

HỌC SÂU

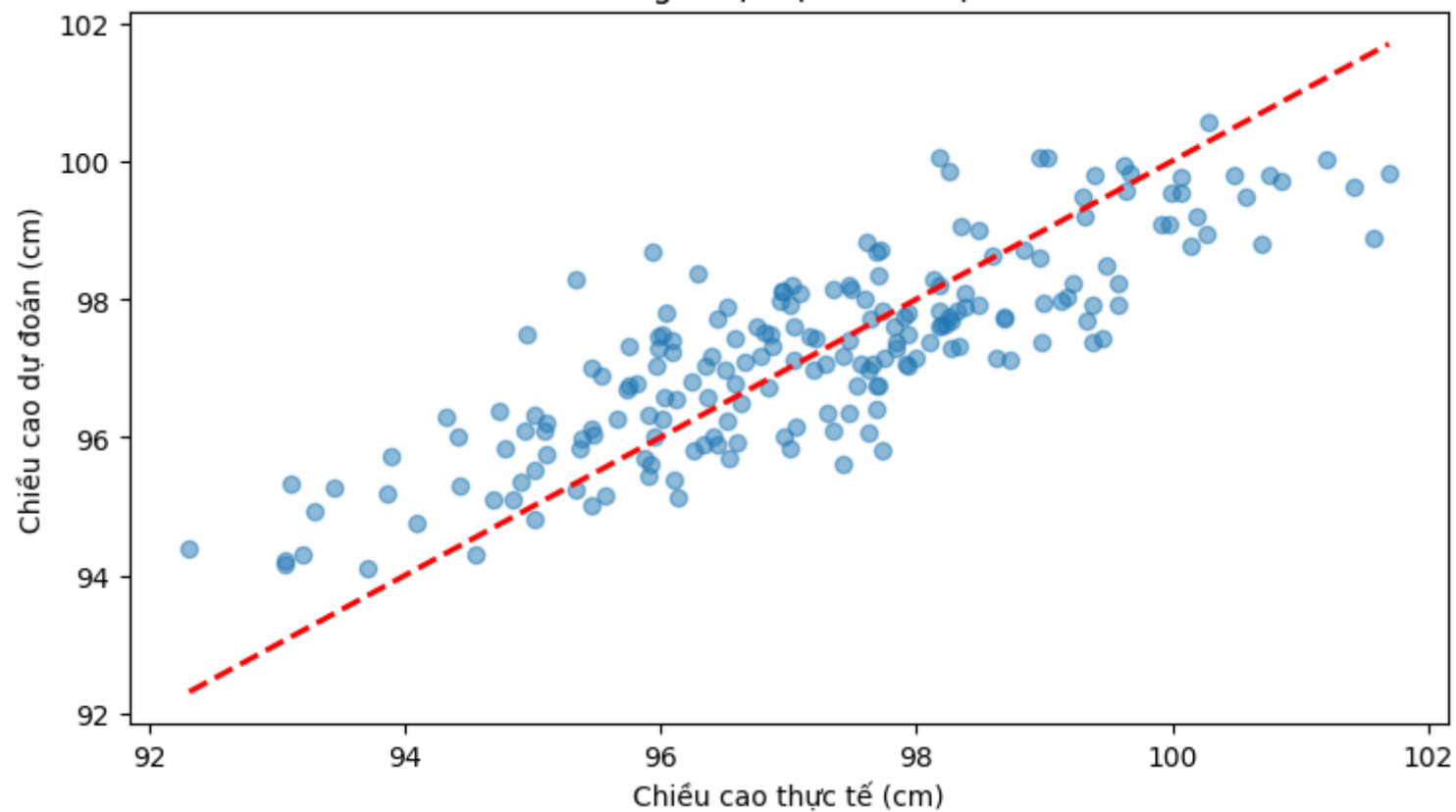
BÀI 2. HỒI QUY TUYẾN TÍNH VÀ HỒI QUY LOGISTIC

Hồi quy tuyến tính

Linear Regression là một thuật toán **học có giám sát (supervised learning)** trong Machine Learning, nó là một phương pháp thống kê dùng để ước lượng mối quan hệ giữa các biến độc lập (input features) và biến phụ thuộc (output target).

Linear Regression giả định rằng sự tương quan giữa các biến là tuyến tính, từ đó tìm ra hàm tuyến tính tốt nhất để biểu diễn mối quan hệ này. Thuật toán này dự báo giá trị của biến output từ các giá trị của các biến đầu vào.

So sánh giá trị thực tế và dự đoán



Đặc điểm của mô hình

Mục tiêu: Hồi quy tuyến tính hướng đến mô hình hóa và phân tích mối quan hệ giữa một biến phụ thuộc (hay biến mục tiêu) và một hoặc nhiều biến độc lập (hay biến giải thích).

Mục đích của mô hình

Dự đoán và dự báo: Sử dụng biến độc lập để dự đoán giá trị của biến phụ thuộc.

