Effect of Algorithmic Variables in LSTM's Prediction of Constituent Direction of the S&P 500: as Applied to the EMH

AP Research | Aden Goldberg

Introduction

Since the inception of applying computational techniques to the economic sector in the 1950s, the demand for accurate prediction of volatile, nonlinear systems has driven the creation of mathematical models aimed at market forecasting and optimization. Critics of this field of technical analysis (TA), a discipline of evaluating investments based on historical signals, propose the market is random and therefore is unpredictable. Conversely, practitioners of TA assert that the market is not random since researchers who aim computational resources at financial prediction have accurately predicted market trends (Marr, 2016). Although the literature is unresolved, opposition to TA often stems from applying William Sharpe's Efficient Market Hypothesis (EMH).

The EMH states markets are efficient, as they already best reflect all available information for each constituent and therefore the current prices reflect all past prices (Sharpe, 1966). Implications of this theory are evident — predictive signals are already represented through information within the stock price, and unless new financial information is revealed, there is no under or overvalued equity. This perfect representation creates unpredictability as past performance or trends have no further effect on future directions in an efficient market. The project addresses what findings TA can offer to the EMH and explains the degree to which markets are efficient, the causes of their inefficiencies, and the current impact of Machine Learning (ML) on efficiency. Additionally, this project examines algorithmic variables within the Long Short-Term Memory network (LSTM), a recurrent ML algorithm, that optimize the program for nonlinear time-series prediction.

Marr, B. (2016). A Short History of Machine Learning-Every Manager Should Read. Forbes.

Sharpe, W. F. (1966). Mutual fund performance. The Journal of business, 39(1), 119-138.

Zheng, A., & Jin, J (n.d.). Using AI to Make Predictions on Stock Market.

Fama, E. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. The Journal of Finance, 25(2), 383-417.

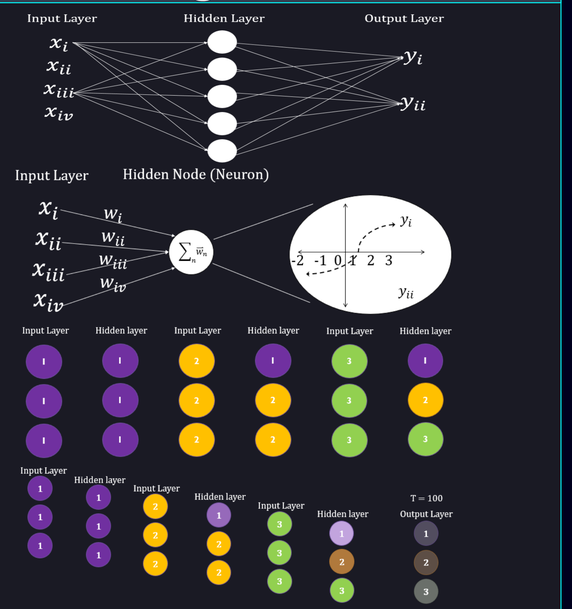
Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Kudlur, M. (2016, November). Tensorflow: a system for large-scale machine learning on heterogenous systems. In OSDI (Vol. 16, pp. 265-283)

Literature Review

There are two main gaps in the research:

- 1) Disparities between academic field and the private sector
  - Academic field lags behind clandestine private sector, so best methods are unpublished
  - Those in the private sector who benefit from fresh research already know what is published as novel, no groundbreaking work is produced.
- 2) As Zheng and Jin (n.d.) describe, the infinite scope of ML projects can discourage research
  - Fool's undertaking to test the infinite combinations

How do Algorithms Work?



Areas of Inquiry

How does the accuracy of a LSTM's prediction of the direction of constituents of the S&P 500 vary in comparison to the fiducial when algorithmic variables are changed? What can be learned about the EMH from the model's outputs?

- Offers replicable standard for fiducial LSTM future researchers could use for testing LSTM optimization.
- Contributes how to best optimize the LSTM to fit specific time-series data
- Adds to field lacking literature, offering encouragement to explore
- Reaches across oppositional perspectives using one to support the other

Model Design

- All trained on daily RIs for all constituents of S&P 500 and compiled in Python 3.7
- Dense input layer shape: 240 timesteps and one feature
- LSTM hidden layer: 25 neurons
- Dense output layer: 2 neurons, SoftMax activation function
- Compiled and optimized: RMSprop (lr = 0.001), binary crossentropy loss function
- Model was evaluated on metrics for accuracy and loss rate
- To avoid overfitting or unlearning, several callback functions were used
- Tensorboard was used for data visualization (See Abadi et al. (2016))

Results

- Fiducial:
  - 55% DA_{avg} daily (SS)
  - ~51% DA_{avg} weekly (!SS)
  - ~60% DA_{avg} monthly/yearly (!SS)
- 5-layer HL: 57% DA_{avg} daily (SS)
- 25-layer HL: 51% DA_{avg} daily (SS)
- Dropout unnecessary and redundant
- 15% increase in performance in volatile markets

Conclusions

- Markets prove semi-efficient on shorter timescales, but pure EMH is disproven
- Markets show efficient tendencies over longer timescales as markets correct anomalies
- Semi-efficiency comes from Fama (1970):
  - Stock prices lag to new information
  - Speculation affects prices
- Machines make market more efficient, less predictable
- LSTM with medium HL predicting near-term daily direction is most accurate predicting American stock data
  - Can be applied to more novel datasets