

TRƯỜNG ĐẠI HỌC BÁCH KHOA

KHOA KHOA HỌC & KỸ THUẬT MÁY TÍNH



ĐỀ CƯƠNG LUẬN VĂN TỐT NGHIỆP

Phát triển công cụ dự đoán xu hướng
giá ngắn hạn các đồng tiền mật mã
bằng kỹ thuật học máy

SVTH:

Vũ Quang Nam

GVHD:

TS Nguyễn An Khương

Nguyễn Lê Thành

Ngày 13 tháng 12 năm 2018

Mục lục

Danh mục hình vẽ	5
0.1 Lời cam đoan	1
1 Giới thiệu	3
1.1 Giới thiệu đề tài	3
1.2 Mục tiêu và phạm vi đề tài	3
1.2.1 Mục tiêu	3
1.2.2 Phạm vi đề tài	3
2 Tổng quan về lĩnh vực nghiên cứu	5
2.1 Những yếu tố tác động đến giá trị đồng tiền mã hóa	5
2.1.1 Cung và cầu của thị trường	5
2.1.2 Tin tức trên các phương tiện thông tin đại chúng	5
2.1.3 Quy định của chính phủ	5
2.1.4 Chính sách của các tổ chức	6
2.1.5 Các vấn đề kỹ thuật	6
2.2 Nhu cầu sử dụng tiền mã hoá của mỗi hệ sinh thái	6
3 Dữ liệu	7
3.1 Chuẩn bị dữ liệu	7
3.2 Mô tả dữ liệu	7
4 Cây hồi quy và phân loại	9
4.0.1 Cấu trúc cây nhị phân cơ bản	9
4.0.2 Các luật tách thường dùng	9
4.0.2.1 Đối với Regression Tree	9
4.0.2.2 Đối với Classification Tree	10
4.0.3 Tiêu chí tách	10
4.0.4 Tỉa cây	11

4.0.4.1	Tập huấn luyện - kiểm tra độc lập	11
4.0.4.2	Cross-Validation	11
4.1	Rừng ngẫu nhiên	11
5	Các khái niệm cơ bản	13
5.1	Các khái niệm về Đại số tuyến tính	13
5.1.1	Các kí hiệu cơ bản thường dùng	13
5.1.2	Phép nhân ma trận	13
5.1.3	Phép nhân vector-vector	13
5.1.3.1	Inner Product	13
5.1.3.2	Outer Product	14
5.1.4	Phép nhân ma trận-vector	14
5.1.5	Ma trận đơn vị và ma trận đường chéo	14
5.1.6	Ma trận dạng toàn phương và ma trận xác định dương	15
5.2	Các khái niệm về Xác suất	15
5.2.1	Likelihood	15
6	Mô hình mạng Markov ẩn (HMM)	17
6.1	Tính sinh	17

Danh mục hình ảnh

0.1 Lời cam đoan

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi dưới sự hướng dẫn của Nội dung nghiên cứu và các kết quả đều là trung thực và chưa từng được công bố trước đây. Các số liệu được sử dụng cho quá trình phân tích, nhận xét được chính tôi thu thập từ nhiều nguồn khác nhau và sẽ được ghi rõ trong phần tài liệu tham khảo. Ngoài ra, tôi cũng có sử dụng một số nhận xét, đánh giá và số liệu của các tác giả khác, cơ quan tổ chức khác. Tất cả đều có trích dẫn và chú thích nguồn gốc. Nếu phát hiện có bất kỳ sự gian lận nào, tôi xin hoàn toàn chịu trách nhiệm về nội dung thực tập tốt nghiệp của mình. Trường đại học Bách Khoa thành phố Hồ Chí Minh không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện.

Chương 1

Giới thiệu

1.1 Giới thiệu đề tài

(Logistic Regression), cây hồi quy và phân loại (CART), rừng ngẫu nhiên (Random Forest), mạng nơ-ron (Neural NetWork), máy vectơ hỗ trợ (SVM). Nhưng ở Việt Nam lại chưa có nhiều nghiên cứu về đề tài này. Vậy nên tôi quyết định chọn đề tài **Dự đoán xu hướng giá ngắn hạn các đồng tiền mật mã bằng kĩ thuật học máy**.

