ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

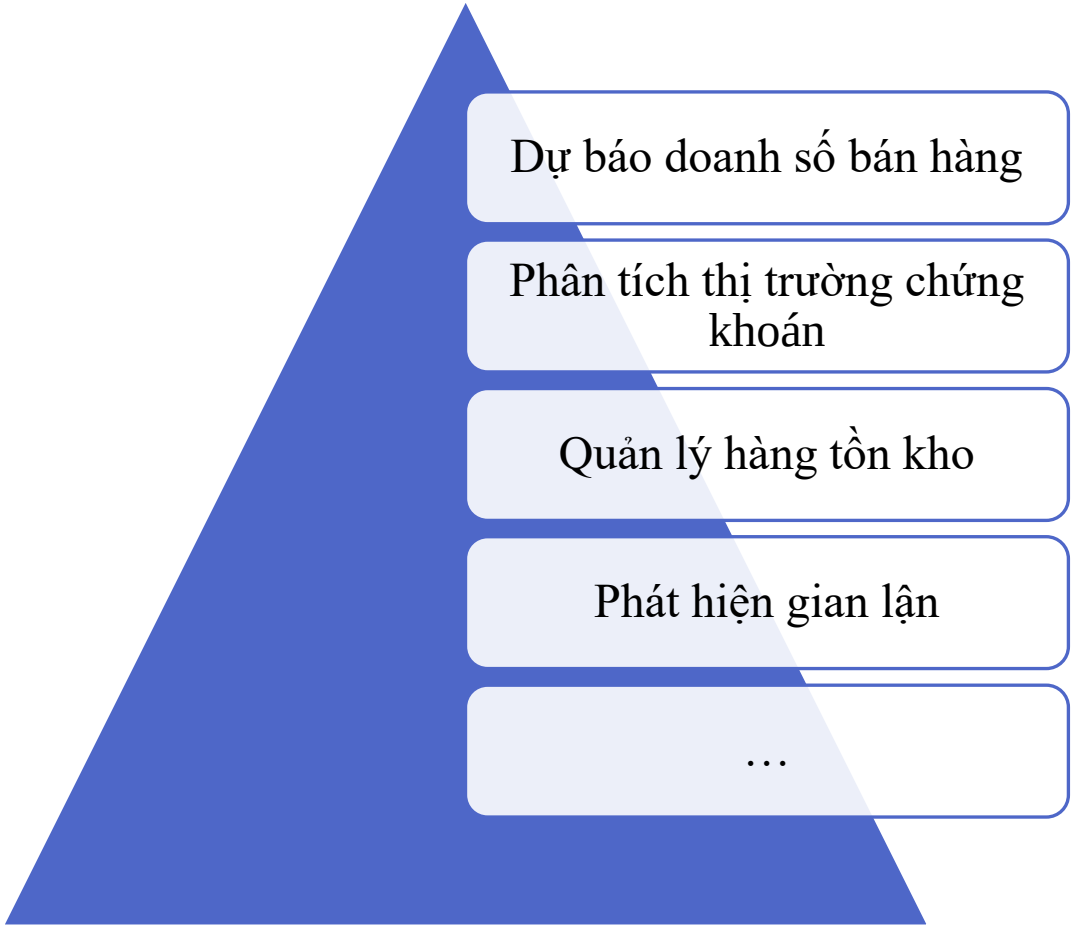
CHƯƠNG 10
Phân tích hồi quy và chuỗi thời gian cho doanh nghiệp

Biên soạn: ThS. Nguyễn Thị Anh Thư

Nội dung

1. Giới thiệu
2. Quy trình triển khai
3. Lựa chọn thuật toán
4. Đánh giá hiệu quả
5. Triển khai hệ thống
6. Giám sát và bảo trì
7. Tổng kết

1. Giới thiệu



Hệ thống phân tích hồi quy và khai phá chuỗi thời gian là những công cụ mạnh mẽ được sử dụng để hiểu và dự đoán các xu hướng trong dữ liệu theo thời gian.

Được sử dụng trong nhiều lĩnh vực khác nhau.

