} > \theta_t$ or R_{t+h} is estimated to be negative and $E_t\left[\left|R_{t+h}\right|\right] + q_\alpha \sqrt{VAR_t\left[\left|R_{t+h}\right|\right]} < -\theta_t$, then we go long or short the asset accordingly, where q_α is an appropriate quantile and θ_t is the return deduction because of the transaction cost. The regression and classification can, of course, be done via deep learning techniques. To reduce computational resource requirement, both the regression and classification can be done by introducing the clustering method described above. We will mainly test this methodology with China A shares.

3.1.2. Clustering-Based Regression

This method is a direct application of the clustering-based regression method introduced in Section 2.1. For each asset, we forecast $E_t\left[R_{t+h}\right]$, $E_t\left[R_{t+h}^2\right]$ and $VAR_t\left[\left|R_{t+h}\right|\right]$. Decide on a percentage α and compute the information ratio $\dfrac{E_t\left[R_{t+h}\right]}{\sqrt{VAR_t\left[\left|R_{t+h}\right|\right]}}$. Rank the information ratio in the cross-section of asset universe, long the top α percent and short the bottom α percent. However, for this methodology, we try not only to predict the forward 1 period return for each asset, but the entire forward *n* period-curve as well. This means at each moment in time *t*, we will predict $\left\{E_t\left[R_{t+ih}\right]\right\}_{i=1}^n$. Then, long at bottom and sell at peak of the curve. Of course, we can consider the forecasting accuracy by looking at the confidence intervals. That is, whenever the accuracy exceeds some predetermined thresholds, we can view the forecasted values as valid. We will test this method using China A-shares.

3.1.3. Choice of Pricing Factors

Because we are forecasting short-term asset returns, we use five technical factors, namely: volatility, skewness, kurtosis of asset returns, past 1 period and T period moving-average of asset returns as our predictive features.

3.2. The Data

The China A-share market data, including the stocks traded in Shanghai and Shenzhen stock exchanges, are downloaded from Wind terminal. Time ranges from 2008-1-2 to 2019-5-23.

3.3. Backtesting

3.3.1. Strategy 1 Performance

For Strategy 1, we use a rolling window of 250 days to compute the factor values. In order to train the clustering and regression model, we use a panel data of past rolling 100 days. To compute the regression model, we use a clustering-based approach with 100 clusters in all and we choose the five clusters with top performance to trade. This strategy is long only. Transaction cost and slippage are assumed to be unilateral 0.15%. Table 1 and Figure 1 document the performance for Strategy 1.

3.3.2. Strategy 2 Performance

For Strategy 2, we use a rolling window of 250 days to compute the factor values. In order to train the regression model, we use panel data of past rolling 100 days. To compute the regression model, we use a clustering-based approach with 100 clusters in all and we choose the five clusters with top performance to trade. This strategy is long-short. Table 2 documents the performance for Strategy 2.

The NAV plot of a long-only strategy based on the methodology in Section 3.1.2 is shown in Figure 2. Transaction cost and slippage are assumed to be unilateral 0.15%.

Table 1. Performance for strategy 1.

Annual Ret	Annual Vol	Info Ratio	Hit Rate	Calmar	MDD
203.56%	36.94%	5.49	72.47%	69.47	2.93%

Table 2. Performance for strategy 2.

Annual Ret	Annual Vol	Info Ratio	Hit Rate	Calmar	MDD
283.79%	46.87%	6.05	73.21%	82.98	3.42%

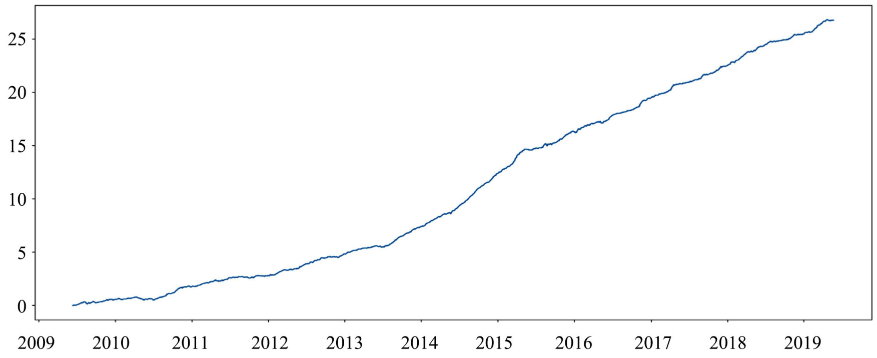

Figure 1. NAV plot for strategy 1.

