TRƯỜNG ĐẠI HỌC BÁCH KHOA

KHOA KHOA HỌC & KỸ THUẬT MÁY TÍNH

LUẬN VĂN TỐT NGHIỆP

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

---

SVTH:
Vũ Quang Nam

GVHD:
Nguyễn An Khương
Nguyễn Lê Thành

Ngày 20 tháng 9 năm 2019

Mục lục

Danh mục hình ảnh

Lời cảm ơn

Tôi xin gửi lời cảm ơn chân thành nhất đến TS. Nguyễn An Khương và anh Nguyễn Lê Thành đã tận tình hướng dẫn trong quá trình chuẩn bị kiến thức để làm luận văn. Tôi cũng xin cảm ơn các bạn trong nhóm Datavisian đã rất nhiệt tình giúp đỡ và góp ý trong quá trình thực hiện các mô hình, các bạn có những phẩm chất của một nhà khoa học dữ liệu mà tôi nên học hỏi. Bên cạnh đó tôi cũng xin cảm ơn anh Nguyễn Xuân Mão, anh Trần Đào Vinh

Lời cam đoan

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi dưới sự hướng dẫn của thầy Nguyễn An Khương và anh Nguyễn Lê Thành. 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 kì 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

Giới thiệu đề tài nghiên cứu

Hiện nay, tiền mã hóa đã trở nên phổ biến, đa dạng với nhiều sàn giao dịch khác nhau. Đồng tiền mã hóa có tỷ giá thay đổi theo thời gian, việc tìm xu hướng giá ngắn hạn tính theo giờ phút không bị ảnh hưởng nhiều bởi các yếu tố bên ngoài như các dự báo, các quy định của chính phủ. Từ dữ liệu cụ thể là tổng hợp của các giao dịch trên các sàn trực tuyến việc tìm ra một giải thuật có thể dự đoán xu hướng giá của các giao dịch tiếp theo với nguyên tắc đề cao khách quan so với kinh nghiệm bản thân là một vấn đề mới mẻ. 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**.

Mục tiêu và phạm vi đề tài

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.

Phạm vi đề tài

- Tìm hiểu và nghiên cứu về lý thuyết học máy thống kê (statistical machine learning)

