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 10 tháng 10 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 Võ Trọng Thư.

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 nhanh 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 chúng 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

Do trên thị trường hiện nay có nhiều sàn giao dịch khác nhau, sàn để nghiên cứu cần phải có tính thanh khoản cao với số lượng giao dịch nhiều, minh bạch về lịch sử giao dịch, ngoài ra sàn phải cung cấp lịch sử giá giữa các cặp đồng với nhau. Thông qua tìm hiểu sàn Binance được thành lập vào tháng 7/2017 là một trong những sàn uy tín với tính thanh khoản cao, số lượng giao dịch luôn nằm trong top 5 những sàn giao dịch trên thế giới, sàn cung cấp api để tra giá giao dịch, lịch sử giao dịch. Ngoài ra Binance còn cung cấp cho một tài khoản có thể tạo tối đa 200 tài khoản phụ (tính năng chỉ áp dụng cho tài khoản đặc biệt theo chính sách của sàn), điều này giúp việc so sánh các chiến lược giao dịch với nhau dựa trở nên dễ dàng hơn khi giao dịch thực tế. Chính vì những lí do trên chúng tôi lựa chọn sàn Binance để nghiên cứu và triển khai các chiến lược trong tương lai.
Về đối tượng đồng mã hóa, hiện nay có hơn 1600 loại đồng mã hóa nên việc lựa chọn các cặp đồng mã hóa được chúng tôi xếp theo lượng giao dịch trong ngày. Bitcoin và Ethereum là hai đồng có sức mua, bán cao nhất trên sàn Binance tính theo hai cặp tương ứng BTC/USDT và ETH/USDT. Ngoài ra sàn còn cung cấp đồng Binance (BNB) nằm trong top 10 khi xét về khối lượng giao dịch, giao dịch các cặp khi có đồng này phí giao dịch sẽ được giảm xuống 25%. Việc lựa chọn 2 đồng cơ bản là BTC, ETH ứng với 2 cặp sẽ được đề cập trong phần nghiên cứu. Trong tương lai khi so sánh thực nghiệm theo lợi nhuận đối với mỗi chiến lược sẽ bổ sung đồng BNB.

Bố cục luận văn

Bố cục luận văn với nội dung tác giả trình bày được chia thành các phần sau đây:

- **Chương 1:** Giới thiệu đề tài.
  Khái quát về vấn đề liên quan đến các chiến lược giao dịch và sự cần thiết của một hệ thống học máy trong mô hình dự đoán.
- **Chương 2:** Tổng quan về lĩnh vực nghiên cứu.
  Tổng quan về thị trường tiền mã hóa, nghiệp vụ giao dịch tiền mã hóa.
- **Chương 4:** Các công trình liên quan.
  Đưa ra một số các công trình xu hướng tăng giảm về giá của các đồng tiền mã hóa.

- **Chương 5:** Chuẩn bị dữ liệu.
  Trình bày quá trình thu thập dữ liệu, quá trình tiền xử lý dữ liệu cho bài toán học có giám sát.
- **Chương 6, 7:** Các phương pháp nghiên cứu, quá trình hiện thực.
  Trình bày cách ứng dụng các chiến lược và các mô hình học máy được sử dụng trong việc dự đoán giá.
- **Chương 8:** Tổng kết.
  Kết quả đạt được, những hạn chế của các chiến lược mô hình đã được sử dụng và hướng phát triển hệ thống trong tương lai.

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 bị 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).

Giao dịch tiền mã hóa

Các sàn tiền mã hóa cung cấp các lệnh giao dịch cơ bản: mua, bán với những cặp (pair) đồng với nhau, ngoài ra còn có thêm phương thức mua, bán tiền ảo bằng tiền mặt thông qua sàn đóng vai trò như bên thứ ba.