Hồi quy (Regression)

Học có giám sát (Supervised Learning)

Hồi quy (Regression)

Thuật toán Hồi quy sử dụng dữ liệu đã được dán nhãn để học mối quan hệ giữa biến phụ thuộc (dependent variable) và các biến độc lập (independent variables), từ đó dự đoán giá trị của biến phụ thuộc cho các dữ liệu mới.

Mục đích:

- Dự đoán giá trị của một biến trong tương lai.
- Tìm kiếm mối quan hệ giữa các biến.
- Xây dựng mô hình để giải thích các hiện tượng.

Chuỗi thời gian (Time series)

Chuỗi thời gian là một tập dữ liệu được thu thập theo thời gian, bao gồm các giá trị được đo lường tại các thời điểm khác nhau.

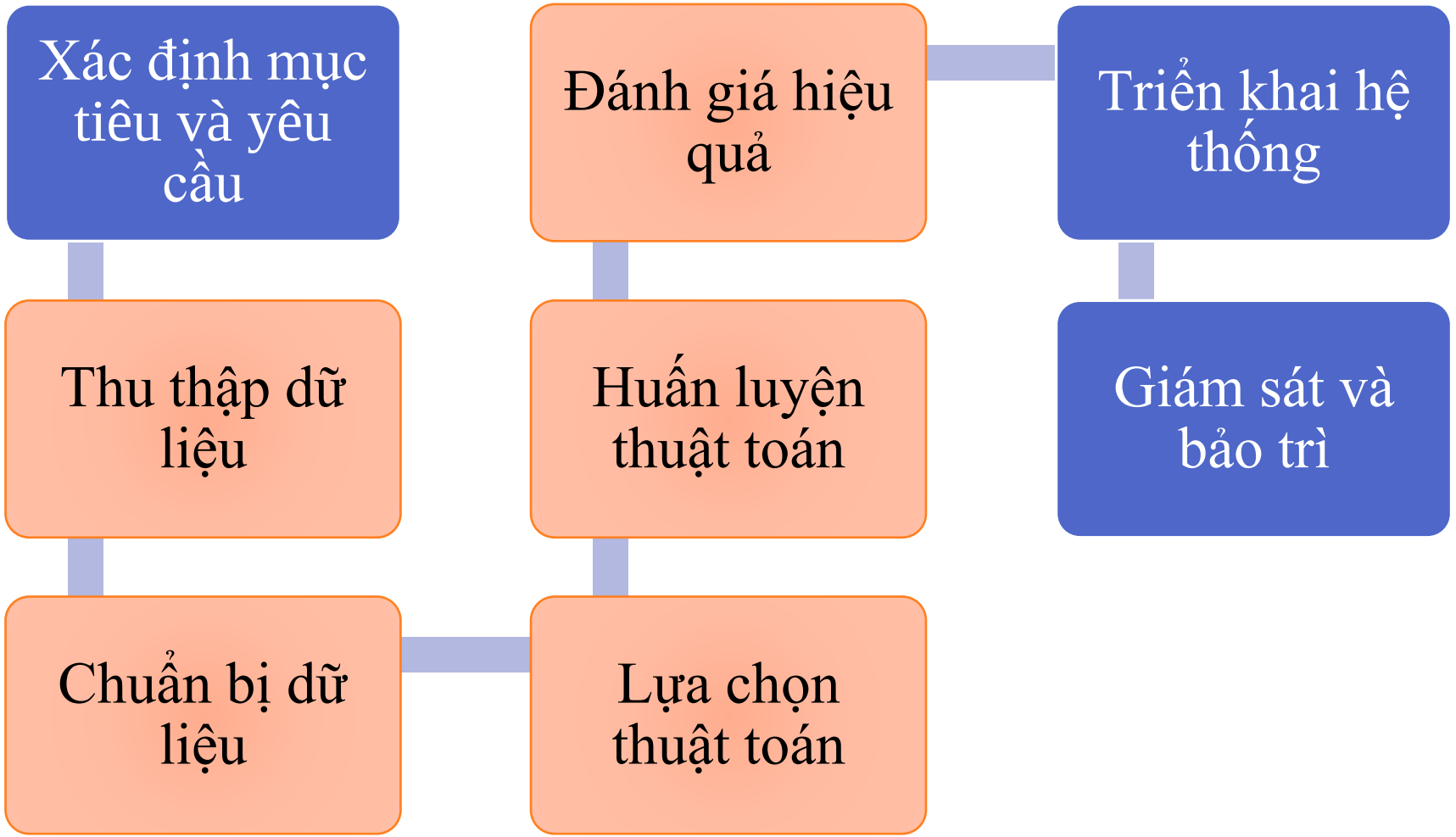
Thuật toán chuỗi thời gian phân tích dữ liệu chuỗi thời gian và trích xuất thông tin hữu ích từ dữ liệu.

