

Predicting S&P 500 index movement

Problem Formulation

ML problem/Introduction

This machine learning problem tries to forecast daily closing prices of one of the most famous stock price indices, Standard and Poor's 500, or better known as S&P 500. The S&P 500 stock index tracks the share prices of the largest public companies in the USA. The S&P tracks U.S stocks with market cap of at least 13.1 billion dollars[1], and since the companies it tracks are that big, it's known to illustrate/correlate the whole status of the U.S economy (and possibly the overall economy around the globe). One could use this model when trying to predict the movement of the index, or to recognize current trend, and possibly do buying/selling decisions based on the model's forecast.

Dataset

The datapoints in this problem are daily prices (opening, high, low, close) of the of the S&P 500 index, the daily volume (the trade volume of the constituents), and exponential moving average from 26 days (formulation explained below). One can find data about the prices from various source, but for this project the data is obtained from the *Yahoo Finance* web page[2]. The data obtained contains the prices for the S&P 500 from the last five years, which covers over thousand datapoints (we are not specifying a certain quantity here in case some of the datapoints are inadequate, that is, has to be ignored).

From the datapoints we can calculate all kinds of indicator values to indicate the movement of the index. Such indicators values are simple moving averages (SMA), exponential moving averages (EMA), relative strength index (RSI), moving average converge (MACD), and the list goes on. The label in this problem is the closing price of the day of the S&P 500 index, and there are five features, where one of them must be calculated from the closing prices (always calculated on the day before the one that we're predicting). The feature that must be calculated is the exponential moving average from the previous 26 days. The other features are the opening, highest, and lowest price of a day, and the volume of a day.

Below is a formula which explains how the EMA is calculated:

EMA:

$$EMA_{Today} = \left(Value_{Today} * \left(\frac{Smoothing}{1 + Num. of Days} \right) \right) + EMA_{Yesterday} \left(1 - \left(\frac{Smoothing}{1 + Num. of Days} \right) \right)$$

We can use the SMA (simple moving average) from the previous day as the first EMA for $EMA_{Yesterday}$ (the "starting point" for this recursive formula). The smoothing factor is a design choice, but the most common value for it is 2 [3], and we will use that.

References

[1] The Balance (2022), *The S&P 500 and How It Works*, <https://www.thebalance.com/what-is-the-sandp-500-3305888>

[2] Yahoo Finance, S&P 500 (^GSPC),
<https://finance.yahoo.com/quote/%5EGSPC/history?p=%5EGSPC>

[3] Investopedia, *Exponential Moving Average (EMA)*,
<https://www.investopedia.com/terms/e/ema.asp>