Các chiến lược giao dịch

Hai chiến lược cơ bản được nghiên cứu và thực hiện trong đề tài như sau:

- Giao dịch cùng một loại cặp với nhau tại hai thời điểm khác nhau nhằm tăng số lượng đồng ban đầu.
- Giao dịch nhiều cặp với nhau theo một vòng dựa theo giá tại cùng một thời điểm. Hai chiến lược sẽ được mô tả chi tiết tại chương 3

Chương 3

Đánh giá thị trường thông qua hai chiến lược

Với dữ liệu được lấy từ sàn, có thể tạo một đánh giá thị trường với giao dịch ngắn hạn có tiềm năng hay không? Từ đó có thể tạo ra công cụ với khả năng dự đoán để tự động giao dịch không? Nhằm trả lời cho hai câu hỏi trên, trong chương này sẽ đề cập tới hai phần chính:

- Các chiến lược cơ bản được đề ra trong phần 2.4 và mô phỏng hai chiến lược trên dữ liệu đã có. Ứng dụng các mô hình học máy để dự đoán xu hướng giá.
- Đánh giá rủi ro của hai chiến lược thông qua giá có trước, từ đó nhận định tiềm năng của thị trường.

Các chiến lược giao dịch ngắn hạn

Chiến lược giao dịch cùng một loại cặp đồng

Khi giao dịch cùng một loại cặp đồng A/B theo thời gian khác nhau, đặt lệnh mua hay đổi đồng B để mua A khi tỷ giá A/B có xu hướng giảm ngược lại đặt lệnh bán khi tỷ giá có xu hướng tăng. Chiến lược này sẽ không hiệu quả khi giá ở mỗi phiên không chênh lệch nhau nhiều đặc biệt có trường hợp lỗ khi mỗi lần giao dịch sẽ mất tiền phí do bên sàn thu. Vì vậy chiến lược nói trên sẽ được thêm một ràng buộc là ngưỡng phí giao dịch ϵ và các biến:

- W_t^a: số đồng A quy ra B theo giá tại thời điểm t.

- W_t^b: số đồng B quy ra A theo giá tại thời điểm t.
- y_t: tỷ giá đồng A/B tại thời điểm t.
- a, b: số đồng A, số đồng B trong ví tại thời điểm đang xét.

Với t là thời điểm gần nhất giao dịch, xét tại thời điểm τ xảy ra sau đó, Việc đặt lệnh mua phải thỏa yêu cầu sau: $W_\tau^a > W_t^a$ hay $\frac{b}{y_\tau}(1-\epsilon) > \frac{b}{y_t}$ hay $y_\tau < y_t(1-\epsilon)$
Tương tự, việc đặt lệnh bán phải thỏa yêu cầu sau: $W_\tau^b > W_t^b$ hay $a(1-\epsilon)y_\tau > ay$ hay $y_\tau(1-\epsilon) > y_t$

Việc mô phỏng chiến lược này cần tuân theo ràng buộc của sàn như sau: đơn vị tối thiểu là 0.000001 BTC và 0.01 USDT ví dụ muốn mua cặp đồng trên khi có 1.234 USDT với số lượng tối đa phí giao dịch sẽ là 0.1% với giá khớp lệnh là 8000 số lượng USDT còn lại trong ví là 0.004 số lượng giao dịch sẽ là 1.23 USDT, số lượng BTC nhận vào ví là $1.23/8000 * (1 - 0.1/100) = 0.00015359625$.

Khi thực hiện khảo sát chiến lược này với dữ liệu thu được trên sàn Binance với cặp đồng BTC/USDT trong khoảng thời gian từ 2017-08-17 đến 2019-09-01, giả sử số tiền trong ví ban đầu là 1.0 USDT các ngưỡng phí ϵ được thay đổi cho ra kết quả được thống kê trong bảng sau:

Table 3.1 – Nonlinear Model Results

Thời gian bắt đầu	Ngưỡng phí(%)	Số lần giao dịch	Tổng USDT đầu	Tổng USDT sau
2017-08-28 13:00:00	0.1	129	4221.04	2193.22
2017-08-28 13:00:00	0.2	129	4221.04	2290.84
2017-08-28 13:00:00	5	17	4221.04	6632.89
2017-08-28 13:00:00	10	7	4221.04	6133.60
2017-12-11 01:00:00	0.1	80	14975.03	9721.09
2017-12-11 01:00:00	0.2	82	14975.03	9884.65
2017-12-11 01:00:00	5	22	14975.03	17373.62
2017-12-11 01:00:00	10	6	14975.03	13452.00