- Có thể thuộc nhiều phương thức học khác nhau, tùy thuộc vào loại bài toán và cách thức áp dụng.

Mục đích:

- Dự đoán giá trị trong tương lai.
- Phát hiện các xu hướng và chu kỳ trong dữ liệu.
- Xác định các điểm bất thường (anomaly) trong dữ liệu.
- Hiểu rõ hơn về hành vi của hệ thống.

2. Quy trình triển khai



2. Quy trình triển khai

1. Xác định mục tiêu và yêu cầu:

- Xác định rõ ràng mục tiêu của việc triển khai hệ thống, ví dụ như dự báo doanh số bán hàng, phát hiện gian lận, hay tối ưu hóa mức tồn kho.
- Xác định các yêu cầu về độ chính xác, hiệu suất và khả năng giải thích của hệ thống.

2. Thu thập dữ liệu:

- Thu thập dữ liệu chuỗi thời gian có liên quan đến mục tiêu phân tích.
- Đảm bảo dữ liệu chất lượng cao, đầy đủ và chính xác.
- Xử lý các giá trị thiếu và bất thường trong dữ liệu.

2. Quy trình triển khai

3. Chuẩn bị dữ liệu:

- Chuẩn bị dữ liệu cho mô hình hồi quy hoặc thuật toán khai phá chuỗi thời gian.
- Có thể bao gồm việc chia dữ liệu thành tập huấn luyện, tập kiểm tra và tập xác thực.
- Áp dụng các kỹ thuật chuẩn hóa dữ liệu nếu cần thiết.

4. Lựa chọn thuật toán:

- Lựa chọn thuật toán hồi quy hoặc kỹ thuật khai phá chuỗi thời gian phù hợp với mục tiêu và loại dữ liệu.

5. Huấn luyện thuật toán:

- Huấn luyện thuật toán được lựa chọn trên tập dữ liệu huấn luyện.
- Điều chỉnh các tham số của thuật toán để đạt được hiệu quả tốt nhất.

2. Quy trình triển khai

6. Đánh giá hiệu quả:

- Đánh giá hiệu quả của thuật toán trên tập dữ liệu kiểm tra.
- Sử dụng các chỉ số hiệu suất phù hợp như độ chính xác, độ sai lệch trung bình, v.v.
- So sánh hiệu quả của các thuật toán khác nhau.

