**VIỆN NGHIÊN CỨU DỮ LIỆU LỚN VIN BIGDATA**

**DỰ ÁN KẾT THÚC MÔN HỌC**

**Mô hình Time Series ứng dụng trong bài toán dự đoán giá chứng khoán**

**Nhóm học viên:**

**Nguyễn Văn Sơn**

**Nguyễn Thế Hiển**

**Đoàn Quang Khải**

**Phạm Ngọc Đông**

**Môn học: Deep learning**

**Giảng viên hướng dẫn: TS. Cao Văn Lợi**

**Hà Nội, tháng 11 năm 2020**

**MỤC LỤC**

[LỜI MỞ ĐẦU 4](#_Toc58816107)

[CHƯƠNG 1: BÀI TOÁN DỰ ĐOÁN GIÁ CHỨNG KHOÁN 6](#_Toc58816108)

[1.1 Bài toán dự đoán giá chứng khoán 6](#_Toc58816109)

[1.1.1 Tổng quan về thị trường chứng khoán 6](#_Toc58816110)

[1.1.2 Bài toán dự đoán giá chứng khoán 7](#_Toc58816111)

[1.2 Mô hình Time Series 8](#_Toc58816112)

[CHƯƠNG 2: CÁC MÔ HÌNH ỨNG DỤNG 10](#_Toc58816113)

[2.1 MÔ HÌNH ARIMA 10](#_Toc58816114)

[2.2 Mô hinh LSTM 10](#_Toc58816115)

[2.2.1 Ý tưởng cốt lõi của LSTM 11](#_Toc58816116)

[2.3 Phương pháp kết hợp giữa ARIMA và LSTM 12](#_Toc58816117)

[2.3.1 Lý do kết hợp 12](#_Toc58816118)

[2.3.2 Ý tưởng kết hợp ARIMA và LSTM 13](#_Toc58816119)

[CHƯƠNG 3: MÔ HÌNH GIẢI QUYẾT BÀI TOÁN 14](#_Toc58816120)

[3.1 Mô hình giải quyết bài toán 14](#_Toc58816121)

[3.2 Các bước thực hiện 14](#_Toc58816122)

[3.2.1 Thu thập dữ liệu 14](#_Toc58816123)

[3.2.2 Tiền xử lý dữ liệu 15](#_Toc58816124)

[3.2.3 Xây dựng mô hình kết hợp ARIMA và LSTM 15](#_Toc58816125)

[CHƯƠNG 4: KẾT QUẢ THỰC NGHIỆM VÀ ĐÁNH GIÁ 16](#_Toc58816126)

[4.1 Kết quả thực nghiệm 16](#_Toc58816127)

[4.2 Đánh giá 17](#_Toc58816128)

[CHƯƠNG 5. KẾT LUẬN 18](#_Toc58816129)

[5.1 Kết luận 18](#_Toc58816130)

[5.2 Hướng phát triển tương lai 18](#_Toc58816131)

[TÀI LIỆU THAM KHẢO 19](#_Toc58816132)

**DANH MỤC BẢNG BIỂU**

**DANH MỤC HÌNH VẼ, ĐỒ THỊ**

[Hình 1.1 Đồ thị về chuỗi nhiệt độ trung bình theo tháng thể hiện yếu tố mùa vụ 9](#_heading=h.4d34og8)

[Hình 1.2 Đồ thị yếu tố xu hướng trong chuỗi thời gian của chuỗi giá 9](#_heading=h.2s8eyo1)

[Hình 2.1 Vòng lặp mạng RNN 11](#_heading=h.26in1rg)

[Hình 2.2 Cấu trúc chuỗi của RNN 12](#_heading=h.1ksv4uv)

[Hình 2.3 Cấu trúc chuỗi của LSTM 12](#_heading=h.44sinio)

[Hình 2.4 Đường trạng thái tế bào (cell state) 13](#_heading=h.z337ya)

[Hình 2.5 Cổng được kết hợp bởi sigmoid và phép nhân 13](#_heading=h.3j2qqm3)

[Hình 2.6 Chi tiết 1 tế bào tại trạng thái thứ t của LSTM 14](#_heading=h.4i7ojhp)

[Hình 3.1 Đồ thị giá chứng khoán trung bình của mã AAL 16](#_heading=h.qsh70q)

[Hình 3.2 Giá dự đoán và giá thực tế 19](#_heading=h.147n2zr)

[Hình 3.3 Giao diện Website dự đoán giá chứng khoán 19](#_heading=h.23ckvvd)

# LỜI MỞ ĐẦU

Thị trường chứng khoán Việt Nam ra đời từ những năm đầu của thế kỷ XXI và nhanh chóng trở thành một kênh đầu tư hấp dẫn, từ nhà đầu tư cá nhân nhỏ lẻ, cho đến các nhà đầu tư chuyên nghiệp, các tổ chức với nhiều phong cách đầu tư, sinh lời khác nhau. Và một điều tất yếu là họ luôn muốn đạt được lợi nhuận, mức sinh lời cao và dĩ nhiên đi kèm cùng việc đó là rủi ro tiềm ẩn cũng hề nhỏ. Do đó, việc dự đoán chính xác giá chứng khoán luôn được các công ty chú trọng đầu tư và đặt lên hàng đầu, từ phân tích kỹ thuật đến phân tích cơ bản để cho các nhà đầu tư có sách lược và phong cách đầu tư phù hợp nhất do mình. Nhưng liệu với một nền kinh tế với số liệu không minh bạch, hàng lang pháp lý còn nhiều sơ hở đề sự thao túng giá cổ phiếu luôn thay đổi hàng phút trên bảng điện tử chứng khoán sau 15 năm hoạt động. Liệu như vậy nhà đầu tư có còn tin vào bản báo cáo đơn thuần về định tính nữa không? Liệu tài chính định lượng có thể được áp dụng vào hoạt động phân tích giá tại thị trường chứng khoán hay không? Không chỉ với thị trường chứng khoán Việt Nam nói riêng, câu hỏi này cũng được đặt ra cho thị trường chứng khoán toàn thế giới. Với sự phát triển của Công nghệ thông tin, đặc biệt thời gian gần đây là Trí tuệ nhân tạo và Khoa học dữ liệu đã tạo ra làn sóng áp dụng trí thông minh nhân tạo vào tất cả các ngành nghề và chứng khoán cũng không là ngoại lệ.

