Feature Extraction

Suryoday Basak

May 12, 2017

Stocks' data is usually acquired as time-series data. In our solution, we consider only the closing price of a stock and we collect these values for many years. Hence, our solution can be considered to be of the form $(date, price_{closing})$. If we plot this data, we may get a graph like this:

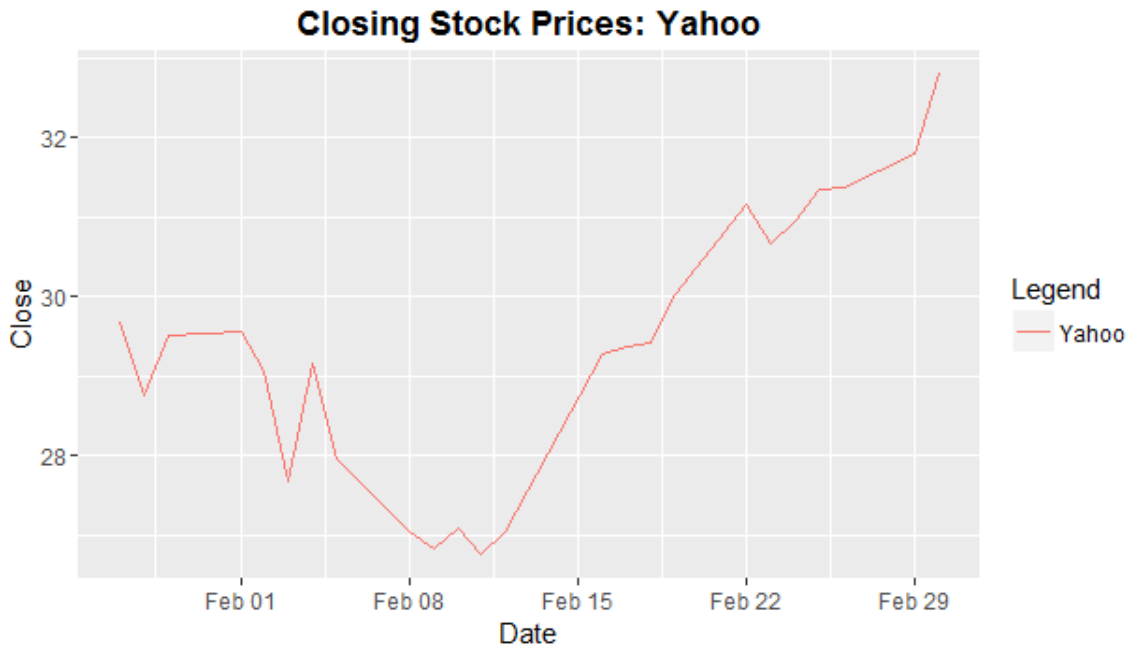

Figure 1: A sample plot of 25 random days of a stock (Yahoo), representative of the fluctuations that are observed.

A keen observer might visually inspect this graph and make an intelligent guess: whether the closing price on a particular day will increase or decrease with respect to the previous day. However, such methods of subjective examination requires a lot of experience in order to predict if an investor can make a profit by investing in a particular stock or not and hence the interest to automate the prediction of the outcome of an investment is ever growing. When this data is fed to a computer, it needs to be filtered and preprocessed through a series of stages in order to be suitable for analysis. As we are dealing with time-series data, this preprocessing eventually leads to an extraction of useful features.

A *feature* in data can be any useful characteristic that can be used to correctly identify the class a data sample belongs to (when we are performing a classification). As an example, consider the closing prices of an arbitrary stock as shown in Table 1. Here, we calculate the *relative strength index* (RSI) (Section 1) as a feature.

The monetary gain or loss of day n is calculated with respect to day $n-1$. Hence, there is a loss on day 6, where the closing price is 53.0, since the closing price of day 5 is 54.0. There is neither a gain nor a loss on day 1 as it is the first day under consideration. As we are calculating RSI over a 14-day period, the average gain and average loss for day 15 is calculated over the period of days 1-14. Similarly, the average gain and average loss for day 16 is calculated over the period of days 2-15, and so on. Accordingly, the RS and RSI are calculated for days 15, 16, ..., till the last day for which the closing price is observed.