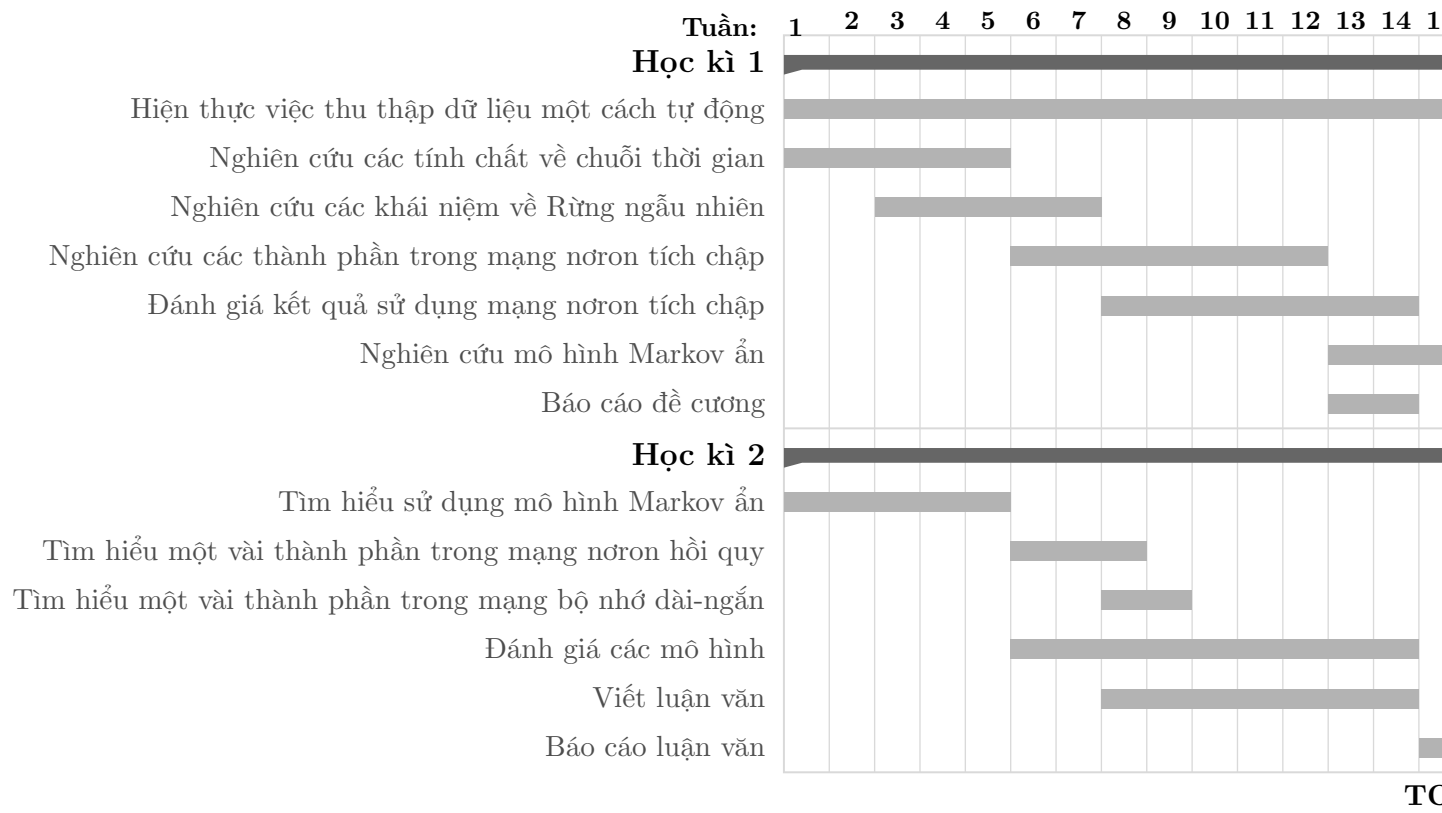
- Xây dựng mô hình dự đoán về xu hướng tăng giảm, dự đoán giá của các đồng trong thời gian ngắn hạn.

Các đối tượng nghiên cứu trong đề tài:

- Tìm hiểu một vài loại đồng tiền mã hóa, và các sàn giao dịch.
- Một vài tài liệu liên quan tới lý thuyết thống kê hiện đại.
- Tìm hiểu một vài mô hình trong học máy: hồi quy logistic, rừng ngẫu nhiên, mạng nơron.
- Sử dụng ngôn ngữ Python, R và một số thư viện để hiện thực mô hình.
- Xây dựng công cụ dự đoán giá một cách tự động.

Tiến độ thực hiện

Trong phần này, tác giả xin trình bày lịch trình công việc đã thực hiện đề tài trong học kỳ I và lịch trình dự kiến hiện thực đề tài trong quá trình làm luận văn chính thức ở học kỳ II dưới dạng biểu đồ Gantt sau đây.


Tuần:
1 2 3 4 5 6 7 8 9 10 11 12 13 14 1
Học kì 1
Hiện thực việc thu thập dữ liệu một cách tự động
Nghiên cứu các tính chất về chuỗi thời gian
Nghiên cứu các khái niệm về Rừng ngẫu nhiên
Nghiên cứu các thành phần trong mạng nơron tích chập
Đánh giá kết quả sử dụng mạng nơron tích chập
Nghiên cứu mô hình Markov ẩn
Báo cáo đề cương
Học kì 2
Tìm hiểu sử dụng mô hình Markov ẩn
Tìm hiểu một vài thành phần trong mạng nơron hồi quy
Tìm hiểu một vài thành phần trong mạng bộ nhớ dài-ngắn
Đánh giá các mô hình
Viết luận văn
Báo cáo luận văn
TC

Chương 2

Tổng quan về lĩnh vực nghiên cứu

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

Cung và cầu của thị trường

Trong nguyên tắc chính của kinh tế nếu nhu cầu mua đối với một đồng tiền tăng, giá trị của đồng tiền sẽ tăng và ngược lại khi nhu cầu bán tăng, giá sẽ giảm.