Hệ thống dự đoán chứng khoán sẽ giúp cho nhà đầu tư có cái nhìn tổng quan về dữ liệu trong quá khứ cũng như những dự đoán giá tương lai mà học máy mang lại. Vấn đề này thu hút rất nhiều sự nghiên cứu của các công ty Công nghệ và nhà Khoa học. Điều này cũng đem lại thách thức cho các nhà nghiên cứu là phải xây dựng ra mô hình dự đoán giá chứng khoán có động chính xác cao nhất để đạt được niềm tin của các nhà đầu tư.

Và một điều không thể thiếu trong bài toán dự đoán giá chứng khoán này mô hình Time Series. Time Series là một lớp mô hình quan trọng trong thống kê, kinh tế lượng và học máy. Và trong báo cáo này, chúng tôi sẽ nói rõ về cách một mô hình Time Series hoạt động và đưa ra kết quả dự đoán.

Báo cáo này được trình thành các chương sau:

**Chương 1 Bài toán dự đoán giá chứng khoán:** Giới thiệu các khái niệm trong thị trường chứng khoán, bài toán dự đoán giá chứng khoán, mô hình Time Series và cách mô hình hoạt động trong bài toán.

**Chương 2 Phương pháp kết hợp ARIMA và LSTM**: Giới thiệu về thuật toán ARIMA, mạng hồi quy LSTM và phương pháp kết hợp ARIMA và LSTM

**Chương 3** **Mô hình giải quyết bài toán:** Trình bày chi tiết các bước xây dựng, đánh giá, triển khai mô hình trên hệ thống Web application.

**Chương 4** **Kết luận**: đưa ra các nội dung chính có trong báo cáo. Đưa ra những điều làm được và chưa làm được, hướng phát triển trong tương lai.

# BÀI TOÁN DỰ ĐOÁN GIÁ CHỨNG KHOÁN

## Bài toán dự đoán giá chứng khoán

### Tổng quan về thị trường chứng khoán

1. **Khái niệm:**

Thị trường chứng khoán là một thị trường mà ở đó người ta mua bán, chuyển nhượng, trao đổi chứng khoán nhằm mục đích kiếm lời. Thị trường chứng khoán trong điều kiện của nền kinh tế hiện đại được quan niệm là nơi diễn ra các hoạt động giao dịch mua bán chứng khoán trung và dài hạn. Việc mua bán này được tiến hành ở thị trường sơ cấp khi người mua mua được chứng khoán lần đầu từ những người phát hành, và ở những thị trường thứ cấp khi có sự mua đi bán lại các chứng khoán đã được phát hành ở thị trường sơ cấp. Như vậy, xét về mặt hình thức, thị trường chứng khoán chỉ là nơi diễn ra các hoạt động trao đổi, mua bán, chuyển nhượng các loại chứng khoán, qua đó thay đổi chủ thể nắm giữ chứng khoán.

**Chức năng của thị trường chứng khoán:**

* Huy động vốn đầu tư cho nền kinh tế
* Cung cấp môi trường đầu tư cho công chúng
* Tạo tính thanh khoản trong chứng khoán
* Đánh giá hoạt động của doanh nghiệp
* Tạo môi trường giúp Chính phủ thực hiện các chính sách kinh tế vĩ mô

**Các chủ thể tham gia thị trường chứng khoán:**

* **Nhà phát hành:** là các tổ chức thực hiện huy động vốn thông qua thị trường chứng khoán. Nhà phát hành có thể là Chính phủ, chính quyền địa phương, công ty.
  + Chính phủ phát hành các loại trái phiếu chính phủ nhằm huy động tiền bù đắp thâm hụt ngân sách hoặc thực hiện các công trình quốc gia lớn.
  + Chính quyền địa phương phát hành trái phiếu địa phương để huy động đầu tư cho các công trình hay chương trình kinh tế, xã hội của địa phương
  + Các công ty muốn huy động vốn đầu tư phát triển sản xuất, phát hành trái phiếu công ty hoặc cổ phiếu.
* **Nhà đầu tư:** 
  + Nhà đầu tư cá nhân: Là công chúng muốn đầu tư vào thị trường chứng khoán.
  + Nhà đầu tư có tổ chức: công ty đầu tư, công ty bảo hiểm, quỹ bảo hiểm xã hội, công ty tài chính, ngân hàng thương mại.
* **Các tổ chức kinh doanh trên thị trường chứng khoán:**
  + Công ty chứng khoán
  + Quỹ đầu tư chứng khoán
  + Các trung gian tài chính

**Các nguyên tắc hoạt động trên thị trường chứng khoán**

* Nguyên tắc công khai
* Nguyên tắc trung gian
* Nguyên tắc đấu giá

### Bài toán dự đoán giá chứng khoán

Dự đoán chứng khoán đã trở thành công việc hết sức quan trọng mà các nhà đầu tư chứng khoán thực hiện để có được quyết định đầu tư thích hợp. Từ kết quả dự báo, người đầu tư chứng khoán sẽ quyết định khi nào mua vào, khi nào bán ra, khi nào thì giữ lại và nên đầu tư vào doanh nghiệp nào.

Qua khảo sát và phân tích có thể thấy rằng giá trong lịch sử và các chỉ số khác có thể tiết lộ mối tương quan và các mô hình biến động giá chứng khoán, do đó có thể dự báo giá trong tương lai. Theo Mackinlay, giá chứng khoán có mối tương quan nhạy cảm với tin tức và các sự kiện mang lại thông tin cho thị trường chứng khoán (kể cả các tin thời sự, kinh tế, chính trị, thời tiết,... đều ảnh hưởng đến thị trường chứng khoán). Với hướng tiếp cận này, sử dụng các kỹ thuật học máy và khai phá dữ liệu để tìm ra mối tương quan giữa giá trong quá khứ và xu hướng trong tương lai.

