Lời đầu tiên, chúng em xin gửi lời cảm ơn chân thành tới PGS. TS Nguyễn Đình Thuân vì sự hướng dẫn và quan tâm tận tâm trong quá trình thực hiện khóa luận tốt nghiệp. Sự tận tâm và sự quan tâm của Thầy đã giúp chúng em tích lũy được nhiều kiến thức quý báu về chuỗi thời gian, mô hình dự báo và các thuật toán máy học và hồi quy.

Chúng em cũng xin chân thành cảm ơn đến các thầy cô giáo trong Khoa Hệ thống Thông tin, những người đã chia sẻ kiến thức chuyên môn của mình để giúp chúng em hoàn thiện khóa luận. Những kiến thức mà chúng em đã nhận được từ quý Thầy, Cô giáo đã giúp chúng em đóng góp vào thành công của khóa luận này.

Chúng em thấu hiểu rằng kiến thức là vô tận, nhưng khả năng tiếp thu kiến thức của mỗi người lại có giới hạn. Do đó, trong quá trình thực hiện khóa luận, chúng em đã gặp một số sai sót nhất định. Chúng em hy vọng được sự góp ý từ Thầy, Cô giáo để khóa luận của chúng em được hoàn thiện hơn, đồng thời giúp chúng em tích lũy thêm kinh nghiệm quý báu trong quá trình học tập và nghiên cứu.

Một lần nữa, chúng em xin chân thành cảm ơn PGS. TS Nguyễn Đình Thuân và tất cả các Thầy, Cô giáo đã giúp đỡ chúng em. Sự hỗ trợ và đóng góp của quý Thầy, Cô giáo là một phần không thể thiếu trong thành công của khóa luận này.

TÓM TẮT ĐỀ TÀI

Cổ phiếu là một phân khúc đầu tư hấp dẫn và đầy tiềm năng trong thị trường tài chính. Đối với nhà đầu tư, dự báo giá cổ phiếu và quyết định mua, bán hoặc giữ cổ phiếu là những vấn đề quan trọng. Trong việc dự báo giá cổ phiếu, thuật toán máy học đã được chứng minh là một công cụ mạnh mẽ.

Thuật toán máy học có khả năng học từ dữ liệu lịch sử và tìm ra các mẫu và mối quan hệ phức tạp giữa các biến số để dự báo giá cổ phiếu trong tương lai. Chúng có khả năng xác định xu hướng và cung cấp cái nhìn tổng quan về tiềm năng của cổ phiếu.

Tuy nhiên, cần lưu ý rằng dự báo giá cổ phiếu không phải là một quá trình hoàn hảo và có thể gặp phải những sai sót. Thị trường tài chính là một môi trường phức tạp và ảnh hưởng của nhiều yếu tố khác nhau có thể làm thay đổi giá cổ phiếu một cách đột ngột. Do đó, việc sử dụng thuật toán máy học chỉ là một phần trong quá trình ra quyết định đầu tư. Kết hợp với các yếu tố khác như phân tích cơ bản, phân tích kỹ thuật và tình hình thị trường, nhà đầu tư có thể có cái nhìn tổng thể và đưa ra quyết định thông minh.

Có một loạt phương pháp dự báo đa dạng được sử dụng hiện nay [2]. Mỗi phương pháp mang lại một cách tiếp cận khác nhau và có ưu điểm riêng. Dưới đây là một số phương pháp dự báo phổ biến:

1. Phương pháp điều tra: Sử dụng thông tin thu thập từ các cuộc điều tra để đưa ra dự báo. Phương pháp này tập trung vào việc thu thập thông tin từ một tập nguồn dữ liệu rộng lớn và xác định các xu hướng và mô hình dự báo từ đó.
2. Phương pháp thực nghiệm: Dựa trên việc tiến hành các thí nghiệm và quan sát kết quả để dự báo. Thông qua việc thực hiện các thí nghiệm trong một môi trường kiểm soát, phương pháp này cố gắng hiểu và dự báo các biến số dựa trên các quan sát thực tế.
3. Phương pháp ngoại suy: Dựa trên quan sát và phân tích các xu hướng lịch sử để dự báo trong tương lai. Phương pháp này giả định rằng các xu hướng đã xảy ra trong quá khứ sẽ tiếp tục trong tương lai và sử dụng chúng để đưa ra dự báo.
4. Phân tích kinh tế lượng: Sử dụng các mô hình kinh tế và các công cụ toán học để dự báo. Phương pháp này tập trung vào việc phân tích các yếu tố kinh tế, tài chính và các biến số khác để xác định các mô hình dự báo.
5. Phân tích chuyên gia: Dựa trên kiến thức và kinh nghiệm của các chuyên gia trong lĩnh vực để đưa ra dự báo. Phương pháp này nhấn mạnh sự hiểu biết chuyên môn và quan điểm của các chuyên gia để đưa ra dự báo chính xác.