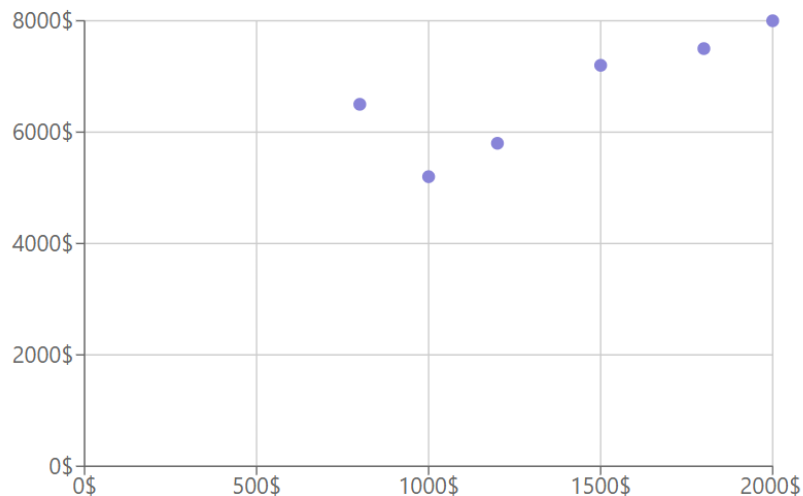
Chẳng hạn như dự đoán giá nhà dựa trên diện tích, dự đoán doanh số bán hàng dựa trên chiến lược tiếp thị.

Số phòng ngủ	Diện tích	Khu đô thị	Giá bán
3	2000	Times City	\$250,000
2	800	Royal City	\$300,000
2	850	Times City	\$150,000
1	550	Times City	\$78,000
4	2000	KDT Linh Đàm	\$150,000

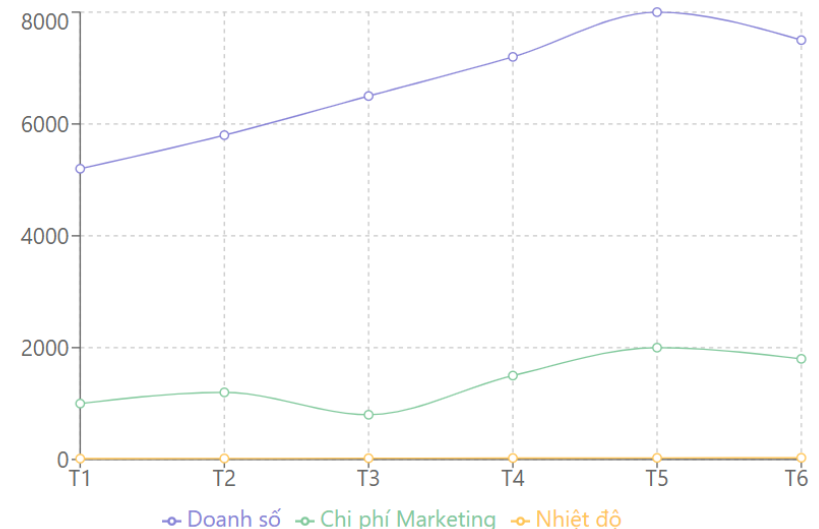
Mục đích của mô hình

Hiểu rõ mối quan hệ: Xác định mức độ ảnh hưởng của các biến độc lập lên biến phụ thuộc. Điều này giúp hiểu rõ các yếu tố nào là quan trọng và cách chúng ảnh hưởng đến kết quả

Mối quan hệ giữa Chi phí Marketing và Doanh số



Xu hướng Doanh số theo thời gian



Mục đích của mô hình

● **Đánh giá tác động:** Trong nghiên cứu và phân tích, hồi quy tuyến tính giúp đánh giá tác động của các yếu tố (biến độc lập) lên một kết quả cụ thể (biến phụ thuộc).