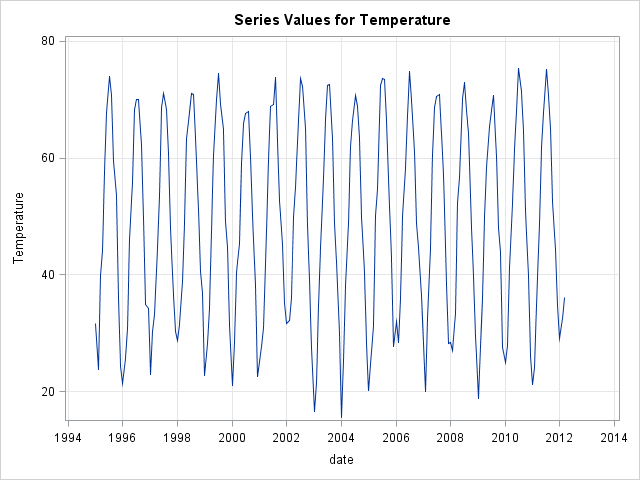
Thông tin của các giao dịch chứng khoán được lưu trữ lại dưới dạng dữ liệu chuỗi thời gian (Time Series). Dữ liệu thời gian hay chuỗi thời gian là một chuỗi các giá trị của một đại lượng nào đó được ghi nhận theo thời gian. Đây là kho dữ liệu khổng lồ để chúng ta có thể khai phá và dự báo xu thế của thị trường chứng khoán. Các dữ liệu lịch sử giao dịch thường được lưu trữ bao gồm các thông tin về thời gian (Date), giá mở cửa (Open), giá đóng cửa (Close), giá cao nhất trong ngày (High), giá thấp nhất trong ngày (Low) và khối lượng giao dịch (Volume).

Để thực hiện các phân tích kỹ thuật, nhà đầu tư thường dựa vào thông tin các chỉ số được thể hiện dưới dạng đồ thị khác nhau để thuận tiện trong việc phân tích.

## Mô hình Time Series

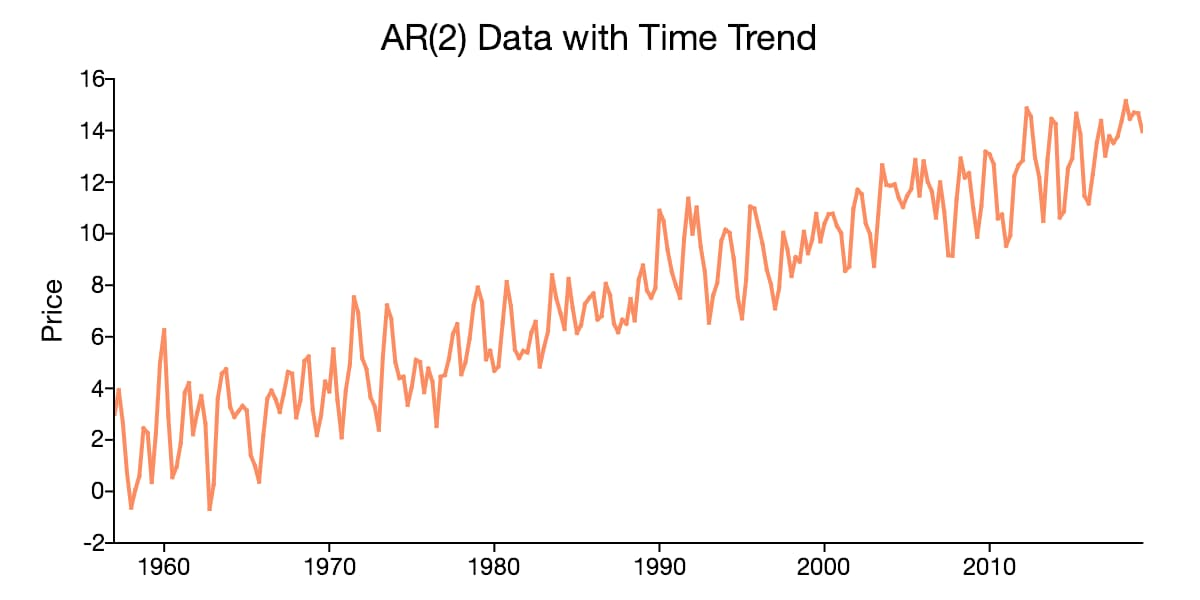
Dự báo Time Series là một lớp mô hình quan trọng trong thống kê, kinh tế lượng và học máy. Sở dĩ chúng ta gọi lớp mô hình này là Time Series là vì mô hình được áp dụng trên các chuỗi đặc thù có yếu tố thời gian. Một mô hình Time Series thường dự báo dựa trên giả định rằng các quy luật trong quá khứ sẽ lặp lại ở tương lai. Do đó xây dựng mô hình Time Series là chúng ta đang mô hình hóa mối quan hệ trong quá khứ giữa biến độc lập và biến phụ thuộc. Dựa vào mối quan hệ này ta có thể dự đoán giá trong tương lai của biến phụ thuộc.

Do dữ liệu chịu ảnh hưởng bởi tính chất thời gian nên Time Series thường xuất hiện những quy tắc như yếu tố chu kỳ, mùa vụ (Cyclical, Seasonal) và yếu tố xu hướng (Trend). Đây là những đặc trưng thường thấy và xuất hiện hầu hết trong các mô hình Time Series



Hình 1.1 Đồ thị về chuỗi nhiệt độ trung bình theo tháng thể hiện yếu tố mùa vụ

* Yếu tố chu kỳ, mùa vụ (Cyclical, Seasonal) là những đặc tính lặp lại chu kỳ. Ví dụ như nhiệt độ trung bình các tháng trong năm sẽ chịu ảnh hưởng bởi các mùa xuân, hạ, thu, đông. Hay xuất nhập khẩu của một quốc gia thường có chu kỳ theo quý.
* Yếu tố xu hướng (Trend) thể hiện đà tăng hoặc giảm của chuỗi thời gian trong tương lai. Chẳng hạn như lạm phát là xu hướng chung của các nền kinh tế, do đó giá cả trung bình của giỏ hàng hóa cơ sở hay còn gọi là chỉ số CPI luôn có xu hướng tăng và xu hướng tăng này đại diện cho sự mất giá của đồng tiền.



