HỘI NGHỊ SINH VIÊN NGHIÊN CỨU KHOA HỌC HUTECH 2020

TÌM HIỂU VỀ LONG SHORT-TERM MEMORY

Tên thí sinh

Khoa Công nghệ Thông tin, Trường Đại học Công Nghệ TP. Hồ Chí Minh (HUTECH)

Giảng viên hướng dẫn:

TÓM TẮT

Trong các giao dịch hàng ngày của thị trường chứng khoáng, các nhà đầu tư chiến lược và cả những người tham gia giao dịch luôn tìm cách để dự đoán ra những chuyển biến tiếp theo của giá trị hàng hoá mà họ đang tham gia theo dõi, giao dịch. Những cá nhân, tổ chức có khả năng đưa ra những dự đoán gần chính xác nhất đối với thị trường sẽ có khả năng dẫn đầu trong việc nắm bắt lợi thế thị trường, tăng cơ hội thành công trong giao dịch. Trong quá trình dự đoán, các thuật toán dự đoán bằng máy tính được ứng dụng bên cạnh việc phán đoán dựa trên kĩ thuật và kinh nghiệm của cá nhân. Máy học (marchine learning) là một trong những kĩ thuật tiềm năng luôn được nghiên cứu và ứng dụng trong ngành này. Trong bài viết này sử dụng kĩ thuật có tên “Long short-term memory” (bộ nhớ dài-ngắn hạn), để tìm hiểu khả năng dự đoán các bước chuyển biến của thị trường giao dịch tài chính.

*Từ khoá: finance, long short-term memory, lstm, marchine learning, python, time series forecasting.*

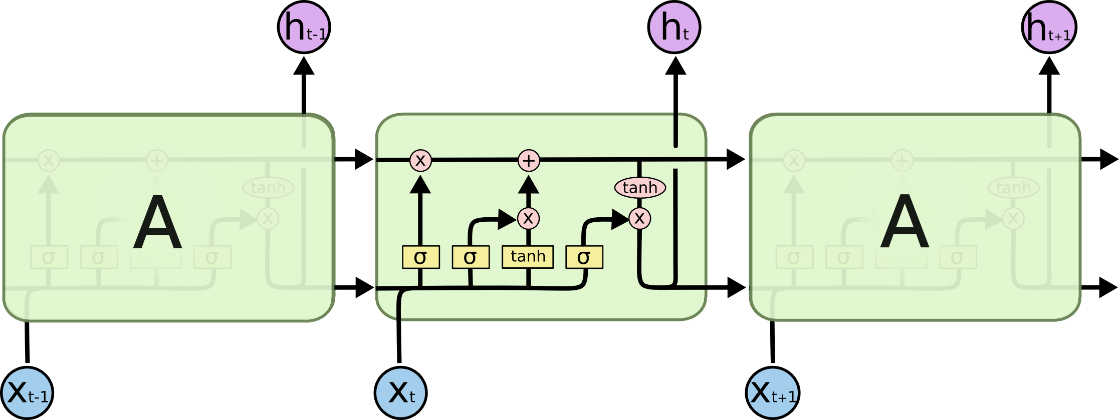
1. GIỚI THIỆU
   1. Long short-term memory
      1. Lý thuyết

Mạng bộ nhớ dài-ngắn (Long Short Term Memory networks), thường được gọi là LSTM - là một dạng đặc biệt của RNN, nó có khả năng học được các phụ thuộc xa. LSTM được giới thiệu bởi Hochreiter & Schmidhuber (1997), và sau đó đã được cải tiến và 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.

LSTM được thiết kế để tránh được vấn đề phụ thuộc xa (long-term dependency). Việc nhớ thông tin trong suốt thời gian dài là đặc tính mặc định của chúng, chứ ta không cần phải huấn luyện nó để có thể nhớ được. Tức là ngay nội tại của nó đã có thể ghi nhớ được mà không cần bất kì can thiệp nào.

