

Bài tập thực hành 3:

Dự đoán chuỗi thời gian sử dụng Machine Learning models

MỤC TIÊU:

- Chuyển bài toán chuỗi thời gian \rightarrow học có giám sát
- Huấn luyện mô hình XGBoost Regressor
- Đánh giá chất lượng dự báo

BÀI TẬP THỰC HÀNH:

Cho một chuỗi thời gian đơn biến y_t , hãy dự đoán giá trị tại thời điểm tiếp theo dựa trên các giá trị trong quá khứ. Thực hiện theo các bước dưới đây:

Bước 1. Chuẩn bị dữ liệu

Bước 2. Tạo lag features (Supervised Learning)

Bước 3. Chia $X - y$

Bước 4. Chia train / test theo thời gian

Bước 5. Huấn luyện mô hình XGBoost

Bước 6. Dự đoán và đánh giá

Bước 7. Phân tích feature importance để tìm hiểu lag nào ảnh hưởng mạnh nhất đến dự báo

BÀI TẬP NÂNG CAO:

1. Thêm rolling mean / rolling std
2. Dự báo multi-step
3. Tune hyperparameter bằng TimeSeriesSplit
4. So sánh với Random Forest

Câu hỏi cần trả lời:

1. So sánh XGBoost với Linear Regression
2. Điều gì xảy ra khi tăng số lag p ?
3. Khi nào XGBoost phù hợp hơn ARIMA?

DATASET: Hourly Energy Consumption

\rightarrow Dữ liệu tiêu thụ năng lượng hàng giờ trong hơn 10 năm từ PJM, tính bằng Megawatt.

\rightarrow link: https://www.kaggle.com/datasets/robikscube/hourly-energy-consumption?select=AEP_hourly.csv