Hình 1.2 Đồ thị yếu tố xu hướng trong chuỗi thời gian của chuỗi giá

Các mô hình Time Series có tính ứng dụng cao và được sử dụng rất nhiều trong các lĩnh vực và chứng khoán cũng không là ngoại lệ.

# CÁC MÔ HÌNH ỨNG DỤNG

### MÔ HÌNH ARIMA

Mô hình ARIMA về cơ bản là một mô hình hồi quy tuyến tính được sử dụng để theo dõi các xu hướng tuyến tính trong dữ liệu chuỗi thời gian tĩnh. Mô hình được biểu diễn dưới dạng ARIMA (p, d, q). Trong đó, các tham số p, d và q là các giá trị nguyên quyết định cấu trúc của mô hình Time Series, tham số p, q là bậc của mô hình AR và mô hình MA và tham số d là mức độ khác biệt được áp dụng cho dữ liệu. Biểu diễn toán học của mô hình ARMA với tập tham số (p, q) như sau:

=

Box & Jenkins đã giới thiệu phương pháp chuẩn hóa để xây dựng mô hình ARIMA. Phương pháp này bao gồm ba bước lặp đi lặp lại. (1) Xác định mô hình và lựa chọn mô hình, loại mô hình, AR (p) hoặc MA (q), hoặc ARMA (p, q) được xác định. (2) Ước lượng các tham số của mô hình. (3) Mô hình kiểm định phân tích phần dư được thực hiện để cải thiện mô hình.

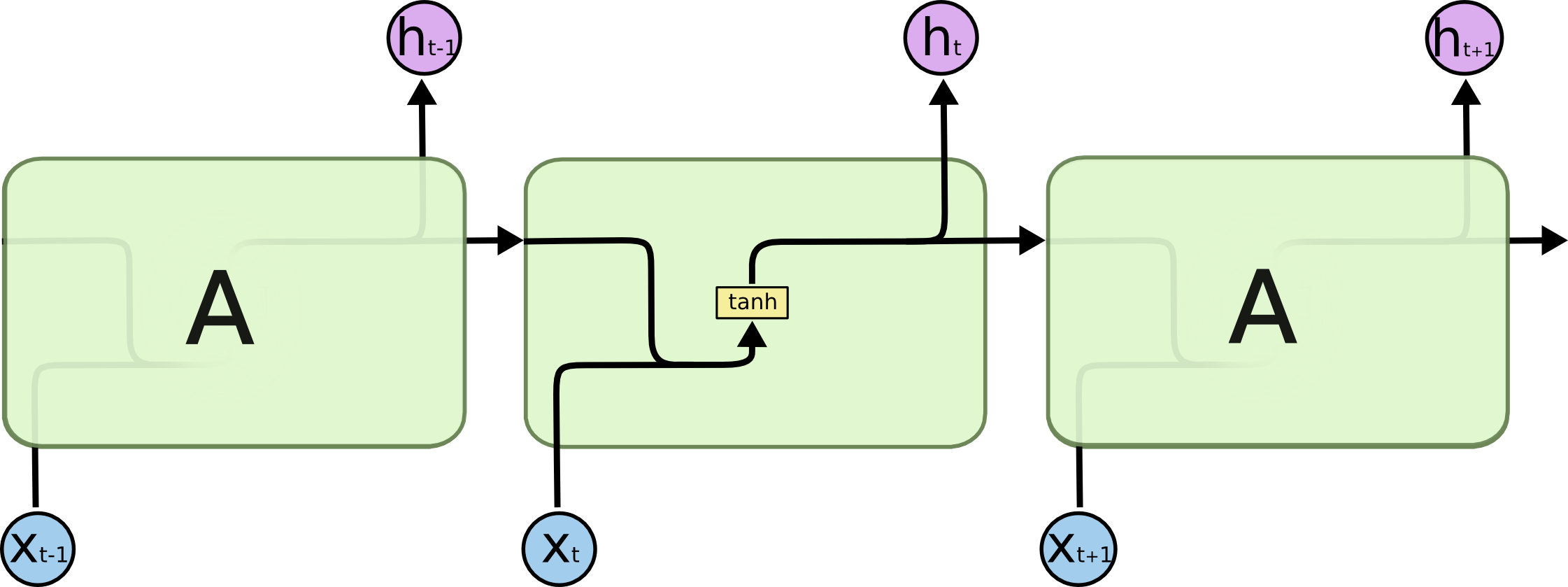
Ở bước xác định tham số đầu vào (p, d, q), sử dụng một số phương pháp ước lượng liên quan đến quá trình tối ưu hóa sử dụng các chỉ số lỗi toán học như the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) hoặc the Hannan-Quinn Information Criterion (HQIC). Trong dự án này, chúng tôi sử dụng AIC để ước lượng các tham số.

Ký hiệu là giá trị của hàm khả năng và k là bậc tự do, tức là tham số được sử dụng, Một mô hình có giá trị AIC nhỏ thường được coi là một mô hình tốt. Có nhiều cách khác nhau để tính toán hàm khả năng. Trong bước kiểm tra mô hình, phân tích phần dư được thực hiện để hoàn thiện mô hình ARIMA. Nếu phân tích phần dư kết luận rằng giá trị còn lại không đủ tiêu chuẩn, ba bước được lặp đi lặp lại cho đến khi đạt được mô hình ARIMA tối ưu nhất.