Có rất nhiều phương pháp dự báo khác nhau, mỗi phương pháp đều có ưu điểm và hạn chế riêng. Quan trọng nhất là lựa chọn phương pháp phù hợp với tình huống cụ thể và kết hợp nhiều phương pháp để đảm bảo tính chính xác và đáng tin cậy của dự báo.

1.1.3. Tính chất và đặc điểm của dự báo

Dự báo có tính chất là đưa ra xu hướng để giúp tăng cường sự cạnh tranh và thu lợi nhuận dựa trên cơ sở khoa học. Tuy nhiên, các phương pháp dự báo thường không đạt độ hoàn hảo và thường có sai số. Vì vậy, cần có một mức độ sai số cho phép để đảm bảo tính khả thi và đáng tin cậy của dự báo.

Dự báo cho một nhóm đối tượng thường chính xác hơn so với dự báo cho từng đối tượng riêng lẻ. Điều này là do việc dự báo cho một nhóm đối tượng giúp loại bỏ các yếu tố ngẫu nhiên và tạo ra một mức độ chính xác trung bình.

Tuy nhiên, độ chính xác của dự báo sẽ giảm dần khi thời gian dự báo kéo dài. Điều này là do các yếu tố bất ngờ và không thể kiểm soát có thể xảy ra trong tương lai, dẫn đến sự khác biệt giữa dự báo và kết quả thực tế.

1.1.4. Các ứng dụng của dự báo

Dự báo có nhiều ứng dụng trong các lĩnh vực khác nhau, như sau:

* Dự báo thời tiết: Đây là ứng dụng phổ biến nhất của dự báo, dựa trên các đặc điểm thời tiết của quá khứ để dự báo cho những ngày tiếp theo.
* Dự báo giá vàng, cổ phiếu: Ứng dụng này lấy những đặc điểm xu hướng về giá của thời điểm trước đó để dự báo cho tương lai.
* AirVisual Air Quality Forecast: Đây là ứng dụng theo dõi dự báo về chất lượng không khí.
* Dự báo còn được sử dụng trong nhiều lĩnh vực khác như kinh tế, chính trị, văn hóa, xã hội, du lịch, giáo dục, và nhiều lĩnh vực khác nhằm giúp cho các quyết định có thể được đưa ra dựa trên các dự báo chính xác và đáng tin cậy.

# **cơ sở lý thuyết cho dự báo chuỗi thời gian đa biến**

## **1. Giới thiệu**

Cơ sở lý thuyết cho dự báo chuỗi thời gian đa biến là một lĩnh vực quan trọng trong thống kê và khoa học dữ liệu. Nó tập trung vào việc dự báo giá trị tương lai của một chuỗi thời gian đa biến, trong đó mỗi thời điểm có nhiều biến đầu vào liên quan đến nhau. Dự báo chuỗi thời gian đa biến đòi hỏi sử dụng các phương pháp và mô hình dự báo phù hợp để xử lý các yếu tố sau.

## **2. Chuỗi thời gian**

Chuỗi thời gian đa biến được xem như một tập hợp các chuỗi thời gian riêng biệt, với mỗi biến trong chuỗi đại diện cho một thuộc tính cụ thể. Điều này có nghĩa là các mô hình dự báo chuỗi thời gian đơn biến có thể được áp dụng cho mỗi biến riêng lẻ trong chuỗi thời gian đa biến. Các mô hình dự báo chuỗi thời gian đơn biến phổ biến bao gồm mô hình ARIMA (Autoregressive Integrated Moving Average) và mô hình Holt-Winters.

## **3. Tương quan giữa các biến**

Các biến trong chuỗi thời gian đa biến có thể có mối tương quan với nhau. Điều này có thể được mô hình hóa bằng các mô hình dự báo đa biến như VAR (Vector Autoregression) hoặc VECM (Vector Error Correction Model). Các mô hình này giúp xác định mối tương quan giữa các biến và dự báo chúng cùng nhau. Mô hình VAR sử dụng các biến đầu vào trong quá khứ để dự báo các biến trong tương lai, trong khi mô hình VECM thêm vào khả năng sửa lỗi sai số của các biến.

