TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN

———————— * ————————

BÁO CÁO MÔN DỰ ÁN CÔNG NGHỆ THÔNG TIN 1

TÌM HIỂU VỀ TIME SERIES FORECASTING TRONG DEEP LEARNING (PHẦN 2)

Cán bộ hướng dẫn: **PGS. TS. Nguyễn Thanh Hiên**
Giảng viên giám sát: **TS. Huỳnh Ngọc Tú**
Người thực hiện: **Trần Quốc Lĩnh - 51703124**

TP. Hồ Chí Minh, ngày 20 tháng 08 năm 2020

LỜI CẢM ƠN

Em xin chân thành cám ơn khoa Công Nghệ Thông Tin và trường Đại Học Tôn Đức Thắng và Công ty TNHH Tin học Đại Phát. Với sự giúp đỡ của trường, khoa và công ty, em đã hoàn thành báo cáo môn dự án công nghệ thông tin 1 với đề tài: Tìm hiểu về time series forecasting trong deep learning (phần 2)

Trong quá trình soạn thảo không thể tránh khỏi những thiếu sót. Em rất mong nhận được ý kiến đóng góp của các thầy cô để hoàn thiện tốt hơn báo cáo của em. Và cũng như để nâng cao kiến thức, kinh nghiệm của bản thân. Xin chân thành cám ơn!

TP. Hồ Chí Minh, Ngày 20 tháng 08 năm 2020

Trần Quốc Lĩnh

BÀI BÁO CÁO ĐƯỢC HOÀN THÀNH TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

Em xin cam đoan đây là sản phẩm nghiên cứu của riêng em. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu ,hình ảnh được chính em thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong bài tiểu luận còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào em xin hoàn toàn chịu trách nhiệm về nội dung bài tập lớn của mình. Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do em gây ra trong quá trình thực hiện (nếu có).

TP. Hồ Chí Minh, Ngày 20 tháng 08 năm 2020

Tác giả

Trần Quốc Lĩnh

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

Phần xác nhận của GV hướng dẫn

.....

.....

.....

.....

.....

.....

TP. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

Phần đánh giá của GV chấm bài

.....

.....

.....

.....

.....

.....

TP. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

Tóm tắt

Ở bài báo cáo trước em đã trình bài một cái tổng quát và khái lượt các nội dung về Time series forecasting. Thế nên, trong bài báo cáo này em sẽ trình bày một chút cụ thể hơn, chi tiết hơn thông qua các mô hình dự báo được ưa chuộng và hiệu quả nhất ở thời điểm hiện tại. Sau đó, đánh giá các mô hình trong một ví dụ thực tiễn.

Mục lục

Danh sách hình vẽ

Chương 1

Một số mô hình time series forecasting

Để thực hiện, xây dựng các mô hình chuỗi thời gian, hiện tại có hai ngôn ngữ hỗ trợ tốt cho việc này là R (R có các packages như forecast và lmtest) và python. Mặc dù python dễ sử dụng và có cộng đồng lớn mạnh hơn nhiều so với R, nhưng R lại hỗ trợ tốt hơn python trong thống kê và hiện thực hóa mô hình chuỗi thời gian. Đó cũng là một trong những lý do mà các nhà thống kê và kinh tế lượng ưa chuộng sử dụng R. Tuy nhiên, ở bài báo cáo này chúng ta chỉ dừng lại với python trong việc tìm hiểu về cách xây dựng các mô hình.

1.1 Autoregressive Integrated Moving Average (ARIMA)

Thông qua bài báo cáo trước, ắt hẳn chúng ta đã biết rằng chuỗi thời gian là các thông số được ghi lại theo thời gian, theo các mốc thời gian xác định. Và giá trị của hiện tại có sự tương quan đến các giá trị trong quá khứ, dễ hiểu hơn là để xác định giá trị trong thời điểm hiện tại cần có các giá trị trong quá khứ.

Mô hình ARIMA là một mô hình dựa trên ý tưởng dự đoán giá trị tương tai bằng các thông số đã ghi nhận được trong quá khứ. Và giả thuyết chuỗi thời gian là chuỗi dừng và phương sai của sai số không đổi

ARIMA gồm AR (Auto regression), I (Intergrated) kết hợp với MA (Moving Average). Trong đó:

- **Auto regression**: Có nghĩa là tự hồi qui. Thành phần hồi qui này gồm một tập hợp các giá trị lùi về p bước thời gian của chuỗi. Được biểu diễn dưới dạng:

$$AR(p) = \sum_{i=0}^{p} \phi x_{t-0}^{0} = \phi_0 + \phi x_{t-1} + \phi x_{t-2} + ... + \phi x_{t-p}$$

- **Moving average**: Có nghĩa là trung bình trượt. Đây là quá trình dịch chuyển - quá trình thay đổi giá trị trung bình của chuỗi. Tuy nhiên chuỗi này phải thõa mãn các tính chất sau:

$$\begin{cases} \mathrm{E}(\varepsilon_t) = 0 & (0) \\ \sigma(\varepsilon_t) = \alpha & (2) \\ \rho(\varepsilon_t, \varepsilon_{t-s}) = 0, \forall s <= t & (3) \end{cases}$$

  Trong đó: (1) nghĩa là kỳ vọng phải bằng không và (2) phương sai không đổi để đảm bảo tính dừng của chuỗi.