$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

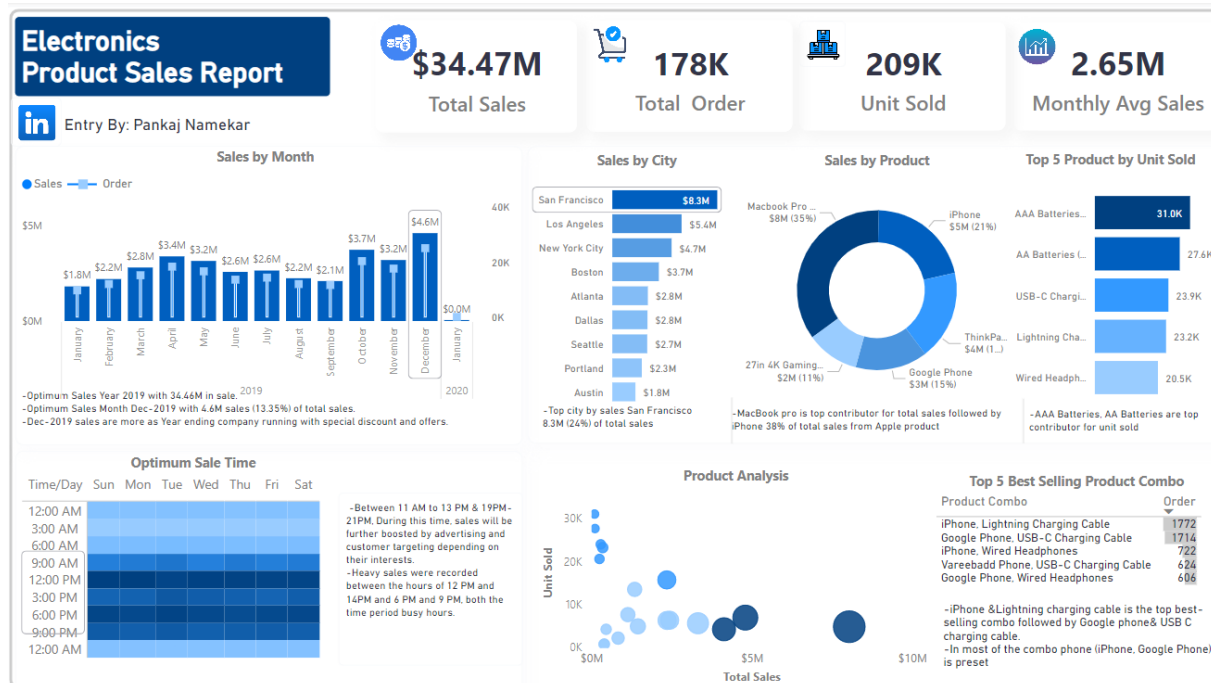
Mục đích của mô hình

Kiểm soát các biến: Trong mô hình hồi quy tuyến tính đa biến, có thể kiểm soát ảnh hưởng của các biến nhiễu để tập trung vào mối quan hệ cụ thể giữa một số biến chọn lọc.



Mục đích của mô hình

Phân tích xu hướng và mẫu: Hồi quy tuyến tính giúp phân tích xu hướng và mẫu trong dữ liệu, qua đó cung cấp thông tin hữu ích cho việc ra quyết định và lập kế hoạch.



Giả định cơ bản:

- **Tính tuyến tính:** Mỗi quan hệ giữa biến phụ thuộc và các biến độc lập được giả định là tuyến tính, tức là có thể được mô tả thông qua một đường thẳng.
- **Độc lập của các sai số (Residuals):** Các sai số từ mô hình được giả định là độc lập với nhau, không có sự phụ thuộc hoặc mẫu định hình nào.
- **Phân phối chuẩn của sai số:** Các sai số được giả định tuân theo một phân phối chuẩn.
- **Homoscedasticity:** Phương sai của sai số được giả định là nhất quán qua tất cả các giá trị của biến độc lập, không biến đổi theo mức độ của biến dự đoán.
- **Không có hoặc hạn chế đa cộng tuyến:** Giả định rằng không có mối quan hệ tương quan cao giữa các biến độc lập, hay nói cách khác, các biến độc lập không được phụ thuộc lẫn nhau một cách mạnh mẽ.

Tham số

- **Hệ số hồi quy (*regression coefficients*):** Đây là các trọng số được gán cho mỗi biến độc lập. Chúng xác định mức độ mà mỗi biến độc lập ảnh hưởng đến biến phụ thuộc.
- **Điểm chặn (*intercept*):** Đây là giá trị của biến phụ thuộc khi tất cả các biến độc lập bằng 0. Nói cách khác, đây là điểm bắt đầu của đường hồi quy trên trục tung.
- **Lỗi (*error term*):** Còn được gọi là dư lượng, lỗi là phần của dữ liệu mà không được giải thích bởi mô hình hồi quy. Nó bao gồm cả ảnh hưởng của các yếu tố ngoại lai và sai số ngẫu nhiên.

$$\text{giá nhà} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

β_0 : Intercept

X_1 : Diện tích

X_2 : Số phòng

X_3 : Khoảng cách trung tâm

$$\text{Điểm}^2 = 50 + 2 * \text{số giờ học} + \varepsilon$$

Tham số

- **Hệ số xác định (*R-squared*):** Mặc dù không phải là một tham số cần thiết lập trước khi xây dựng mô hình, *R-squared* là một chỉ số quan trọng để đánh giá mức độ phù hợp của mô hình với dữ liệu. Nó thể hiện tỷ lệ phần trăm biến thiên của biến phụ thuộc được giải thích bởi mô hình.
- **Tham số chuẩn hóa (*Regularization parameters*):** Đối với các biến thể của hồi quy tuyến tính như Ridge (L2 regularization) hoặc Lasso (L1 regularization), tham số chuẩn hóa được sử dụng để kiểm soát mức độ phạt đối với độ lớn của hệ số hồi quy, nhằm giảm overfitting.
- **Tiêu chí dừng (*stopping criteria*):** Trong các phương pháp học máy, tiêu chí dừng xác định khi nào quá trình tối ưu hóa nên dừng lại, thường dựa trên sự cải thiện của hàm mất mát hoặc số lần lặp tối đa.

Mô hình hồi quy tuyến tính đơn giản (một biến độc lập)

Trong trường hợp đơn giản nhất với chỉ một biến độc lập, mô hình hồi quy tuyến tính có dạng:

$$y = \beta_0 + \beta_1 x + \epsilon$$

trong đó:

- y là biến phụ thuộc.
- x là biến độc lập.
- β_0 là hệ số chặn (intercept).
- β_1 là hệ số hướng (slope).
- ϵ là sai số ngẫu nhiên (không quan sát được).