## **4. Mô hình hóa độ trễ**

Trong dự báo chuỗi thời gian đa biến, các biến có thể có sự trễ đối với nhau. Điều này có nghĩa là giá trị của một biến tại một thời điểm có thể phụ thuộc vào giá trị của các biến khác ở các thời điểm trước đó. Để mô hình hóa độ trễ này, chúng ta có thể sử dụng các mô hình dự báo độ trễ như VARIMA (Vector Autoregressive Integrated Moving Average) hoặc VARMA (Vector Autoregressive Moving Average). Các mô hình này kết hợp cả yếu tố độ trễ và sự tương quan giữa các biến để dự báo giá trị tương lai.

## **5. Đặc điểm đa biến**

Chuỗi thời gian đa biến có thể có các đặc điểm đa biến đặc biệt cần được xem xét. Ví dụ, có thể có hiệu ứng lẫn nhau (cross-correlation) giữa các biến hoặc hiệu ứng giữa các thành phần chu kỳ. Để mô hình hóa các đặc điểm này, chúng ta có thể sử dụng các phương pháp dự báo đa biến như phân tích thành phần chính đa biến (Multivariate Principal Component Analysis) hoặc phân tích cấu trúc phụ thuộc thời gian đa biến (Multivariate Time Series Dependence Structure Analysis).

## **6. Kết luận**

Cơ sở lý thuyết cho dự báo chuỗi thời gian đa biến cung cấp các phương pháp và mô hình để dự báo giá trị tương lai trong một chuỗi thời gian đa biến. Việc hiểu và áp dụng các phương pháp này có thể giúp chúng ta dự báo và phân tích dữ liệu chuỗi thời gian đa biến một cách hiệu quả. Tuy nhiên, lựa chọn phương pháp và mô hình phù hợp phụ thuộc vào các đặc điểm cụ thể của dữ liệu và mục tiêu dự báo của chúng ta.

## **1. Giới thiệu**

Thuật toán VAR (Vector Autoregression) là một phương pháp dự báo trong lĩnh vực dự báo chuỗi thời gian đa biến. Nó được sử dụng để mô hình hóa và dự báo sự tương quan giữa các biến trong một chuỗi thời gian đa biến. Trong thuật toán VAR, chúng ta giả định rằng mỗi biến trong chuỗi thời gian đa biến phụ thuộc vào giá trị của các biến khác trong quá khứ.

## **2. Mô hình VAR**

Mô hình VAR xem xét một chuỗi thời gian đa biến với p biến (p > 1). Để xây dựng mô hình VAR, chúng ta sẽ sử dụng dữ liệu lịch sử của các biến để ước lượng các hệ số của mô hình.

Mô hình VAR(p) có dạng như sau:

arduinoCopy code

Y\_t = c + A\_1\*Y\_{t-1} + A\_2\*Y\_{t-2} + ... + A\_p\*Y\_{t-p} + ε\_t

Trong đó:

* Y\_t là vector chứa giá trị của các biến tại thời điểm t.
* c là vector hằng số.
* A\_1, A\_2, ..., A\_p là các ma trận hệ số.
* ε\_t là vector nhiễu ngẫu nhiên tại thời điểm t.VA

Mô hình VAR(p) cho phép mô hình hóa tương quan giữa các biến trong quá khứ p thời điểm trước. Sử dụng mô hình này, chúng ta có thể dự báo giá trị của các biến trong tương lai bằng cách sử dụng các giá trị của các biến trong quá khứ.

## **3. Ước lượng mô hình VAR**

Để ước lượng mô hình VAR, chúng ta thường sử dụng phương pháp bình phương tối tiểu (Least Squares Method). Phương pháp này tìm ra các hệ số A\_1, A\_2, ..., A\_p và c sao cho tổng bình phương sai số giữa dữ liệu thực và dữ liệu được dự báo từ mô hình là nhỏ nhất.

Sau khi ước lượng được các hệ số, chúng ta có thể sử dụng mô hình VAR để dự báo giá trị của các biến trong tương lai. Dự báo VAR có thể được thực hiện bằng cách sử dụng các giá trị của các biến trong quá khứ đã biết để tính toán giá trị dự báo cho thời điểm tiếp theo.