  Quá trình Moving average sẽ tìm mối liên hệ về mặt tuyến tính giữa các phần tử ngẫu nhiên ε_t của giá trị hiện tại và quá khứ (stochastic term). Do kì vọng và phương sai không đổi nên chúng ta gọi phân phối của nhiễu trắng là phân phối xác định (identical distribution) và được kí hiệu là $\varepsilon_t \sim \mathrm{WN}(0, \sigma^2)$. Qúa trình trung bình trượt được biểu diễn như sau:

$$\mathrm{MA}(q) = \mu + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

- **Intergrated**: Là quá trình đồng tích hợp hoặc lấy sai phân. Hầu hết các chuỗi thời gian không có tính dừng thường. Do đó ta cần biến đổi nó sang chuỗi dừng bằng sai phân. Khi biến đổi sang chuỗi dừng, các nhân tố ảnh hưởng thời gian được loại bỏ và chuỗi sẽ dễ dự báo hơn. Qúa trình sai phân bậc d của chuỗi được thực hiện như sau:

$$\mathrm{I}(1) = \Delta(x_t) = x_t - x_{t-1} \text{ (sai phân bậc 1)}$$
$$\mathrm{I}(d) = \Delta^d(x_t) = \underbrace{\Delta(\Delta(\ldots\Delta(x_t)))}_{\text{d times}} \text{ (sai phân bậc d)}$$

Như vậy, tham số đặc trưng của mô hình được đặc tả bởi 3 tham số ARIMA(p, d, q). Và có thể được biểu diễn dưới dạng:

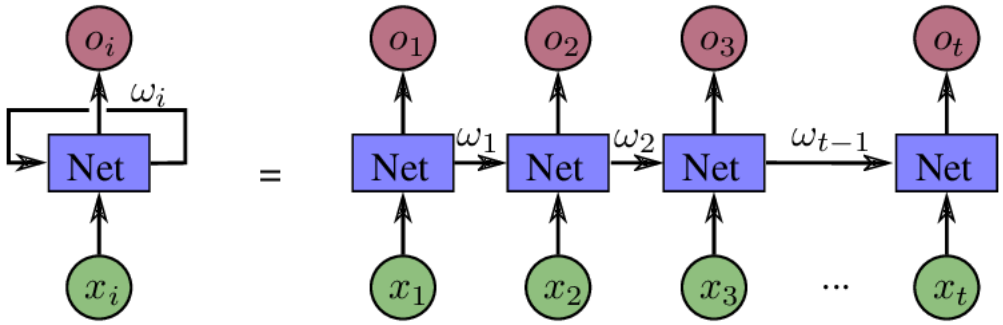
$$\Delta x_t = \phi_1 \Delta x_{t-1} + \phi_2 \Delta x_{t-2} + ... + \phi_p \Delta x_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q}$$