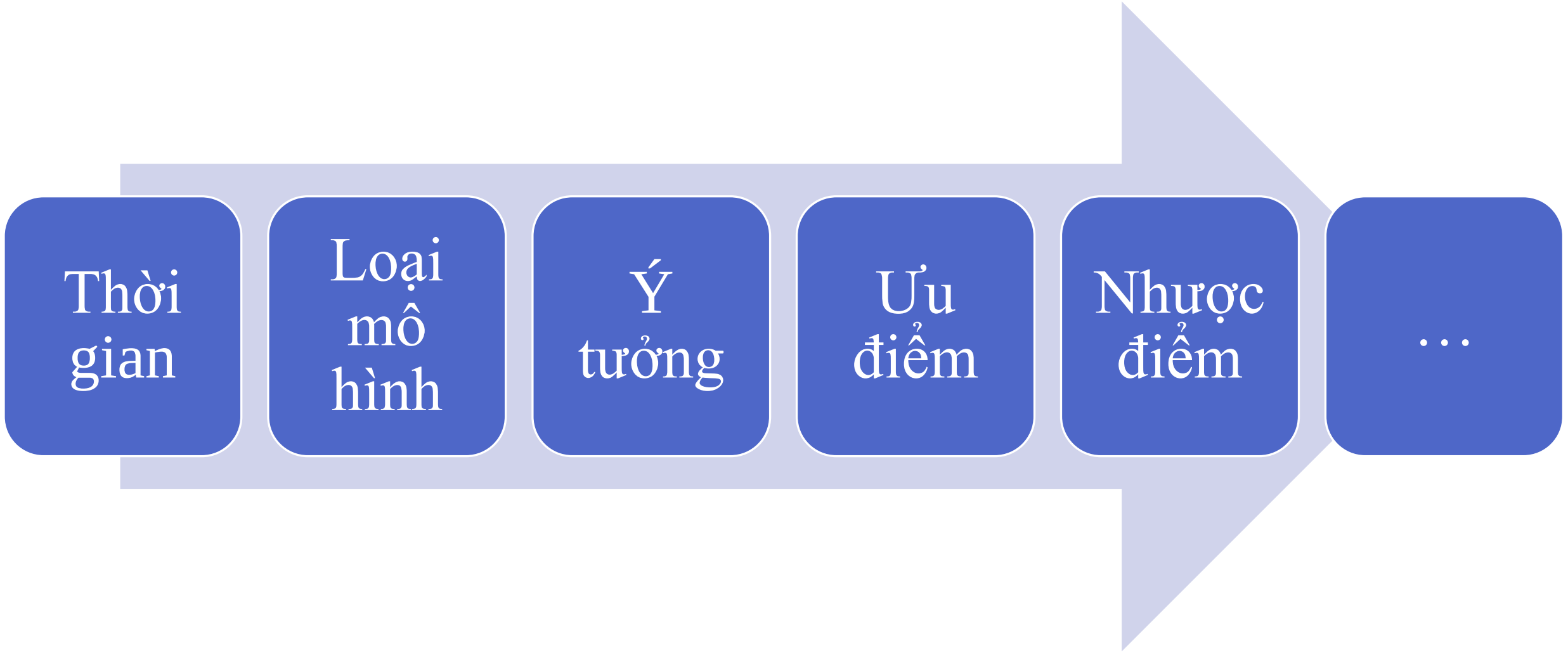
7. Triển khai hệ thống:

- Triển khai hệ thống vào môi trường thực tế.
- Cung cấp giao diện người dùng để tương tác với hệ thống.

8. Giám sát và bảo trì:

- Giám sát hiệu suất của hệ thống theo thời gian.
- Cập nhật dữ liệu và thuật toán khi cần thiết.

3. Lựa chọn thuật toán



Hồi quy (Regression)

Thuật toán	Ý tưởng	Ưu điểm	Nhược điểm	Trường hợp áp dụng
Hồi quy tuyến tính	Sử dụng một đường thẳng để mô tả mối quan hệ giữa biến phụ thuộc và biến độc lập.	Đơn giản, dễ hiểu, dễ giải thích kết quả.	Không phù hợp với dữ liệu phi tuyến tính.	Dự báo giá cả, phân tích mối quan hệ giữa các biến kinh tế.
Hồi quy logistic	Sử dụng hàm logistic để mô tả mối quan hệ giữa biến phụ thuộc nhị phân (0 hoặc 1) và biến độc lập.	Phù hợp với dữ liệu nhị phân, có khả năng giải thích kết quả.	Có thể gặp khó khăn trong việc tối ưu hóa mô hình.	Phân loại email rác, phân tích dữ liệu y tế.
Hồi quy lũy kế	Sử dụng cây quyết định để chia dữ liệu thành các nhóm con và xây dựng mô hình hồi quy riêng cho từng nhóm.	Có thể xử lý dữ liệu phi tuyến tính và tương tác, hiệu quả cao với dữ liệu lớn.	Có thể khó giải thích kết quả, mô hình có thể bị quá phức tạp.	Dự báo rủi ro tín dụng, phân tích hành vi khách hàng.
Hồi quy mạng nơ-ron nhân tạo	Sử dụng mạng nơ-ron nhân tạo để học hỏi mối quan hệ phức tạp giữa biến phụ thuộc và biến độc lập.	Có thể xử lý dữ liệu phi tuyến tính và tương tác phức tạp, hiệu quả cao với dữ liệu lớn.	Có thể khó giải thích kết quả, mô hình có thể bị quá phức tạp, đòi hỏi nhiều dữ liệu để huấn luyện.	Nhận dạng hình ảnh, xử lý ngôn ngữ tự nhiên, dự báo thị trường chứng khoán.