Chiến lược trên tuy đảm bảo khi đổi giữa hai lần xen kẽ nhau, số đồng được tăng lên,

nhìn thấy của chiến lược này là không biết trước giá của phiên giao dịch tiếp theo. Tuy nhiên cụ thể với ngày xxxx giá đạt ngưỡng cao nhất khi đó theo chiến lược, đổi hết đồng BTC sang thành USDT, tiếp theo giá giảm đều và qua ngưỡng phí và tiếp tục giảm, khi

này số đồng USDT đã được chuyển sang BTC số lượng đồng BTC so với thời điểm trước khi bán ban đầu là nhiều hơn, khi giá tiếp tục giảm lệnh mua sẽ không được thực hiện do đã hết đồng USDT phải chờ đến khi giá tăng so với lần mua tại ngày xxxxx. Điều này dẫn tới việc tính theo giá USDT tổng giá trị BTC là giảm từ 14975.03 USDT xuống 13452 USDT. Để giảm rủi ro này, ta có thể dự đoán xu hướng giá đóng của phiên giao dịch kế tiếp tức xxxxx, nếu giá có xu hướng giảm, lệnh mua sẽ được giữ lại tới khi giá có xu hướng tăng. Đây chính là ý tưởng chính cho việc hình thành bài toán dự đoán xu hướng giá ngắn hạn, các mô hình sẽ được học từ các phiên giao dịch trước và dự đoán xu hướng giá của phiên giao dịch sau. Việc đánh nhãn cho dữ liệu sẽ được trình bày trong mục 5.3.

Chiến lược giao dịch nhiều loại cặp đồng

Với 3 đồng là A, B, C, việc chuyển đồng A chuyển sang đồng B, chuyển đồng B sang đồng C và cuối cùng chuyển lại đồng C sang đồng A tạo thành một vòng lặp, việc đồng A tăng lên hoặc giảm đi có thể xảy ra. Trong cùng một phiên giao dịch, việc tìm vòng lặp như trên sao cho số lượng đồng A được tăng lên so với trước đòi hỏi các lần chuyển đổi giữa các cặp diễn ra liên tục và có thứ tự nói cách khác tất cả các lần giao dịch đều phải được hoàn thành, đây cũng là nhược điểm của chiến lược này vì trong khi biết giá của các giao dịch trước, giá của các cặp sẽ đổi khi thực hiện giao dịch đòi hỏi giao dịch phải diễn ra nhanh. Lấy ví dụ ở thời điểm lúc 8 giờ ngày 04/01/2018 xét 3 cặp đồng là BTC/USDT, ETH/BTC, ETH/USDT có tỉ giá tương ứng là 15172.12, 0.060893, 920.08 với phí giao dịch cho mỗi lần trao đổi mặc định là 0.1% đối với sàn Binance, với 1.0 USDT lần lượt đổi các cặp là USDT sang ETH, ETH sang BTC và BTC sang USDT số đồng USDT thu về trong ví là 1.0011162574497365 với giả thiết ở mỗi giao dịch đều đổi hết (bỏ qua ràng buộc về số đồng tối thiểu). Trong ví dụ này với 3 đồng trên ta có thể thấy một cách trực quan rằng số đồng USDT tăng sau một vòng chuyển đổi. Việc tìm các vòng chuyển đổi tại mỗi thời điểm như trên có thể được mô hình hóa bằng bài toán như sau:

Data: T phiên giao dịch gồm tổng quát $\frac{N(N-1)}{2}$ cặp tương ứng với N đồng
Result: Số lần xuất hiện vòng có xu hướng làm tăng số lượng đồng ban đầu
Khởi tạo:

- Đồ thị hai phía đầy đủ.
- Tensor M kích thước TxNxN chứa giá trị của T phiên giao dịch.
- Phí giao dịch ϵ.
- t = 0

for *t trong T* **do**
  Cập nhật trọng số của đồ thị
  Khi không có giao dịch giữa hai đồng A/B trọng số cạnh $d(A, B) \leftarrow 1e - 20$;
  $d(B, A) \leftarrow 1e - 20$
  for *u, v trong* M_t **do**
    $d(u, v) \leftarrow -log(d(u, v)) - log(\epsilon)$
  end
  ▷ Chuyển giá sang giá trị logarit
  if *Đồ thị có trọng số âm* **then**
    $t \leftarrow t + 1$
    Tìm chu trình nhỏ nhất không lặp.
  end
end

Algorithm 1: Tìm số lượng phiên giao dịch có thể làm tăng số đồng ban đầu.