1.2 Long Short Term Memories (LSTM)

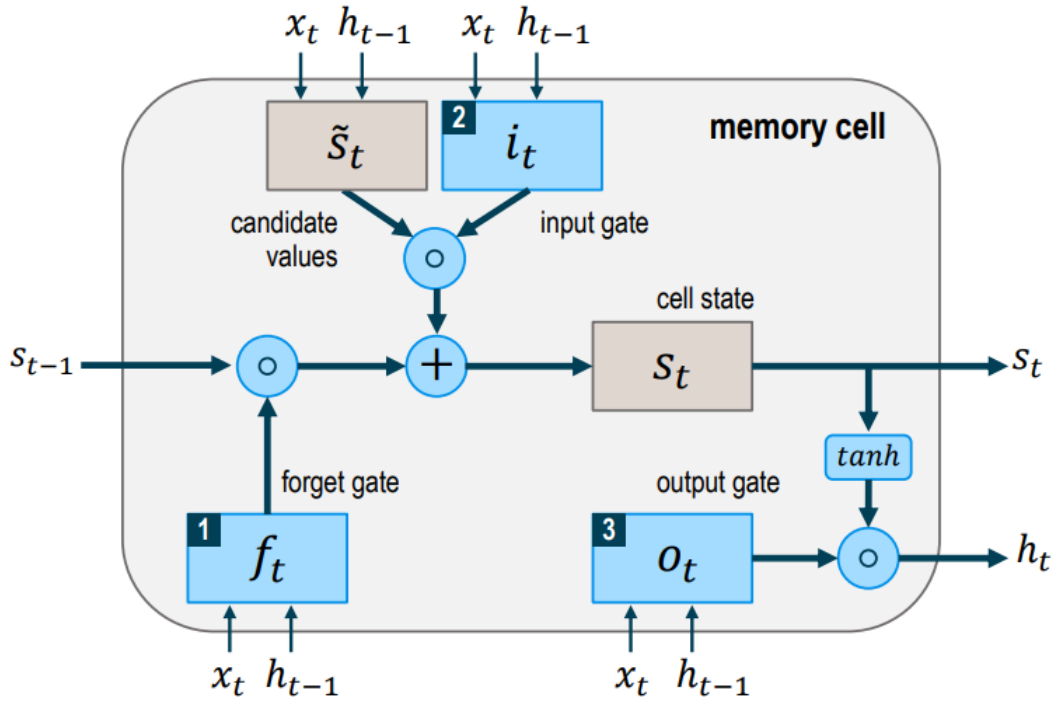
1.2.1 Định nghĩa

Long Short Term Memories (LSTM) là một biến thể của mô hình Recurrent neural network RNN (hình 1) với cấu trúc mạng được tổng hợp từ nhiều đơn vị Long short-term memory. Một đơn vị LSTM gồm có

nhân (cell), một cổng vào (input gate), một cổng ra (output gate) và một cổng quên (forget gate). Sơ đồ minh hoạt cho một đơn vị LSTM được thể hiện ở hình 1.2.



Hình 1.1: Một dạng mô hình RNN



Hình 1.2: Cấu trúc của một nhân LSTM

Vì có khả năng nhớ nên LSTM rất phù hợp với bài toán phân loại và dự báo dựa trên dữ liệu dạng chuỗi thời gian. Hơn nữa LSTM còn có đặt trưng loại bỏ được hiện tượng triệt tiêu (vanishing) và bùng nổ (exploding) gradient.

1.2.2 Cơ chế hoạt động

Như đã giới thiệu ở trên, mỗi đơn vị LSTM có nhiều cổng khác nhau, tương ứng với các chức năng khác nhau:

- **Forget gate:** Cổng này quyết định xem thông tin nào trong bộ nhớ hiện tại được giữ và thông tin nào bị bỏ lại. Thông tin đầu vào được cho vào hàm signoid. Đầu ra của hàm này đóng vai trò là mask để lọc thông tin từ trạng thái nhân.
- **Input gate:** Cổng này dùng để cập nhật bộ nhớ với các thông tin mới. Ở đây có xuất hiện 2 hàm sigmoid và hàm tanh. Tác dụng của chúng cũng như trên. Output từ hàm sigmoid sẽ có tác dụng lọc thông tin đã qua xử lý từ output hàm tanh.
- **Output gate:** cổng này quyết định output của từ hiện tại là gì. Nó được lấy thông tin từ 2 nguồn: trạng thái nhân và input hiện tại. Trạng thái cell sau khi chỉnh sửa sẽ đi qua hàm tanh và input hiện tại thì được đi qua hàm sigmoid. Từ đây ta kết hợp 2 kết quả trên để có được kết quả đầu ra. Chú ý rằng cả kết quả đầu ra và cả trạng thái cell đều được đưa vào bước tiếp theo.

Các biến số

Trước khi đi vào chi tiết quá trình hoạt động của LSTM, ta cần hiểu ý nghĩa của các giá trị dưới đây:

- x_t là vector đầu vào tại bước thời gian t
- $W_{f,x}, W_{f,h}, W_{\tilde{s},x}, W_{\tilde{s},h}, W_{i,x}, W_{i,h}, W_{o,x}, W_{o,h}$ là các ma trận trọng số trong mỗi tế bào.
- $b_f, b_{\tilde{s}}, b_i, b_o$ là các véc tơ bias.
- f_t, i_t, o_t là các hàm kích hoạt tại forget gate, input gate và output gate.
- $s_t, \tilde{s}$ là véc tơ trạng thái của nhân và giá trị đề cử (candidate value).
- h_t là tham số đầu vào của mỗi tế bào LSTM.

