Practice 4 – Multivariate Time Series Forecasting
Electricity Load + Weather

1. Mục tiêu của bài thực hành

Trong bài thực hành này, sinh viên sẽ xây dựng và so sánh các mô hình dự báo chuỗi thời gian đa biến dựa trên dữ liệu điện năng và thời tiết. Cụ thể:

- Hiểu cách mô hình hóa chuỗi thời gian đa biến
- Triển khai và so sánh ba nhóm mô hình:
  - VAR (mô hình thống kê)
  - XGBoost (Machine Learning)
  - LSTM (Deep Learning)
- Đánh giá mô hình bằng các chỉ số:
  - RMSE
  - MAE
  - MAPE

2. Dataset và bài toán

Sử dụng **Electricity Load Diagrams Dataset (UCI)**, bao gồm các biến:

- load – điện năng tiêu thụ
- temperature – nhiệt độ
- humidity – độ ẩm
- wind_speed – tốc độ gió

Tại mỗi thời điểm t, vector quan sát được định nghĩa là:

$$X_t = [load_t, temp_t, humidity_t, wind_t]$$

Mục tiêu dự báo: Sử dụng lịch sử 24 giờ gần nhất để dự đoán điện năng tiêu thụ ở bước tiếp theo:

$$X_{t-24:t} \Rightarrow load_{t+1}$$

3. Data Understanding

Trước khi xây dựng mô hình, cần khám phá và hiểu dữ liệu, đặc biệt là mối quan hệ giữa điện năng và thời tiết. Mục tiêu của bước này là xác định các biến thời tiết có ảnh hưởng đến điện năng như thế nào.

Thực hiện các bước sau:

- Kiểm tra:
  - Missing values
  - Outliers
  - Seasonality (chu kỳ ngày, chu kỳ tuần)
- Trực quan hóa:
  - Biểu đồ load theo thời gian
  - Biểu đồ load vs temperature
  - Ma trận tương quan (correlation matrix)

4. Preprocessing

4.1 Chuẩn hóa dữ liệu

Tất cả các biến được chuẩn hóa bằng **StandardScaler** để:

- Giúp LSTM học ổn định
- Tránh bias do khác đơn vị đo

- Cải thiện hiệu quả của VAR và XGBoost

4.2 Chia tập dữ liệu theo thời gian

Dữ liệu phải được chia **theo thứ tự thời gian**, không được shuffle, theo tỷ lệ (train 70%, validation 15%, Test 15%)

5. Feature Engineering cho ML và DL

5.1 Cửa sổ trễ (Lag Window)

Chọn: $L = 24$(24 giờ)

Mỗi mẫu đầu vào tại thời điểm $t$bao gồm:

$$[load_{t-1}, \ldots, load_{t-24},\ temp_{t-1}, \ldots, temp_{t-24},\ humidity_{t-1}, \ldots, wind_{t-24}]$$

5.2 Rolling Features (chỉ dùng cho XGBoost)

Ngoài lag features, tạo thêm:

- Rolling mean 6 giờ
- Rolling standard deviation 24 giờ
- Load trend: $load_t - load_{t-24}$

Các đặc trưng này giúp XGBoost học được xu hướng và độ biến động.

6. Mô hình 1 – VAR (Statistical Baseline)

1- Kiểm tra tính dừng

- Dùng **ADF test**
- Nếu chuỗi không dừng → thực hiện **differencing**

2- Chọn bậc trễ

- Dựa trên **AIC** hoặc **BIC**

3- Huấn luyện VAR

Mô hình:

$$X_t = \sum_{i=1}^{p} A_i X_{t-i} + \varepsilon_t$$

VAR dự báo toàn bộ vector $[load, temp, humidity, wind]$.

Khi đánh giá, chỉ lấy cột **load**.

7. Mô hình 2 – XGBoost (Machine Learning)

1- Input và Output

Đầu vào là vector phẳng từ cửa sổ lag:

$$X_t = [load_{t-24:t}, temp_{t-24:t}, hum_{t-24:t}, wind_{t-24:t}]$$

Đầu ra:

$$y_t = load_{t+1}$$

2- Huấn luyện

Điều chỉnh:

- max_depth
- n_estimators
- Sử dụng **early stopping** trên validation set

8. Mô hình 3 – LSTM (Deep Learning)

1- Dạng dữ liệu đầu vào

Tensor 3D: (batch_samples, 24 bước thời gian, 4 biến)

2- Kiến trúc mạng

LSTM(64) → LSTM(32) → Dense(1)

- Loss: **MSE**
- Optimizer: **Adam**
-

9. Backtesting (Walk-forward Validation)

Áp dụng cho cả 3 mô hình:

Train → predict t+1 → thêm vào chuỗi → predict t+2 → ...

Điều này mô phỏng quá trình dự báo trong thực tế.

10. Đánh giá mô hình

Sử dụng các chỉ số đánh giá:

- RMSE
- MAE
- MAPE

So sánh các mô hình từ kết quả thu được (RMSE, MAE, MAPE)

11. Câu hỏi phân tích

1. Biến thời tiết nào ảnh hưởng mạnh nhất đến điện năng?
2. Khi nào VAR thất bại?
3. XGBoost học được thông tin gì từ thời tiết?
4. LSTM học được chu kỳ nào (ngày, tuần)?
5. LSTM có bị overfitting không?
6. Sự khác biệt giữa Machine Learning và Deep Learning trong bài toán này?

12. Bài tập nâng cao – Traffic Forecasting

Sử dụng dataset **METR-LA hoặc PEMS-BAY** (California Highway Sensors).

Dữ liệu:

- Tốc độ xe tại **207 trạm**
- Mỗi **5 phút**
- Trong nhiều tháng

Mỗi thời điểm:

$$X_t = [speed_1, speed_2, \dots, speed_{207}]$$

Bài toán:

$$[X_{t-12}, \dots, X_t] \Rightarrow X_{t+3}, X_{t+6}, X_{t+12}$$

(15, 30, 60 phút phía trước)

Nguồn:

- Paper: *Diffusion Convolutional Recurrent Neural Network (DCRNN)*
- Dataset: METR-LA, PEMS-BAY
  Chuẩn benchmark cho deep learning time series.