Thống kê với phí giao dịch mặc định là 0.1% đối với sàn Binance từ 2018-01-01 đến 2019-09-21 gồm 15000 phiên giao dịch với khoảng thời gian mỗi phiên là 1 giờ xét trên 3 đồng BTC, ETH, USDT được thu thập từ 3 cặp: USDT/ETH, ETH/BTC, BTC/USDT đưa ra 75 lần đồ thị tồn tại chu trình âm. Khi thêm đồng BNB đồ thị với 4 loại đồng gồm 6 cặp, con số này đạt 1027 lần.

Với chiến lược đổi trên nhiều đồng này, khi tăng số lượng đồng, việc giao dịch theo chiến lược này trở lên khó khăn hơn vì khi duyệt chu trình, tất cả các cạnh đều phải đi qua, nói cách khác các lần đổi đều phải hoàn thành. Tuy nhiên khi đổi, giá của hai đồng sẽ không giữ nguyên như giá hiện tại, việc khớp giá sẽ khó xảy ra, vậy nên bổ sung mô hình dự đoán xu hướng giá của phiên giao dịch tiếp theo có thể hỗ trợ thêm cho chiến lược này.

Rủi ro và tiềm năng của thị trường

Trong phần 3.1.1 với giao dịch trên một cặp đồng, rủi ro và tiềm năng của chiến lược này phụ thuộc vào thị trường và ngưỡng phí cho trước, lệnh giao dịch được đưa ra dựa trên giá hiện tại, bổ sung cho chiến lược này khi biết chính xác xu hướng giá cần và

Với giao dịch trên nhiều đồng các thông số

Chương 4

Các công trình liên quan

Trong chương này, chúng tôi sẽ trình bày các công trình liên quan tới việc sử dụng các mô hình học máy trong việc đoán xu hướng giá.
Isaac Madan, Shaurya Saluja và Aojia Zhao đã ứng dụng các mô hình học máy để dự đoán giá của đồng Bitcoin với kết quả có độ chính xác vào khoảng 50-55% tín hiệu giá tăng hay giảm với khoảng thời gian tiếp theo là 10 phút. Nhóm đã hiện thực việc thu thập thông qua api của sàn Coinbase và sàn OKCoin và cho ra dữ liệu gồm 25 đặc trưng liên quan tới đồng Bitcoin. Nhón tác giả đã hiện thực các mô hình như SVM, rừng quyết định, và Binomial GLM

Chương 5

Dữ liệu

Thu thập dữ liệu

Hiện nay đa số các sàn giao dịch lớn như Binance, Huobipro, OKCoin đều cung cấp api hỗ trợ cho phép xem lại các OrderBook, tỷ giá giao dịch các phiên trước đó, đặt lệnh giao dịch. Như đã đề cập ở chương 1, trong phạm vi sàn nghiên cứu của đề tài, chúng tôi chọn sàn Binance và các đồng cơ bản là BTC, ETH, USDT, BNB để nghiên cứu; Thông qua tìm hiểu, có hai thư viện hỗ trợ thu thập thông qua api trên sàn biannce là python-binance do sàn viết và ccxt (đã được Binance chứng nhận) đều được viết bằng ngôn ngữ python. Ccxt với khả năng hỗ trợ hơn 120 sàn khác nhau, hỗ trợ với nhiều ngôn ngữ lập trình, chính vì vậy ccxt được chọn làm thư viện chính để thu thập dữ liệu và tạo các lệnh giao dịch để có thể mở rộng đề tài trong tương lai.
Để dễ hình dung về chức năng của api, api được ccxt chia thành hai loại là:

- Public api: hỗ trợ lấy tickers, OrderBook; tỉ giá cặp tại thời điểm trước; các giao dịch trong khoảng thời gian trước đó.
- Private api: hỗ trợ tạo, hủy lệnh giao dịch; lấy các lệnh giao dịch trước đây; xem số đồng trong các ví của tài khoản sở hữu api này.