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.

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, Nga).
- Cảnh báo rủi ro đối với người sử dụng, đầu tư.
- Chấp nhận như một phương tiện thanh toán (các nước chấp nhận đồng bitcoin gồm có Mỹ, Canada, Úc, Liên minh châu Âu, Phần Lan [1]).

Chính sách của các tổ chức

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

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.

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

Chuẩn bị dữ liệu

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

symbol	market	timeIndicator	minTimestamp	maxTimestamp
SRN/BTC	huobipro	2018-12-13 08:26:00 UTC	1544689603614	1544689615119
WTC/BTC	huobipro	2018-12-13 08:26:00 UTC	1544689570921	1544689590036
EOS/PAX	binance	2018-12-13 08:26:00 UTC	1544689566328	1544689618905
EKT/BTC	huobipro	2018-12-13 08:26:00 UTC	1544689562604	1544689611453
NEO/USDT	huobipro	2018-12-13 08:26:00 UTC	1544689561044	1544689618173
OMG/BTC	bitfinex2	2018-12-13 08:26:00 UTC	1544689588624	1544689588624
XRP/BTC	binance	2018-12-13 08:26:00 UTC	1544689568501	1544689619519
BTG/BTC	binance	2018-12-13 08:26:00 UTC	1544689577770	1544689577770
HB10/USDT	huobipro	2018-12-13 08:26:00 UTC	1544689569087	1544689617665
ICX/BTC	huobipro	2018-12-13 08:26:00 UTC	1544689563889	1544689614203

openPrice	closePrice	highPrice	lowPrice	volume
1.331e-05	1.325e-05	1.331e-05	1.331e-05	0.0085977935000000009
0.00026948	0.00027049	0.00027049	0.00027049	5.3997e-06
1.9152	1.9152	1.9207	1.9207	204.88715399999998
1.24e-06	1.25e-06	1.25e-06	1.25e-06	0.0109388594
5.86	5.85	5.87	5.87	2191.216003
0.00036052	0.00036052	0.00036052	0.00036052	0.026111466704516802
8.898e-05	8.896e-05	8.9e-05	8.9e-05	1.1474145999999998
0.003396	0.003396	0.003396	0.003396	0.030564
0.2421	0.2419	0.2421	0.2421	56.687566
6.607e-05	6.625e-05	6.625e-05	6.625e-05	0.005767279581

Mô tả dữ liệu

Với 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

Cơ sở lý thuyết

Cây hồi quy và phân loại

Cây hồi quy và phân loại (CART) là một cây quyết định nhị phân được đề xuất bởi Breima [**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 cấn chọn ra một thuộc tính và một giá trị sao cho giảm được "nhiễu" 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).

Các luật tách thường dùng

Ta có thể chia thành 2 loại theo:
Đối với Cây hồi quy

- 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 đếu 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).

Đối với Cây phân loại

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

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 $S_1, S_2, \ldots, S_t$, 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.

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, sau đó tỉa các nhánh đưa cây về độ sâu đã chọn.

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.

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.

