Machine Learning Engineer Nanodegree

Capstone Project

Guanyi Yang
September 16, 2019

I. Domain Background

Looking at various fields of quantitative investment, we find that multi-factor stock selection is the most suitable framework for transforming into machine learning. In particular, massive increases in data, access to low cost computing power, and advances in machine learning have significantly changed the financial industry. This paper will explore how several machine and deep learning algorithms can be applied in China A-share market.

II. Problem Statement

A common method for price prediction are regression-based strategies. Such strategies use regression analysis to extrapolate a trend to derive a financial instruments direction of future price movement. The problem to be solved, then, is understanding the relationship between historical data and future price prediction. The classical multi-factor model is expressed as:

$$\tilde{r} = \sum_{k=1}^{K} x_{ik} * \tilde{f_k} + \mu_i$$

X_{ik}: The factor exposure of the stock i on the factor k

$\tilde{f_k}$: Factor return of factor k

μ_i: The residual yield of stock i

The essence of classical multi-factor model is linear regression. Factor exposure x_{ik} can be observed as an independent variable of linear regression. Stock return--$\tilde{r}$, is equivalent to the dependent variable of linear regression. By fitting the historical data of x_{ik} and $\tilde{r}$, we can calculate the regression coefficient $\tilde{f_k}$, which is the estimator of factor return. Without taking risk into consideration, we can select stocks and allocate their weight only based on $\tilde{r}$ and constraints. For example, by ranking industries, selecting the top N stocks in each industry, allocate weights according to the principle of industry neutrality, and get the portfolio of the next phase.

Support vector machine (SVM) is one of the most widely used machine learning methods. Linear support vector machine can solve linear classification problems, kernel support vector machine is mainly for nonlinear classification problems. In this report, we apply support vector machine to multi-factor stock selection, focusing on the following issues:

1. Firstly, model selection. Is there any improvement in the performance of support vector machines compared with linear regression models? Does the nonlinear classifier such as polynomial kernel, Sigmoid kernel and Gaussian kernel support vector machine, have

advantages in classification performance compared with the linear classifier represented by linear support vector machine? Is there any difference between the predictive ability of support vector regression and support vector classifier?

2. Secondly, parameter optimization. SVM depends much more on parameters than generalized linear models. The SVM contains two important parameters: the penalty coefficient C and gamma (γ) value. In the context of the problem of combining multiple factors, what is the most reasonable value of parameters? Which indicators should be used to determine the optimal parameters?
3. Finally, portfolio construction. After measuring the performance of different support vector machine models, how to use the prediction results of the model to construct a strategy combination for back-test? What are the similarities and differences of stock selection effect of each model in CSI 300, CSI 500 and all A-share pools?

III. Datasets & Inputs

i. Data acquisition

CSI 300, CSI 500 and all A-share constituent stocks, excluding. The ST shares, stocks suspended in the next trading day of each cross-section period and the stocks listed within 3 months were excluded.

ii. Feature and label extraction

On the last trading day of each month, factor exposure was calculated as the original feature of the sample. Calculate the excess return of individual stocks for the next month as the sample label

iii. Eigenvalue preprocessing

1. Median Absolute Deviation (MAD). Let the exposure sequence of a certain factor on all stocks in period T be D_i. D_M is the median of the sequence. D_{M1} is the median of $|D_i - D_M|$. Then, reset all the numbers in sequence D_i that are greater than $D_M + 5D_{M1}$ to $D_M + 5D_{M1}$, and all the numbers in sequence D_i that are less than $D_M - 5D_{M1}$ to $D_M - 5D_{M1}$.

2. Filling missing data. Fill the missing value in the factor exposure sequence with the average value of the first-class industry CITIC after the new factor exposure sequence is obtained.

3. Neutral market value of the industry. Linear regression was made on the dummy variable of the industry and the log market value, and the residual was taken as the new factor exposure.

IV. Solution Statement

The SVM includes feature and label extraction, feature preprocessing, in-sample training, cross validation and out-of-sample testing. Finally, at the end of each month, the predicted value of the next rise probability of all stocks can be generated. Then the model is evaluated according to the accuracy, AUC and other indicators as well as the strategy test results. Based on the prediction

results of the model, we also constructed stock selection strategies within CSI 300, CSI 500 and all A-share constituent stocks, and comprehensively evaluated the effect of the strategy through annualized return rate, information ratio and maximum withdrawal.

V. Benchmark Model

At first, linear regression is the most common routine in the traditional multi-factor model and the most basic supervised learning method.

Expressed in machine learning, we "train" a model to reflect the linear relationship between the known "feature" x1 and "label" y. If this relationship can sustain over a period of time in the future, then we can "predict" the rise or fall of a stock in the future ($\hat{y} = w_0 + w_1 x_1$) given the EP factor x1 of the current moment. Training model based on existing feature and labels and predicting by new features constitute the two most core procedures of supervised learning.

Secondly, many times, we don't need to predict the specific rise or fall rate of the stock in the next month, we want to predict whether the stock will rise or fall. In other words, we have a classification problem, not a regression problem. The following logistic regression, although which contains the word “regression” in its name, is a machine learning method often used to solve classification problems.

On the other hand, the machine learning methods discussed above is essentially combining the original features into new features by means of linear combination, which is similar to the idea of dimensionality reduction. What is the effect of adding dimensions, and how is it achieved? Support Vector Machine (SVM) is a method to add new dimensions to view problems, which has been widely used in the field of machine learning. We would discuss and apply SVM in detail in this capstone.

VI. Evaluation Metrics

Evaluation indicators include two aspects: one is the accuracy of test set, AUG and other indicators to measure the performance of the model; Secondly, the performance of the portfolio strategy constructed in the previous step, such as annualized excess return, information ratio etc.

VII. Project Design

After data cleaning and preparing, we use SVM or SVR to train the training set. SVM selects five different kernel functions: linear kernel, 3-order polynomial kernel, 7-order polynomial kernel, Sigmoid kernel and Gaussian kernel. The SVR selects Gaussian kernel. Then, after the model training, using this model to predict the cross validation set. Parameters with the highest AUC (SVM) or IC (SVR) in the cross validation set are selected as the optimal parameters of the model. Use these optimal parameters to test out of the samples.