Mô hình hồi quy tuyến tính đa biến

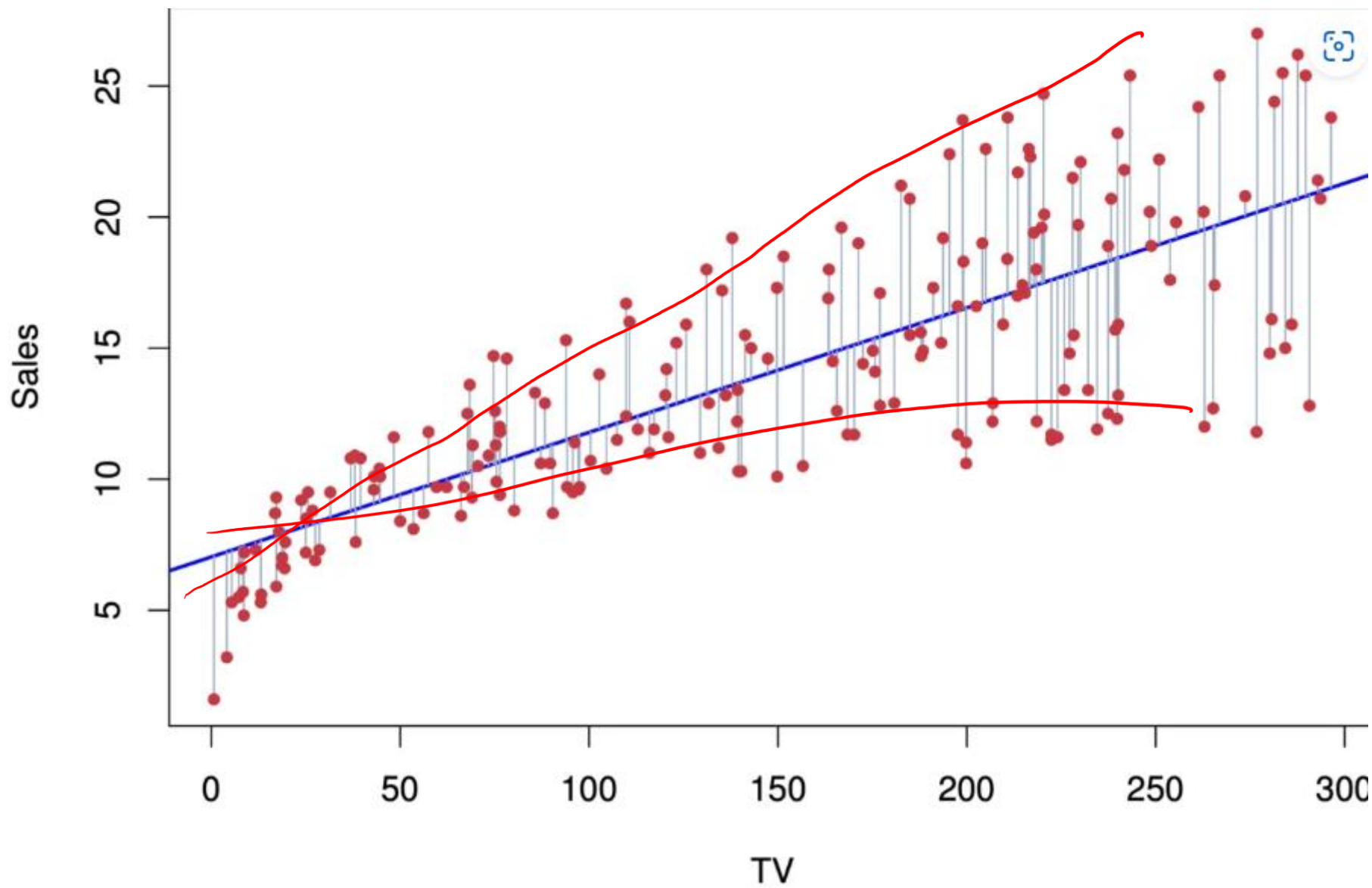
Trong trường hợp có nhiều biến độc lập, mô hình mở rộng thành:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

trong đó, mỗi x_i đại diện cho một biến độc lập khác nhau, và β_i là hệ số tương ứng với mỗi biến độc lập đó.

Tìm hệ số mô hình

Mục tiêu của hồi quy tuyến tính là ***tìm ra các giá trị của hệ số β*** sao cho ***tổng bình phương sai số (sum of squared errors)*** giữa giá trị dự đoán và giá trị thực tế là nhỏ nhất. Phương pháp phổ biến để tìm ra các hệ số này là phương pháp bình phương tối thiểu (least squares method).



Đánh giá mô hình

Mô hình hồi quy tuyến tính thường được đánh giá dựa trên các chỉ số:

- **R-squared**,
- **Root Mean Squared Error (RMSE)**, hoặc
- **Mean Absolute Error (MAE)**, cho biết mức độ chính xác của mô hình trong việc dự đoán dữ liệu.

Quy trình thực hiện của mô hình

Bước 1:

Thu thập dữ liệu: Bước đầu tiên trong quá trình hồi quy tuyến tính là thu thập dữ liệu. Dữ liệu này có thể đến từ nhiều nguồn khác nhau như khảo sát, ghi chép, cơ sở dữ liệu, v.v. Dữ liệu phải bao gồm cả biến độc lập (predictors) và biến phụ thuộc (target) muốn dự đoán.