Chuỗi thời gian (Time series)

Thuật toán	Ý tưởng	Ưu điểm	Nhược điểm	Trường hợp áp dụng
Trung bình di động (MA)	Tính toán trung bình của một cửa sổ cố định các giá trị gần đây trong chuỗi thời gian.	Đơn giản, dễ hiểu, hiệu quả với dữ liệu có xu hướng rõ ràng.	Không phù hợp với dữ liệu có xu hướng thay đổi nhanh chóng hoặc có nhiễu cao.	Dự báo giá cả ngắn hạn, làm mịn dữ liệu chuỗi thời gian.
Làm mịn theo cấp số mũ (EWMA)	Gán trọng số cao hơn cho các giá trị gần đây trong cửa sổ so với các giá trị cũ hơn.	Phù hợp hơn với dữ liệu có xu hướng thay đổi nhanh chóng so với MA.	Có thể nhạy cảm với nhiễu trong dữ liệu.	Dự báo giá cả ngắn hạn, theo dõi hiệu suất thị trường chứng khoán.
Phân tích ARIMA	Sử dụng mô hình thống kê để mô tả mối quan hệ giữa các giá trị hiện tại, giá trị quá khứ và lỗi dự đoán trong chuỗi thời gian.	Có thể mô tả các mối quan hệ phức tạp giữa các giá trị trong chuỗi thời gian.	Có thể khó lựa chọn mô hình ARIMA phù hợp và giải thích kết quả.	Dự báo giá cả dài hạn, phân tích dữ liệu kinh tế.
Mạng nơ-ron nhân tạo RNN	Sử dụng mạng nơ-ron nhân tạo để học hỏi các mối quan hệ phụ thuộc thời gian trong chuỗi thời gian.	Có thể xử lý dữ liệu phi tuyến tính và phụ thuộc thời gian phức tạp.	Có thể khó huấn luyện và giải thích kết quả.	Nhận dạng giọng nói, dịch máy, dự báo thị trường chứng khoán.
Transformer	Sử dụng kiến trúc mạng nơ-ron nhân tạo để xử lý các chuỗi dữ liệu dài.	Hiệu quả cao với dữ liệu chuỗi dài, có thể học hỏi các mối quan hệ phụ thuộc thời gian dài hạn.	Có thể khó huấn luyện và giải thích kết quả.	Xử lý ngôn ngữ tự nhiên, dịch máy, tóm tắt văn bản.

4. Đánh giá hiệu quả

Các tham số:

- **y_i:** Giá trị thực tế của biến phụ thuộc tại điểm dữ liệu thứ i.
- **ŷ_i:** Giá trị dự đoán của biến phụ thuộc tại điểm dữ liệu thứ i do mô hình đưa ra.
- **n:** Số điểm dữ liệu.
- **$\bar{y}$:** Giá trị trung bình của biến phụ thuộc.
- **$\bar{\hat{y}}$:** Giá trị trung bình của các giá trị dự đoán.

4. Đánh giá hiệu quả

Sai số tuyệt đối trung bình (MAE):

- Công thức: **MAE = 1/n * Σ |y_i - ŷ_i|**
- Ý nghĩa: MAE đo lường mức độ sai lệch trung bình giữa giá trị dự đoán (ŷ_i) và giá trị thực tế (y_i) cho tất cả các điểm dữ liệu (n là số điểm dữ liệu).
- Ví dụ: Giả sử ta có mô hình để dự đoán giá nhà. Giá trị MAE là 10 triệu đồng cho thấy mô hình dự đoán trung bình sai lệch 10 triệu đồng so với giá thực tế.