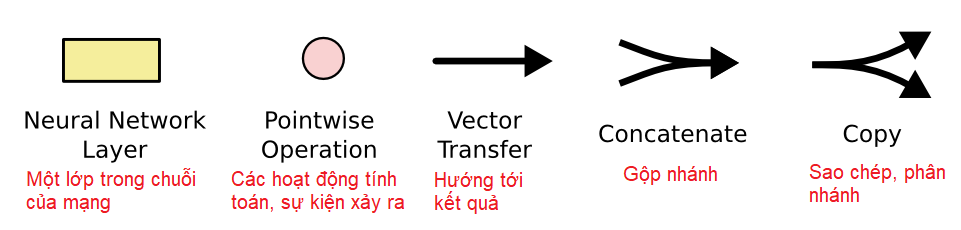
Mọi mạng hồi quy đều có dạng là một chuỗi các mô-đun lặp đi lặp lại của mạng nơ-ron. Với mạng RNN chuẩn, các mô-dun này có cấu trúc rất đơn giản, thường là một tầng *tanh*.

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



*Hình 1. Cấu tạo của một modul trong kiến trúc chuỗi LSTM*

Các kí hiệu dùng trong LSTM:

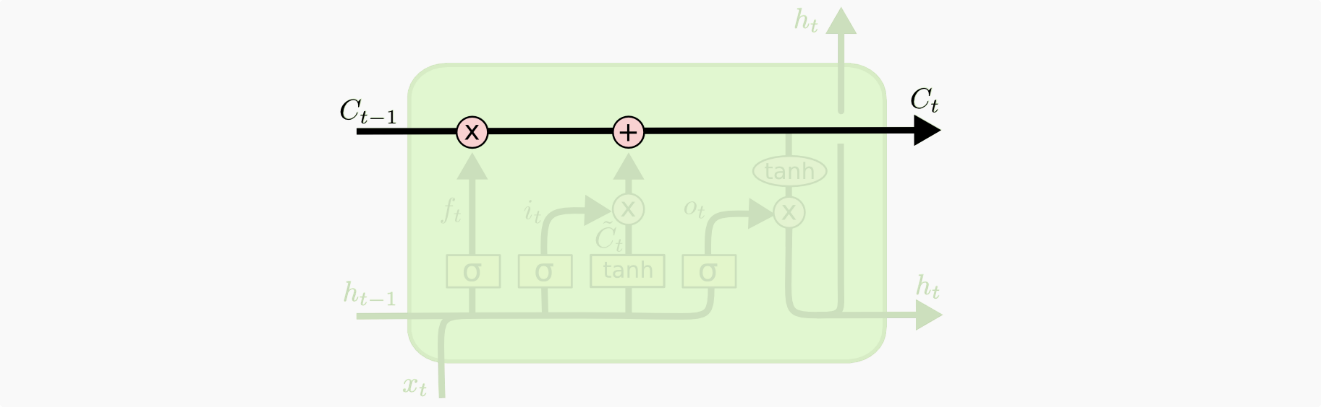


*Hình 2. Các kí hiệu dùng trong LSTM*

* + 1. Ý tưởng cốt lõi của LTSM

Chìa khóa của LSTM là trạng thái tế bào (cell state)

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 3. Trạng thái tế bào*

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.