← linear regression dataset

<> Notebooks 70,838 ← Comments 1,316 💬 Topics 1,097 📁 Datasets 911 👤 Models 29 🏆 Competitions 11

Filter by 74,202 Results Relevance ▾

DATE

☐ Last 90 days 4,732

☐ This week 322

☐ Today 41

CREATOR

☐ You 0

☐ Others 74,202


DATASET SIZE

☐ small 845


☐ medium 62

☐ large 4

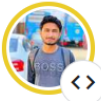
DATASET FILE TYPES




Simple Linear Regression Algorithm Dataset on Sales Data
Discussion Topic · 2y ago · by [Shruti Pandit](#)
[dataset](#) that you can use to apply Simple [Linear](#) and Advanced [Linear Regression](#) Techniques and check underlying 15 comments




Linear Regression - Salary Dataset
Notebook · 3y ago · by [Shubham](#)
/input/salary-data-[dataset-for-linear-regression](#)/Salary_Data.csv") **EDA** df.head() #check number of 7 comments



Tip Prediction on Tips dataset ([Linear Regression](#))
Notebook · 9mo ago · by [Atif Ali AK](#)
using [Linear Regression](#) - **Task** : - To predict the tip amount on the basis of total bill, sex 34 comments



Comment on: Simple Linear Regression Algorithm Dataset on Sales Data
Discussion Comment · 2y ago · by [Shruti Pandit](#)
In the [General](#) forum 1 reply



Stacked Regressions : Top 4% on LeaderBoard
Notebook · 7y ago · by [Serigne](#)
[A study on [Regression](#) applied to the Ames [dataset](#)][2] by **Julien Cohen-Sola** : Thorough features 1,134 comments

THS. LÊ NHẬT TÙNG

Quy trình thực hiện của mô hình

Bước 2:

Chuẩn bị dữ liệu: Sau khi thu thập, dữ liệu cần được làm sạch và chuẩn bị. Điều này bao gồm việc loại bỏ hoặc xử lý dữ liệu thiếu hoặc nhiễu, chuẩn hóa hoặc tiêu chuẩn hóa các biến độc lập, và chuyển đổi dữ liệu (nếu cần).

Quy trình thực hiện của mô hình

Bước 3:

Huấn luyện mô hình: Dựa trên dữ liệu huấn luyện, mô hình hồi quy tuyến tính được xây dựng bằng cách ước lượng các hệ số cho mỗi biến độc lập. Quá trình này thường bao gồm việc sử dụng phương pháp bình phương nhỏ nhất (least squares method) để tìm ra đường thẳng (hoặc mặt phẳng, siêu phẳng) phù hợp nhất với dữ liệu.

Quy trình thực hiện của mô hình

Bước 4:

Đánh giá mô hình: Mô hình được đánh giá dựa trên hiệu suất của nó trên dữ liệu kiểm thử. Các chỉ số đánh giá thường dùng bao gồm R-squared (đo lường mức độ “phù hợp” của mô hình với dữ liệu), Mean Squared Error (MSE), hoặc Root Mean Squared Error (RMSE).

Quy trình thực hiện của mô hình

Bước 5:

Tinh chỉnh mô hình: Dựa trên kết quả đánh giá, mô hình có thể cần được tinh chỉnh để cải thiện hiệu suất. Điều này có thể bao gồm việc điều chỉnh các biến đầu vào, sử dụng các kỹ thuật regularization (như Ridge hoặc Lasso), hoặc thử nghiệm các mô hình hồi quy khác nhau.

Quy trình thực hiện của mô hình

Bước 6:

Sử dụng mô hình: Cuối cùng, mô hình được sử dụng để thực hiện dự đoán trên dữ liệu mới. Kết quả của quá trình này cho phép ta ứng dụng những phát hiện từ mô hình hồi quy vào thực tế, dự đoán kết quả hoặc hiểu rõ hơn về mối quan hệ giữa các biến. Mỗi bước trong quá trình này đều quan trọng và cần được thực hiện cẩn thận để đảm bảo tính chính xác và hiệu quả của mô hình hồi quy tuyến tính.

Ví dụ

Giả sử muốn dự đoán giá nhà dựa trên diện tích (m^2) của nó. Trong trường hợp này, giá nhà là biến phụ thuộc (y), và diện tích nhà là biến độc lập (x).

Bước 1 (Thu thập dữ liệu): Thu thập dữ liệu về các ngôi nhà bao gồm giá bán và diện tích của chúng như sau:

Diện tích (m^2)	Giá (nghìn USD)
50	200
70	270
80	300
100	370
120	450

Ví dụ

Bước 2 (Chuẩn bị dữ liệu): Trong trường hợp này, dữ liệu đã khá sạch và không cần xử lý nhiều.

Bước 3 (Phân chia dữ liệu): Sử dụng 80% dữ liệu để huấn luyện mô hình và 20% còn lại để kiểm thử mô hình.