Rừng ngẫu nhiên

Tuy trực quan, các cây quyết định đơn lẻ thường có phương sai cao khi tập kiểm tra khác với tập huấn luyện dẫn đến kết quả dự đoán kém trên dữ liệu thực tế. Mục đích của rừng ngẫu nhiên nhằm giảm overfitting. Các bước thực hiện của rừng ngẫu nhiên gồm[**1**]:

- Bootstrap: Chọn lựa các dữ liệu ngẫu nhiên .
- Chọn lựa các thuộc tính: chọn các thuộc tính ngẫu nhiên đối với mỗi dữ liệu vừa tạo.
- Xây dựng cây.
- Tính toán tỉ lệ lỗi của mỗi cây dựa trên dữ liệu kiểm thử không nằm trong dữ liệu vừa tạo.
- Tổng hợp các cây.

Lớp tích chập trong mô hình mạng nơron tích chập (Convolution neural network)

Mô hình mạng nơron truyền thống với nhiều lớp truyền ngược có thể học được dữ liệu phức tạp, nhiều chiều. Tuy nhiên khi sử dụng nhiều lớp kết nối đầy đủ(fully connected) với đầu vào dạng véc-tơ với kích thước lớn gây ra hiện tượng overfitting do quá nhiều tham số (high-variance) làm bùng nổ mô hình liên quan trực tiếp tới vấn đề hiệu năng, bộ nhớ, khả năng tính toán của phần cứng. Để tránh hiện tượng này, ta có thể xây dựng bộ trích xuất đặc trưng trước khi đưa vào các lớp kết nối đầy đủ để giảm số lượng trọng số này. Một cách tổng quát hơn ta có thể xây dựng những lớp gồm các bộ lọc để giảm được số chiều của đầu vào trước đó, trong quá trình huấn luyện mô hình, lớp này có khả năng tự điều chỉnh các trọng số của các bộ lọc tại bước truyền ngược. Đây chính là lý do chính của lớp tích chập, kiến trúc của phép tích chập gồm các tham số:

- Kích thước của bộ lọc (cửa sổ)
- Hệ số lề (padding)
- Hệ số bước trượt (stride)

Chương 5

Các khái niệm cơ bản có liên quan tới đề tài

Các khái niệm về tài chính

Tính thanh khoản (Liquidity)

Khái niệm về tính thanh khoản dùng để chỉ mức độ mà một tài sản có thể được mua hoặc bán trên thị trường mà không làm ảnh hưởng nhiều đến giá thị trường. Khái niệm tính thanh khoản được chia thành 2 loại: tính thanh khoản thị trường (liquid market) và tính thanh khoản về tài sản (liquid asset). Thị trường có tính thanh khoản cao ám chỉ rằng trong thị trường thường xuyên có các nhà đầu tư sẵn sàng giao dịch. Một tài sản có tính thanh khoản cao đồng nghĩa với việc tài sản đó có thể chuyển đổi sang tiền mặt một cách dễ dàng. Đối với thị trường tiền mã hóa, để so sánh tính thanh khoản giữa các sàn trong cùng một thời điểm hoặc tính thanh khoản của một sàn tại những thời điểm khác nhau có 3 yếu tố quan trọng:

- Lượng đồng giao dịch trong ngày.
- Số lượng lệnh mua/bán dựa trên danh sách lệnh (order book) được công khai dựa theo các sàn như Coinbase Pro [5], Binance, Bittrex, ...
- Lượng chênh lệch giữa giá yêu cầu của bên bán và giá đặt của bên mua (bid/ask spread).