Hence, in this example, RSI can be considered to be a characteristic of the data and we calculate it for day 15 onwards. This process of determining and calculating relevant and important characteristics is known as *feature extraction*. Since the acquired data is a time-series data, feature extraction gives us a matrix of the important characteristics. In cases where tabulated data already exists (with a lot of features, some of which are noisy), algorithms like principal component analysis (PCA), independent component analysis (ICA), factor analysis, etc. helps to further reduce the number of important features for analysis or

Day	Price ($)	Gain ($)	Loss ($)	Average Gain_{14} ($)	Average Loss_{14} ($)	RS_{14}	RSI_{14}
1	50.0	-	-	-	-	-	-
2	52.0	2	-	-	-	-	-
3	51.0	-	1	-	-	-	-
4	55.0	4	-	-	-	-	-
5	54.0	-	1	-	-	-	-
6	53.0	-	1	-	-	-	-
7	56.0	3	-	-	-	-	-
8	57.0	1	-	-	-	-	-
9	56.0	-	1	-	-	-	-
10	57.0	1	-	-	-	-	-
11	55.0	-	2	-	-	-	-
12	58.0	3	-	-	-	-	-
13	56.0	-	2	-	-	-	-
14	59.0	3	-	-	-	-	-
15	58.0	-	1	1.21	0.57	2.12	67.95
16	60.0	2	-	1.21	0.64	1.89	65.38

Table 1: Sample Closing Prices

classification.

Feature selection, on the other hand, is how a certain classification algorithm handles data which results in the best classification result. The *Gini impurity criteria* is a method used to select the best features for classification (see Section **??** for details on the use of Gini impurity criteria).

1 Relative Strength Index (RSI)

RSI is a popular momentum indicator which determines whether the stock is overbought or oversold. A stock is said to be overbought when the demand unjustifiably pushes the price upwards. This condition is generally interpreted as a sign that the stock is overvalued and the price is likely to go down. A stock is said to be oversold when the price goes down sharply to a level below its true value. This is a result caused due to panic selling. RSI ranges from 0 to 100 and generally, when RSI is above 70, it may indicate that the stock is overbought and when RSI is below 30, it may indicate the stock is oversold.

The formula for calculating RSI is:

$$RSI = 100 - \frac{100}{1 + RS} \tag{1}$$

$$RS = \frac{\textit{Average Gain Over past 14 days}}{\textit{Average Loss Over past 14 days}} \tag{2}$$

2 Stochastic Oscillator

Stochastic Oscillator follows the speed or the momentum of the price. As a rule, momentum changes before the price changes. It measures the level of the closing price relative to low-high range over a period of time.

The formula for calculating Stochastic Oscillator is:

$$\%K = 100 \times \frac{(C - L_{14})}{(H_{14} - L_{14})} \tag{3}$$

where,

C = current closing price

L_{14} = lowest price over the past 14 days

H_{14} = highest price over the past 14 days

3 Williams Percentage Range

Williams Percentage Range or Williams %R is another momentum indicator, similar in idea to stochastic oscillator. The Williams %R indicates the level of a market's closing price in relation to the highest price for the look-back period, which is 14 days. It's value ranges from -100 to 0. When its value is above -20, it indicates a *sell signal* and when its value is below -80, it indicates a *buy signal*.

Williams %R is calculated as follows:

$$\%R = -100 \times \frac{(H_{14} - C)}{(H_{14} - L_{14})} \tag{4}$$

where,

C = current closing price

L_{14} = lowest price over the past 14 days

H_{14} = highest price over the past 14 days

4 Moving Average Convergence Divergence

the moving average convergence-divergence (MACD) is a momentum indicator which compares two moving averages of prices. MACD is calculated by deducting the 26-day exponential moving average (EMA) from the 12-day EMA. A 9-day EMA of the MACD is considered as the *signal line*, which serves as the threshold for the buy or sell signals

The formula for calculating MACD is:

$$MACD = EMA_{12}(C) - EMA_{26}(C) \tag{5}$$

$$SignalLine = EMA_9(MACD) \tag{6}$$

where,

C = closing price

EMA_n = n-day exponential moving average

When the MACD goes below the signal line, it indicates a sell signal. When it goes above the signal line, it indicates a buy signal.

5 Price Rate of Change

The Price Rate of Change (PROC) is a technical indicator which reflects the percentage change in price between the current price and the price over the window that we consider to be the time period of observation.

It is calculated as follows:

$$PROC_t = \frac{C_t - C_{t-n}}{C_{t-n}} \tag{7}$$

where,

$PROC_t$ = price rate of change at time t

C_t = closing price at time t

6 On Balance Volume

On balance volume (OBV) measures the most recent change in price with respect to the price n days ago. This technical indicator is used to find buying and selling trends of a stock. The formula for calculating OBV is:

$$OBV(t) = \begin{cases} OBV(t-1) + Vol(t) & \text{if } C(t) > C(t-1) \\ OBV(t-1) - Vol(t) & \text{if } C(t) < C(t-1) \\ OBV(t-1) & \text{if } C(t) = C(t-1) \end{cases} \tag{8}$$

where,

$OBV(t)$ = on balance volume at time t

$Vol(t)$ = trading volume at time t

$C(t)$ = closing price at time t