Bước 4 (Huấn luyện mô hình): Sử dụng phương pháp bình phương nhỏ nhất để tìm ra hệ số cho mô hình hồi quy tuyến tính. Giả sử mô hình tìm được là:

$$\text{Giá} = 50 + 3 \times \text{Diện tích}$$

Bước 5 (Đánh giá mô hình): Kiểm tra mô hình với 20% dữ liệu còn lại và tính toán các chỉ số như R-squared, MSE để đánh giá hiệu suất.

Bước 6 (Tinh chỉnh mô hình): Điều chỉnh mô hình dựa trên kết quả đánh giá.

Bước 7 (Sử dụng mô hình): Sử dụng mô hình để dự đoán giá của nhà dựa trên diện tích. Chẳng hạn, với một ngôi nhà có diện tích 85 m², mô hình sẽ dự đoán giá là: $50 + 3 \times 85 = 305$ (nghìn USD).

Bài tập

- Thực hành với bộ dữ liệu giả định
- Phân tích dữ liệu bằng mô hình hồi quy tuyến tính trên bộ dữ liệu thực IRIS
- Phân tích dữ liệu bằng mô hình hồi quy tuyến tính trên bộ dữ liệu giả định MTCARS

MÔ HÌNH HỒI QUY LOGISTIC

Hồi quy logistic là một phương pháp thống kê được sử dụng rộng rãi trong việc phân tích và dự đoán dữ liệu phân lớp. Đặc biệt hiệu quả với dữ liệu có biến phụ thuộc nhị phân, hồi quy logistic mô hình hóa xác suất của một sự kiện dựa trên một hoặc nhiều biến độc lập.

MÔ HÌNH HỒI QUY LOGISTIC

Phương pháp này **sử dụng hàm logistic** để chuyển đổi các giá trị dự đoán thành xác suất, giúp dễ dàng diễn giải và áp dụng trong nhiều lĩnh vực như y học, kinh tế, khoa học xã hội và nhiều ngành khác. Hồi quy logistic không chỉ giúp xác định các yếu tố ảnh hưởng đến một kết quả nhất định mà còn cung cấp khả năng hiểu rõ mối quan hệ giữa các biến và kết quả, làm cho nó trở thành công cụ quan trọng trong việc phân tích dữ liệu và ra quyết định.

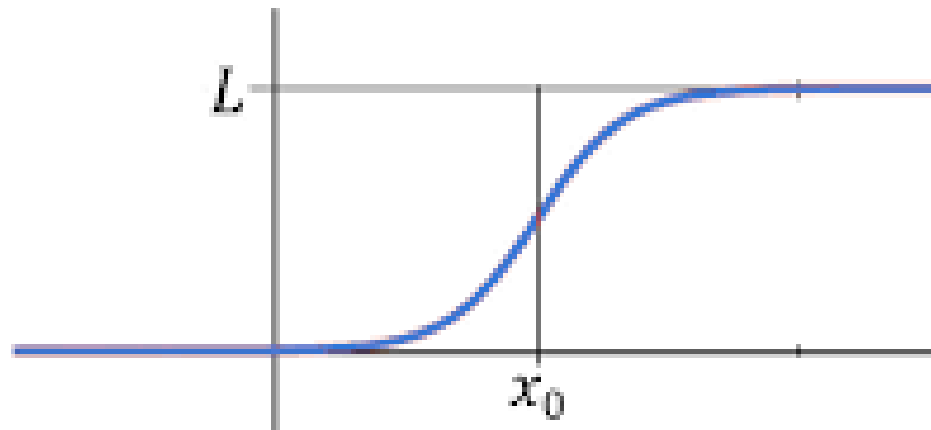
Logistic Function

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

x_0 = x value of midpoint

L = maximum value

k = growth rate



MÔ HÌNH HỒI QUY LOGISTIC

Phương pháp này **sử dụng hàm logistic** để chuyển đổi các giá trị dự đoán thành xác suất, giúp dễ dàng diễn giải và áp dụng trong nhiều lĩnh vực như y học, kinh tế, khoa học xã hội và nhiều ngành khác. Hồi quy logistic không chỉ giúp xác định các yếu tố ảnh hưởng đến một kết quả nhất định mà còn cung cấp khả năng hiểu rõ mối quan hệ giữa các biến và kết quả, làm cho nó trở thành công cụ quan trọng trong việc phân tích dữ liệu và ra quyết định.

Đặc điểm của mô hình

- ***Phân lớp và dự đoán:*** Dự đoán biến phụ thuộc nhị phân hoặc danh mục từ một hoặc nhiều biến độc lập.
- ***Xác định mức độ ảnh hưởng của biến độc lập:*** Xác định cách thức và mức độ mà các biến độc lập ảnh hưởng đến xác suất của sự kiện hoặc lớp mục tiêu.
- ***Tính toán xác suất sự kiện:*** Cung cấp ước lượng xác suất cho một sự kiện xảy ra dựa trên biến độc lập.

Đặc điểm của mô hình

- ***Đánh giá rủi ro và khả năng xảy ra:*** Đánh giá rủi ro hoặc khả năng xảy ra của một sự kiện cụ thể trong các lĩnh vực như y học, tài chính, nghiên cứu xã hội, và hơn thế nữa.
- ***Phân tích mối quan hệ tuyến tính giữa logit của kết quả và biến độc lập:*** Phân tích mối quan hệ tuyến tính giữa logit của kết quả (log odds) và các biến độc lập.

