Predicting Market Fluctuations via Machine Learning

Michael Lim,Yong Su

December 9, 2010

Abstract

Much work has been done in stock market prediction. In this project we predict a 1% swing (either direction) in the next day's closing price of $S\&P$ 500 from historical data. We implement our model with support vector machines.

We capture market sensation with three features, namely the volume, the daily momentum indicator and the Chicago Board Options Exchange Market Volatility Index.

Our model does not work well enough to be of practical use.

1 Introduction

We apply machine learning algorithms to predict the magnitude of percentage swings in Standard & Poor's (S&P) 500 index of large publicly traded companies. Historically, large swings have occurred during periods of high volatility and unusually large trading volumes [1]. If the past is any guide to the future, one might reasonably hope to predict price movements in the near future based on current and past observations of market volatility and trading volume, and hence profit from it by placing appropriate trades.

We thus have a binary classification problem: on day D, given volatility and volume data for the past d days (including day D), will the market move by 1% or more? This is an example of a supervised learning problem. We are given training examples (historical data), and we would like to predict the next day's outcome. There is a wide variety of machine learning algorithms to tackle such tasks[2], and we employ a support vector machine (SVM) in this paper.

We propose to capture volatility, trading volume, and momentum with three datasets. We will use the VIX (VIX) as a proxy for market volatility, daily closing volume (VOL), and the Commodity Research Bureau's Stock Market Momentum Indicator (MI). We also have the percentage change in the S&P index for each day in this period (PC).

1. Volume reflects the buying and selling of shares and so is an important factor to analyze market activity. Here we scale it so it is a number between 0 and 1.

2. Daily Momentum Indicator Published by the Commodity Research Bureau, the daily momentum indicator is quoted as a value in between 0 and 1 and analyzes the price strength of the $S\&P$ 500 stocks. Momentum indicators generally work because empirical studies have shown that [3] stock return is positively autocorrelated, so its precious path can be used to predict future returns. Roughly, MI measures the moving 25-day average of the proportion of stocks on the S&P 500 that are moving up.

3. Chicago Board Options Exchange Market Volatility Index (VIX)

   Also known as the fear index, VIX is a measure of the implied volatility of $S\&P$ 500 index call and put options. VIX reflects the market expectation of how volatile $S\&P$ 500 might be over the next 30 days. VIX is quoted as a percentage, for instance, if VIX is 10, it means that one expects $S\&P$ 500 index to move up or down $\frac{10\%}{\sqrt{12}}$ over the next 30 days.

2 Data

We took VIX, VOL, and MI data (values at market close) for the period 16 Jan 1992 - 5 Nov 2010 (roughly 19 years). We normalized VOL to lie in $[0, 1]$, simply by dividing VOL by its maximum value.

We begin by performing some exploratory data analysis, primarily to search for stable statistical relationships between PC and each of the three datasets.

Figure 1: Histogram of PC vs VIX

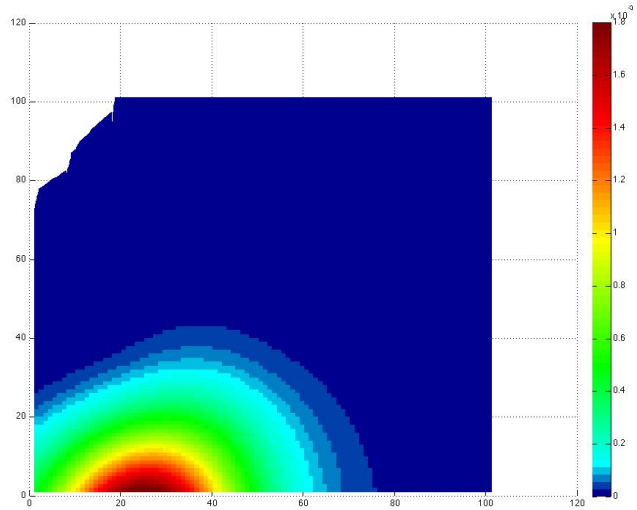
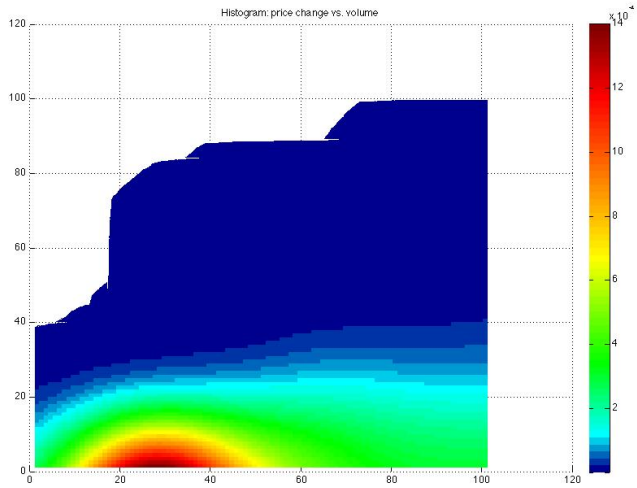


Figure 1 shows the relationship between PC and VIX, and was obtained with a moving window of size $d = 10$ days. In particular, VIX values (x-axis) for the past 10 days were