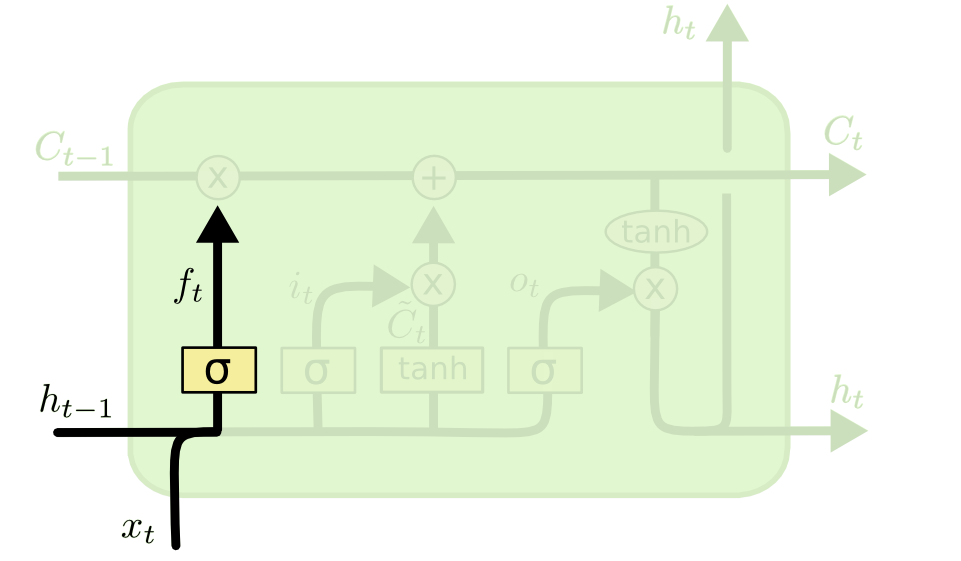
*Hình 4. Cổng sàng lọc*

Tầng sigmoid sẽ cho đầu ra là một số trong khoản [0, 1][0,1], mô tả có bao nhiêu thông tin có thể được thông qua. Khi đầu ra là 00 thì có nghĩa là không cho thông tin nào qua cả, còn khi là 11 thì có nghĩa là cho tất cả các thông tin đi qua nó.

Một LSTM gồm có 3 cổng như vậy để duy trì và điều hành trạng thái của tế bào.

* 1. Hoạt động
     1. Bước 1

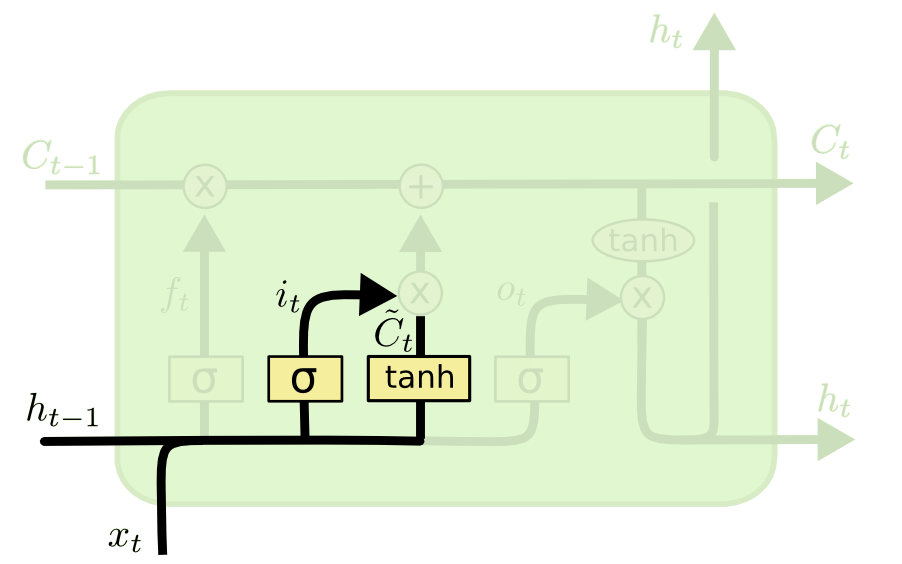
Bước đầu tiên của LSTM là quyết định xem thông tin nào cần bỏ đi từ trạng thái tế bào. Quyết định này được đưa ra bởi tầng sigmoid - gọi là “tầng cổng quên” (forget gate layer). Nó sẽ lấy đầu vào là *ht*−1 và xt ds rồi đưa ra kết quả là một số trong khoảng [0, 1][0,1] cho mỗi số trong trạng thái tế bào C\_{t-1}*Ct*−1​. Đẩu ra là 11 thể hiện rằng nó giữ toàn bộ thông tin lại, còn 00 chỉ rằng toàn bộ thông tin sẽ bị bỏ đi.