1.2 Mục tiêu và phạm vi đề tài

1.2.1 Mục tiêu

Mục tiêu của luận văn này là xây dựng một công cụ dự đoán xu hướng giá ngắn hạn các đồng tiền mật mã bằng kĩ thuật học máy. Dữ liệu đầu vào là các thông tin về lịch sử giá các đồng tiền ảo trong các phiên giao dịch.

1.2.2 Phạm vi đề tài

- Tìm hiểu và nghiên cứu về lý thuyết học máy thống kê (statistical machine learning)
- Xây dựng mô hình dự đoán về xu hướng tăng giảm, dự đoán giá của các đồng trong thời gian ngắn hạn.

Các đối tượng nghiên cứu trong đề tài:

- Các tài liệu liên quan tới lý thuyết thống kê hiện đại
- Các mô hình trong học máy: hồi quy logistic, rừng ngẫu nhiên, mạng nơ-ron

- Sử dụng ngôn ngữ Python, R và một số thư viện để hiện thực mô hình.

Chương 2

Tổng quan về lĩnh vực nghiên cứu

2.1 Những yếu tố tác động đến giá trị đồng tiền mã hóa

2.1.1 Cung và cầu của thị trường

Trong nguyên tắc chính của kinh tế nếu người ta mua một đồng tiền, giá trị của đồng tiền sẽ tăng lên và nếu người ta bán đồng tiền, giá sẽ giảm.

2.1.2 Tin tức trên các phương tiện thông tin đại chúng

Các sự kiện chính trị và kinh tế trên toàn thế giới ảnh hưởng đến cách mà con người phản ứng với các dự đoán giá, tin tức cảnh báo về rủi ro tác động chính lên cung-cầu.

2.1.3 Quy định của chính phủ

Có 4 cấp độ quản lý tiền ảo hiện nay đang được các nước thực thi, cụ thể:

- Cấm trên diện rộng.
- Cấm trong lĩnh vực tài chính ngân hàng (trong đó có Trung Quốc).
- Cảnh báo rủi ro đối với người sử dụng, đầu tư,.
- Chấp nhận như một phương tiện thanh toán (trong đó có Hàn Quốc, Nhật Bản và Mỹ).

cập nhật ngày 14/4/2018.

2.1.4 Chính sách của các tổ chức

Facebook, Google và Twitter đã ngăn chặn khách hàng và người dùng sử dụng dịch vụ cryptocurrency.

2.1.5 Các vấn đề kỹ thuật

Vì đồng tiền mã hóa có thể bị hack thành công vào tài khoản hoặc tấn công máy chủ, có thể làm giảm tỷ giá hối đoái, dẫn đến giá giảm.

2.2 Nhu cầu sử dụng tiền mã hoá của mỗi hệ sinh thái

- Số thành viên tham gia vào hệ sinh thái (Số người đến khu vui chơi mua vé tham gia các trò chơi trong đó bằng tiền A).
- Số lượng dịch vụ trong hệ sinh thái (Khu vui chơi có càng nhiều trò chơi thì nhu cầu sử dụng tiền A càng tăng); Và các nền tảng như Ethereum luôn mở cho các đối tác tạo các dịch vụ gia tăng trên đó giống như khu vui chơi cho phép đối tác bên ngoài vào tổ chức trò chơi ở trong.
- Số người đầu cơ: Những người nhận thấy nhu cầu tiền mã hoá của một hệ sinh thái tăng dần sẽ mua để nắm giữ chờ tăng giá thì bán ra. (Giống như phe vé bóng đá ngày trước mua vé chờ sát trận nhu cầu tăng vọt thì bán ra. Khu vui chơi thì ít có nhóm này vì lượng vé không bị giới hạn).
- Số người bán bên ngoài chấp nhận tiền mã hoá: Một số người bán nhận thấy tính thanh khoản của tiền mã hoá và giá trị tăng dần của nó nên đã chấp nhận khách hàng thanh toán các hàng hoá dịch vụ của mình bằng loại tiền này (Nhà hàng bên cạnh khu vui chơi có thể chấp nhận khách hàng thanh toán bằng tiền A).