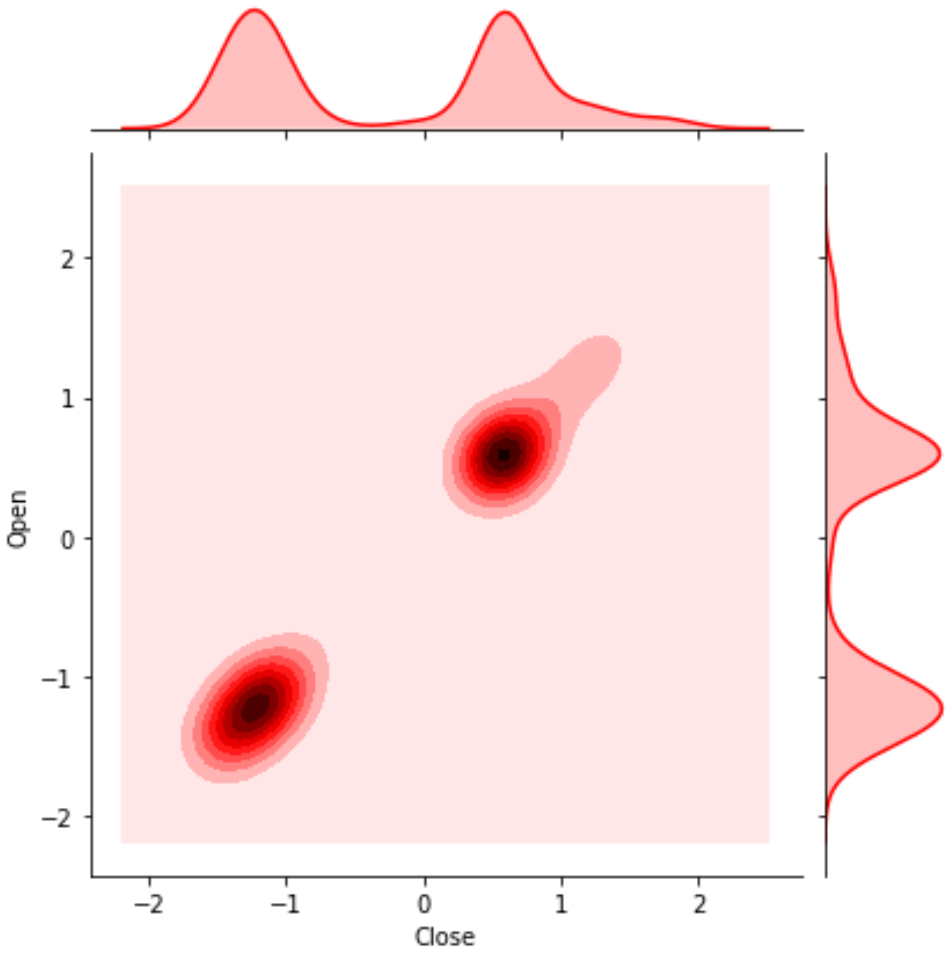
Nhiễu (Noise)

Khái niệm nhiễu có quan hệ đối lập với khái niệm thông tin (information), cụ thể với dữ liệu giá cung cấp đầy đủ thông tin, việc dự đoán dễ dàng và ngược lại với dữ liệu có nhiễu cao do bị ảnh hưởng bởi các yếu tố khác như phần đề cập tại phần 2.1. Nhiễu khiến những dự đoán của các nhà đầu tư không được hoàn hảo, điều này dẫn thị trường có khả năng lưu động[3].

Các khái niệm về xác suất

Hàm mật độ xác suất (Probability density function)

Với các phiên giao dịch có các thành phần như giá mở, giá đóng, số lượng đồng giao dịch,..., ta có thể coi như các biến ngẫu nhiên liên tục tương ứng. Khái niệm hàm mật độ xác suất trong văn cảnh trên được hiểu như một hàm gồm các tham số thể hiện được mật độ phân bố của các biến ngẫu nhiên.



Hình 5.1 – Phân phối biên giá mở/đóng dữ liệu đã xử lý

Hình 5.1 thể hiện mật độ của phân phối đồng thời giữa giá đóng và giá mở của các khối nến được biểu diễn dưới dạng $p_{data}(Open, Close)$.

