Predicting Market Direction with Machine Learning

Arvind Rao
tech@arvindrao.in

October 29, 2016

Abstract
Predicting market direction is a key aspect of trading and risk management. A number of tools and techniques exist that use historical time-series data to build models for prediction. Some of these techniques involve machine learning, a sub-field of artificial intelligence. Neural networks are a popular machine learning algorithm. This paper presents an approach for training a feed-forward neural network to predict a range for the closing value of a stock market index, the NIFTY 50. Ideas for feature construction and conversion of a time-series sequence to a set of independent observations are presented here.

1. Introduction

In the field of trading and risk management, a number of tools and techniques are used for predicting market direction. Many of these techniques use historical time-series data such as stock price-volume, index values and FX rates to build models. Strategies are built upon these models with the aim of building profitable positions and minimizing risk.

Leading players in the global markets, such as buy-side firms with quantitative trading strategies, often use “machine learning” based trading models built using historical market data.

Machine learning, or ML, is a sub-field of artificial intelligence that enables computers to learn to perform a task without being explicitly programmed. It is commonly used today in the fields of speech and image recognition, autonomous vehicles and recommender systems.

ML algorithms can be broadly categorized as either supervised – where a function is approximated using labelled training data; or unsupervised – where structure is determined from unlabeled data. The supervised learning techniques can further be categorized as solving regression or classification problems.

Linear regression, support vector machines, random forests and neural networks are examples of commonly used ML algorithms.

This paper describes an approach for training a neural network to predict a specific aspect of market direction – a range for the closing value of a stock market index, the NIFTY 50.

2. NIFTY 50 index

NIFTY 50, also referred to as “the NIFTY”, is the benchmark index for the National Stock Exchange of India (NSE). It is a diversified 50 stock index accounting for a number of sectors in the Indian economy.

For trading on the equity segment, NSE conducts a pre-open order entry session between 09:00 and 09:08 IST, followed by pre-open order matching. The NIFTY opening value is published by 09:10. Regular trading session begins at 09:15 and ends at 15:30.

3. Feed-forward neural network

A neural network is a machine learning algorithm that is based on a collection of neural units modelled similar to neurons in the human brain. Information flows from the input layer of neural units through the hidden layers and then to the output layer.

A "feed-forward" neural network is one in which information flows in the forward direction only and the connections between the neural units do not form a cycle.

The structure of the neural network needs to be defined by identifying:

- Number of input units
- Number of output units
- Number of hidden layers and units in each layer
- Addition of bias units
- Choice of activation function

The network is initialized with random weights and trained on the training data set for a number of epochs using an algorithm - most commonly, backpropagation. The trained network is then evaluated on a test data set.

4. Training a neural network to predict NIFTY closing value range

Using historical data, a feed-forward neural network was trained to predict a range for the NIFTY's closing value for the current trading day.

The procedure for feature construction and training the network are described in the following sections.

4.1. Selection of input data points

Choosing an optimal set of features is a key aspect of building a machine learning model.

Apart from historical data for the NIFTY index itself, certain other markets and indices with which the NIFTY has some degree of correlation needed to be considered.

The following data points were chosen as important factors that would influence the current day's (T) closing value of the index:

- NIFTY closing value and total number of shares traded for previous trade date (T-1)
- India VIX (volatility index) closing value for T-1
- United States S&P 500 index closing value for T-1
- USD-INR FX rate for T-1
- Simple moving averages (SMA) for the NIFTY as of T-1 close
- The NIFTY opening value for the current day (T) was added as an additional input data point for training the model. In the absence of an after-hours trading session on the NSE, the other data points alone would not account for news events that may have occurred post market close on T-1. The NIFTY opening value is published by 09:10, a

few minutes before regular trading begins, and would typically reflect those news events.

Historical data for these data points was collected for the period between Oct 21st 2010 and Sep 6th 2016, covering 1445 trading days.

4.2. Data cleansing

Historical data was cleansed to account for the following:

- Handle India trading holidays - Sept 13th 2016 was a trading holiday in India, but not in the United States. For predicting the NIFTY closing on Sept 14th, the model would need the previous NIFTY close on Sept 12th (the previous India trade date), but the S&P 500 previous closing value to be considered would be Sept 13th (the previous US trade date).
- Remove "special" trading sessions from historical data - Usually conducted by the NSE for testing system upgrades on some weekends and on the festival day of Diwali, these trading sessions last for approximately an hour with low volumes. They do not necessarily reflect the historical patterns that the model would require to predict future state.

4.3. Conversion of time-series into independent observations and feature construction

The data points mentioned above form a time-series. However, the standard feed-forward neural network does not work well with time-series data as it is not designed for sequence dependence. While other neural network architectures such as recurrent networks exist that can handle time-series, they are relatively more complex to train.

To build a model using the feed-forward network architecture, a new set of 68 features were constructed from the "raw" input data points with the following ideas:

- Represent the change from the previous value, rather than the point-in-time value. For example, to predict the T closing value, instead of using the NIFTY T-1 closing value as a feature, construct a new feature that represents the percentage change in this value from the T-2 closing value.
- Represent the difference between data points – such as the difference between T-1 closing and T opening value - as a percentage difference. The new feature should specify, for example, that the NIFTY opened 0.4% higher that the previous close.
- Use the above difference between data points to represent crossovers. For example, when the 5-day SMA is crosses over higher than the 20-day SMA, identify this event as day 1 of the 5-day SMA exceeding the 20-day and construct a feature with value 1 (positive 1). A consecutive day of a similar nature would result in the feature having value 2. The feature would thus represent the "number of consecutive days the 5-day SMA exceeded the 20-day SMA". Similar features would be constructed for other data points as well.