Chương 3

Dữ liệu

3.1 Chuẩn bị dữ liệu

Có nhiều nguồn cung cấp dữ liệu cho bài toán dự đoán giá đồng tiền mã hóa, trong phạm vi nghiên cứu có sử dụng các lịch sử giao dịch các đồng với nhau (trading pair) được lấy từ API có sẵn từ 8 sàn giao dịch với cấu trúc bảng:

symbol	market	timeIndicator	openPrice	closePrice	highPrice	lowPrice	volume	minTimestamp	maxTimestamp
SRN/BTC	huobipro	2018-12-13 08:26:00 UTC	1.331e-05	1.325e-05	1.331e-05	1.331e-05	0.0085977935000000009	1544689603614	1544689615119
WTC/BTC	huobipro	2018-12-13 08:26:00 UTC	0.00026948	0.00027049	0.00027049	0.00027049	5.3997e-06	1544689570921	1544689590036
EOS/PAX	binance	2018-12-13 08:26:00 UTC	1.9152	1.9152	1.9207	1.9207	204.88715399999998	1544689566328	1544689618905
EKT/BTC	huobipro	2018-12-13 08:26:00 UTC	1.24e-06	1.25e-06	1.25e-06	1.25e-06	0.0109388594	1544689562604	1544689611453
NEO/USDT	huobipro	2018-12-13 08:26:00 UTC	5.86	5.85	5.87	5.87	2191.216003	1544689561044	1544689618173
OMG/BTC	bitfinex2	2018-12-13 08:26:00 UTC	0.00036052	0.00036052	0.00036052	0.00036052	0.026111466704516802	1544689588624	1544689588624
XRP/BTC	binance	2018-12-13 08:26:00 UTC	8.898e-05	8.896e-05	8.9e-05	8.9e-05	1.1474145999999998	1544689568501	1544689619519
BTG/BTC	binance	2018-12-13 08:26:00 UTC	0.003396	0.003396	0.003396	0.003396	0.030564	1544689577770	1544689577770
HB10/USDT	huobipro	2018-12-13 08:26:00 UTC	0.2421	0.2419	0.2421	0.2421	56.687566	1544689569087	1544689617665
ICX/BTC	huobipro	2018-12-13 08:26:00 UTC	6.607e-05	6.625e-05	6.625e-05	6.625e-05	0.005767279581	1544689563889	1544689614203

3.2 Mô tả dữ liệu

Phiên giao dịch dòng 1 SRN/BTC được ghi lại thành một dòng với thông tin như sau:

- symbol: Tên giao dịch giữa hai đồng với nhau cụ thể là đồng SRN so với đồng BTC
- Market: Tên sàn giao dịch cụ thể là sàn huopipro
- timeIndicator: Thời điểm mở phiên giao dịch 8 giờ 13/12/2018 UTC
- openPrice: Tỷ giá thời điểm mở phiên
- openPrice: Tỷ giá thời điểm đóng phiên

- `highPrice`: Tỷ giá cao nhất phiên giao dịch
- `lowPrice`: Tỷ giá thấp nhất phiên giao dịch
- `volume`: Khối lượng giao dịch (Ví dụ symbol là SRN/BTC `volume` có nghĩa là số đồng SRN)
- `minTimestamp`, `maxTimestamp`: do mỗi giao dịch cần một thời gian nhất định nên cần có thời gian bắt đầu và kết thúc giao dịch tính theo POSIX time.

Chương 4

Cây hồi quy và phân loại

Cây hồi quy và phân loại (CART) là một cây quyết định nhị phân được đề xuất bởi Breiman [23].

4.0.1 Cấu trúc cây nhị phân cơ bản

Ứng với một tập data ta cần tạo một cây nhị phân có đầu ra thành một chuỗi các lá, mục tiêu các lá có giá trị đầu ra tương đồng nhiều nhất. Khi bắt đầu từ nút cần chọn ra một thuộc tính và một giá trị sao cho giảm được "nhiều" nhiều nhất có thể. Ta có thể lựa chọn các độ đo khác nhau nhằm sinh ra cây nhị phân với ý tưởng này, với mỗi độ đo khác nhau tương ứng với một luật tách (splitting rule).