Forward

Quá trình lan truyền xuôi (forward) của mạng LSTM được thực hiện như sau:

1. Đầu tiên, một giá trị h_{t-1} ngẫu nhiên (hoặc được đặt sẵn) được tạo ra và làm tham số đầu vào cho nhân LSTM đầu tiên.
2. Kế đến, nhân LSTM sẽ quyết định thông tin nào cần loại bỏ trong s_{t-1} thông qua hàm f_t dựa trên ba giá trị x_t, h_{t-1} và bias b_f của f_t.

$$f_t = \sigma\left(W_{f,x}x_t + W_{f,h}h_{t-1} + b_f\right)$$

$$i_t = \tanh(W_{i,x}x_t + W_{i,h}h_{t-1} + b_i)$$

3. Sau đó, nhân LSTM sẽ quyết định thông tin nào sẽ được thêm vào s_t thông qua phép toán:

$$\widetilde{s_t} = \tanh\left(W_{\widetilde{s},x}x_t + W_{\widetilde{s},h}h_{t-1} + b_{\widetilde{s}}\right)$$

4. Tiếp theo, giá trị s_t sẽ được tính toán với phép nhân Hadamard theo từng phần tử $\circ$ (Hadamard prodcut).

$$s_t = f_t \circ s_{t-1} + i_t \circ \widetilde{s_t}$$

5. Ở bước cuối cùng của một nhân, giá trị đầu ra h_t của nhân hiện tại và là đầu vào của nhân kế đó được tính bằng:

$$h_t = o_t \circ \tanh\left(s_t\right)$$

   Trong đó:

$$o_t = \sigma\left(W_{o,x}x_t + W_{o,h}h_{t-1} + b_o\right)$$

6. Quá trình 1-5 sẽ được thực hiện tuần tự lặp đi lặp lại cho đến nhân LSTM cuối cùng.

Chương 2

Đánh giá các mô hình dự báo

2.1 Mục tiêu

Phần này sẽ trình bày về kết quả dự báo của hai mô hình đã trình bày ở phần trước, thông qua bài toán thực tế. Với mục tiêu dự đoán chính xác nhất có thể doanh số bán dầu gọi trong những tháng sắp tới dựa trên bộ dữ liệu đã cho.

2.2 Dữ liệu

Shampoo Sales dataset. Đây là bộ dữ liệu về doanh số bán dầu gội đầu theo từng tháng trong vòng 3 năm (36 quan sát)

2.3 Các gói cài đặt cần thiết

- Python3
- SciPy
- Keras
- scikit-learn
- Pandas
- Numpy
- Matplotlib

2.4 Kết quả

2.4.1 ARIMA

2.4.2 LSTM

Chương 3

Tổng kết

Kết quả

Kết quả đạt được

Như vậy chúng ta đã đi qua hai phần của bài báo cáo time series forecasting trong deep learning. Hy vọng qua hai bài báo cáo này thì người đọc có thể nắm chắc được nền tảng kiến thức căn bản của mô hình time series và ứng dụng nó vào thực tế, nhằm phục vụ cho các nghiên cứu của cá nhân hoặc phục vụ mục đích tìm hiểu của bản thân.

Hạn chế

Bài báo cáo tuy đã trình bài chi tiết về các mô hình (ARIMA và LSTM), tuy nhiên vẫn còn một vài mô hình dự báo đáng chú ý khác vẫn chưa được đề cập trong bài báo cáo này. Người đọc có thể tìm hiểu thêm: RNN, ResNet, ode_gru_bayes,...

Hướng phát triển

Từ những kiến thức đã có, hy vọng người đọc có thể tự xây dựng một mô hình phù hợp với nhu cầu và mục đích sử dụng của riêng mình. Hoặc có thể nghiên cứ, đề ra một phát kiến mới đóng góp cho sự phát triển chung của ngành AI nói chung và lĩnh vực time series forecasting nói riêng.

Tài liệu tham khảo

[1] Jason Brownlee. *Deep Learning for Time Series Forecasting.*

[2] Jason Brownlee. How to create an arima model for time series forecasting in python.

[3] Jason Brownlee. Time series prediction with lstm recurrent neural networks in python with keras.

[4] Minh-Hoang Bui. Hướng dẫn chi tiết về cơ chế của lstm và gru trong nlp.

[5] Alexiei Dingli and Karl Sant Fournier. Financial time series forecasting – a deep learning approach.

[6] Piotr Fryzlewicz. Financial time series, arch and garch models.

[7] ADAM HAYES. Simple moving average (sma).

[8] Pham Dinh Khanh. Mô hình arima trong time series.

[9] Bryan Lim and Stefan Zohren. Time series forecasting with deep learning: A survey.

[10] Nguyễn Trường Long. Giải thích chi tiết về mạng long short-term memory (lstm).

[11] Nttuan8. Long short term memory (lstm).

[12] Wikipedia. Time series.

[13] Tony Zhou. Deep learning for time series and why deep learning?

[14] Phạm Đình Khánh. Lstm.

[15] Hoàng Đức Quân. Time-series data.