Academic Research Review

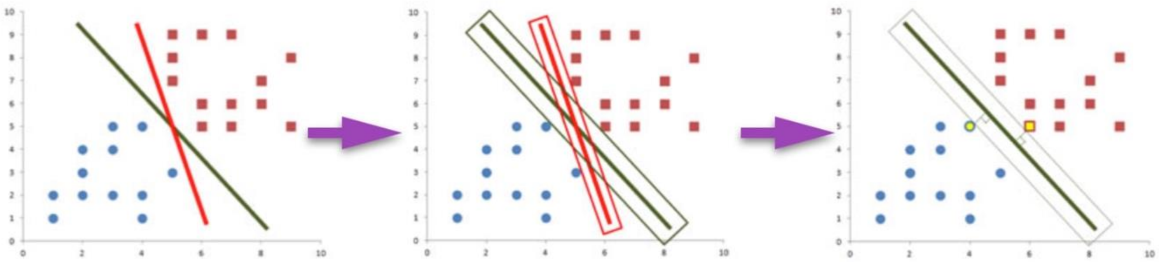
Sentiment Analysis Using SVM

EXECUTIVE SUMMARY

A key factor contributing to the market price of a stock is the expectation of the future price. With the advances of technologies and social media, such expectations are increasingly expressed by individuals on various social media platforms. Sentiment analysis have developed to evaluate the overall strength of such sentiments, using natural language processing. Furthermore, machine learning methods can be used to learn how to trade based on such quantified sentiments, which are produced by natural language processing programs. In this paper, we first introduce the Support Vector Machine model as our chosen machine learning method. We then discuss how the model can be trained to identify up or down trends. Finally, we will present the back-testing results and share possible improvements and extensions.

INTRODUCTION AND MODEL

We use SVMs as our machine learning method, to learn how to identify the pattern between consecutive sentiments and an up-or-down trend movement in the market. Let us do a quick introduction to SVM first. SVM is a supervised ML algorithm that takes in pre-specified output tags and a vector of predictors. It then "trains" the algorithm to make best prediction based on the predictors. The following graphs illustrates how a SVM with a linear kernel and 2 output categories adjusts its optimal boundary.



Sentiment Data

The sentiment data is from the PscyhSignal on Quandl platform. PscyhSignal takes in general expressions individuals add to the online conversion, on Twitter, StockTwits and chat rooms. It scores each emotion or attitude to determine the degree of sentiment present in the words. The data set contains 2 numerical sentiment indicators each day for AAPL, from 2009 September to 2015 April. Data for 2009 September to 2013 June is used for training and data for 2013 June to 2015 April is used for testing.

APPL stock prices are pulled from Yahoo Finance using R.

Our SVMs takes in 5-year daily price and sentiment data of APPL. For each day, SVM is trained with the sentiment data and stock returns data in the past 5 days, to learn how to make predictions for today's stock returns.

Tuning

The parameter *C* must be chosen to indicate trade-off between training error versus complexity of the model. A smaller C encourages 'flatness' of the model while increasing potential training error, a larger C reduces training error, but causes more fluctuation in the model which hampers its generalizability.

The *tuning* procedure is used to determine the ideal choice of *C* among a range of possible values.

METHODOLOGY

The model is run periodically, and attempts to predict an expected future price for the next period. Each period, data from a fixed interval (i.e. a rolling window) in the past is collected and used as a training set for the SVM. For our demonstration, the period is *days*, to conform with our data source. The method can however be done as frequently as per tick, depending on the availability of data and computational speed.

Our model uses a *linear* kernel for simplicity. The method can also be done for other choices of kernel, perhaps one that will give a better fit in theory or in practice.

The choice of training size is important. There has to be enough data for the SVM model to work reasonably well; on the other hand, too much data would over-emphasize past events and take long to tune. We chose a training set of 50 days, and a input size of 5 days.