Quá trình chuẩn bị dữ liệu được thực hiện thông qua public api.
Khác với sàn giao dịch truyền thống (sàn chứng khoán), sàn giao dịch tiền mã hóa hoạt động liên tục vì vậy việc chia phiên giao dịch được sàn định nghĩa theo các khoảng thời gian cụ thể theo mặc định như 1 phút, 1 giờ, 1 ngày,... Điều này giúp cho người tham gia có thể biết tỷ giá trong các phiên giao dịch trước. Tuy nhiên trong phiên giao dịch có thể gồm nhiều các giao dịch với giá, số lượng đồng khác nhau. Nhằm dễ thống kê, public api cung cấp để lấy giá OHLCV(Open, High, Low, Close, Volume) tương ứng với giá mở sàn;

giá cao nhất, thấp nhất trong phiên; giá đóng sàn và lượng đồng trao đổi. Đây là các đặc trưng cơ bản cho một phiên giao dịch. Dữ liệu sẽ được lưu dạng bảng với mỗi dòng tương ứng với một phiên giao dịch gồm thời gian mở phiên theo dạng Unix time, và 5 đặc trưng nêu trên, các đặc trưng khác sẽ được trình bày tóm tắt như sau:

- Tổng số giao dịch mua; bán.
- Số lượng đồng mua; bán trung bình
- Độ lệch chuẩn số lượng đồng mua; bán.
- Giá trung bình của các giao dịch mua, bán và độ lệch chuẩn tương ứng.

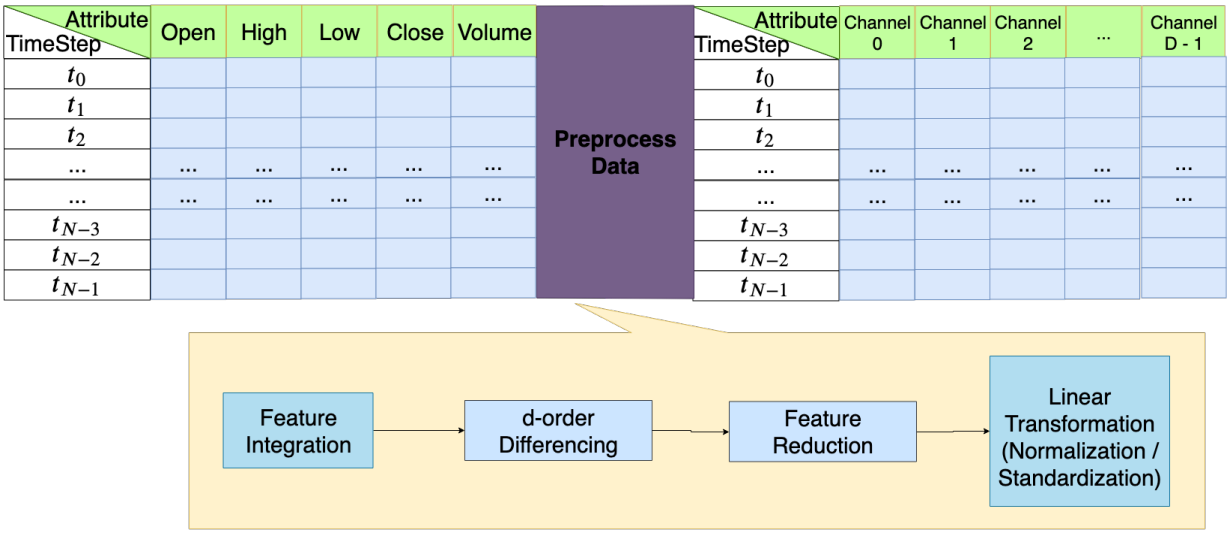
Trong trường hợp một phiên giao dịch không có giao dịch mua nào bán hoặc không có giao dịch bán, tổng giao dịch sẽ bằng 1 và các trường trung bình, độ lệch chuẩn sẽ là 0. Trong thực nghiệm, hiếm khi xảy ra trường hợp này, cụ thể với khoảng thời gian từ 2017-08-17 đến 2019-09-01 có 16 phiên giao dịch như trên trong tổng số 17875 phiên giao dịch với khoảng thời gian phiên là 1 giờ cho cặp BTC/USDT.

