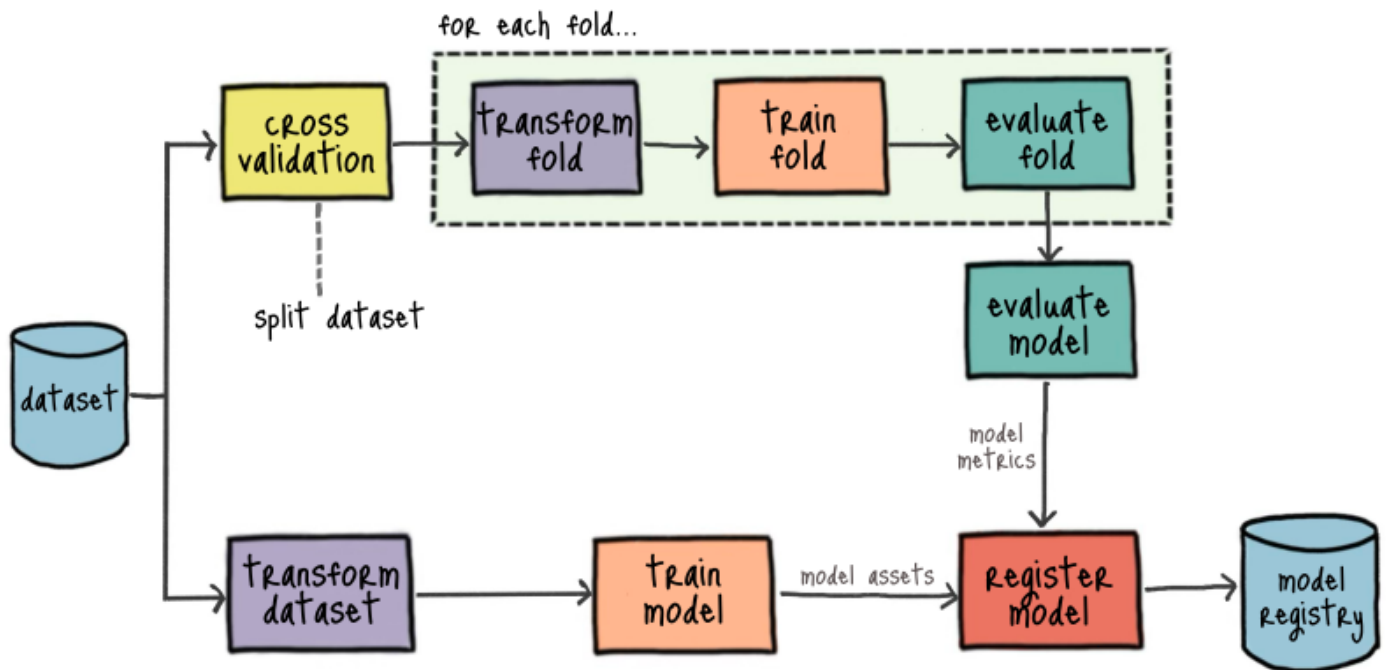


Tổng Quan Quy Trình Pipeline Huấn Luyện Mô Hình



Đây là một **quy trình (pipeline) huấn luyện mô hình học máy theo phương pháp Cross-Validation** (kiểm tra chéo), thường được sử dụng để đánh giá hiệu suất mô hình một cách ổn định. Dưới đây là giải thích chi tiết từng bước:

1. Tổng quan

Pipeline này mô tả quy trình **K-Fold Cross-Validation**, trong đó:

- Tập dữ liệu (**dataset**) được chia thành **K phần (folds)** bằng nhau.
- Mô hình được huấn luyện **K lần**, mỗi lần sử dụng **K-1 folds để train** và **1 fold còn lại để validation**.
- Mục tiêu: Đảm bảo mô hình không bị overfitting và đánh giá khách quan trên toàn bộ dữ liệu.

