Project Report for CIS 419/519
Top Stock Recommendations using Machine Learning

Abstract

In this report, we present a stock recommendation system based on patterns found from scanning the price and volume history of a list of equities then utilizing various machine learning techniques to try to predict which stocks should be bought or sold to maximize return on investment going forward. We chose to employ Neural Networks, SVM with a gaussian kernel and Online naive Bayes in order find these patterns in order to make our prediction and then choose from the top candidates to make a recommendation for trading.

1. Introduction

Many financial institutions and investors are wary of trading in the Stock Market since the financial collapse in 2008. Almost all financial institutions have turned to machine learning or other algorithmic techniques in order to reduce the risk in trading. These techniques are used both in high-frequency trading and longer term trades based on a buy and hold strategy. Our project attempts to be somewhere in the middle in order to solve a small subset of this larger financial landscape, making good short-term predictions for buying or selling stocks that the average investor might be able to utilize.

1.1. Trading model

The principle of our trading strategy is based on the idea that price movements in a stock form technical patterns that represent the sentiment of a particular equity over time and can help predict the future price movement of a stock improving overall returns in the Stock Market. In order to simplify our recommendation model we have limited our selection to the stocks in the Dow 30 and only look at closing price for the stocks in our list as the basis of our predictions. Also, in order to avoid the complication of taxes and trading costs we decided to subtract a flat 0.5% from each trade made.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1.2. Definitions

Return on investment and Portfolio Present Value (TBD)

2. Methodology

2.1. Source of data

The dataset used was downloaded from the Wharton Research Data Services (WRDS) with price and volume as the primary features and comprised the 30 stocks from the Dow.

It was necessary to normalize our data across different stocks so volume and price are comparable to each other. That way, a stock that costing $100 does not dominate a stock that costs only $10. Volume was normalized by subtracting Average Daily Volume over the slice of time within our pattern's time frame.

2.2. Trading strategy

The strategy is to buy or sell a stock from the collection of stocks we follow within the Dow 30 and then based on the pattern our system finds, executes a trade for a given period of time, or until our predicted price target is hit. Alternatively, our prediction is wrong and we exit our position with a stop loss set to 2%. This is to minimize our potential loss when we are wrong and maximize our profit when we're correct.

2.3. Stock selection

Table 1. Classification accuracies for naive Bayes vs SVM

STOCK TICKER	NAIVE	SVM	NN?
GS	95.9± 0.2	96.7± 0.2	√
DIS	83.3± 0.6	80.0± 0.6	×
IBM	74.8± 0.5	78.3± 0.6	
CAT	73.3± 0.9	69.7± 1.0	×
MSFT	61.9± 1.4	83.8± 0.7	√
JNJ	44.9± 0.6	61.5± 0.4	√

3. Results (Progress for far)

3.1. Back testing simulation

We have written software to look at different slices of time in our data of historical trades and then store those patterns for future comparisons using different machine learning algorithms. The software then steps forward one day at a time finding successful patterns to be used in making a recommendation for the current pattern of say, the last 30 days within our testing set. We're also tracking of how successful our strategy has been given our current machine learning algorithm.

Below are some example patterns that our system found given the current pattern in cyan. The other lines are past patterns that our system found that are potential matches and the dots to the right represent those outcomes for each past pattern. The most right dots are the actual outcome and the average of the past patterns.

3.2. In Summary

At this point, weve also set up another framework that allows us to easily evaluate different classifiers, different amounts of data, and different features easily. This will help us make an educated guess on which classifiers, features, and data to use in our final report by plugging this into our backtesting system.

Weve also reviewed numerous published papers to understand what existing work has been done that has been both successful and unsuccessful which will ultimately aid in narrowing the scope on our final evaluation and help us determine best practices.

3.3. Classifier Evaluation

So far weve evaluated a Boosted Decision Tree as well as SVMs with polynomial, gaussian, and cosine similarity kernels. Using our framework, weve evaluated small subsets of data and different combinations of features for our preliminary trials in order to get some idea of performance without investing too much time.

3.4. Future Work

Next, we plan to evaluate some neural network and online naive Bayes classifiers with similar baseline variables for data and features.

We are also still trying to find a better source of data where we can incorporate additional economic features into our dataset.

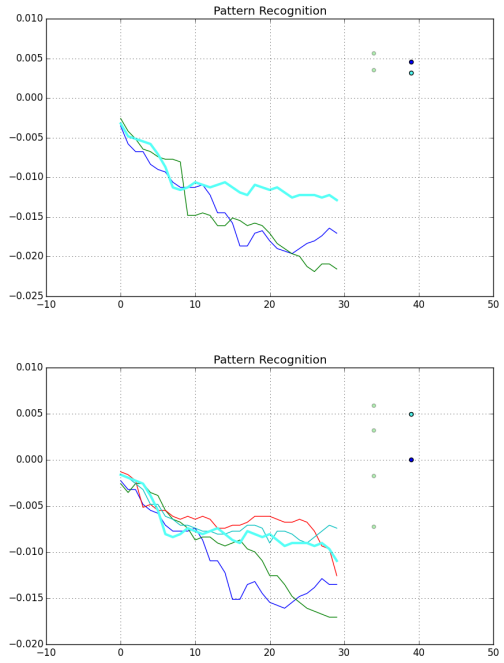

Figure 1. Examples of downward trend reversals

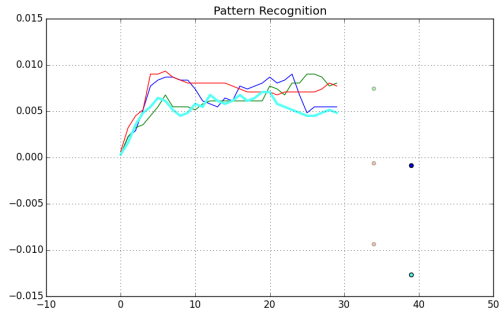

Figure 2. Examples of upward trend reversal

3.5. Figures

3.6. Citations and References

Acknowledgments

I