Hàm phân phối biên (Marginal distribution)

Với dữ liệu liên tục như trên, hàm phân phối biên đối với giá mở được biểu diễn dưới dạng: $p_{data}(Open) = \int_y p_{data}(Open, Close = y)dy = \int_y p_{data}(Open \mid Close = y)p_{data}(Close = y)dy$
Một cách trực quan, hàm phân phối trên được biểu diễn bởi đường biên bên trái Hình 5.1

Nhiễu trắng (White noise)

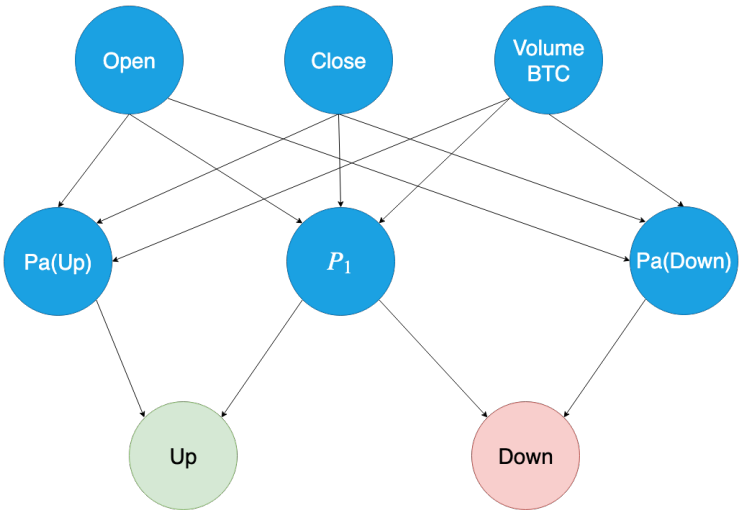
Nhiễu trong dữ liệu như phần đề cập trong phần 5.1.2 có thể được giảm thiểu bằng cách tìm hàm phân phối của nhiễu bằng thống kê, nếu phân phối của nhiễu có dạng phân phối chuẩn với trung bình là 0, nhiễu này được gọi là nhiễu trắng Gauss (white Gaussian noise). Việc giảm thiểu nhiễu trong dữ liệu làm mô hình trở nên dễ tìm được mẫu đặc trưng (pattern) hơn.

Biến ẩn (Latent variable)

Biến ẩn được hiểu theo cách trừu tượng là biến không thể quan sát trực tiếp[2, trang 264] mà được suy luận từ biến quan sát được trong dữ liệu. Cụ thể hơn, với dữ liệu là giá của 100 ngày đầu một mô hình có khả năng tìm được quan hệ giữa giá ngày thứ 50 phụ thuộc nhiều vào giá ngày thứ 49 hơn so với ngày thứ 99, mô hình này được gọi là mô hình biến ẩn (latent variable model) với quan hệ được biểu diễn bằng phép toán có giá trị được lưu trong các biến ẩn.

Mô hình đồ thị có hướng (Directed graphical model)

Trong mô hình đồ thị có hướng hay mạng Bayes (Bayes network) việc suy diễn từ các trạng thái trước sang các trạng thái sau. Cụ thể với mô hình được được trực quan theo như Hình



Hình 5.2 – Mô hình mạng Bayes

5.2, một giao dịch BTC/USD vào 6 giờ sáng ngày 31/7/2017 có giá mở là 2439.97$, giá đóng là 2415.19$, lượng giao dịch là 138.82 đồng BTC với xu hướng giao dịch tiếp theo có xác suất được kí hiệu là: $P(Up \,|\, Open = 2727.26, Close = 2740.01, VolumeBTC = 385.41)$ với xác suất đồng thời của giao dịch và được tính:

$$\begin{aligned} & p(Up, Open = 2727.26, Close = 2740.01, VolumeBTC = 385.41) \\ = {} & p(Open = 2727.26) \cdot p(Close = 2740.01) \cdot p(VolumeBTC = 385.41) \\ \cdot {} & p(Pa(Up) \mid Open = 2727.26, Close = 2740.01, VolumeBTC = 385.41) \\ \cdot {} & p(P_1 \mid Open = 2727.26, Close = 2740.01, VolumeBTC = 385.41) \\ \cdot {} & p(Up \mid Pa(Up), P_1) \end{aligned}$$

Một cách tổng quát xác suất đồng thời của giao dịch và xu hướng tăng giảm về giá của giao dịch tiếp theo được biểu diễn dưới dạng: $p_\theta(x_1, x_2, \ldots, x_M) = \prod_{i=1}^{M} p_\theta(x_j, Pa(x_j))$ với $Pa(x_j)$ giá trị của nút mạng trước đó (parent variable) của x_j.

