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.

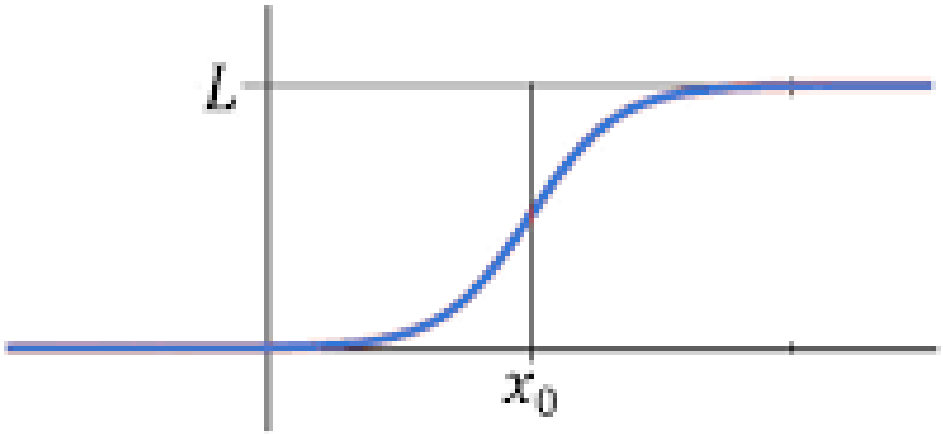
MÔ HÌNH HỒI QUY LOGISTIC

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:

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

trong đó,

- $P(Y = 1)$ là xác suất để sự kiện $Y = 1$ xảy ra (ví dụ: sự kiện thành công, lớp 1, …).
- $X_1, X_2, \ldots, X_k$ là các biến độc lập.
- $\beta_0, \beta_1, \ldots, \beta_k$ 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:

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

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 tuyển 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 hô hình:** Diễn giải kết quả và áp dụng mô hình vào tình huống thực tế, giống như trong hồi quy tuyến tính.

Projects

[Search | Kaggle]

Logistic Regression (Titanic Dataset)
Notebook · 4y ago · by Mohamed Hany
Logistic Regression (Titanic **Dataset**) We will be working with the [Titanic Data Set from Kaggle
57
9 comments

Credit Fraud || Dealing with Imbalanced Datasets
Notebook · 6y ago · by Janio Martinez Bachmann
regression from sklearn.metrics import accuracy_score # **Logistic Regression** with Under-Sampling y_pred
5632
662 comments

Logistic Regression-Customer Churn for Telecom Domain Dataset
Discussion Topic · 3y ago · by Shruti Pandit
In the General forum
45
3 comments

Why logistic regression works better then LightGBM on this dataset?
Discussion Topic · 3y ago · by The Devastator
As it turns out, **logistic regression** works much better on this **dataset** than any other algorithm.
23
15 comments

Deep Learning Tutorial for Beginners
Notebook · 6y ago · by DATAI
logistic regression. * However, in deep learning tutorial what to do with **logistic regression** there??
2874
279 comments

Logistic Regression-Titanic Dataset
Notebook · 2y ago · by SandhyaKrishnan02
Why **Logistic Regression** ?
65
12 comments

Bài tập

- Làm lại các bài tập với mô hình đa biến