2. Giải thích từng thành phần

a. Split dataset

- **Ý nghĩa:** Chia tập dữ liệu thành **K folds** (phần).
- **Chi tiết:**
 - Ví dụ: Với $K=5$, dữ liệu được chia thành 5 phần, mỗi phần chứa 20% dữ liệu.
 - Có thể chia ngẫu nhiên hoặc theo tỷ lệ lớp (stratified splitting) cho bài toán phân loại.

b. For each fold...

- **Ý nghĩa:** Lặp lại quy trình train/validation trên từng fold.
- **Chi tiết:**
 - Mỗi vòng lặp chọn **1 fold làm validation set, K-1 folds còn lại làm training set**.
 - Ví dụ: Lần 1: Fold 1 là validation, Folds 2-5 là train; Lần 2: Fold 2 là validation, Folds 1,3-5 là train...

c. Transform dataset

- **Ý nghĩa:** Tiền xử lý dữ liệu trước khi huấn luyện.
- **Chi tiết:**
 - Chuẩn hóa dữ liệu (scaling, normalization).
 - Xử lý giá trị thiếu (imputation), mã hóa categorical features.
 - **Lưu ý:** Quy trình transform phải được áp dụng **riêng cho train/validation** để tránh data leakage.

d. Train model

- **Ý nghĩa:** Huấn luyện mô hình trên tập train.
- **Chi tiết:**
 - Sử dụng thuật toán như Random Forest, SVM, Neural Network...
 - Có thể tinh chỉnh hyperparameter (nếu dùng kết hợp với GridSearch/RandomSearch).

e. Evaluate model

- **Ý nghĩa:** Đánh giá mô hình trên tập validation.
- **Chi tiết:**
 - Tính các metrics: Accuracy, Precision, Recall (bài toán phân loại) hoặc MAE, RMSE (bài toán hồi quy).
 - Lưu lại kết quả để tổng hợp sau K lần chạy.

f. Model assets

- **Ý nghĩa:** Các tài nguyên liên quan đến mô hình sau huấn luyện.
- **Chi tiết:**
 - File trọng số (weights), kiến trúc mô hình (architecture), logs.

- Metadata: Hyperparameters, thời gian huấn luyện.

g. Register model

- **Ý nghĩa:** Lưu trữ mô hình vào **Model Registry**.
- **Chi tiết:**
 - Dùng công cụ như MLflow, AWS SageMaker Model Registry.
 - Quản lý versioning (phiên bản), đánh dấu mô hình tốt nhất.

h. Model registry

- **Ý nghĩa:** Kho lưu trữ tập trung các mô hình đã huấn luyện.
- **Chi tiết:**
 - Cho phép triển khai (deploy) mô hình từ registry lên production.
 - Hỗ trợ rollback nếu mô hình mới có vấn đề.

3. Luồng hoạt động của pipeline

1. **Chia dữ liệu** → K folds.
2. **Với mỗi fold:**
 - Transform dữ liệu train/validation.
 - Train mô hình trên train set.
 - Evaluate trên validation set.
3. **Tổng hợp kết quả** từ K lần evaluate để tính **độ ổn định** của mô hình (ví dụ: mean accuracy \pm độ lệch chuẩn).
4. **Lưu mô hình tốt nhất** vào Model Registry để triển khai.

4. Ứng dụng thực tế

- **Cross-Validation** đặc biệt hữu ích khi:
 - Dữ liệu ít, cần tận dụng tối đa để đánh giá mô hình.
 - Tránh overfitting do chia ngẫu nhiên 1 lần (train-test split thông thường).
- **Ví dụ:** Dự đoán lưu lượng giao thông dựa trên dữ liệu cảm biến, với K=5 để đảm bảo mô hình hoạt động tốt trên mọi khu vực.

Tác giả: Đỗ Ngọc Tú
Công Ty Phần Mềm VHTSoft

Phiên bản #3
Được tạo 19 tháng 4 2025 15:18:50 bởi Đỗ Ngọc Tú
Được cập nhật 22 tháng 4 2025 17:42:25 bởi Đỗ Ngọc Tú