Suy luận biến phân (Variational inference)

Phương pháp suy luận biến phân được sử dụng trong mô hình đồ thị có hướng[4] với các biến ẩn z và các quan sát x với mục tiêu ước lượng được phân bố của x được xấp xỉ bằng phân phối $Q(Z)$: $P(X \mid Z) \approx Q(Z)$ với $Q(Z)$ là phân phối tiên nghiệm đơn giản hơn $P(Z|X)$.

Sử dụng độ đo bất đồng Kullback–Leibler nhằm thể hiện sự khác nhau giữa phân phối Q so với phân phối P:

$D_{KL}(Q \parallel P) \triangleq \mathbb{E}_{Q(Z)}[\log \frac{Q(Z)}{P(Z \mid X)}]$

Sử dụng luật Bayes: $D_{KL}(P \parallel Q) = \mathbb{E}_{Q(Z)}[\log \frac{Q(z)}{P(z,x)} + \log P(x)]$
hay:
$\log(P(x)) = D_{KL}(P \parallel Q) - \mathbb{E}_{Q(z)}[\log \frac{Q(z)}{P(z,x)}] = D_{KL}(P \parallel Q) + \mathcal{L}(Q)$

Kiến trúc các mô hình đã tham khảo

Rừng ngẫu nhiên (Random forest)

Máy học véctơ hỗ trợ (Support vector machines)

Giải thuật Máy học véctơ hỗ trợ

Mô hình Variational autoencoder

Chương 6

Thí nghiệm với mô hình

Phân tích hiện thực mô hình tham khảo

Xử lý dữ liệu bài toán

Yêu cầu của bài toán là dự đoán giá của các đồng với nhau. Bài toán có thể chia ra làm 2 loại như sau:

- Phân loại giá tăng hoặc giảm trong một khoảng thời gian kế tiếp.
- Dự đoán giá của các đồng xác định trong phạm vi liên tục tại thời gian kế tiếp.

Các mô hình đã tham khảo thuộc bài toán học có giám sát, dữ liệu được chia thành 3 ph

Áp dụng dữ liệu vào mô hình rừng ngẫu nhiên

Đề tài có sử dụng thư viện hỗ trợ scikit-learn để xây dựng rừng ngẫu nhiên. Rừng ngẫu nhiên có đầu vào là dữ liệu ở dạng bảng với mỗi thuộc tính (mỗi cột) có thể ở dạng loại rời rạc (category) hoặc liên tục (numerical). Khi áp dụng dữ liệu thô, với mỗi quan sát là một dòng thuộc bảng và nhãn là giá lên hoặc xuống, đối với giá trong 1 phút sau ta có thể phân ra 2 loại tăng hoặc giảm. Tuy nhiên với mỗi quan sát trước có thể phụ thuộc vào nhiều quan sát trước đó, việc sử dụng các cây CART chỉ với một giao dịch là không phù hợp,ta có thể thêm các thuộc tính như giá trung bình trong 1 giờ trước đó, trong một ngày trước đó, ...