## Mô hinh LSTM

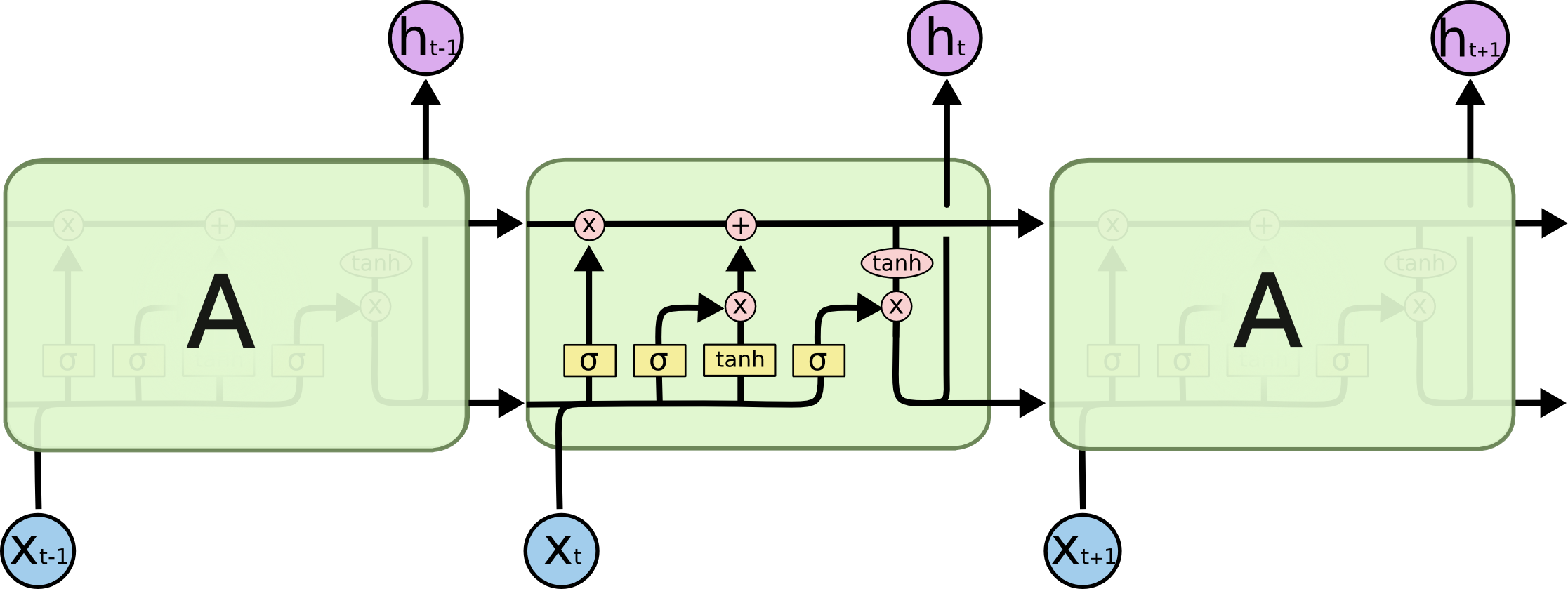
LSTM là một dạng đặc biệt của RNN, nó có khả năng xử lý các phụ thuộc xa. LSTM được giới thiệu bởi Hochreiter & Schmidhuber (1977) và sau đó đã được cải tiến, phổ biến bởi rất nhiều người trong ngành. Chúng hoạt động cực kì hiệu quả trên nhiều bài toán khác nhau nên dần đã trở nên phổ biến như hiện nay.

Mọi mạng RNN đều có dạng chuỗi các mô-đun lặp đi lặp lại của mạng nơ-ron. Với mạng RNN chuẩn, các mô-đun này có cấu trúc rất đơn giản, thường là một tầng *tanh*.



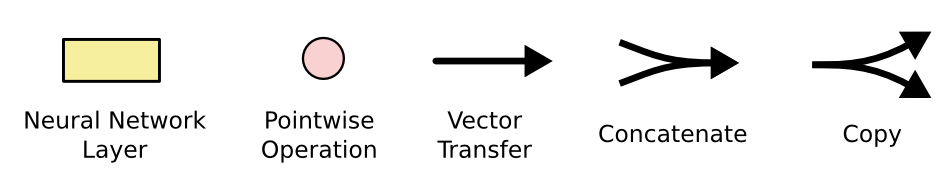
Hình 2.2 Cấu trúc chuỗi của RNN

LSTM cũng có kiến trúc dạng chuỗi như vậy, nhưng các mô-đun trong nó có cấu trúc khác nhau với mạng RNN chuẩn. Thay vì chỉ có một tầng mạng nơ-ron, chúng có tới 4 tầng tương tác với nhau một cách đặc biệt.



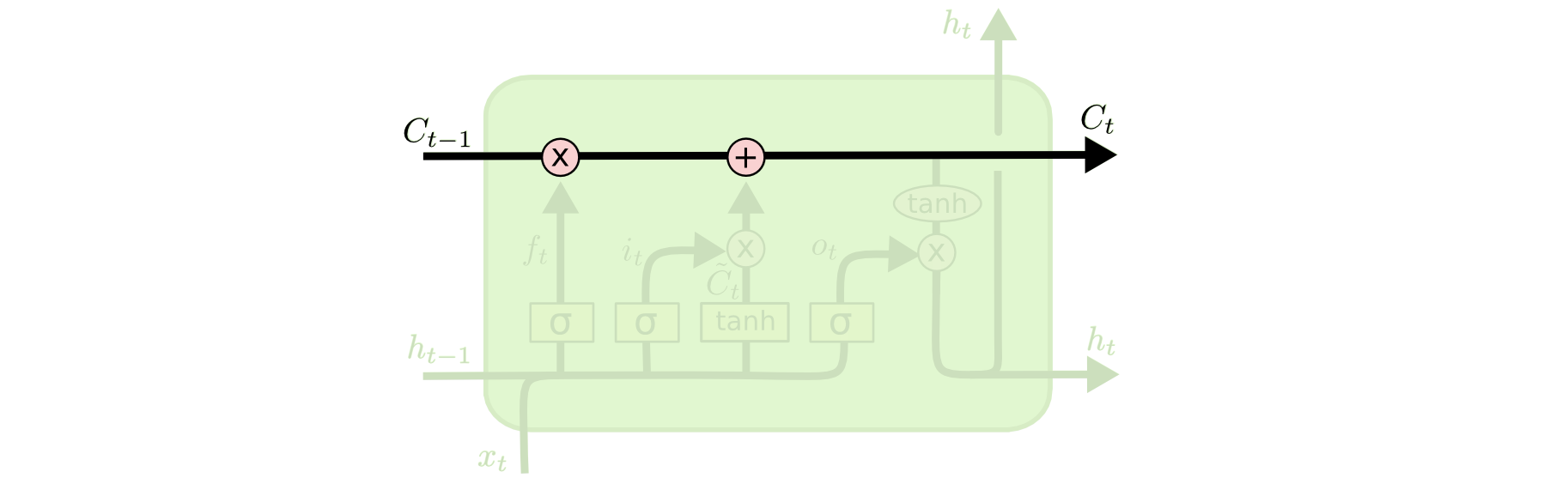
Hình 2.3 Cấu trúc chuỗi của LSTM

Ý nghĩa của các ký hiệu trong hình trên là:



