Hồi quy logistic là gì?

Hồi quy logistic là một kỹ thuật phân tích dữ liệu sử dụng toán học để tìm ra mối quan hệ giữa hai yếu tố dữ liệu. Sau đó, kỹ thuật này sử dụng mối quan hệ đã tìm được để dự đoán giá trị của những yếu tố đó dựa trên yếu tố còn lại. Dự đoán thường cho ra một số kết quả hữu hạn, như có hoặc không.

Ví dụ: giả sử bạn muốn đoán xem khách truy cập trang web của bạn sẽ nhấp vào nút thanh toán trong giỏ hàng của họ hay không. Phân tích hồi quy logistic xem xét hành vi của khách truy cập trước đây, chẳng hạn như thời gian dành cho trang web và số lượng các mặt hàng trong giỏ hàng. Quá trình phân tích này xác định rằng, trước đây, nếu khách truy cập dành hơn năm phút trên trang web và thêm hơn ba mặt hàng vào giỏ hàng, họ sẽ nhấp vào nút thanh toán. Nhờ vào thông tin này, sau đó, hàm hồi quy logistic có thể dự đoán hành vi của một khách mới truy cập trang web.

Uư điểm:

* Tính đơn giản

Các mô hình hồi quy logistic ít phức tạp về mặt toán học hơn các phương pháp ML khác. Do đó, bạn có thể triển khai chúng ngay cả khi đội ngũ của bạn không ai có chuyên môn sâu về ML.

* Tốc độ

Các mô hình hồi quy logistic có thể xử lý khối lượng lớn dữ liệu ở tốc độ cao bởi chúng cần ít khả năng điện toán hơn, chẳng hạn như bộ nhớ và sức mạnh xử lý. Điều này khiến các mô hình hồi quy logistic trở nên lý tưởng đối với những tổ chức đang bắt đầu với các dự án ML để đạt được một số thành tựu nhanh chóng.

* Sự linh hoạt

Bạn có thể sử dụng hồi quy logistic để tìm đáp án cho các câu hỏi có hai hoặc nhiều kết quả hữu hạn. Bạn cũng có thể sử dụng phương pháp này để xử lý trước dữ liệu. Ví dụ: bạn có thể sắp xếp dữ liệu với một phạm vi giá trị lớn, chẳng hạn như giao dịch ngân hàng, thành một phạm vi giá trị hữu hạn, nhỏ hơn nhờ hồi quy logistic. Sau đó, bạn có thể xử lý tập dữ liệu nhỏ hơn này với các kỹ thuật ML khác để phân tích chính xác hơn.

* Khả năng hiển thị

Phân tích hồi quy logistic cung cấp cho nhà phát triển khả năng nhìn nhận các quy trình phần mềm nội bộ rõ hơn so với các kỹ thuật phân tích dữ liệu khác. Khắc phục sự cố và sửa lỗi cũng trở nên dễ dàng hơn do các phép toán ít phức tạp hơn.

Hồi quy logistic có những ứng dụng nào?

Sản xuất

Các công ty sản xuất áp dụng phân tích hồi quy logistic để ước tính xác suất xảy ra sự cố ở bộ phận trong máy móc. Sau đó, họ sẽ lên lịch bảo trì dựa trên xác suất đã ước tính này để giảm thiểu sự cố trong tương lai.

Chăm sóc sức khỏe

Các nhà nghiên cứu y khoa lên kế hoạch điều trị và chăm sóc dự phòng bằng cách dự đoán khả năng mắc bệnh ở bệnh nhân. Họ sử dụng các mô hình hồi quy logistic để so sánh tác động của tiền sử gia đình hoặc của bộ gen lên bệnh tật.

Tài chính

Các công ty tài chính phải phân tích các giao dịch tài chính để đề phòng gian lận, xem xét các đơn xin vay và đơn bảo hiểm để đề phòng rủi ro. Những vấn đề này phù hợp với mô hình hồi quy logistic bởi chúng có kết quả cụ thể, chẳng hạn như rủi ro cao hoặc rủi ro thấp và gian lận hoặc không gian lận.

Bộ phận Tiếp thị:

Các công cụ quảng cáo trực tuyến sử dụng mô hình hồi quy logistic để dự đoán xem người dùng sẽ nhấp vào một quảng cáo hay không. Kết quả là, các nhà tiếp thị có thể phân tích phản ứng của người dùng đối với những từ ngữ và hình ảnh khác nhau, tạo ra các quảng cáo hiệu suất cao có khả năng thu hút khách hàng.