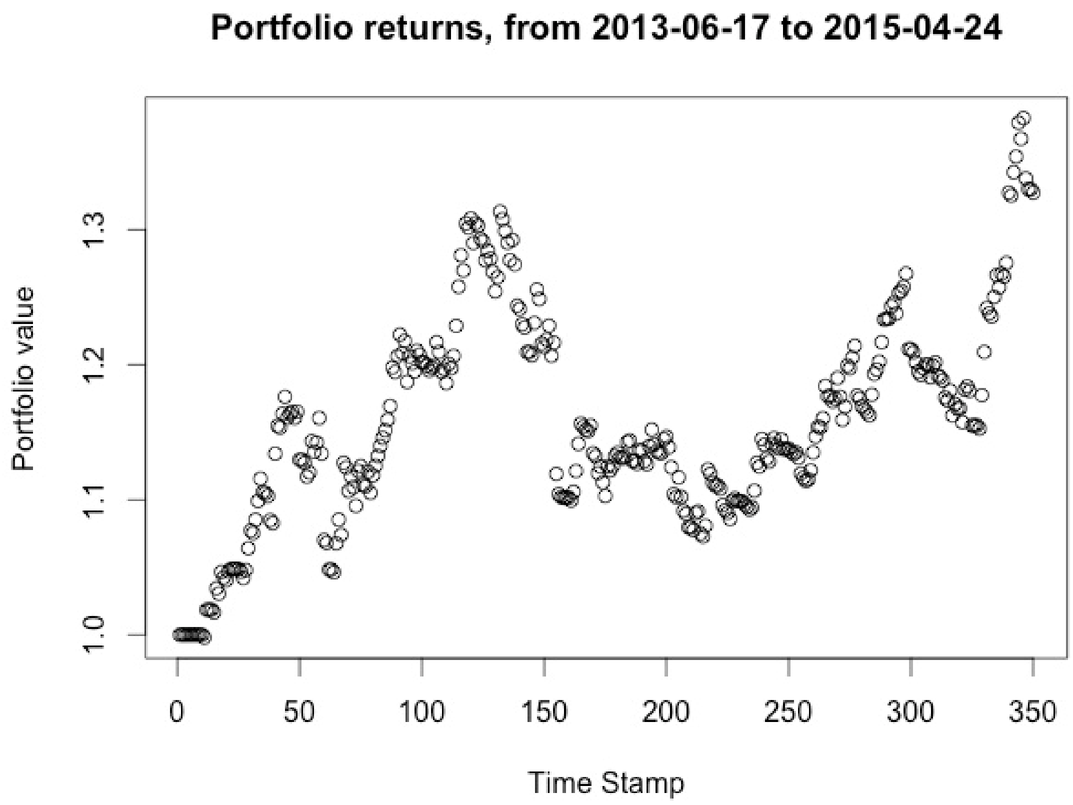
Once training and tuning is complete, our inputs are the close price and sentiment for the past 5 days, along with today's sentiment. The output is today's predicted close price. This figure is then used to change our trading position accordingly.

TRADING STRATEGY

When the SVM outputs a positive return prediction, long APPL; otherwise, short APPL. Interest rate of 0.1% and transaction cost of 0.2% are used.

BACK-TESTING RESULTS

The results of back-testing are displayed below in terms of the cumulative portfolio value over the testing period (2013 June – 2015 April), with key statistics displayed on the left side of the curve. A Sharpe Ratio of 1.14 and annualized return of 22.7% are recorded.



IMPROVEMENT AND EXTENSIONS

With the increase in market efficiencies and investments in new technologies in the financial markets, sentiment analysis has been explored a lot by hedge funds in recent years. One key factor to outperform the market is faster processing time and minimum lag in retrieving data. If we could access to some of the online real-time sentiment data and let our algorithm run on the server 24/7.

Another improvement would be o reduce the training time for SVM, especially when we are tuning parameters. SVM would get very slow when the input vector is too large. Right now we only train each day's return with past 5-day data. Possible solution to this problem is to do parallel computing in R, which we will explore further in the future publications.

RESEARCH ANALYSTS

Lim Huan Hock

Henry Jefferson Co Morco

Cai Juntao