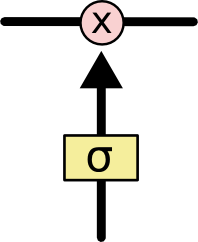
### Ý tưởng cốt lõi của LSTM

Chìa khóa của LSTM là trạng thái tế bào (cell state) - chính là đường chạy ngang phía trên của sơ đồ hình vẽ. Trạng thái tế bào là một dạng giống như băng truyền. Nó chạy xuyên suốt tất cả các mắt xích (các nút mạng) và chỉ tương tác tuyến tính đôi chút. Vì vậy mà các thông tin có thể dễ dàng truyền đi thông suốt mà không sợ bị thay đổi.



Hình 2.4 Đường trạng thái tế bào (cell state)

LSTM có khả năng bỏ đi hoặc thêm vào các thông tin cần thiết cho trạng thái tế báo, chúng được điều chỉnh cẩn thận bởi các nhóm được gọi là cổng (gate). Các cổng là nơi sàng lọc thông tin đi qua nó, chúng được kết hợp bởi một tầng mạng sigmoid và một phép nhân.



Hình 2.5 Cổng được kết hợp bởi sigmoid và phép nhân

Tầng sigmoid sẽ cho đầu ra là một số trong khoảng [0, 1], mô tả có bao nhiêu thông tin có thể được thông qua. Khi đầu ra là 0 thì có nghĩa là không cho thông tin nào qua cả, còn khi là 1 có nghĩa là cho tất cả các thông tin đi qua nó.

Một LSTM gồm có 3 cổng như vậy với mục đích để duy trì và điều hành trạng thái của tế báo.

## Phương pháp kết hợp giữa ARIMA và LSTM

### Lý do kết hợp

Time Series là chuỗi dữ liệu thời gian giả định bao gồm cả phần tuyến tính và phi tuyến tính. Do đó có thể biểu diễn dưới dạng như sau:

[4]

Trong đó, đại diện cho dữ liệu tuyến tính trong thời điểm *t,* đại diện cho dữ liệu phi tuyến tính và là giá trị sai số.

Autoregressive Integrated Moving Average (ARIMA) là mô hình thống kê truyền thống trong dự đoán dữ liệu chuỗi thời gian. ARIMA phù hợp với việc giải quyết những vấn đề tuyến tính. Còn Long Short-Term Memory (LSTM) phù hợp tính toán dữ liệu xu thế phi tuyến. Vậy nên, khi kết hợp hai mô hình sẽ đưa ra được mô hình bao gồm cả xử lý tuyến tính và phi tuyến.

### Ý tưởng kết hợp ARIMA và LSTM

Từ những ưu điểm của hai mô hình ARIMA và LSTM chúng tôi đã lên ý tưởng thiết kế kết hợp cả hai mô hình để đưa ra dự đoán đúng nhất cho giá chứng khoán.

Dữ liệu chứng khoán sau khi được thu thập được làm sạch rồi đưa vào mô hình ARIMA. Trước khi tiến hành huấn luyện mô hình, các tham số mô hình phải được xác định trước. Chúng tôi sử dụng biểu đồ Autocorrelation và Partial Autocorrelation để hỗ trợ việc đưa ra quyết định lựa chọn tham số. Ngoài ra chúng tôi cũng sử dụng thêm phương pháp tìm kiếm tham số AIC để đưa ra quyết định lựa chọn của mình.

Sau khi đưa dữ liệu chạy qua ARIMA, chúng tôi phân chia dữ liệu thành các tập nhỏ chiều dài dựa vào độ tương quan trong Autocorrelation, điểm dữ liệu cuối cùng trong tập con sẽ được sử dụng để làm biến mục tiêu Y và phần trước sẽ là X. Tập dữ liệu mới được phân tách sẽ là giá trị đầu vào cho mô hình LSTM tiếp theo.

Kiến trúc của LSTM là một mạng hồi quy do đó việc lựa chọn các tham số mạng cũng là một khó khăn. Chúng tôi thực hiện lựa chọn các tham số dựa trên tham khảo một số tài liệu với phương pháp tương tự.

# MÔ HÌNH GIẢI QUYẾT BÀI TOÁN

## Mô hình giải quyết bài toán

Quy trình xây dựng mô hình bài gồm 4 bước chính:

* Thu thập dữ liệu
* Tiền xử lý dữ liệu
* Xây dựng mô hình dự đoán bằng phương pháp kết hợp ARIMA và LSTM
* Thực nghiệm và đánh giá

Trong quá trình huấn luyện, các bước được thực hiện lặp đi lặp lại cho đến khi đạt được mô hình tốt nhất.

## Các bước thực hiện

### Thu thập dữ liệu

Dữ liệu trong bài toán Time Series với yếu tố quan trọng nhất là chuỗi thời gian. Yếu tố thời gian sẽ quyết định, hình thành lên mô hình Time Series. Vì vậy, để xây dựng lên mô hình chân thực nhất, chúng tôi đã sử dụng dữ liệu được lấy từ API Alpha Vantage. Chi tiết về API Alpha Vantage được để ở đường link dưới đây: <https://www.alphavantage.co/>.