Áp dụng dữ liệu vào mô hình mạng nơron tích chập

Dề tài có sử dụng thư viện hỗ trợ keras trên nền tensorflow một thư viện mã nguồn mở có hỗ trợ khả năng tính toán của các bộ xử lý đồ họa. Trong bước tiền xử lý dữ liệu để đưa vào trong mạng có sử dụng mỗi cửa sổ trượt làm một ảnh một chiều với số kênh là số thuộc tính của mỗi giao dịch, độ dài của mỗi ảnh được định nghĩa trước là số giao dịch liên tục. Nhãn của những ảnh này là xu hướng tăng hoặc giảm của giao dịch cuối cùng với mỗi ảnh.
Khi dữ liệu được đưa vào mạng nơron cần được chuẩn hóa để tránh hiện tượng các nơron không cập nhật được khi hàm kích hoạt có họ ReLU hoặc khó kích hoạt khi các hàm kích hoạt phi tuyến khác như hàm sigmoid hoặc hàm tanh.
Các bước **chuẩn hóa dữ liệu** được mô tả như sau:

- Các thuộc tính về thời gian đổi về dạng số nguyên theo chuẩn UNIX.Các số này khá lớn nên được chuẩn hóa dạng logarit, sau đó chuẩn hóa theo standard score.
- Các thuộc tính còn lại như lượng giao dịch, giá mở, giá đóng theo dạng standard score

Dữ liệu được đưa vào mạng nơron tích chập với ý tưởng chính như sau:

Data: 507918 giao dịch bitcoin/Yên
Result: Mô hình với các tham số, độ chính xác khi test
Khởi tạo:

- Chuẩn hóa dữ liệu theo từng cột.
- Tập train: 480000 bộ cửa sổ đầu, mỗi cửa sổ chứa 100 giao dịch liên tục nhau, mỗi cửa sổ liên tiếp nhau 1 giao dịch.
- Các tham số của mô hình tại mỗi lớp.
- $\ell = \infty$

while *Khi số lượng bộ test nhỏ hơn 1024* **do**
    Tập test : sau 100 giao dịch tiếp theo so với tập train lấy 1024 bộ cửa sổ liên tiếp.;
    Lấy tập test làm tập kiểm định; Tính ℓ là Loss của mô hình đối với tập kiểm định;
    if $\ell_1 < \ell$ **then**
        Cập nhật tham số;
        $\ell = \ell_1$;
    end
    Tập train: tăng kích thước của tập train thêm lên 1024 cửa sổ tiếp theo.
end

Algorithm 1: Áp dụng kĩ thuật rolling window

Khi áp dụng standard score, với mỗi thuộc tính có giá trị trung bình về 0 và phương sai về 1. Điều này ảnh hưởng tốt cho mạng nơron tích chập khi hội tụ nhanh và khó bị 'kẹt lại' ở những điểm thung lũng hơn.

Kết quả

Các số liệu kết quả trong phần này được lấy từ các lần thực nghiệm huấn luyện trên google colab, một nền tảng dịch vụ có hỗ trợ phần cứng và thư viện miễn phí, với dữ liệu đầu vào từ 2012-01-01 đến 2018-11-11 và độ chính xác theo bài toán phân loại lên hoặc xuống.

Mô hình rừng ngẫu nhiên Kết quả cho độ chính xác là 53%.
Mô hình mạng nơron tích chập Kết quả cho độ chính xác là 71% được tính bằng trung bình của 65 lần kiểm thử cuối.

Tài liệu tham khảo

[1] Prableen Bajpai. *Countries Where Bitcoin Is Legal & Illegal*. https://www.investopedia.com/articles/forex/041515/countries-where-bitcoin-legal-illegal.asp. cited May 2019.

[2] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. URL: http://research.microsoft.com/en-us/um/people/cmbishop/prml/.

[3] Fischer Black. "Noise". In: *Journal of Finance, Volume 41* (1986). DOI: https://doi.org/10.1111/j.1540-6261.1986.tb04513.x.

[4] Ghahramani Z. Jaakkola T.S. Jordan M.I. *An Introduction to Variational Methodsfor Graphical Models*. Springer, 1999. DOI: https://doi.org/10.1023/A:1007665907178.

[5] Trade Volume. *Cryptometer Live Order Book*. https://www.cryptometer.io/data/coinbase_pro/btc/usd. cited April 2019.