Sai số bình phương trung bình gốc (RMSE):

- Công thức: **RMSE = √(1/n * Σ (y_i - ŷ_i)^2)**
- Ý nghĩa: RMSE tương tự như MAE nhưng nhạy cảm hơn với những sai lệch lớn. RMSE cao cho thấy mô hình dự đoán kém chính xác hơn.
- Ví dụ: Giả sử ta có mô hình để dự đoán giá nhà. Giá trị RMSE là 20 triệu đồng cho thấy mô hình dự đoán trung bình sai lệch 20 triệu đồng so với giá thực tế, với những sai lệch lớn có thể cao hơn 20 triệu đồng.

4. Đánh giá hiệu quả

Hệ số tương quan (R):

- Công thức: **$R = \Sigma ((y\_i - \bar{y})(\hat{y}\_i - \bar{\hat{y}})) / \sqrt{(\Sigma (y\_i - \bar{y})^2 * \Sigma (\hat{y}\_i - \bar{\hat{y}})^2)}$**
- Ý nghĩa: R đo lường mức độ liên hệ tuyến tính giữa giá trị dự đoán và giá trị thực tế. Giá trị R nằm trong [-1, 1], với 1 là hoàn toàn tương quan và -1 là hoàn toàn ngược tương quan. R cao cho thấy mô hình dự đoán có mối liên hệ mạnh với giá trị thực tế.
- Ví dụ: Giả sử ta có mô hình để dự đoán giá nhà. Giá trị R là 0.8 cho thấy mô hình dự đoán có mối liên hệ mạnh với giá nhà thực tế.

Hệ số xác định (R^2):

- Công thức: **$R^2 = R * R$**
- Ý nghĩa: R^2 thể hiện tỷ lệ biến đổi của biến phụ thuộc được giải thích bởi mô hình. R^2 cao cho thấy mô hình giải thích được nhiều biến đổi của biến phụ thuộc.
- Ví dụ: Giả sử ta có mô hình để dự đoán giá nhà. Giá trị R^2 là 0.64 cho thấy mô hình giải thích được 64% biến đổi của giá nhà.

5. Triển khai hệ thống

1. Chuẩn bị:

- **Mô hình:** Mô hình đã được huấn luyện và đánh giá hiệu quả.
- **Ứng dụng web / Ứng dụng di động:** Đã được phát triển và sẵn sàng để tích hợp mô hình.

2. Triển khai mô hình trên server:

- **Môi trường:** Cloud Platforms / On-Premise Servers
- Mô hình được triển khai trên server và ứng dụng sẽ gửi yêu cầu đến server để thực hiện tác vụ.
  - **Lưu trữ mô hình:** Mô hình được lưu trữ trên server.
    - File format: .h5 (HDF5), .pb (Protocol Buffer), .pkl (Pickle), v.v.
  - **Tạo API:** API được tạo ra để nhận yêu cầu từ ứng dụng và gửi kết quả.
    - REST API: Tạo RESTful API bằng các frameworks như Flask, FastAPI, v.v.
  - **Gọi API:** Ứng dụng gửi yêu cầu đến API và nhận kết quả.

6. Giám sát và bảo trì

Giám sát:

- **Theo dõi các chỉ số chính:** Theo dõi thường xuyên các chỉ số phản ánh hiệu suất và tình trạng của mô hình.
- **Phát hiện dữ liệu bất thường:** Phân phối dữ liệu thực tế được cung cấp cho mô hình bắt đầu khác với dữ liệu mà nó được đào tạo.
- **Cảnh báo và ghi nhật ký:** Theo dõi hành vi của mô hình và chẩn đoán các vấn đề dễ dàng hơn.

Bảo trì:

- **Đào tạo lại:** Thường xuyên đào tạo lại mô hình với dữ liệu mới, đặc biệt nếu phát hiện dữ liệu bất thường. Điều này giúp mô hình thích ứng với những thay đổi trong thế giới thực và duy trì hiệu suất tối ưu.

Công cụ và tài nguyên: TensorBoard, Mlflow, Cloud Monitoring Services.

7. Tổng kết

Hồi quy (Regression)

Chuỗi thời gian (Time series)

Question & Answer