Dữ liệu kéo về từ API Alpha Vantage là dữ liệu lịch sử chứng khoán tại thời điểm gần nhất (Dữ liệu là thông tin về giá chứng khoán tại ngày trước đó). Mỗi mã chứng khoán sẽ được phân loại dựa trên mã của mã chứng khoán đó, nó sẽ trả về một file định dạng JSON với tất cả thông tin về giá chứng khoán đó trong vòng 20 năm kể từ ngày hôm trước. Chúng tôi lấy dữ liệu từ URL định dạng file JSON rồi biểu diễn dữ liệu đó bằng pandas DataFrame và lưu nó dưới định dạng CSV . Chúng tôi nhận được dữ liệu của mã chứng khoán đó với các trường thông tin:

* Date: Trường thời gian ngày mã chứng khoán mở cửa
* Open: Giá mở cửa của mã chứng khoán trong ngày
* Close: Giá của mã chứng khoán khi phiên giao dịch trong ngày đóng lại
* High: Giá cao nhất trong ngày của mã
* Low: Giá thấp nhất trong ngày của mã

Ngoài ra để so sánh kết quả mô hình với bài báo của thầy Huỳnh Quyết Thắng (HUST), chúng tôi sử dụng thêm tập dữ liệu VN-index là chỉ số chứng khoán đầu tiên và chính thức của Sở Giao dịch chứng khoán TP HCM (HSX). Chúng tôi thu thập dữ liệu từ ngày 02/02/2015 đến hết 09/08/2016, tập dữ liệu bao gồm kết quả chứng khoán của 377 ngày liên tiếp.

### Tiền xử lý dữ liệu

Dữ liệu thu thập về được định dạng dưới dạng CSV và đọc lại bằng pandas DataFrame xong dữ liệu vẫn chưa thể đưa trực tiếp vào mô hình để huấn luyện.

Chúng tôi sẽ sử dụng giá trị trung bình (Close + Open)/2 của mã. Sau đó chúng tôi phân chia dữ thành hai tập train và test, tập train sẽ được sử dụng để huấn luyện mô hình bằng LSTM tập test sẽ dùng để đánh giá mô hình. Phân chia theo tỉ lệ 70:30 với 70% tổng dữ liệu là dữ liệu train và 30% tổng dữ liệu là dữ liệu test. Dữ liệu dùng trong bài toán này chỉ gồm chuỗi các giá trị trung bình của giá chứng khoán trong 1 ngày, dạng dữ liệu như vậy được gọi là single value. Với các bài toán dự báo giá chứng khoán trong thực tế, dữ liệu đầu vào của mô hình phải gồm nhiều thuộc tính khác mới có thể đưa ra những dự đoán chính xác và khách quan.

Ở từng bước chúng tôi lựa chọn các phương pháp tiền xử lý cho từng mã khác nhau. Trước khi đưa vào ARIMA, chúng tôi kiểm tra tính dừng của dữ liệu, nếu dữ liệu không có tính dừng, thực hiện khử tính dừng bằng cách lấy sai phân bậc 1. Sau khi đưa qua ARIMA, chúng tôi lấy giá trị nhãn trừ đi dữ liệu khi qua ARIMA để lấy được phần dư sai số. Sau đó đưa vào mạng LSTM.

### Xây dựng mô hình kết hợp ARIMA và LSTM

Với ý tưởng ban đầu chúng tôi thực hiện việc cài đặt kết hợp hai mô hình ARIMA và LSTM. Chúng tôi sử dụng AIC để thực hiện lựa chọn các tham số (p, d, q) trước khi đưa vào mô hình. Thực hiện tiền xử lý rồi tiếp tục đưa vào mạng LSTM.

Kiến trúc mô hình sử dụng trong đề tài bao gồm 3 mạng LSTM xếp chồng lên nhau, giữa các mạng có sử dụng dropout nhằm giảm độ phức tạp tính toán cũng như giảm overfit. Hàm mất mát “mean square error” và hàm tối ưu là “adam” được sử dụng để đánh giá và tối ưu mô hình. Cả quá trình được thực hiện trên google colab, sử dụng thư viện chính là tensorflow để xây dựng và huấn luyện mô hình.

Sau đó chúng tôi thực hiện đóng gói module từng phần và chạy thực nghiệm.

# CHƯƠNG 4: KẾT QUẢ THỰC NGHIỆM VÀ ĐÁNH GIÁ

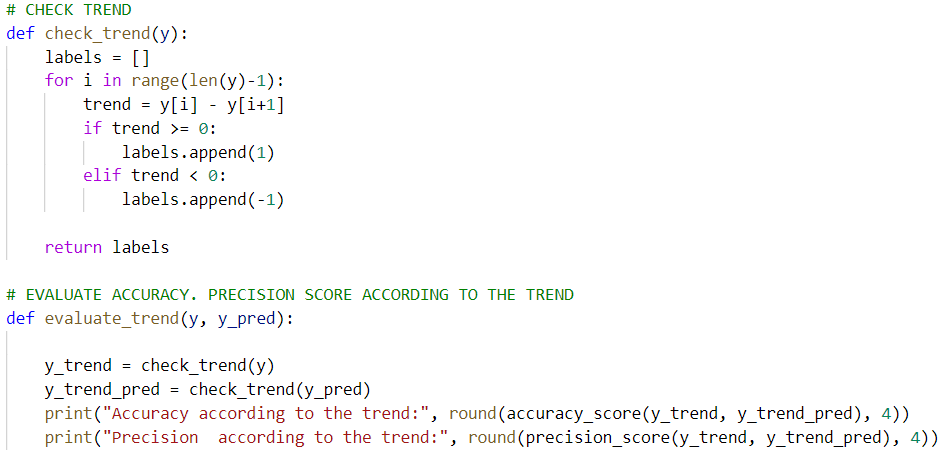
## Kết quả thực nghiệm

+ Chúng tôi sử dụng 3 mô hình: ARMIA, LSTM, ARMIA kết hợp LSTM.

