XÂY DỰNG VÀ ĐÁNH GIÁ MÔ HÌNH LINEAR REGRESSION CHO BÀI TOÁN DỰ ĐOÁN GIÁ NHÀ

Nguyễn Tiến Mạnh

Ngày 26 tháng 1 năm 2026

Tóm tắt nội dung

Tóm tắt: Linear Regression là kỹ thuật nền tảng trong thống kê và học máy để mô hình hóa mối quan hệ giữa các biến số. Báo cáo này trình bày cơ sở lý thuyết về mô hình hồi quy tuyến tính đa biến và thuật toán tối ưu Gradient Descent.

Mục lục

1 Giới thiệu

Trong báo cáo này, chúng tôi xây dựng một mô hình Hồi quy tuyến tính (Linear Regression) từ đầu (from scratch) để tìm hiểu sâu về cơ chế hoạt động nội tại của thuật toán, thay vì chỉ sử dụng các thư viện có sẵn.

2 Cơ sở lý thuyết

2.1 Mô hình toán học

Mô hình Linear Regression giả định mối quan hệ giữa biến mục tiêu y (giá nhà) và các đặc trưng đầu vào $x = [x_1, x_2, ..., x_n]^T$ là tuyến tính:

$$\hat{y} = w_0 + w_1 x_1 + ... + w_n x_n = w^T x + b \tag{1}$$

Trong đó w là vector trọng số và b là hệ số chệch (bias).

2.2 Hàm mất mát (Loss Function)

Để tìm được bộ tham số tối ưu, ta cần cực tiểu hóa sai số giữa giá trị dự đoán và thực tế. Giả định rằng nhiễu của dữ liệu tuân theo phân phối chuẩn (Gaussian Noise), nguyên lý Ước lượng hợp lý cực đại (MLE) dẫn dắt ta đến việc cực tiểu hóa hàm Bình phương sai số trung bình (Mean Squared Error - MSE):

$$J(w, b) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)})^2 \tag{2}$$

Trong đó m là số lượng mẫu dữ liệu.

2.3 Thuật toán tối ưu Gradient Descent

Hàm mất mát MSE là một hàm lồi (convex function), đảm bảo tồn tại cực trị toàn cục. Chúng ta sử dụng thuật toán Gradient Descent để cập nhật tham số iterively:

$$w_j := w_j - \alpha \frac{\partial J}{\partial w_j} \tag{3}$$

Với α là tốc độ học (learning rate). Chi tiết các bước biến đổi đạo hàm được trình bày tại **Phụ lục A**.

3 Thực nghiệm

3.1 Mô tả dữ liệu

Thực nghiệm được tiến hành trên tập dữ liệu `Housing.csv`, bao gồm 545 mẫu với 13 thuộc tính như: *price* (biến mục tiêu), *area, bedrooms, bathrooms, stories, mainroad, guestroom, ...*

3.2 Tiền xử lý dữ liệu

Dữ liệu thô cần được chuyển đổi để phù hợp với mô hình toán học:

- **Mã hóa nhị phân:** Các biến phân loại (Yes/No) như *mainroad, guestroom, basement* được chuyển về dạng số (1/0).
- **Mã hóa One-Hot:** Biến *furnishingstatus* được tách thành các biến giả (dummy variables).
- **Chuẩn hóa đặc trưng:** Áp dụng Min-Max Scaling cho các biến số thực (*area, bedrooms...*) để đưa về khoảng $[0, 1]$, giúp Gradient Descent hội tụ nhanh hơn.

3.3 Thiết lập thí nghiệm

- **Phân chia dữ liệu:** 80% Huấn luyện - 20% Kiểm thử.
- **Tham số:** Learning rate $\alpha = 0.1$, Số vòng lặp = 1500.

