Stock price prediction based on machine learning methods

Kovrizhnykh Dmitrii kovrizhnykh.diu@phystech.edu

Optimization Class Project. Moscow Institute of Physics and Technology

Introduction

The project consists of the implementation of three methods and their work on a fixed-sum portfolio: the adaptive filtering algorithm [1], the random forest algorithm [2], and the Markowitz algorithm [3] , whose result will be considered the main one, and from which the other results will be measured. For each campaign we will take statistics in the form of a csv file. As a result, the article solves the problem of making an optimal portfolio and comparing the final value of the portfolios relative to each other as a percentage.

Markowitz algorithm

The optimization problem for a portfolio of assets with the vector of average return r by the covariance matrix V can be formulated as follows.

$$\begin{cases} \sigma_p^2 = d^T V d \to min \\ \mathrm{d}^T r = r_p \\ \mathrm{d}^T e = 1 \end{cases}$$ To these conditions in the problem of asset portfolio optimization we should add the condition of portfolio positivity (shares). If we denot:

$$A = \begin{pmatrix} \mathrm{r}^T \\ \mathrm{e}^T \end{pmatrix} V^{-1}(r,e) = \begin{pmatrix} r^T V^{-1} r & r^T v^{-1} e \\ e^T V^{-1} r & e^T v^{-1} e \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} (1)$$

Then the solution of the problem has the form:

$$d^* = V^{-1}(r,e) A^{-1} \begin{pmatrix} r_p \\ 1 \end{pmatrix} \quad (2)$$

Then the dependence of the variance of the optimized (effective) portfolio on the required return will be:

$$\sigma_p^2 = (r_p, 1) A^{-1} \begin{pmatrix} r_p \\ 1 \end{pmatrix} = \frac{a_{22} r_p^2 - a_{12} r_p + a_{11}}{a_{11} a_{22} - a_{12}^2} = \sigma_0^2 \frac{(r_p - r_0)^2}{r_0 (r_1 - r_0)} + \sigma_0^2 \quad (3)$$

where $\sigma_0^2 = \frac{1}{a_{22}}, r_0 = \frac{a_1 2}{a_2 2}$ - the minimum possible variance of the portfolio return and its corresponding average return.
$r_1 = \frac{a_1 1}{a_1 2}$ - portfolio returns, with a risk-return ratio the same as the minimum risk portfolio

Adaptive filtering algorithm:

The parameter controlling the fit and the duration of the observation window is called the forgetting coefficient (it is applied to the memory of the algorithm), 0 « < 1 and the object function looks like this

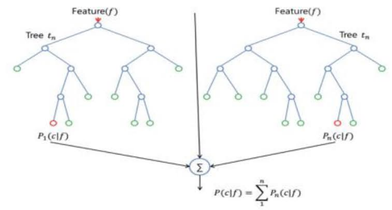
$$F(e(k)) =^d (k) = \sum_{i=0}^{k} \lambda^{k-i} e^2(k) = \sum_{i=0}^{k} \lambda^{k-i} [d(i) - x^T(i) w(k)]^2$$

This function is convex in the multidimensional space w(k), that is, $\xi^d(k)$ has only one global minimum and no local minimum. Finding this minimum is our problem. Thus, we can reach this point by setting the gradient of $\xi^d(k)$ to zero, which leads to this formula:

$$w(k) = [\sum_{i=0}^{k} \lambda^{k-1} x(i) x^T(i)]^{-1} \sum_{i=0}^{k} \lambda^{k-1} x(i) d(i)$$

Random Forest algorithm

Random forest can be represented as an ensemble of of decision trees, which are trained independently. Each of decision trees are usually made simple. In order to make the value voted for by the majority of trees is chosen for the final decision. of the trees is chosen. Each tree individually gives a low quality, but due to their large number the result is improved. The The figure below shows an example of a random forest of two trees.



There are peculiarities in the construction of trees on data. Let us describe the traditional approach to tree building. Let N be the number of examples in training. Then: 1. a subsample of the training sample of size N is selected (possibly with returns) 2. The tree is built. When searching for partitioning in the construction of each node not all attributes are considered, but only some of them (the root of number of attributes). realization - the root of the number of attributes) 3. the tree is built until complete exhaustion of a subsample, or on the basis of some other criteria Each of the trees tries to make it as simple as possible. When The class voted for by the majority of decision trees is chosen. majority of the decision trees.

CSV statistics

The data will be csv files with quotation statistics - stock price graphs from the trading date. They can be found on the git-hubs of the corresponding courses or on specialized sites. Let's attach some of them. https://www.finam.ru/profile/moex-akcii/gazprom/export/ https://www.moex.com/ru/index/MRSV/archive/?from=2022-03-15till=2022-04-14sort=TRADEDATEorder=desc. Subsequently, a transition was made to the library y.finance in order to simplify the operation of the algorithm.

Python notebook colab links

Random forest:

https://colab.research.google.com/drive/1HGKOY1XYoWXQA0C0Z3gShqE88V1t-os3scrollTo=n-RvRLEsqtDy

Markowitz:

https://colab.research.google.com/drive/1UwBeOLvcxu5AB33TibxPRDJAwNKh4LfxscrollTo=mu6jBoCD2rCm

Adaptive filtration:

https://colab.research.google.com/drive/1UwBeOLvcxu5AB33TibxPRDJAwNKh4LfxscrollTo=mu6jBoCD2rCm

Results

The random forest algorithm is depicted as a black line, adaptive filtering is orange, and purple is Markowitz.

Conclusion

As you can see in the graphs, the adaptive filtering algorithm showed a drop of 25 percent, Markowitz almost 20 precent, and random forest 14 percent. Given that the chart showed a 12 percent drop over the same time, the result can be considered more or less fair. In the end, the clear winner is the random forest algorithm, which saw a 2 percent drop below the market.

Acknowledgements

This material is based upon work supported by the X Fellowship and my mom.

References

[1] H. M. de Oliveira J. E. Wesen, V. Vermehren V. Adaptive filter design for stock market prediction using a correlation-based criterion. *Adaptive filtering, Stock market*, 21(16):330–333, 2017.

[2] Шульга Валентин Александрович. Сравнительный анализ алгоритмов машинного обучения в задачах исследования фондового рынка. *https://dspace.spbu.ru/bitstream/11701/26368/1/Diplomnaa$_r$abota.pdf*, 21(1 30 − −51, 2020.

[3] В.Ю. Попов В.А. Бабайцев, А.В. Браилов. БЫСТРЫЙ АЛГОРИТМ МЕТОДА КРИТИЧЕСКИХ ЛИНИЙ МАРКОВИЦА. Финансовый университет при правительстве РФ, *21(20), 2011.*