Tiền xử lý dữ liệu

Sau công đoạn thu thập dữ liệu thô từ sàn, các phiên giao dịch được vectơ hóa và sắp xếp theo thời gian thành bảng được lưu dạng csv. Bước tiền xử lý dùng dữ liệu trên với các bước được tóm tắt theo hình sau:

Thêm đặc trưng

Ngoài các đặc trưng được lấy trực tiếp từ sàn, Trong một phiên giao dịch với các giá High, Low chênh lệch nhau không rõ rệt khi thời gian phiên ngắn như 1 phút, 1 giờ; cần chọn thêm các đặc trưng như giá chênh lệch giữa High-Low và giá chênh lệch giữa Close-Open sẽ hiệu quả, đặc biệt khi chuẩn hóa z-score, min-max; hệ số chuẩn hóa được tăng lên giúp tăng độ lệch chuẩn đối với đặc trưng mới này khi so sánh các phiên giao dịch với nhau.



Hình 5.1 – Tiền xử lý dữ liệu

d order difference

Sử dụng 1d order difference, thể hiện sự chênh lệch giá hiện tại với giá trước.

Loại bớt đặc trưng

Ngoài các đặc trưng được lấy trực tiếp từ sàn, Trong một phiên giao dịch với các giá High, Low chênh lệch nhau không rõ rệt khi thời gian phiên ngắn như 1 phút, 1 giờ; cần chọn thêm các đặc trưng như giá chênh lệch giữa High-Low và giá chênh lệch giữa Close-Open sẽ hiệu quả, đặc biệt khi chuẩn hóa z-score, min-max; hệ số chuẩn hóa được tăng lên giúp tăng độ lệch chuẩn đối với đặc trưng mới này khi so sánh các phiên giao dịch với nhau.

Chuẩn hóa dữ liệu

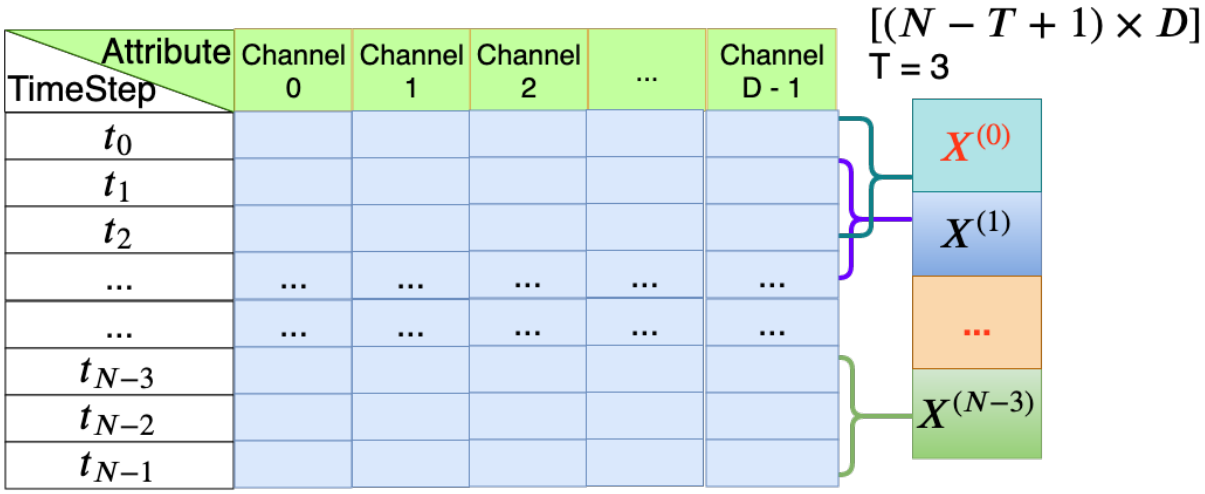
Với dữ liệu dạng bảng mỗi dòng tương ứng một phiên giao dịch, các dòng được sắp xếp với thời gian tăng dần. Tập dữ liệu được chia thành tập huấn luyện và tập kiểm thử. Khi chuẩn hóa dữ liệu tập huấn luyện tạo ra hệ số chuẩn hóa, hệ số này sẽ chuẩn hóa tập dữ liệu kiểm thử với giả thiết khi có tập huấn luyện, một giao dịch mới sẽ được chuẩn hóa theo hệ số trước đây. Điều này có hạn chế khi chuẩn hóa theo min-max với khoảng $[0, 1]$ hoặc $[-1, 1]$ giá trị của các giao dịch trong tập kiểm thử có thể vượt ngoài 1, để tránh