## **4. Đánh giá và cải thiện mô hình VAR**

Sau khi xây dựng mô hình VAR, chúng ta cần đánh giá và cải thiện mô hình. Các phương pháp thông thường để đánh giá mô hình VAR bao gồm:

* Kiểm tra các giả định về nhiễu của mô hình, chẳng hạn như tính tuân thủ phân phối chuẩn và tính không tương quan.
* Kiểm tra tính ổn định của mô hình để đảm bảo các giá trị dự báo không bị sai lệch theo thời gian.
* Đánh giá hiệu suất dự báo bằng các độ đo như sai số trung bình bình phương (Mean Squared Error) hoặc hệ số tương quan.

Nếu mô hình VAR không đạt được kết quả dự báo tốt, chúng ta có thể cải thiện nó bằng cách thay đổi cấu trúc mô hình (sử dụng số lượng biến, cấu trúc độ trễ, v.v.) hoặc áp dụng các biến đổi dữ liệu và phương pháp dự báo khác.

# **LSTM (Long Short-Term Memory)**

## **1. Giới thiệu**

Thuật toán LSTM (Long Short-Term Memory) là một loại mạng nơ-ron hồi quy (RNN - Recurrent Neural Network) được sử dụng rộng rãi trong dự báo chuỗi thời gian. Nó được thiết kế để giải quyết vấn đề mất mát thông tin dài hạn trong mô hình RNN truyền thống và cho phép xử lý các chuỗi dữ liệu dài và phức tạp.

## **2. Mạng nơ-ron hồi quy (RNN)**

Mạng nơ-ron hồi quy (RNN) là một loại mạng nơ-ron nhân tạo có khả năng xử lý dữ liệu chuỗi thời gian hoặc dữ liệu có mối quan hệ thời gian. RNN có khả năng lưu trữ thông tin từ quá khứ và sử dụng nó để dự báo các giá trị trong tương lai. Tuy nhiên, mô hình RNN truyền thống gặp vấn đề mất mát thông tin dài hạn khi chuỗi dữ liệu quá dài.

## **3. LSTM (Long Short-Term Memory)**

LSTM là một biến thể của mạng nơ-ron hồi quy (RNN) được thiết kế để vượt qua vấn đề mất mát thông tin dài hạn trong RNN truyền thống. LSTM sử dụng các cấu trúc gọi là "cổng" để quyết định thông tin nào được lưu trữ và thông tin nào bị loại bỏ. Điều này giúp LSTM lưu trữ thông tin quan trọng trong một khoảng thời gian dài và giảm mất mát thông tin.

## **4. Cấu trúc LSTM**

LSTM bao gồm các đơn vị LSTM (LSTM cells) được xếp chồng lên nhau để tạo thành một mô hình dự báo chuỗi thời gian. Mỗi đơn vị LSTM bao gồm các thành phần sau:

* Cổng quên (Forget gate): Xác định thông tin nào trong quá khứ nên bị loại bỏ.
* Cổng đầu vào (Input gate): Xác định thông tin mới nào nên được lưu trữ.
* Cổng đầu ra (Output gate): Xác định thông tin nào sẽ được sử dụng để tạo ra đầu ra.

Các cổng này sử dụng hàm sigmoid để quyết định xem thông tin nào được chuyển qua và thông tin nào bị bỏ qua. Đồng thời, LSTM sử dụng các "cơ cấu ô nhớ" (memory cell) để lưu trữ và truyền tải thông tin trong quá trình thời gian.

## **5. Ước lượng và dự báo với LSTM**

Để ước lượng mô hình LSTM, chúng ta cần huấn luyện mạng bằng cách sử dụng các dữ liệu chuỗi thời gian đã biết và thuật toán lan truyền ngược (backpropagation). Quá trình huấn luyện LSTM bao gồm việc tối ưu hóa các tham số mô hình để giảm sai số dự báo giữa giá trị thực và giá trị dự báo.

Sau khi ước lượng được các tham số, chúng ta có thể sử dụng mô hình LSTM để dự báo giá trị trong tương lai. Dự báo LSTM thường được thực hiện bằng cách sử dụng các giá trị đầu vào đã biết để tính toán giá trị dự báo cho thời điểm tiếp theo.