*Hình 5. Chọn lọc thông tin đầu vào cho lstm*

* + 1. Bước 2

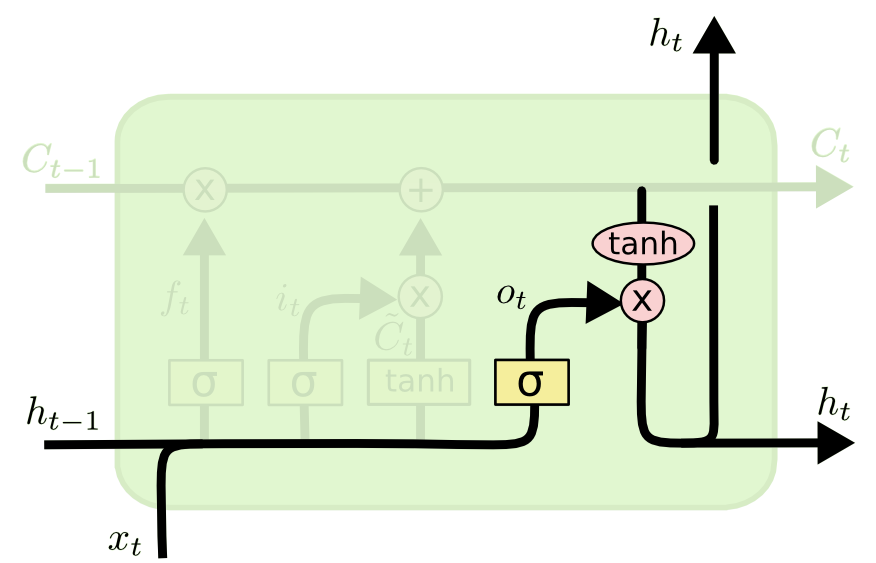
Bước tiếp theo là quyết định xem thông tin mới nào ta sẽ lưu vào trạng thái tế bào. Việc này gồm 2 phần. Đầu tiên là sử dụng một tầng sigmoid được gọi là “tầng cổng vào” (input gate layer) để quyết định giá trị nào ta sẽ cập nhập. Tiếp theo là một tầng *tanh* tạo ra một véc-tơ cho giá trị mới  nhằm thêm vào cho trạng thái. Trong bước tiếp theo, ta sẽ kết hợp 2 giá trị đó lại để tạo ra một cập nhập cho trạng thái.



*Hình 6. Chọn lọc thông tin để lưu vào tế bào*

* + 1. Bước 3

Giá trị đầu ra sẽ dựa vào trạng thái tế bào, nhưng sẽ được tiếp tục sàng lọc. Đầu tiên, ta chạy một tầng sigmoid để quyết định phần nào của trạng thái tế bào ta muốn xuất ra. Sau đó, ta đưa nó trạng thái tế bảo qua một hàm *tanh* để có giá trị nó về khoảng [-1, 1][−1,1], và nhân nó với đầu ra của cổng sigmoid để được giá trị đầu ra ta mong muốn.



*Hình 7. Cho ra kết quả sau quá trình phân loại*

1. THỬ NGHIỆM
   1. Dữ liệu thử nghiệm

Dữ liệu thử nghiệm được dùng trong bài được trích từ nguồn dữ liệu giao dịch của Yahoo Finance.

Có hai luồng dữ liệu chính:

* Luồng dữ liệu thứ nhất gồm giá trị giao dịch của 2 công ty: Google và Apple với khung thời gian từ 01-01-1998 đến 24-04-2020. Phục vụ việc mô phỏng quá trình dự đoán ngắn hạn hàng ngày.
* Luồng dữ liệu thứ hai gồm giá trị giao dịch được phép tải về của Yahoo Finance do công ty Carriage Services, Inc. (CSV) cung cấp.Có khung thời gian từ 08/09/1996 đến 12/03/2020. Phục vụ việc mô phỏng quá trình dự đoán trong 30 ngày tiếp theo của tháng trước.
  1. Môi trường thử nghiệm
* Môi trường thực hiện: Google Colab
* Ngôn ngữ thực hiện: Python 3
* Các thư viện (dùng cho python):
  + Math
  + Pandas\_datareader
  + Numpy
  + Pandas
  + MinMaxScaler (sklearn.preprocessing)
  + Sequential (keras.models)
  + Dense, LSTM (keras.layers)
  + Mathplotlib.pyplot
  + Plotly.express
  + Plotly.graph\_objects
  + TimeseriesGenerator (keras.preprocessing.sequencec)
* Cài đặt của lstm:

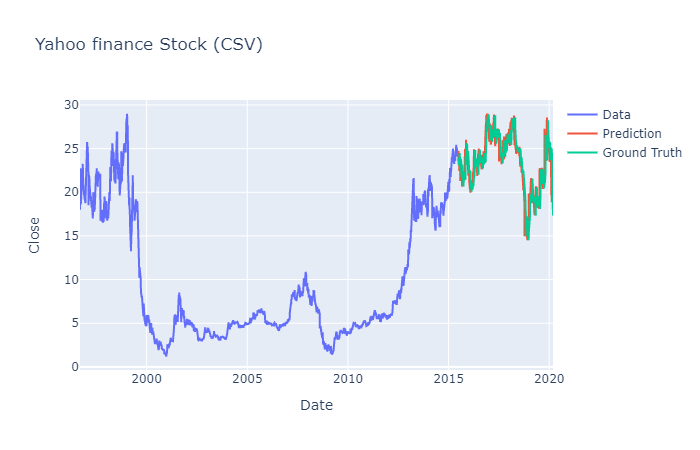
Bảng 1. Giá trị cài đặt của luồng dữ liệu thứ nhất