3.4 Kết quả thực nghiệm

3.4.1 Quá trình huấn luyện

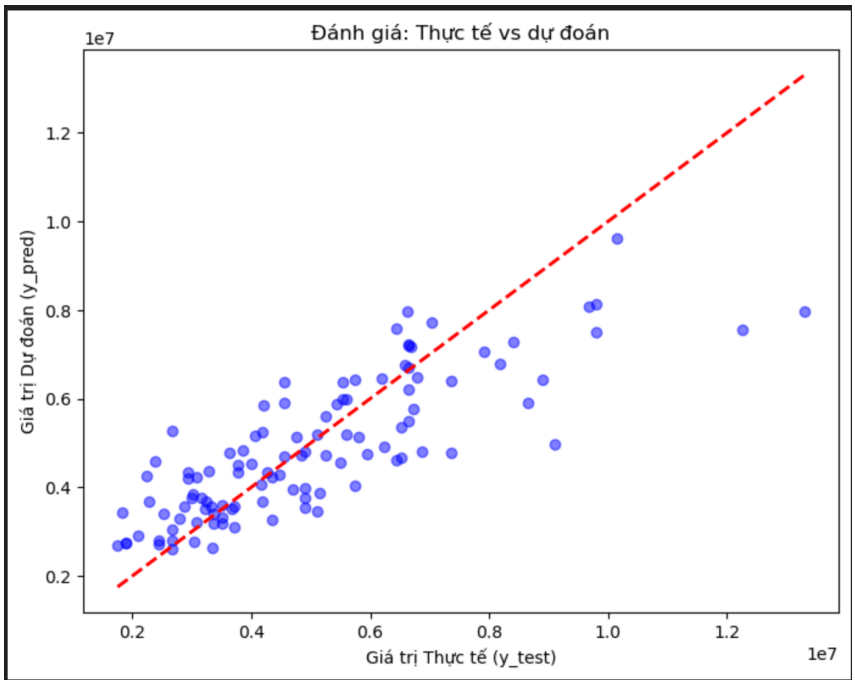
Hàm mất mát giảm đều đặn từ 1.26×10^{13} xuống 4.85×10^{11}, cho thấy thuật toán hoạt động ổn định.

3.4.2 Đánh giá và Phân tích lỗi

Mô hình được đánh giá trên tập kiểm thử (Test set) với các chỉ số sau:

Độ đo	Giá trị
MAE (Sai số tuyệt đối trung bình)	959,496.99
RMSE (Sai số bình phương trung bình)	1,320,285.15
R^2 Score	0.6551

Bảng 1: Kết quả đánh giá trên tập Test



Hình 1: Biểu đồ phân tán: Giá thực tế vs Giá dự đoán

Quan sát biểu đồ phân tán (Hình 1), ta nhận thấy:

- Các điểm dữ liệu tập trung khá sát đường chéo chính ở vùng giá trị thấp và trung bình, cho thấy mô hình dự đoán tốt ở phân khúc này.
- **Hạn chế của phương pháp:** Ở vùng giá trị cao (phía trên bên phải biểu đồ), sai số dự đoán tăng lên đáng kể. Điều này phản ánh nhược điểm cố hữu của Linear Regression là nhạy cảm với các giá trị ngoại lai (outliers) và không nắm bắt được các mối quan hệ phi tuyến phức tạp trong định giá bất động sản cao cấp.

4 Thảo luận

4.1 Tổng kết

Nghiên cứu đã xây dựng thành công quy trình dự đoán giá nhà từ tiền xử lý đến mô hình hóa. Mức độ giải thích $R^2 \approx 65.5\%$ là kết quả khả quan cho một mô hình tuyến tính cơ

bản không sử dụng các thư viện tối ưu sẵn.

4.2 Hạn chế nghiên cứu

Trong phạm vi báo cáo này, một số khía cạnh vẫn chưa được thực hiện:

- Chưa thực hiện tinh chỉnh siêu tham số (Hyperparameter tuning) để tìm ra learning rate tối ưu nhất.
- Chưa áp dụng các kỹ thuật trích xuất đặc trưng (Feature Engineering) như tạo biến đa thức hay tương tác biến để cải thiện độ chính xác.
- Chưa xử lý triệt để các giá trị ngoại lai (Outliers) trong bước tiền xử lý.

4.3 Phương hướng phát triển

Các nghiên cứu tiếp theo có thể tập trung vào:

- Áp dụng Regularization (L1/L2) để kiểm soát trọng số và giảm thiểu overfitting.
- Thử nghiệm các mô hình phi tuyến như Polynomial Regression hoặc Random Forest để so sánh hiệu quả.

A Phụ lục: Đạo hàm Gradient

Gradient của hàm mất mát J đối với tham số w được tính như sau:

$$\frac{\partial J}{\partial w} = \frac{1}{m} X^T (\hat{y} - y) \tag{4}$$