trường hợp này có thể xóa các dữ liệu bất thường này hoặc dùng phép chuẩn hóa khác như z-score:
$Z_{scale} = \frac{Z-\mu}{\sigma}$,
trong đó:

- z là giá trị trước khi chuẩn hóa.
- μ, σ lần lượt là giá trị trung bình, độ lệch chuẩn sau khi đã hiệu chỉnh.
- z là giá trị sau khi chuẩn hóa.

Đánh nhãn dữ liệu

Nhãn được chia thành hai loại là xu hướng tăng và xu hướng giảm của giá đóng phiên thời điểm hiện tại. Thống kê với dữ liệu BTC/USDT thời gian 1 giờ có 9051 nhãn xu hướng tăng và 8452 nhãn xu hướng giảm, trường hợp giá không đổi tại phiên giao dịch sau là không có. Công việc đánh nhãn được thực hiện sau khi gộp các phiên giao dịch thành các điểm dữ liệu như hình sau: Với các tham số:



Hình 5.2 – Đánh nhãn dữ liệu

- D: Số thuộc tính của mỗi phiên giao dịch
- T: Tổng số phiên giao dịch

Chương 6

Phương pháp nghiên cứu

Chương 7

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).

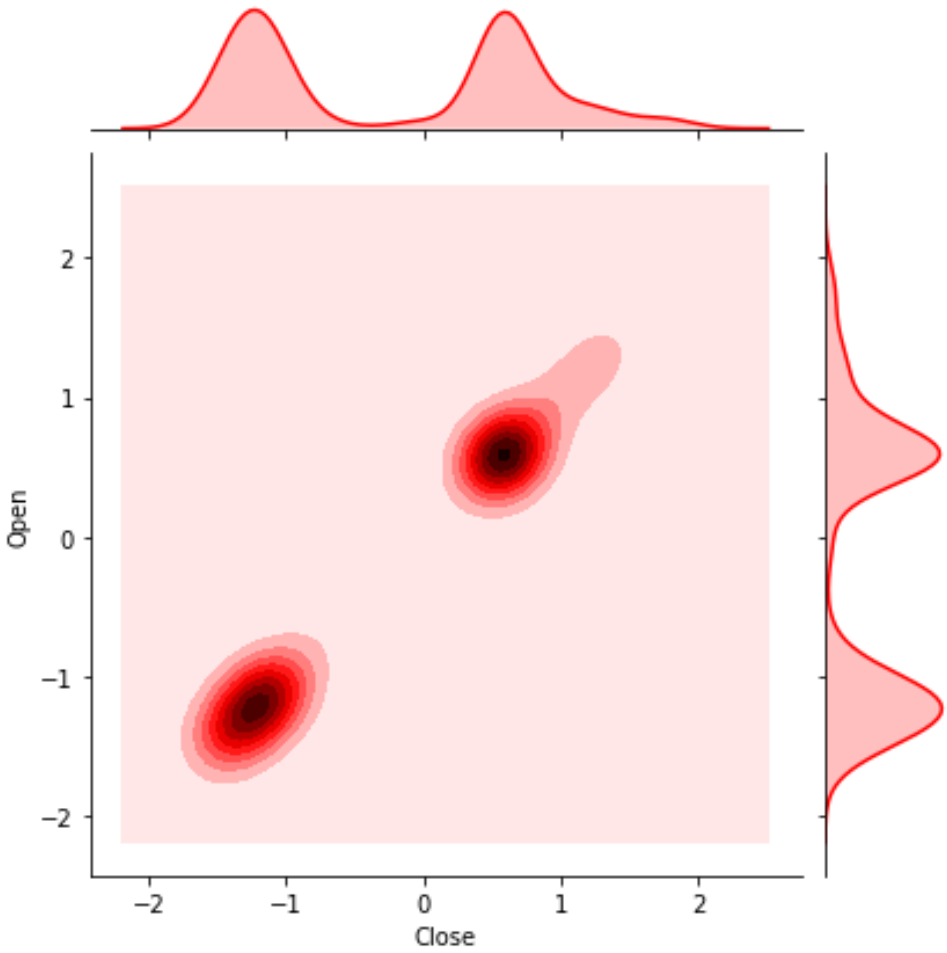
Nhiễu (Noise)