4.0.2 Các luật tách thường dùng

Ta có thể chia thành 2 loại theo:

4.0.2.1 Đối với Regression Tree

- Least squares: phương pháp chọn tổng bình phương lỗi (SSE) nhỏ nhất giữa các quan sát với giá trị trung bình. Giá trị này tốt nhất khi đạt tới 0 nghĩa là tất cả các giá trị quan sát đều như nhau.
- Least absolute deviations: phương pháp chọn tổng trị tuyệt đối nhỏ nhất giữa các quan sát với giá trị trung bình, so với Least squares thì phương pháp này ít nhạy hơn đối với các dữ liệu ngoại lai (outlier).

4.0.2.2 Đối với Classification Tree

- Misclassification error: là tỉ lệ của các quan sát không cùng loại với loại chính.
- Gini index Entropy:
- Entropy index: hay cross-entropy
- Twoing:

4.0.3 Tiêu chí tách

CART sử dụng chỉ số Gini để làm tiêu chí tách với mô hình phân loại. Gọi $RF(C_j, S)$ biểu diễn tần suất xuất hiện của lớp C_j trong các phần tử của tập S . Chỉ số Gini được xác định bằng công thức:

$$I_{gini}(S) = 1 - \sum_{j=1}^x RF(C_j, S)^2$$

Sau khi tập S được chia thành nhiều tập con S_1, S_2, \dots, S_t , bởi phép chia B , độ lợi thông tin $G(S, B)$ được tính bằng công thức:

$$G(S, B) = I(S) - \sum_{i=1}^t \frac{|S_i|}{|S|} I(S_i)$$

Ta chọn phép chia B nào làm tối đa hóa độ lợi $G(S, B)$. Sau đó CART sẽ xây dựng các mô hình trên các tập S_i . Một cây phân loại sẽ dự đoán phân phối của một mẫu trên một lớp nhất định. Hiệu quả của mỗi cây phân loại sẽ được tính dựa trên sai số toàn phương trung bình. Với mỗi lớp j , gọi $C_j(e)$ là chỉ báo có giá trị bằng 1 nếu mẫu e thuộc lớp j và bằng 0 nếu không. Sai số toàn phương trung bình MSE được tính bằng công thức:

$$MSE = E_e \left[\sum_{j=1}^x (C_j(e) - P_j(e))^2 \right]$$

với kì vọng trên toàn bộ các mẫu, $P_j(e)$ đại diện cho xác suất mẫu e thuộc lớp j . Đối với cây hồi quy, độ lệch $R(S_i)$ là sai số toàn phương trung bình:

$$R(S) = \frac{1}{n} \sum_i (y_i - h(t_i))^2$$

với y_i là giá trị thực của biến mục tiêu trong mẫu t_i và $h(t_i)$ là giá trị dự đoán của mô hình.

4.0.4 Tỉa cây

Khi xây dựng cây bằng cách "vét cạn", tối ưu tất cả các mẫu trong tập huấn luyện, dẫn đến các node lá trong cây mang ít các quan sát. Điều này làm kết quả xấu khi thử ở tập kiểm tra mặc dù tập huấn luyện có kết quả tốt. Nếu một cây được xây dựng quá nhỏ tức độ sâu quá ngắn thì chưa trích xuất được hết thông tin.

Ta có thể tùy chỉnh kích thước của cây theo các cách sau đây:

- Không nhất thiết node lá hoàn toàn đồng nhất, ta nên dừng việc tách nhánh khi độ đồng nhất trên mức chấp nhận được.
- Một cách khác là "vét cạn" cây đến khi đạt đến node lá nhỏ nhất (thường chỉ có một quan sát). Xác định độ sâu thích hợp dựa trên tập kiểm tra độc lập với tập huấn luyện hoặc dùng cross-validation (TODO: trình bày mục sau). Sau đó tỉa các nhánh đưa cây về độ sâu đã chọn.

4.0.4.1 Tập huấn luyện - kiểm tra độc lập

Khi tập mẫu đủ lớn, ta chia tập thành 2 phần riêng, độc lập với nhau.

- Tập huấn luyện: dùng để sinh cây có độ dài lớn đủ để có thể tỉa cây.
- Tập kiểm thử: từ cây đã sinh ở trên ngẫu nhiên tỉa các nhánh để tạo ra nhiều cây con, thử các quan sát ở tập kiểm thử trên những cây con này từ đó xác định được số lỗi nhỏ nhất theo bài toán regression hoặc classification.

