DM final

2016 Spring Semester

Name:Jingyu Cui Student ID:5130309758 Email: 1040899365@sjtu.edu.cn

1 Moving Average

In statistics, a moving average is a calculation to analyse data points by creating series of averages of different subsets of the full data set. It is also called a moving mean or rolling mean. In our project, two kinds of moving average are applied and will be discussed in the following sections respectively.

1.1 Simple Moving Average

A simple moving average (SMA) is the unweighted mean of the previous n data in financial applications. This is similar to the definition of sliding-window smoothing, which is widely applied in electronic engineering. An SMA of n days is defined below:

$$\begin{aligned} SMA_M &= \frac{p_M + p_{M-1} + \ldots + p_{M-(n-1)}}{n} \\ &= \frac{1}{n} \sum_{i=0}^{n-1} p_{M-i} \end{aligned}$$

In science and engineering, the mean is usually derived from an equal number of data on both side, and can be also regarded as a kind of convolution. As an approach of smoothing, it is widely applied in noise cancelling. In financial terms, moving-average levels can be interpreted as support in a falling market, or resistance in a rising market. We use SMA in our project to approximate the trend in a following period. With a smoother curve, we may get a better global trend rather than a short-term one.

1.2 Exponential moving average

An exponential moving average (EMA), is also called an exponentially weighted moving average (EWMA). In a engineering view, it can also be regarded as an infinite impulse response filter that applies weighting factors which decrease exponentially. The weighting for older data points decrease exponentially, but will never reach zero. This exponentially-decreasing property gives it a wide application in financial analysis.

The EMA for a series of data X is given by the following recursive form:

$$EMA_t = \begin{cases} X_1, & t = 1 \\ \alpha X_t + (1-\alpha) X_{t-1}, & t \leq 2. \end{cases}$$

Where α is a parameter within $(0, 1)$ indicating the decreasing speed. Usage in financial indicators will be discussed in below.

2 Financial Indicators

In finance, technical analysis is a security analysis methodology for forecasting the direction of prices through the study of past market data, primarily price and volume. The analysis includes several important indicators, which will be introduced in the remaining section.

2.1 Average Convergence / Divergence

MACD, short for moving average convergence/divergence, is a trading indicator used in technical analysis of stock prices. It is supposed to reveal changes in the strength, direction, momentum, and duration of a trend in a stock's price.

The MACD series is defined as the difference between a "fast" (short period) exponential moving average (EMA), and a "slow" (longer period) EMA of the price series. A mathematical interpretation goes like:

$$\begin{aligned}
EMA_{12}(t) &= \tfrac{11}{13}EMA_{12}(t-1) + \tfrac{2}{13}p(t) \\
EMA_{26}(t) &= \tfrac{25}{27}EMA_{26}(t-1) + \tfrac{2}{27}p(t) \\
DIF(t) &= EMA_{12}(t) - EMA_{26}(t) \\
DEM(t) &= \tfrac{4}{5}DEM(t-1) + \tfrac{1}{5}DIF(t) \\
MACD(t) &= 2(DIF(t) - DEM(t))
\end{aligned}$$

Where DIF is the differential between the fast average and the slow one. DEM is the EMA of DIF and MACD is the difference between these two values.

2.2 Stochastic Indicator (KDJ)

Stochastic oscillator is a momentum indicator used to qualify the divergence from a ordinary range. It take the closing price as well as the peak and bottom into consideration. This has a relatively complex calculation since it actually includes 3 indicators. It is derived from raw stochastic value (RSV), which it related to the closing, highest and lowest price:

$$\begin{aligned}
RSV_\tau(t) &= \tfrac{p(t)-\max_\tau(p(t))}{\max_\tau(p(t))-\min_\tau(p(t))} \\
K(t) &= \tfrac{2}{3}K(t-1) + \tfrac{1}{3}RSV_\tau(t) \\
D(t) &= \tfrac{2}{3}D(t-1) + \tfrac{1}{3}K(t) \\
J(t) &= 3D(t) - 2k(t)
\end{aligned}$$

Where $\max_\tau(p(t))$ and $\min_\tau(p(t))$ is the peak and bottom for $p(t)$ in a time period of τ respectively. We can see that the K indicator is an EMA of RSV, and D is an EMA of K.

2.3 Relative Strength Index

The relative strength index (RSI) is intended to chart the current and historical strength or weakness of a stock or market based on the closing prices of a recent trading period.

RSI is defined as the ratio of rising points and the sum of rising and falling points. Mathematically, it can be written as:

$$\begin{aligned}
RS_\tau(t) &= \tfrac{\sum_{i=t-\tau+1}^{t}\Delta_+p(i)}{\sum_{i=t-\tau+1}^{t}\Delta_-p(i)} \\
RSI(t) &= \tfrac{RS}{1+RS}
\end{aligned}$$

Where $\Delta_+p(i) = \max(0, p(i) - p(i-1))$ and $\Delta_-p(i) = \max(0, p(i-1) - p(i))$ are the rising and falling price respectively.

3 Model Assumptions

We base our model on the hypothesis that the trend of our financial data is relevant to a series of previous data. This is called a time-sequential model. Furthermore, the financial indicators are derived from our sequential data. Hence we can suppose that these indicators also include important informations about the trend in the future. Our hypothesis can be represented by:

$$T(t) = f(MACD(t-1), RSI(t-1), KDJ(t-1), p(t-1))$$

Where $T(t)$ is the trend at time slot t, and $p(t-1)$ is the price at time slot $t-1$. This assumption is also called a causal model. In our problem, a "time slot" is defined to be a time instance with at least a deal. Those without any deal should be picked out from our data. We base our training process on this causal model and the detail will be discussed in the following sections.

4 Data Process

In this section, we will mainly show how we derive the indicators introduced above. Our processing is finished by a python library named pandas, which is designed for ? .

After reading a csv file, we first select our "time slot". This is done by deleting rows without a deal. Then we pick out the dealing price to form a sequence. This sequence is used to calculate all the indicators above, as well as the trend of the stock.

For the trend, we first use a window smoothing with size 50 to draw a moving average. Then for each time slot, the trend is defined as follows:

$$T(t) = \begin{cases} 1, & p(t+\tau) - p(t) \geq thresh \\ 0, & -thresh \leq |p(t+\tau) - p(t)| \leq thresh \\ -1, & p(t) - p(t+\tau) \geq thresh. \end{cases}$$

Where τ is a parameter indicating the number of time slots. The larger τ is, the further we are predicting from the given data. In our case, we do not want τ to be too large, which may go over some peak. τ should be chosen such that it can approximately represent a monotonic period.

For the other indicators, the calculation is based on their definition. Notice that in the previous section, the indicators are derived based on the closing price. We apply a similar process calculating them based on each time slot. The parameters are chosen likewise to the original definition.

5 Training and Analysis

With the indicators, we are able to train our model and make predictions. The learning take the indicators as input and the output is the trend of the moving average. Notice that we turn this prediction problem into a 3-classification problem by dividing the trend into three classes: rise, fall and keep. We choose the threshold between them, making them have approximately equal number of samples.

We use a neural network with 1 hidden layer to train our data. The number of neurons in the hidden layer is chosen to be 10. The evaluation process will give an error rate at around 42%. For a 3-classification problem, this result is fairly a good one.

The result confirmed our causal model with the hypothesis that the trend can be revealed by looking into the financial indicators. Since our indicators are only derived from the time sequence of price, we can also conclude that these indicators are a good representation for the potential trend behind the price sequence.