Giải Thích Quy Trình Huấn Luyện Mô Hình Dự Đoán Giá Trị Chuyển Nhượng trong bài 4

Người làm: Nguyễn Hải Nam

Ngày: 3 tháng 5 năm 2025

Tài liệu này giải thích chi tiết quy trình huấn luyện mô hình hồi quy tuyến tính để dự đoán giá trị chuyển nhượng ước tính (ETV) cho các vị trí trong bóng đá, sử dụng dữ liệu từ tệp result.csv.

Mục lục

1. Giải Thích Quy Trình Huấn Luyện Mô Hình

Quy trình huấn luyện mô hình hồi quy tuyến tính để dự đoán giá trị chuyển nhượng ước tính (ETV) bao gồm các bước chuẩn bị dữ liệu, chia tập dữ liệu, chọn siêu tham số, huấn luyện, đánh giá và tối ưu hóa. Mô hình sử dụng các đặc trưng được chọn (như `Key Passes`, `Gls`, `GA90`, v.v.) từ tệp `result.csv` để dự đoán ETV, với mục tiêu giảm thiểu sai số dự đoán.

1.1. Chuẩn Bị Dữ Liệu

Dữ liệu được tiền xử lý để đảm bảo phù hợp với mô hình hồi quy tuyến tính:

- **Xử lý giá trị thiếu**: Các giá trị thiếu trong các đặc trưng (như `Key Passes`, `Gls`, `GA90`) được điền bằng trung vị của cột tương ứng. Nếu trung vị là `NaN`, giá trị 0 được sử dụng.
- **Chuyển đổi kiểu dữ liệu**: Tất cả đặc trưng được chuyển thành kiểu số (`float` hoặc `integer`) để đảm bảo tính toán chính xác.
- **Biến đổi log**: Áp dụng hàm `np.log1p` ($\log(1+x)$) cho các đặc trưng có phân phối lệch (như `Gls`, `Key Passes`) để giảm độ lệch và cải thiện tính tuyến tính.
- **Tăng trọng số cho đặc trưng quan trọng**: Các đặc trưng quan trọng (`important_features`, ví dụ: `Key Passes` cho tiền vệ, `Gls` cho tiền đạo, `Save%` cho thủ môn) được nhân với hệ số 2.0 để nhấn mạnh vai trò trong dự đoán ETV.
- **Chuẩn hóa**: Sử dụng `StandardScaler` để chuẩn hóa các đặc trưng về trung bình 0 và độ lệch chuẩn 1, đảm bảo các đặc trưng có thang đo tương đương.
- **Đảo ngược đặc trưng tiêu cực**: Đối với các đặc trưng mà giá trị thấp là tốt (như `GA90` cho thủ môn), giá trị được đảo ngược (ví dụ: lấy nghịch đảo hoặc trừ đi giá trị tối đa) để đảm bảo tương quan đúng với ETV.

1.2. Chia Tập Dữ Liệu

Dữ liệu được chia thành ba tập:

- **Tập huấn luyện (70%)**: Dùng để huấn luyện mô hình, học các hệ số β_i của mô hình hồi quy tuyến tính:

  $$\text{ETV} = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_n \cdot x_n$$

  trong đó x_i là các đặc trưng (như `Key Passes`, `Gls`, `Save%`).
- **Tập xác nhận (15%)**: Dùng để điều chỉnh siêu tham số và đánh giá hiệu suất mô hình trong quá trình huấn luyện.
- **Tập kiểm tra (15%)**: Dùng để đánh giá cuối cùng hiệu suất của mô hình trên dữ liệu chưa thấy.

Việc chia dữ liệu được thực hiện ngẫu nhiên với `train_test_split` từ thư viện `scikit-learn`, đảm bảo phân phối đồng đều giữa các vị trí (`MF`, `FW`, `GK`).