Hồi quy logistic thường được ưu tiên sử dụng trong các bài toán phân lớp và dự đoán nơi mà biến phụ thuộc không liên tục mà là nhị phân hoặc danh mục, giúp cung cấp cái nhìn sâu sắc và chính xác về mối quan hệ giữa các biến.

Giả định cơ bản

- **Biến phụ thuộc nhị phân hoặc danh mục:** Biến phụ thuộc phải là nhị phân (ví dụ: có/không, thành công/thất bại) hoặc danh mục.
- **Mối quan hệ tuyến tính giữa logit và các biến độc lập:** Cần có mối quan hệ tuyến tính giữa logit của kết quả (log của tỷ lệ xác suất) và các biến độc lập.
- **Không có đa cộng tuyến:** Các biến độc lập không nên có mối quan hệ tuyến tính mạnh mẽ với nhau.
- **Không có nhiễu đặc biệt trong biến phụ thuộc:** Mỗi trường hợp trong dữ liệu phải rõ ràng thuộc về một trong hai danh mục của biến phụ thuộc, không có trường hợp nhiễu.
- **Kích thước mẫu đủ lớn:** Cần một kích thước mẫu đủ lớn để đảm bảo độ tin cậy của các ước lượng.

Tham số

- **Hệ số hồi quy (coefficients):** Các hệ số hồi quy, hay trọng số, xác định mức độ ảnh hưởng của mỗi đặc trưng (feature) đối với xác suất dự đoán của mô hình.
- **Điểm chặn (intercept):** Điểm chặn là hệ số hằng số trong mô hình, điều chỉnh xác suất dự đoán khi tất cả các đặc trưng có giá trị bằng không.
- **Hàm liên kết (link function):** Trong hồi quy logistic, hàm sigmoid (hoặc hàm logit) được sử dụng làm hàm liên kết để chuyển đổi giá trị dự đoán sang dạng xác suất.

Tham số

- **Chuẩn hóa (regularization):** Các phương pháp như L1 (lasso), L2 (ridge), hoặc kết hợp của cả hai (elastic net) được sử dụng để chuẩn hóa mô hình, giúp tránh overfitting và cải thiện khả năng tổng quát hóa.
- **C (penalty parameter trong chuẩn hóa):** Đối với mô hình có chuẩn hóa, C là tham số điều chỉnh mức độ mạnh của chuẩn hóa. Một giá trị C thấp tăng cường hiệu ứng của chuẩn hóa, trong khi một giá trị C cao giảm bớt hiệu ứng đó.
- **Tiêu chí dừng (stopping criteria):** Điều kiện để dừng quá trình học của mô hình, thường dựa trên sự cải thiện của hàm mất mát hoặc đạt đến số lượng lần lặp tối đa.

Cách thức hoạt động của mô hình

Xác định hàm logistic: Hồi quy logistic sử dụng hàm logistic (còn gọi là hàm sigmoid) để chuyển đổi giá trị dự đoán thành xác suất. Hàm logistic có dạng:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k)}}$$

trong đó,

- $P(Y = 1)$ là xác suất để sự kiện $Y = 1$ xảy ra (ví dụ: sự kiện thành công, lớp 1, ...).
- X_1, X_2, \dots, X_k là các biến độc lập.
- $\beta_0, \beta_1, \dots, \beta_k$ là hệ số mô hình cần được ước lượng.
- e là cơ số của logarit tự nhiên.

Cách thức hoạt động của mô hình

- **Ước lượng hệ số mô hình:** Hệ số β của mô hình được ước lượng thông qua quy trình tối ưu hóa, thường là phương pháp Maximum Likelihood Estimation (MLE). MLE tìm cách tối đa hóa xác suất của dữ liệu quan sát dựa trên hệ số β .
- **Phân lớp:** Dựa vào xác suất được dự đoán từ hàm logistic, quyết định phân loại một quan sát vào lớp 1 nếu $P(Y = 1)$ **vượt quá một ngưỡng cụ thể** (thường là 0.5) và ngược lại là lớp 0. Ví dụ: Nếu $P(Y = 1) > 0.5$, quan sát được phân loại là lớp 1.
- **Đánh giá mô hình:** Mô hình hồi quy logistic thường được đánh giá thông qua các chỉ số như độ chính xác (accuracy), precision, recall, điểm số F1, hoặc thông qua ROC và AUC.

Ví dụ

Ví dụ: Giả sử bạn là một nhà phân tích tại một ngân hàng và muốn sử dụng hồi quy logistic để dự đoán liệu một khách hàng có khả năng vay vốn thành công hay không. Bộ dữ liệu bao gồm các đặc trưng như "Thu nhập hàng năm", "Điểm tín dụng", và "Số năm làm việc".

Ví dụ

- *Xác định biến độc lập và phụ thuộc:*
 - Biến phụ thuộc (Y): Khả năng vay vốn (1: Thành công, 0: Thất bại).
 - Biến độc lập (X): "Thu nhập hàng năm", "Điểm tín dụng", "Số năm làm việc".