+ Để kểm tra chất lượng của các mô hình sử 3 các mã chứng khoán: AAL (American Airlines), AAPL (Apple), SEV (Samsung), AMZN (Amazon), FB (facebook)

+ Các độ đo được sử dụng: MAE, MSE, độ đo xu thế (Accuracy, precision)

Với độ đo xu thế được tính như sau:



+ Đối với mô hình: LSTM có kiến trúc mạng:

- lớp 1: 50 neurons LSTM, dropout: 0.2

- lớp 2: 20 neurons LSTM, dropuout: 0.2

- Dữ liệu khi đưa vào mạng được chuẩn hóa về dạng 0 – 1

- Optimizer = ‘adam’, loss = ‘MSE’

- early stopping: monitor = ‘val\_loss’

+ Đối với mô hình ARIMA kết hợp LSTM:

- lớp 1: 100 neurons LSTM, dropout: 0.2

- lớp 2: 20 neurons LSTM, dropuout: 0.2

- Dữ liệu khi đưa vào mạng được chuẩn hóa về dạng 0 – 1

- Optimizer = ‘adam’, loss = ‘MSE’

- early stopping: monitor = ‘val\_loss’

+ Đối với mô mã chúng tôi chia là 2 tập train và test với tỷ lệ là: 700/300 bản ghi (tương ứng ngày)

Kết quả thực nghiệm trên tập test được thể hiện trong các bảng sau:

Mã: AAL (American Airlines)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | RMSE | MAE | Accuracy | Precision |
| ARIMA | 0.6952 | 0.4708 | 0.5986 | 0.5918 |
| LSTM | 0.7572 | 0.571 | 0.5173 | 0.5331 |
| ARIMA&LSTM | 0.8639 | 0.6066 | 0.517 | 0.5108 |

Mã: AAPL (Apple)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | RMSE | MAE | Accuracy | Precision |
| ARIMA | 0.5193 | 0.3739 | 0.5612 | 0.5503 |
| LSTM | 0.6291 | 0.571 | 0.5173 | 0.5331 |
| ARIMA&LSTM | 0.8639 | 0.4714 | 0.5306 | 0.5211 |

Mã: SEV (Samsung)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | RMSE | MAE | Accuracy | Precision |
| ARIMA | 0.1001 | 0.0727 | 0.5 | 0.6306 |
| LSTM | 0.5326 | 0.5077 | 0.568 | 0.6943 |
| ARIMA&LSTM | 0.1547 | 0.1118 | 0.5476 | 0.673 |

Mã: AMZN (Amazon)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | RMSE | MAE | Accuracy | Precision |
| ARIMA | 0.9878 | 0.7081 | 0.5272 | 0.4643 |
| LSTM | 2.5578 | 1.9668 | 0.5408 | 0.4727 |
| ARIMA&LSTM | 1.3155 | 0.9379 | 0.4966 | 0.4296 |

Mã: FB (Facebook)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | RMSE | MAE | Accuracy | Precision |
| ARIMA | 2.3129 | 1.5469 | 0.5204 | 0.4667 |
| LSTM | 7.3034 | 5.8825 | 0.4966 | 0.4403 |
| ARIMA&LSTM | 3.0812 | 2.0998 | 0.415 | 0.3571 |

## Đánh giá

Đối với các trường hợp kiểm tra: mô hình ARMIA cho các chỉ sổ tốt nhất tuy nhiên các dự đoán có su thế dự đoán cao hơn giá trị thực.

Mô hình kết hợp ARIMA và LSTM: cho kết quả mô hình tốt thứ hai, khắc phục được nhược điểm của mô hình ARIMA, tuy nhiên độ trễ lớn (có thể hiệu chỉnh được). Mô hình này cần tối ưu lại cấu trúc và các tham số mô hình.

# CHƯƠNG 5. KẾT LUẬN

## Kết luận

Nối tiếp đề tài của môn học máy cơ bản, nhóm chúng tôi tiếp tục giải quyết bài toán dự báo giá chứng khoán bằng cách điều chỉnh lại tham số của các mô hình cũ, cụ thể là ARIMA và LSTM,sau đó kết hợp hai mô hình vào để dự báo. Về mặt lý thuyết, mô hình kết hợp có thể giải quyế t thành phần tuyến tính và phi tuyến tính trong dữ liệu time series.

Tuy nhiên, mô hình ở đây mới chỉ giải quyết được bài toán single value. Thực tế giá chứng khoán bị ảnh hưởng, liên quan đến nhiều thông tin khác chứ không chỉ phụ thuộc vào giá của các ngày trước đó. Bài toán thực tế này rất khó giải quyết, đặc biệt là ở thị trường Việt Nam, nơi mà thị trường chứng khoán được đánh giá là ít tuân theo các quy luật của thị trường kinh tế.

## Hướng phát triển tương lai

Để có thể đưa mô hình time-series vào dự báo giá chứng khoán trong thực tế thì chúng ta cần phải quan tâm thêm nhiều thông tin ảnh hưởng đến thị trường chứng khoán, giả dụ như giá vàng, giá dầu, các sự kiện xã hội, ….Khi đó sẽ cần phải sử dụng những mô hình, kiến trúc hay những kỹ thuật phức tạp hơn để giải bài toán. Nhóm hy vọng sau đề tài này, sẽ nhận được sự góp ý của các thầy cô để có thể giải quyết được các bài toán thực tế như trên, hướng tới việc ứng dụng các mô hình time-series vào thực tế không chỉ trong dự báo giá chứng khoán mà còn trong nhiều bài toán khác, góp phần mang lại giá trị tốt đẹp cho xã hội nói chung, đất nước con người Việt Nam nói riêng.

# TÀI LIỆU THAM KHẢO

1. <https://nttuan8.com/bai-14-long-short-term-memory-lstm/>
2. <https://www.alphavantage.co/documentation/>
3. <https://www.researchgate.net/publication/329123511_Du_doan_xu_the_chi_so_chung_khoan_Viet_Nam_VN-Index_su_dung_phan_tich_hoi_quy_Gaussian_Process_va_mo_hinh_tu_hoi_quy_trung_binh_dong_ARMA>
4. <https://www.researchgate.net/publication/326852308_Stock_Price_Correlation_Coefficient_Prediction_with_ARIMA-LSTM_Hybrid_Model>

**PHỤ LỤC**

**Bảng nhiệm vụ:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Khải** | **Đông** | **Hiển** | **Sơn (leader)** |
| EDA data | Data preprocessing | Data preprocessing | Data collection |
| Research | Research | Research | Research |
| collect, code utils.py, preprocesing.py, model\_train\_test.py, test1,2,3,4,5.ipynb | Research, code function utils and LSTM | Research, code ARIMA & LSTM | Document writing |
| Evaluate, deploy model | Document writing | Document writing | Project management |
| Presentation, document writing |  |  |  |