## **6. Đánh giá và cải thiện mô hình LSTM**

Để đánh giá hiệu suất của mô hình LSTM, chúng ta có thể sử dụng các độ đo như sai số trung bình bình phương (Mean Squared Error) hoặc hệ số tương quan giữa giá trị thực và giá trị dự báo. Ngoài ra, chúng ta có thể cải thiện mô hình bằng cách điều chỉnh kiến trúc mạng, số lượng đơn vị LSTM, cách thức xử lý dữ liệu đầu vào, v.v.

# **thuật toán XGBoost**

## **1. Giới thiệu**

XGBoost (eXtreme Gradient Boosting) là một thuật toán học máy được phát triển dựa trên phương pháp gradient boosting. Nó là một trong những thuật toán dự báo mạnh mẽ và phổ biến nhất hiện nay. XGBoost kết hợp cả gradient boosting và các cải tiến đặc biệt để cải thiện hiệu suất và tốc độ của mô hình dự báo.

## **2. Gradient Boosting**

Gradient boosting là một phương pháp học máy tập trung vào việc xây dựng một mô hình dự báo mạnh mẽ bằng cách kết hợp nhiều mô hình yếu thành một mô hình tổng hợp. Phương pháp này tạo ra một chuỗi các mô hình học máy tuần tự, mỗi mô hình học từ sai số còn lại của mô hình trước đó. Gradient boosting sử dụng các kỹ thuật tối ưu hóa gradient để tìm kiếm các mô hình tốt nhất cho từng giai đoạn.

## **3. Cải tiến của XGBoost**

XGBoost giới thiệu một số cải tiến quan trọng để cải thiện hiệu suất và tốc độ của mô hình dự báo. Các cải tiến này bao gồm:

* **Regularization**: XGBoost sử dụng hai thành phần chính là regularized learning objective và regularization term để tránh overfitting và cải thiện khả năng tổng quát hóa của mô hình.
* **Gradient-based optimization**: XGBoost tận dụng các phép tính gradient và hessian để tối ưu hóa hàm mất mát. Điều này giúp tăng tốc độ huấn luyện và tối ưu hóa hiệu suất.
* **Parallelization**: XGBoost hỗ trợ việc xử lý song song trên nhiều bộ xử lý (CPU) để giảm thời gian huấn luyện và dự báo.
* **Sparsity-aware split finding**: XGBoost hiểu và tận dụng tính thưa của dữ liệu để tìm kiếm cách chia cây hiệu quả hơn, đồng thời giảm độ phức tạp tính toán.

## **4. Cấu trúc XGBoost**

XGBoost có một cấu trúc chính gồm các thành phần sau:

* **Boosting rounds**: XGBoost xây dựng mô hình bằng cách kết hợp nhiều vòng lặp (boosting rounds), trong mỗi vòng lặp, mô hình tối ưu hóa hàm mất mát dựa trên các gradient và hessian.
* **Weak learners**: Mỗi vòng lặp trong XGBoost sử dụng một mô hình học yếu, thường là cây quyết định (decision tree). Các cây quyết định được xây dựng theo phương pháp tăng cường gradient (gradient boosting) và được kết hợp lại để tạo thành mô hình dự báo cuối cùng.
* **Loss function**: XGBoost sử dụng một hàm mất mát (loss function) để đánh giá sai số giữa giá trị dự báo và giá trị thực tế. Các thuật ngữ trong hàm mất mát phản ánh mức độ sai lệch giữa dự báo và thực tế.

## **5. Đánh giá mô hình XGBoost**

Để đánh giá hiệu suất của mô hình XGBoost, chúng ta có thể sử dụng các phương pháp sau:

* **Tập dữ liệu kiểm tra**: Chia tập dữ liệu ban đầu thành tập huấn luyện và tập kiểm tra. Mô hình được huấn luyện trên tập huấn luyện và sau đó được đánh giá trên tập kiểm tra để đo lường hiệu suất dự báo.
* **Độ đo đánh giá**: Sử dụng các độ đo như sai số trung bình bình phương (Mean Squared Error - MSE), hệ số tương quan (Correlation Coefficient), hay các độ đo khác để đánh giá độ chính xác và sự phù hợp của dự báo từ mô hình XGBoost.
* **Cross-validation**: Áp dụng phương pháp cross-validation để đánh giá hiệu suất của mô hình XGBoost trên nhiều tập dữ liệu kiểm tra khác nhau.