Naive Bayes Selecting time strategy

Ruiyang You

January 2022

1 Aims

Using machine learning and multi factors to select trading time for stock.

2 Introduction to Gaussian Naives Bayes

Naive Bayes is a method for statistical inference using conditional probability. It assumes that features are independent of each other and uses Bayesian formula to convert prior probability to posterior probability for statistical inference. The main advantages are the model assumption is relatively simple and fewer parameters are needed to be adjusted and the training efficiency is high. Secondly, the model is relatively stable and the risk of overfitting is small. As a result, the algothrithm is suitable for relatively data. Thirdly, it supports incremental training. The main disadvantage is that the assumption of mutual independence between features is harsh, which is generally not satisfied and will affect the classification result.In addition, it requires a prior probability distribution to follow normal distribution. Thus for certain data, incorrect assumptions may affect model performance.

3 Introduction to Strategy

3.1 How to choose sample data

Naive Bayes is generally used for classification problems. Therefore, in the timing strategy, we take volatility of the index for a week (5 trading days is a week) (calculated by the closing price) as the label that the model needs to learn. When the decline in five trading days exceeds 2%, it is labelled as -2. when the decline is between 0 and 2%, it is labelled as -1. When the increase is between 0 and 2%,we label 1 and it's same for increase larger than 2%. The reason for choosing five trading days as a cycle is that there are a lot of daily volatilities in the index and the current economic theory cannot explain the noise. Using the weekly rate of return can reduce these unexplainable noises to a certain extent, so that the model will not be misled by the noise. The choice of labelling in four intervals is also for the consideration of noise control. The weekly rise and fall of a simple index will have many random phenomena that we cannot explain. But generally speaking, the larger fluctuatation may due to some particular reason that we can concentrated on which may reduce the impact of noise on the learning process

3.2 Features used

Naive Bayes requires no correlation between features. So from a logical point of view, factors from four fields valuation macroeconomics momentum and volatility are selected. Among them, the evaluation features are described by the quantile of the historical pb value where the current pb value is located. The macroeconomic features include three indicators: year-on-year change in CPI, year-on-year change in PMI and year-on-year change in imports and exports. Momentum characteristics are described by Everbright RSRS. The volatility characteristics are described by the ATR indicator. At the same time, considering that predicting direction of the next week itself is a time series problem. Therefore, the model also include the data of the week in previous week which is equivalent to the autoregressive term in the time series model. To sum up, there are 7 features. Quantile of pb value, CPI year-on-year change, PMI year-on-year change, import and export year-on-year change, Everbright RSRS, ATR and return of the last week index

4 Research

4.1 Introduction

The research uses the Naive Bayes model to predict the signal of the CSI 300 Index. In the selection of test set and training set, the method of incremental training and testing is adopted. In the week T, all data before the current time point is used for training and the trained model is obtained to predict the index in week t+1. The first training week is week 70, which guarantees 70 samples for the initial training. The sample period is from 2010 to 2020.

4.1.1 Holding strategy

When the predicted range of price rise and fall in week t+1 is -2, we sell it. When the range of predicted price change in week t+1 is not -2 (ie -1, 1, 2), we buy it. The reason for this design is that the timing strategy basically has the problem of low accuracy. Only selling at -2 can reduce the transaction frequency and reduce transaction costs. At the same time, it can also avoid forecast errors and miss some opportunities.

4.1.2 Data preprocessing

The Naive Bayes algorithm does not require data scaling. In addition, the factors we selected also have a mismatch in update frequency. For example,

Result	RSRS	CPI	ATR
-2	0.0039	0.0277	145.3794
-1	0.1382	0.0245	118.1582
1	0.1502	0.0242	117.4235
2	0.2518	0.0247	140.6117

Table 1: Outputs of the model

valuation indicators are updated daily while macroeconomic indicators such as CPI are updated monthly. In order not to introduce future data, each phase of training data adopts the latest data that can be observed at the training time point.

4.1.3 Feature Selection

Naive Bayes has many assumptions about features and the features selected according to my opinion in previous part may not be able to satisfy the assumptions of Naive Bayes well. Therefore, this paper adopts both lasso regression and forward selection to perform feature impotance checking and feature selection by using all subsets of the 7 features (valuation, CPI, PMI, import and export, RSRS, ATR) proposed as feature sets to train the model separately. Finally, it is found that training with feature subsets (RSRS, CPI, ATR) is the best.