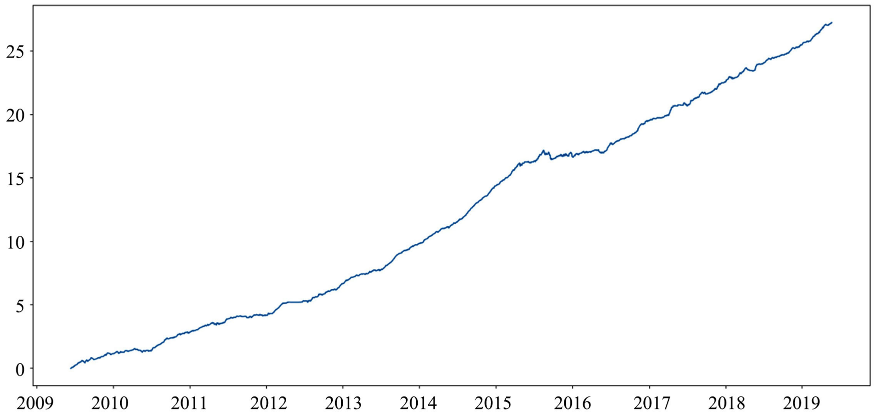

Figure 2. NAV plot-long only for china a-share market.

4. Conclusion

In this paper, we propose a clustering-based methodology to compute the expected asset returns and create trading strategies based on it. Numerical results show superior performance in China A-Share markets. Future research includes applying the proposed approach in LSTM or reinforcement learning contexts.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Ang, A. (2013) Factor Investing. https://doi.org/10.2139/ssrn.2277397

[2] Homescu, C. (2015) Better Investing through Factors, Regimes and Sensitivity Analysis. https://doi.org/10.2139/ssrn.2557236

[3] Gu, S., Kelly, B. and Xiu, D. (2018) Empirical Asset Pricing via Machine Learning. https://doi.org/10.3386/w25398

[4] Bianchi, D., Buchner, M. and Tamoni, A. (2018) Bond Risk Premia with Machine Learning. https://doi.org/10.2139/ssrn.3232721

[5] Chen, L., Pelger, M. and Zhu, J. (2019) Deep Learning in Asset Pricing. https://doi.org/10.2139/ssrn.3350138

[6] Ye, T. and Zhang, L. (2019) Derivatives Pricing via Machine Learning. *Journal of Mathematical Finance*, **9**, 561-589. https://doi.org/10.4236/jmf.2019.93029

Appendix

We will only consider the case where the functions are continuously defined on a compact domain of R^r. The extension to general functions which are defined in R^r is straightforward with mollifiers and the assumption that the distributions of asset returns are exponentially decaying at tails. We first need the following assumption.

Assumption A.1 (On Function Representation). For any asset return *R*, we have $\varphi(t, X_t) = E_t[R_{t+h}]$, *i.e.*, the conditional expected asset returns can be expressed as functions of state variables.

Lemma A.2 (On Lead-Lag Regression). Suppose that Φ is an appropriate function space. Then, we have

$$\arg\min_{\varphi \in \Phi} E\left[\left|\psi(X_T) - \varphi(t, X_t)\right|^2\right] = \arg\min_{\varphi \in \Phi} E\left[\left|\varphi^*(t, X_t) - \varphi(t, X_t)\right|^2\right]$$

where $\varphi^*(t, X_t) = E_t[\psi(X_T)]$.

Proof of Lemma A.2. The proof of this lemma follows from Theorem 8 of [6].

Theorem A.3 (On Polynomial Regression). Assume that ψ is a continuous function defined on a compact domain *U*, $\{U_t^k\}_{k=1}^K$ is a partition of domain *U* and

$$\hat{\varphi}_{k,J}(t, X_t) = \arg\min_{p_J \in P_J(U_t^k)} E\left[\left|\psi(X_T) - p_J(X_t)\right|^2\right]$$

where $P_J(U_t^k)$ is the space of all polynomials, whose coefficients depend on time *t* and *T*, with degree less or equal to *J*. Then, we have

$$\varphi(t, X_t) = \lim_{\max_{1 \le k \le K} d(U_t^k) \to 0} \sum_{k=1}^K \hat{\varphi}_{k,J}(t, X_t) 1_{X_t \in U_t^k}$$

here distance $d(U) = \sup_{x,y \in U} |x - y|$.

Proof of Theorem A.3. The proof of this theorem follows from Lemma A.1, Theorem 23 of [6] and Cauchy-Schwarz inequality.

Under Assumption A.1 and Theorem A.3, increasing the computational budget will ensure that we will obtain the true solution asymptotically.