4.0.4.2 Cross-Validation

Nếu dữ liệu chưa đủ cho việc tách riêng biệt thành hai tập với tỉ lệ như trên

nói cách khác chúng ta cần giữ lại tập train càng nhiều càng tốt nhưng vẫn cần sự độc lập giữa hai tập này.

4.1 Rừng ngẫu nhiên

Chương 5

Các khái niệm cơ bản

5.1 Các khái niệm về Đại số tuyến tính

5.1.1 Các kí hiệu cơ bản thường dùng

Bài báo cáo này có sử dụng những kí hiệu:

$A \in \mathbb{R}^{m \times n}$: ma trận A có m hàng và n cột.

$x \in \mathbb{R}^n$: vector gồm n phần tử hay vector n chiều.

5.1.2 Phép nhân ma trận

Phép nhân ma trận $A \in \mathbb{R}^{m \times n}$ và $B \in \mathbb{R}^{n \times p}$ là một ma trận $C = AB \in \mathbb{R}^{m \times p}$ sao cho

$$C_{ij} = \sum_{k=1}^n A_{ik} B_{kj}$$

5.1.3 Phép nhân vector-vector

5.1.3.1 Inner Product

Cho hai vector $x, y \in \mathbb{R}^n$ đại lượng $x^\top y$ được gọi là inner product hay dot product của hai vector có giá trị là một số thực sao cho:

$$x^\top y \in \mathbb{R}^n = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^n x_i y_i$$

Nhận xét inner products là trường hợp đặc biệt của phép nhân ma trận dễ nhận thấy $x^\top y = y^\top x$.

5.1.3.2 Outer Product

Cho hai vector $x \in \mathbb{R}^n, y \in \mathbb{R}^m$ đại lượng $xy^\top \in \mathbb{R}^{n \times m}$ được gọi là outer product của hai vector sao cho:

$$xy^\top \in \mathbb{R}^n = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \begin{bmatrix} y_1 & y_2 & \dots & y_m \end{bmatrix} = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \dots & x_1 y_m \\ x_2 y_1 & x_2 y_2 & \dots & x_2 y_m \\ \vdots & \vdots & \ddots & \vdots \\ x_n y_1 & x_n y_2 & \dots & x_n y_m \end{bmatrix}$$

5.1.4 Phép nhân ma trận-vector

Cho ma trận $A \in \mathbb{R}^{n \times m}$ và vector $y \in \mathbb{R}^m$ phép nhân $y = Ax \in \mathbb{R}^n$.

Ta có thể biểu diễn phép nhân khi nhìn A thành dạng hàng như sau:

$$y = Ax = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_n \end{bmatrix} x = \begin{bmatrix} A_1 x_1 \\ A_2 x_2 \\ \vdots \\ A_n x_n \end{bmatrix}$$

5.1.5 Ma trận đơn vị và ma trận đường chéo

Ma trận đơn vị (Identity matrix) $I \in \mathbb{R}^{n \times n}$ các phần tử trong I thỏa:

$$I_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \quad (5.1)$$

$$(5.2)$$

Ma trận đường chéo

5.1.6 Ma trận dạng toàn phương và ma trận xác định dương

Cho một ma trận vuông $A \in \mathbb{R}^{n \times n}$ và vector $x \in \mathbb{R}^n$, giá trị $x^\top Ax$ được gọi là có dạng toàn phương:

$$x^\top Ax = \sum_{i=1}^n x_i^\top (Ax)_i = \sum_{i=1}^n x_i^\top \left(\sum_{j=1}^n A_{ij} x_j \right) = \sum_{i=1}^n \sum_{j=1}^n A_{ij} x_i x_j$$

5.2 Các khái niệm về Xác suất

5.2.1 Likelihood

Chương 6

Mô hình mạng Markov ẩn (HMM)

6.1 Tính sinh

Mô hình Markov ẩn mô tả phân phối tiên nghiệm ở mỗi trạng thái, HMM là mô hình sinh mẫu có nghĩa từ mô hình và dữ liệu sẵn có, ta có thể sinh dữ liệu mới có cùng phân phối so với dữ liệu sẵn có.

Bibliography

- [1] L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen (1984). *Classification and regression trees*. CRC press.