4.1.4 Model outputs

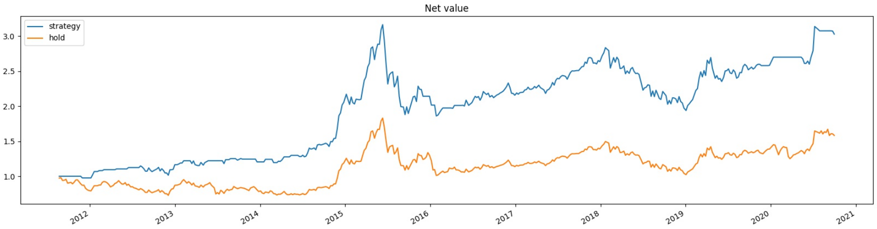
The output of the Naive Bayes model is the mean and variance of each feature (if you are confused about this, you can read the derivation in the first part). Table 1 presents the mean of the model outputs. The mean value of rsrs shows a simple upward trend in the rate of return for category -2 2, indicating that rsrs and the rate of return category have a monotonic relationship. The larger the rsrs, the higher the interval category of the predicted rate of return. Cpi and atr are a little more complicated and the relationship is not simply explained by some particular functions. when the values of cpi and atr are large, the probability of the rate of return falling into the range of -2 and 2 is high. When the values of cpi and atr are small, the probability of falling into -1 and 1 is greater. These two indicators can be thought of as indicators describing the volatility of returns. The larger these two indicators are, the greater the volatility of the yield and the easier it is to fall into the two larger volatilities groups -2 and 2.

4.2 Backtest results

Due to the need of training, the backtest of the strategy was not involved in the beginning of the sample period. The complete backtest period began on August 23, 2011, and a total of 444 samples 444 weeks of data) are involved in the evaluation.

Matrics	Strategy	Holding all the time
Annul return	13%	5%
Sharp ratio	0.73	0.34
Max draw	41.21%	44.77%
Accuracy	57%	54%

Table 2: Evaluation table



The static indicators of the Naive Bayesian timing model compared with HS300 are as follows:

It can be seen that for each indicator the strategy is ahead of holding all the time. The annualized return exceeds 8%.

It can be seen that the model effectively captured the drawdown during this period from 2011 to the middle of 2012. And curve of net value for the strategy has always been higher than that of holding. Therefore,it played a snowball effect when the stock market rose sharply in 2014-2015 where the strategy equity started to be significantly higher than simply holding. However, it also experienced a larger drawdown in the 2015 flash decline.

Due to the strict conditions for short positions in the strategy (only when the predicted weekly rate of return is ¡ -2%), a short position order will be issued. So overall, the curve of the strategy is very similar to holding. In order for readers to better study the timing effect of the strategy, Table 2 lists the backtest indicators of the strategy by year:

Since the original intention of the strategy itself is to control the index retracement. The short position conditions are relatively strict. It can be seen that the quality of each indicator is closely related to whether the current year is in the rising stage. The biggest problem with the whole strategy is that it did nothing in the fall in 2018. The reasons may be as follows: at that time, the market was in a volatile decline and the overall atr indicator was small, making the forecast fall into the two ranges of -1 and 1. Additionally, the decline in 2018 was mainly due to the trade war and the explanatory power of cpi for the stock market was weakened. In 2018, the cpi showed a moderate upward trend and the value was small, which also made the forecast results more likely to fall into the range of -1 and 1.

Year	Return	Sharp ratio	Max draw	Accuracy	Short ratio
2011	-0.0657	-1.6667	0.0242	0.5000	0.9444
2012	0.1981	1.1897	0.1160	0.6735	0.4286
2013	0.0363	0.3482	0.0562	0.5745	0.5957
2014	0.7251	2.9698	0.0377	0.6735	0.3265
2015	0.0432	0.3002	0.4060	0.5306	0.0816
2016	0.0070	0.1222	0.0769	0.6122	0.2245
2017	0.2273	2.1790	0.0511	0.6122	0.0000
2018	-0.2604	-1.1851	0.3042	0.4167	0.0000
2019	0.3163	1.4765	0.1270	0.5918	0.1020
2020	0.2662	1.5181	0.0370	0.4595	0.6757

Table 3: Annual Strategy Backtest Indicators

5 Conclusion

In general, the Naive Bayes model has a relatively strong enhancement effect on the index return and can obtain an excess annualized return of more than 8%.

In this paper, the purpose of classifying returns as four intervals instead of two is to enhance the stability of the model. Because the market trend is noisy and has many unpredictable elements, low winning rate is a common problem of timing strategies. Four classifications can make the model effect more stable, and avoid too much short position time to miss the opportunity of a big rise. Of course, there will be a corresponding price of facing more drawdowns.

At the same time, market logic is not static and some effective features will face the problem of failure. For example, the failure of the feature cpi and atr in 2018 caused the model to do nothing for the whole year. Continuously adding new features to the model according to market changes is the fundamental guarantee for long-term profitability