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

Hình 7.1 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 7.1

Nhiễu trắng (White noise)

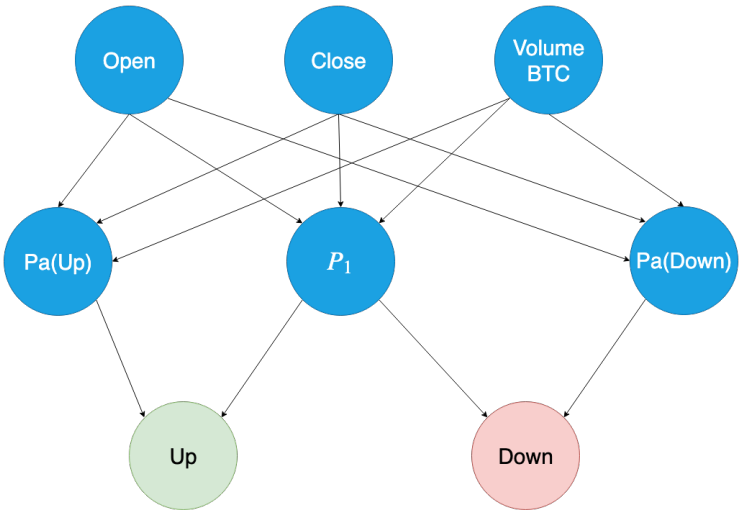
Nhiễu trong dữ liệu như phần đề cập trong phần 7.1.2 có thể được giảm thiểu bằng cách tìm hàm phân phối của nhiễu bằng thống kê, nếu phân phối của nhiễu có dạng phân phối chuẩn với trung bình là 0, nhiễu này được gọi là nhiễu trắng Gauss (white Gaussian noise). Việc giảm thiểu nhiễu trong dữ liệu làm 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 7.2 – Mô hình mạng Bayes

7.2, 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:
$\log(P(x)) = D_{KL}(P \parallel Q) - \mathbb{E}_{Q(z)}[\log \frac{Q(z)}{P(z,x)}] = D_{KL}(P \parallel Q) + \mathcal{L}(Q)$

Chương 8

Thí nghiệm với mô hình

Phân tích hiện thực mô hình tham khảo

Áp dụng dữ liệu vào mô hình rừng ngẫu nhiên

Đề tài có sử dụng thư viện hỗ trợ scikit-learn để hiện thực rừng ngẫu nhiên. Mô hình 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

Đề 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 2: 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 xxxxxxxxx và độ chính xác theo bài toán phân loại lên hoặc xuống.

Mô hình SVM Kết quả được tính bằng trung bình của 3 tháng cuối tương ứng với 2160 phiên liên tục. Kết quả được mô tả như hình sau:

		Dự đoán	
		Giảm	Tăng
Thực tế	Giảm	445	663
	Tăng	396	716

Mô hình hồi quy logistic Kết quả cho độ chính xác là 55.56%.

		Dự đoán	
		Giảm	Tăng
Thực tế	Giảm	606	443
	Tăng	517	594

Mô hình rừng ngẫu nhiên Kết quả cho độ chính xác là 54.54%.

		Dự đoán	
		Giảm	Tăng
Thực tế	Giảm	553	516
	Tăng	446	645

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

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. URL: https://papers.nips.cc/paper/2020-on-discriminative-vs-generative-classifiers-a-comparison-of-logistic-regression-and-naive-bayes.pdf.

[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.