The newly constructed input feature set would then state, for example:

- On T-1, the NIFTY closed 0.4% higher than T-2
- Volume was 0.5% higher
- 5-day SMA was 0.2% higher than the 20-day SMA
- On T, NIFTY opened up 0.2% from T-1 close

- This was the 4th consecutive day of the NIFTY closing higher than the previous day
- This was the 3rd consecutive day of higher volumes
- This was the 2nd consecutive day of the 5-day SMA exceeding the 20-day SMA

The "sequence" of the time-series was therefore transformed into a feature set that represented point-in-time statements and events. Each observation would then be treated as independent from the others.

4.4. Labeling the observed output values

Treating this as a classification problem, the following 8 ranges were defined for the T closing value.

0 – Up 0 to 0.5% (from T-1)
1 – Up 0.5 to 1%
2 – Up 1 to 1.5%
3 – Up more than 1.5%
4 – Down 0 to 0.5%
5 – Down 0.5 to 1%
6 – Down 1 to 1.5%
7 – Down more than 1.5%

The number on the left is the label used by the network for classifying the prediction. For a given input observation, if the observed NIFTY closing value was 0.4% higher than the previous day, the label "0" was used as the observed output.

4.5. Splitting and normalizing training and test data

The data set, now comprising of independent observations, was shuffled randomly. 80% of the data, representing 1156 trading days, were used as the training data set for the model. The mean and standard deviation of each feature was captured and used to normalize the training data set.

The remaining 20% - 289 trading days - were the test data set. The mean and standard deviation of the training data set were also used to normalize the test data set.

4.6. Determining hyper-parameters for the neural network

A number of trials were carried out to determine the optimal "hyper-parameters" of the neural network. For example, architectures with more than one hidden layer, varying the number of hidden units and random initializations of weights.

Optimal performance on the test data set was given by a 3-layer architecture (a single hidden layer) with bias units added. The hidden layer used a Sigmoid activation function and the number of hidden units was 1.5 times the number of input units. The output layer had 8 units representing each label with a Softmax activation function that specifies probability of the prediction.

4.7. Training the neural network

The neural network was trained using the backpropagation algorithm on the training data set. After a certain number of epochs, the trained network's performance was evaluated against the test data set. When the error rate on the training data set approached convergence to near 0, and the accuracy of prediction on test data set reached a reasonable level, the training was halted and the state of the network was captured.

4.8. Evaluating performance of the neural network

Against the training data set, the neural network would have near 100% accuracy of prediction as the model exactly fits the training data. It was important to evaluate the model against a test data set that is distinct from the training data set.

The trained network was run against the test data set of 289 trading days. The output of the network was the set of probabilities of each label (representing the predicted closing value range). The probabilities totaled to 1. The label with the highest probability was chosen as the predicted output.

By comparing the the actual and predicted output labels, a multi-class F1-score was computed to evaluate the model. The weighted F1-score was 0.84547 and the unweighted score was 0.82583, indicating a reasonable level of performance in predicting the closing value range.

A better F1-score was obtained on the test data set by ignoring trading days where the certainty of the prediction is less than a certain threshold. For example, ignoring trading days where the predicted label had a probability of less than 80% resulted in a weighted F1-score of 0.92219 and an unweighted score of 0.91617, albeit at the cost of not making a prediction for 30 trading days.

5. Explaining the predictive ability of the neural network

The weights of the trained neural network do not specifically convey why the model makes a particular prediction for a given set of input features, because there is no clear link between the weights of the network and the function it has approximated. For this reason, a neural network is considered a "black box" in terms of its inability to explain its decision based on a given input. This is unlike certain other machine learning algorithms such as regression which are more "interpretable" because the importance of a parameter can be determined.

It can merely be stated that the neural network has approximated a function based on the training data set and the learning algorithm used for training. When provided with an input observation, the model uses this function to solve a problem of the following nature:

"If the NIFTY closed yesterday up 0.4%, and volumes were higher by 0.5%, and the 5-day SMA was 0.2% higher than the 20-day SMA, and this was the 4^{th} consecutive day of the NIFTY closing higher, and this was the 3^{rd} consecutive day of higher volumes, and the NIFTY opened today up 0.2% from yesterday's close...";

"Then determine the probabilities of the NIFTY closing today in each of the 8 ranges – up 0 to 0.5%, up 0.5 to 1%, …" etc.

6. Future work

There is scope for improvement in the performance of the model by tuning the hyper-parameters of the network and tweaking the input feature set. Given the large number of possible variations in these parameters, an automated approach to optimal parameter selection such as a genetic algorithm would be more appropriate than manual trials.

Amongst all the input features available to the model, the most recent data point available to it is the current day's opening value for the index. The model's prediction tends to be inaccurate if, during the course of the trading day, new information is made available to market players that can significantly alter the "pattern" of market behavior. An "online" machine learning model that is based on intra-day tick data would potentially be more accurate in making short term predictions of market movement.

7. Conclusion

Automation in decision making in the field of trading and risk management can be achieved using machine learning algorithms such as neural networks. With the vast amount of information including high-frequency tick data available in most developed markets, and cheap computing power, complex models can be built that accurately predict market direction. Machine learning algorithms are an important element in the toolset of leading players in global markets today.