Feature Selection in Finance

Benjamin A. Schifman
Justin J. Siekmann
Department of Electrical and Computer Engineering
University of Arizona
Tucson, AZ 85719
bschifman@email.arizona.edu
jsiekmannemail.arizona.edu

Abstract

The abstract paragraph should be indented ½ inch (3 picas) on both the left- and right-hand margins. Use 10 point type, with a vertical spacing (leading) of 11 points. The word **Abstract** must be centered, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

1 Introduction

There exist technical analysis indicators traditionally used by analysts to evaluate and predict market and equity performance, as they "can provide a unique perspective on the strength and direction of the underlying price action" (**<–what is this quote from??**).
Feature selection is used to determine relevant indicators while identifying those that are irrelevant and redundant. Different implementations of algorithms based on these indicators could be used to predict performance of individual equities, sectors, or overall markets. They could also be used to classify and identify the correlation and interdependencies between equities, sectors, and markets. Our goal is to implement various approaches to determine efficacy of technical indicators as enablers to financial analysis. This project presented a couple challenges during implementation including developing an accurate testing method as well as handling and computing such large volumes of data.

Possibly reword and keep/move:
From this project we hope to deepen our understanding of the usage cases for applying specific machine learning algorithms as well as expanding upon our technical analysis of the stock market and which indicators play a role in successful market analysis.

2 Related Work

This is optional. If wanted, save til last.

3 Methods/Approach

The following subsections present details and explanations of the methods and functions implemented as part of this project.

3.1 Data and Technical Analysis Indicators

The Quandl platform was used to fetch 11 years of market data from Dec 31, 2006 to Dec 31, 2017 on various identified US tickers across different sectors. Tickers used in the project can be found in Table 1 categorized by sector. As the amount of fetched and calculated data is very large, pickle files

ECE523: Engineering Applications of Machine Learning and Data Analytics (Sp2018).

were used to save all data to be quickly reimported instead of refetching data through the Quandl platform. TA-Lib: Technical Analysis Library was used to calculate features on the market data for each ticker obtained using Quandl. The features incorporated into this project are found with in Table 2.

Financials	Utilities	Energy	Healthcare	Technology	Real Estate
JPM	T	XOM	JNJ	AAPL	ECL
BAC	VZ	CVX	UNH	GOOGL	DWDP
WFC	NEE	BP	PFE	MSFT	FMC
C	TMUS	GE	MRK	FB	IP
MS		SLB	ABBV	INTC	PPG
			MMM	CSCO	VMC
			AMGN	ORCL	BMS
			MDT	IBM	
				NVDA	

Table 1: Tickers

Indicators
SMA
RSI
OBV
EMA
BBAND_Upper
BBAND_Middle
BBAND_Lower
ATR
MOM

Table 2: Technical Analysis Indicators

3.2 Normalization

As each technical analysis indicator produces values applicable based on the way the indicator was calculated, normalization of the indicators makes correlations between them during feature selection more accurate and applicable. Each value in a specified time period is normalized using Equation (1). If there is extra data not consisting of a full time period, the extras are thrown out at the beginning of the data as data near the end may be more relevant and thus more desirable to keep. The start indices are computed for each ticker and return to ease future handling.

$$x_n = \frac{x - min}{max - min} \tag{1}$$

3.3 Maximal Information Coefficient (MIC)

The Maximal Information Coefficient is "a measure of two-variable dependence designed specifically for rapid exploration of many-dimensional data sets" - http://www.exploredata.net/. A benefit to MIC correlations between two variables is that it can be described regardless of linear or non-linear relationships. The MIC yields a single vale $0 \leq MIC \leq 1$ with a value closer to 1 representing that the variables are more closely correlated, and a value near 0 indicates statistically independent variables that have neither linear nor nonlinear relationships. The $minepy$ library was used in python to rank the features according to their MIC with the target variable. The MIC was calculated for each feature in each ticker, and then a final MIC value for each feature was calculated by taking the mean of the values.

3.4 Recursive Feature Elimination (RFE)

3.5 Principle Component Analysis (PCA)

4 Results

5 Conclusion

References

References follow the acknowledgments. Use unnumbered first-level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to `small` (9 point) when listing the references. **Remember that you can use a ninth page as long as it contains *only* cited references.**

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.