Ví dụ

- Xây dựng mô hình hồi quy logistic: Sử dụng dữ liệu để ước lượng hệ số của mô hình:

$$P(Y = 1) = \frac{1}{1 + e^{-\beta_0 + \beta_1 \times \text{Th(u Nhập)} + \beta_2 \times \text{Điểm Tin Dụng} + \beta_3 \times \text{Số Năm Làm Việc}}}$$

Ví dụ

- *Ước lượng và dự đoán:*
 - Ước lượng hệ số β thông qua quá trình huấn luyện mô hình.
 - Dùng mô hình để dự đoán xác suất vay vốn thành công cho khách hàng mới dựa trên các đặc trưng của họ.
- *Phân loại:* Đặt một ngưỡng xác suất, ví dụ 0.5. Nếu mô hình dự đoán xác suất vay thành công lớn hơn 0.5, phân loại khách hàng vào lớp "Vay thành công"; ngược lại, phân loại vào lớp "Vay thất bại".

Thu Nhập	Điểm Tín Dụng	Số Năm Làm Việc	Xác Suất Vay Thành Công	Phân Loại
50k	600	5	0.7	Thành Công
30k	500	2	0.3	Thất Bại
80k	700	10	0.9	Thành Công

Đánh giá ưu điểm và hạn chế của mô hình

Ưu điểm: Hồi quy logistic mang lại nhiều ưu điểm, làm cho nó trở thành một công cụ phân tích dữ liệu mạnh mẽ, đặc biệt khi xử lý với dữ liệu phân lớp

Đánh giá ưu điểm và hạn chế của mô hình

- ***Phù hợp với biến phụ thuộc nhị phân hoặc danh mục:*** Hiệu quả trong việc mô hình hóa dữ liệu có biến phụ thuộc là nhị phân hoặc danh mục.
- ***Xác suất trong dự đoán:*** Cung cấp kết quả dưới dạng xác suất, giúp hiểu rõ hơn về khả năng xảy ra của một sự kiện.
- ***Khả năng xử lý biến độc lập không tuyến tính:*** Có khả năng xử lý mối quan hệ phi tuyến giữa các biến độc lập và biến phụ thuộc.

Đánh giá ưu điểm và hạn chế của mô hình

- **Không yêu cầu phân phối chuẩn của biến độc lập:** Không cần biến độc lập tuân theo phân phối chuẩn.
- **Đánh giá mức độ ảnh hưởng của biến độc lập:** Cho phép đánh giá mức độ ảnh hưởng của từng biến độc lập đối với xác suất của sự kiện.
- **Chống nhiễu và đa cộng tuyến:** Kháng nhiễu tốt và ít bị ảnh hưởng bởi đa cộng tuyến so với hồi quy tuyến tính.
- **Linh hoạt và dễ sử dụng:** Có nhiều cách để mở rộng và điều chỉnh mô hình, dễ dàng sử dụng trong nhiều ngữ cảnh khác nhau.

Đánh giá ưu điểm và hạn chế của mô hình

Hạn chế: Hồi quy logistic, mặc dù là một công cụ phân tích mạnh mẽ, nhưng cũng tồn tại một số hạn chế cần được xem xét trong quá trình áp dụng:

- Không thích hợp với biến phụ thuộc liên tục.
- Khó khăn trong việc mô hình hóa mối quan hệ phức tạp hoặc không tuyến tính mà không cần biến đổi dữ liệu.
- Không hiệu quả khi xử lý dữ liệu có nhiều biến độc lập hoặc có sự tương quan cao giữa các biến.

Có thể không phát huy hiệu quả trong tập dữ liệu nhỏ hoặc khi có sự mất cân đối lớn giữa các lớp.

Quy trình thực hiện của mô hình

- **Thu thập dữ liệu:** Giống như hồi quy tuyến tính, bắt đầu bằng việc xác định bài toán và thu thập dữ liệu. Đối với hồi quy logistic, biến phụ thuộc phải là nhị phân.
- **Chuẩn bị dữ liệu:** Giống như hồi quy tuyến tính, làm sạch và chuẩn bị dữ liệu là bước quan trọng bên cạnh việc lựa chọn các biến độc lập và kiểm tra đa cộng tuyến giữa chúng.
- **Phân chia dữ liệu:** Chia dữ liệu thành dữ liệu huấn luyện và dữ liệu kiểm thử, tương tự như trong hồi quy tuyến tính.
- **Huấn luyện mô hình:** Huấn luyện mô hình hồi quy logistic sử dụng dữ liệu huấn luyện. Ở đây, thay vì tìm đường tuyến tính tốt nhất như hồi quy tuyến tính, mục tiêu là tối ưu hóa hàm logistic để ước lượng xác suất.
- **Đánh giá mô hình:** Sử dụng các chỉ số như độ chính xác, AUC-ROC để đánh giá mô hình trên tập kiểm tra, tương tự như cách đánh giá mô hình hồi quy tuyến tính.
- **Tinh chỉnh mô hình:** Tinh chỉnh mô hình dựa trên kết quả đánh giá, có thể bao gồm việc thay đổi ngưỡng phân lớp hoặc thử nghiệm với các biến độc lập khác nhau.
- **Sử dụng mô hình:** Diễn giải kết quả và áp dụng mô hình vào tình huống thực tế, giống như trong hồi quy tuyến tính.

Bài tập

- Làm lại các bài tập với mô hình đa biến