1.3. Chọn Siêu Tham Số

Mô hình hồi quy tuyến tính có ít siêu tham số, nhưng các tham số sau được điều chỉnh:

- **Hệ số chuẩn hóa (Regularization)**: Sử dụng L2 regularization (Ridge Regression) với tham số α để tránh hiện tượng quá khớp (*overfitting*). Giá trị α được chọn dựa trên thử nghiệm.

Quá trình điều chỉnh sử dụng chỉ số đánh giá Mean Squared Error (MSE) trên tập xác nhận để chọn bộ siêu tham số tốt nhất.

1.4. Huấn Luyện Mô Hình

Mô hình hồi quy tuyến tính được huấn luyện trên tập huấn luyện sử dụng thư viện `scikit-learn`:

- **Hàm mất mát**: Mô hình tối ưu hóa hàm mất mát MSE:

$$\text{MSE} = \frac{1}{m} \sum_{i=1}^{m} \left(\text{ETV}_i - \hat{\text{ETV}}_i \right)^2$$

  trong đó ETV_i là giá trị thực tế, $\hat{\text{ETV}}_i$ là giá trị dự đoán, và m là số mẫu.

- **Tối ưu hóa**: Sử dụng phương pháp Gradient Descent hoặc phương pháp giải tích (Normal Equation) để tìm các hệ số β_i tối ưu.

- **Xử lý đặc trưng tiêu cực**: Đối với các đặc trưng như `GA90`, hệ số β_i âm được kiểm tra để đảm bảo ảnh hưởng ngược chiều với ETV.

Quá trình huấn luyện được lặp lại với các giá trị siêu tham số khác nhau, sử dụng tập xác nhận để chọn mô hình tốt nhất.

1.5. Đánh Giá Mô Hình

Mô hình được đánh giá trên tập kiểm tra bằng các chỉ số:

- **Mean Squared Error (MSE)**: Đo lường sai số bình phương trung bình giữa giá trị thực và dự đoán.

- **R-squared (R^2)**: Đo lường mức độ giải thích của mô hình đối với phương sai của ETV:

$$R^2 = 1 - \frac{\sum_{i=1}^{m} \left(\text{ETV}_i - \hat{\text{ETV}}_i \right)^2}{\sum_{i=1}^{m} \left(\text{ETV}_i - \overline{\text{ETV}} \right)^2}$$

trong đó $\overline{\text{ETV}}$ là giá trị trung bình của ETV.

- **Mean Absolute Error (MAE)**: Đo lường sai số tuyệt đối trung bình, dễ diễn giải hơn MSE.

Ví dụ: Nếu mô hình đạt $R^2 = 0.85$ trên tập kiểm tra, điều này có nghĩa là 85% phương sai của ETV được giải thích bởi các đặc trưng.

1.6. Ví Dụ Minh Họa

Giả sử dữ liệu của một tiền đạo có các đặc trưng sau: `Gls` = 20, `Ast` = 8, `xG per 90` = 0.7. Sau tiền xử lý (biến đổi log, chuẩn hóa, tăng trọng số), mô hình dự đoán:

$$\hat{\text{ETV}} = \beta_0 + 2.0 \cdot \beta_1 \cdot \text{Gls}_{\text{scaled}} + 2.0 \cdot \beta_2 \cdot \text{Ast}_{\text{scaled}} + 2.0 \cdot \beta_3 \cdot \text{xG per 90}_{\text{scaled}}$$

Nếu ETV thực tế là 50 triệu euro và ETV dự đoán là 48 triệu euro, sai số tuyệt đối là 2 triệu euro, cho thấy mô hình có độ chính xác cao.

2. Kết Luận

Quy trình huấn luyện mô hình hồi quy tuyến tính bao gồm chuẩn bị dữ liệu kỹ lưỡng, chia tập dữ liệu hợp lý, điều chỉnh siêu tham số, và đánh giá hiệu suất bằng các chỉ số như MSE, R^2, và MAE. Các bước tối ưu hóa như kiểm tra đa cộng tuyến, thử nghiệm mô hình phi tuyến, và kiểm định chéo giúp cải thiện độ chính xác và độ tin cậy của mô hình trong dự đoán ETV.