|  |  |  |
| --- | --- | --- |
| Thứ tự | Chỉ số | Giá trị |
| 1 | Thời gian | 01-01-1998 ~ 24-04-2020 |
| 2 | Batch\_size | 3 |
| 3 | Epochs | 5 |
| 4 | Optimizer | adam |
| 5 | Loss | mean\_squared\_error |
| 6 | Số ngày dự đoán | 1 |

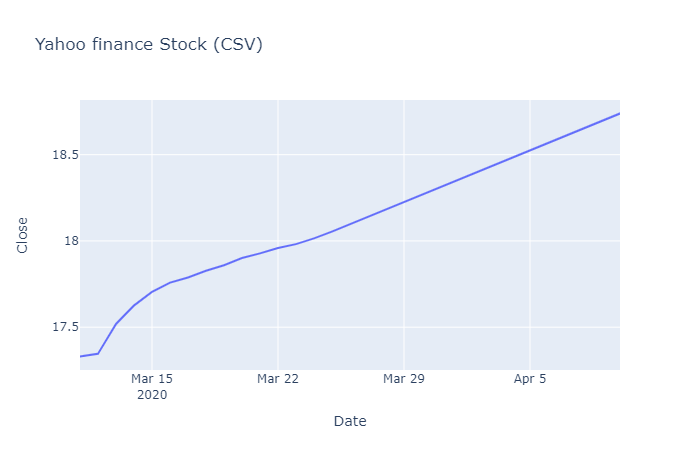
Bảng 2. Giá trị cài đặt của luồng dữ liệu thứ hai

|  |  |  |
| --- | --- | --- |
| Thứ tự | Chỉ số | Giá trị |
| 1 | Thời gian | 08-09-1996 ~ 12-03-2020 |
| 2 | Batch\_size | 2 |
| 3 | Epochs | 25 |
| 4 | Optimizer | adam |
| 5 | Loss | mean\_squared\_error |
| 6 | Số ngày dự đoán | 30 |

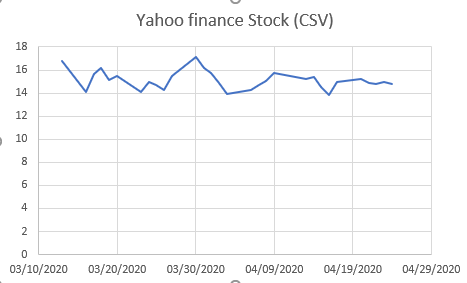
* 1. Kết quả
     1. Luồng dữ liệu thứ 1
     2. Luồng dữ liệu thứ 2



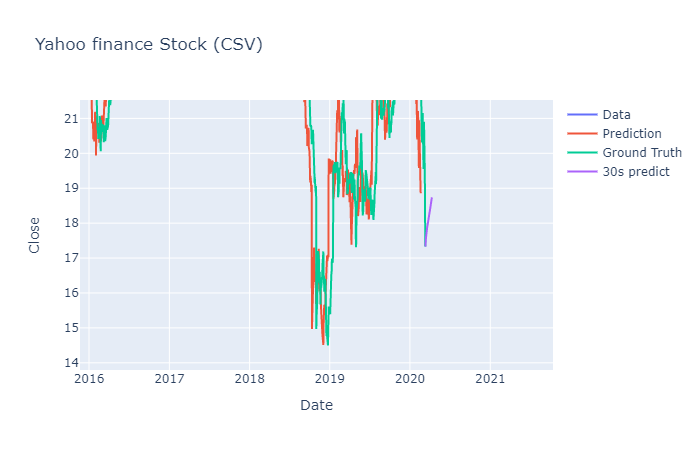
*Hình 8. Kết quả sau quá trình training lstm*



*Hình 9. Dự đoán của lstm cho 30 ngày tiếp theo*

**

*Hình 10. Giá trị thực*



*Hình 11. Tổng hợp của quá trình training và dự đoán*

1. KẾT LUẬN

TÀI LIỆU THAM KHẢO