A note on one Bitcoin statistical project

Cypruian team
`cypruian at protonmail.com`

2021 November

We denote a current state of a system (primarily Bitcoin ledger) as $S(h) \in \mathbf{S}$, where $h \in \mathbb{N} \cup \{0\}$ is a counter for a timestamp or a block height, and $\mathbf{S}$ is some space of all possible current states. The state $S(h)$ includes UTXO set $U = U(S(h))$ and some other data. Then $\mathbf{S}^n$ is a space of all paths (trajectories) of $n \in \mathbb{N}$ states, and

$$(S(h_1), S(h_2), \dots, S(h_n)) \in \mathbf{S}^n, 0 \le h_1 \le h_2 \le \dots \le h_n,$$

is a realization of the path at the points $h_1, h_2, \dots, h_n$. For instance, the full history of the system from the "genesis" point $h = 0$ to the current point $h = h_c$ is:

$$(S(0), S(1), \dots, S(h_c)) \in \mathbf{S}^{h_c+1}.$$

The system exists together with an external variable $p(h) \in \mathbb{R}$ at every point h. It is a price level that does not depend on the system directly, but it interacts with it someway.

Two statistical problems are considered in this paper.

Problem A: determining the nature of this interaction and forecasting the behavior of the price level $p(h)$ by the current system's state based on the interaction's nature.

One of the solutions to the problem A: selection of some dynamic numerical characteristics (metrics) $X(h) \in \mathbb{R}^d$ from the system's state and construction of probabilistic models for the process $Y(h) = (p(h), X(h)) \in \mathbb{R}^{d+1}$. As a result of such construction and statistical identification of the model, for the process $Y(h)$ we can build a transition function in Δ points of a probability distribution with a density

$$\mathfrak{L}\{Y(h+\Delta)|Y(\tau), \tau \le h\} = f_{Y(\tau),\tau\le h}(\cdot).$$

Based on the density, we can build a forecast:

$$\hat{p}(h+\Delta) = \int_{\mathbb{R}^{d+1}} x_1 f_{Y(\tau),\tau\le h}(x)dx, x = (x_1, \dots, x_{d+1}) \in \mathbb{R}^{d+1},$$

or, for instance, an estimate for a probability of the event when $p(h+\Delta)$ differs from $p(h)$ by less than $100 \cdot \varepsilon$ per cent:

$$\hat{\mathbf{P}}\{p(h+\Delta) \in [(1-\varepsilon)p(h), (1+\varepsilon)p(h)]\} = \int_{[(1-\varepsilon)p(h),(1+\varepsilon)p(h)]\times\mathbb{R}^d} f_{Y(\tau),\tau\le h}(x)dx,$$

or an estimate for a probability of some other events.

Selection of the metrics $X(h)$ is performed under mathematical modelling of the process $Y(h)$ or based on Shannon mutual information.

The metrics $X(h)$ can also be discrete, and from the price level $p(h)$ we can allocate a discrete variable, e.g. this variable can store codes for the events of the price level's changing. In this case the discrete process $Y(h)$ can be investigated using the well-developed methods for statistical analysis of Markov models family.

Problem B: recognition (classification) of the current states.

An expert or a group of experts label a finite number of the system's states in which we are interested. A deterministic algorithm for labelling can also be used. For instance, let we have a binary labelling: states before a "specific point" (label 0) and other states (label 1). So, the set of points presents a union of two disjoint sets: $H_+ \subset \mathbb{N} \cup \{0\}$ for the states before the "specific point" and $H_- \subset \mathbb{N} \cup \{0\}$ for the other states. Then a pattern $\Delta = (\Delta_1, \ldots, \Delta_L)$ of a neighborhood is selected, where L is a size of this pattern. For every point h we can construct a "neighborhood profile":

$$Z(h) = (Y(h + \Delta_1), Y(h + \Delta_2), \ldots, Y(h + \Delta_L)).$$

Based on a sample of profiles of states before the "specific point" $\{Z(h), h \in H_+\}$ we construct a probabilistic model $P_+(Z(h))$ for the state's profile before the "specific point". Similarly, based on $\{Z(h), h \in H_-\}$ we construct a probabilistic model $P_-(Z(h))$ for the other state's profile.

Then a classified point h goes to one of the two labels based on a likelihood ratio statistical criteria with a critical region

$$\frac{P_+(Z(h))}{P_-(Z(h))} \leq C,$$

where $C \in \mathbb{R}$ is a threshold. Other supervised machine learning methods can also be applied here.

For instance, a probability distribution of some characteristic $\phi(u)$ of unspent tx outputs $u \in U(S(h))$ can be used as a metric $X(h)$:

$$X(h) = \hat{\mathfrak{L}}\{\phi(u), u \in U(S(h))\},$$

where a corresponding estimate $\hat{\mathfrak{L}}$ is constructed on a sample $\{\phi(u), u \in U(S(h))\}$ as a histogram (non-parametric) or as a result of identification in some class of distributions (parametric). The price level at the point when an output was created can be an example of such function $\phi(u)$:

$$\phi(u) = p(\mathbf{h}(u)),$$

or the value of an output in Satoshi:

$$\phi(u) = \mathbf{v}(u).$$