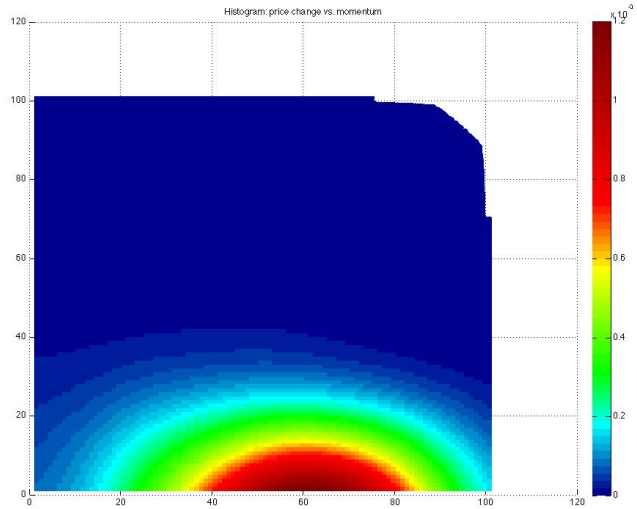
averaged, and plotted against the next day's PC (y-axis). The colors indicate the density of the points. If high volatility is a good predictor of large price swings, one would expect a high density region (red) in the upper right region of the plot. But we see no such trend.

Figure 2: Histogram of PC vs VOL



The same analysis on VOL and MI produced Figure 2 and Figure 3. Once again, keep in mind that, ideally, we would like a high density region in the upper right region.

Figure 3: Histogram of PC vs MI

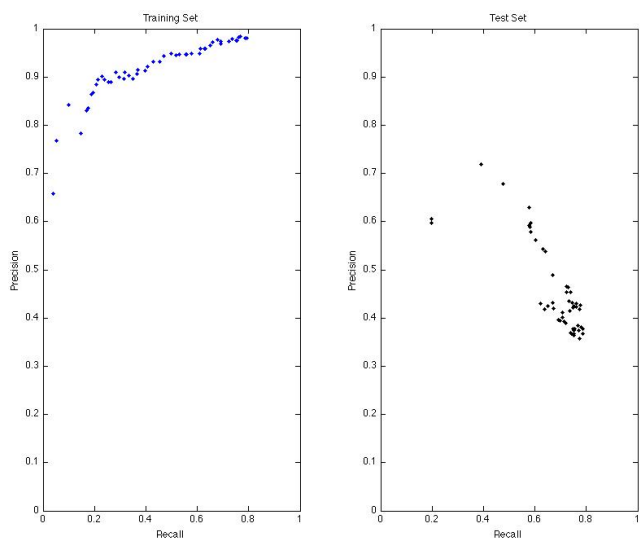


We see that neither VOL nor MI produce the desired trend.

3 Method

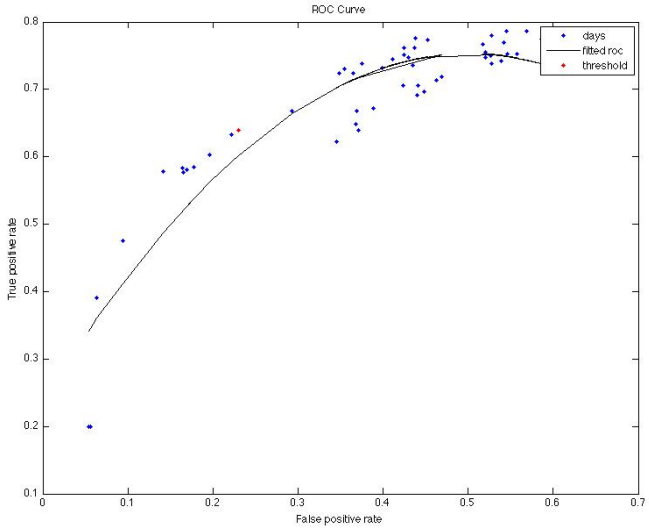
We fixed $PC = 1\%$, and performed a 70-30 split on our dataset to obtain a training set (first 3319 days), and test set (last 1422 days). With historical period $d = 10$, this gives approximately 320 training examples, and 130 test examples. We performed cross validation and the area under precision-recall curves to find optimal values for our SVM, which turned out to be $C = 8192, \sigma = 1.2207 \times 10^{-4}$. The precision-recall curve for our optimized SVM is given in Figure 4.

Figure 4: Precision-Recall



Next, we plotted the ROC curve. Each datapoint corresponds to a choice of d for the historical period.

Figure 5: ROC Curve



To decide how many days of historical data is optimal, we make use of the ROC curve. For simplicity, assume that the profit from a true positive equals the opportunity cost arising from a false positive. We then find the portion of the ROC curve that is tangent to the 45 degree line. Thus, we fit a quadratic polynomial to the curve, and take the datapoint that is closest to the required tangent. This gave us $d = 12$ days.

4 Conclusion

With $PC = 1\%$, $d = 12$, $C = 8192$, and $\sigma = 1.2207 \times 10^{-4}$, we obtained a precision of 0.5374, and recall of 0.6394 on the test set. This is disappointing performance, and more work needs to be done. Our timescale might be too coarse (days), and one might look into using intraday values instead. Alternatively, one might construct better features.

5 Bibliography

1. Saroj Kaushik and Naman Singhal, Pattern Prediction in Stock Market, AI 2009: Advances in Artificial Intelligence, 2009

2. Paul D. Yoo, Maria H. Kim and Tony Jan, Machine Learning Techniques and Use of Event Information for Stock Market Prediction: A Survey and Evaluation, 2005

3. Eugene F. Fama and Kenneth R. French, Dividend yields and expected stock returns, Journal of Financial Economics, Vol 22, Issue 1, October 1988.