Phân tích hồi quy hoạt động như thế nào?

1.Xác định câu hỏi

Bất kỳ quá trình phân tích dữ liệu nào cũng bắt đầu bằng một câu hỏi kinh doanh. Đối với hồi quy logistic, bạn nên giới hạn phạm vi câu hỏi để có được kết quả cụ thể:

* Những ngày mưa có ảnh hưởng đến doanh số hàng tháng của chúng ta không? (có hoặc không)
* Khách hàng đang thực hiện loại hoạt động thẻ tín dụng nào? (ủy quyền, gian lận hoặc có khả năng gian lận)

2.Thu thập dữ liệu lịch sử

Sau khi xác định câu hỏi, bạn cần xác định các yếu tố dữ liệu có liên quan. Sau đó, bạn sẽ thu thập dữ liệu trước đây cho tất cả các yếu tố. Ví dụ: để trả lời câu hỏi đầu tiên ở trên, bạn có thể thu thập dữ liệu doanh số hàng tháng và số ngày mưa mỗi tháng trong ba năm qua.

3.Đào tạo mô hình phân tích hồi quy

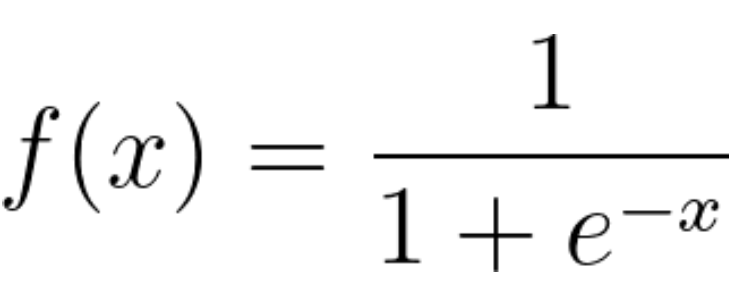
Bạn sẽ xử lý dữ liệu lịch sử bằng phần mềm hồi quy. Phần mềm sẽ xử lý các điểm dữ liệu khác nhau và kết nối chúng theo phương thức toán học bằng cách sử dụng phương trình. Ví dụ: nếu số ngày mưa trong ba tháng là 3, 5 và 8 còn doanh số trong những tháng đó là 8, 12 và 18, thuật toán hồi quy sẽ kết nối các yếu tố này với phương trình:

Doanh số = 2\*(Số ngày mưa) + 2

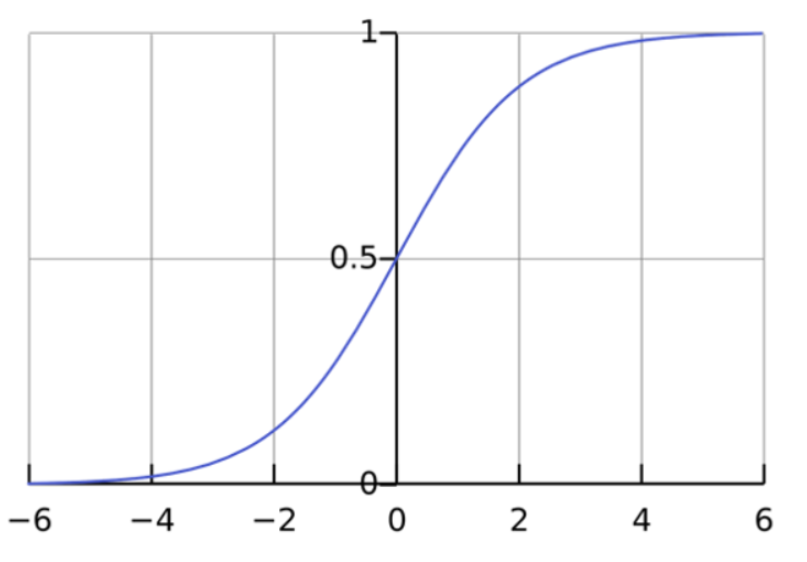
4.Dự đoán các giá trị không xác định

Đối với các giá trị không xác định, phần mềm sẽ áp dụng phương trình để dự đoán. Nếu bạn biết rằng tháng 7 sẽ có sáu ngày mưa, phần mềm sẽ ước tính giá trị doanh số trong tháng 7 là 14.

Hàm hồi quy logistic

Hồi quy logistic là một mô hình thống kê sử dụng hàm logistic, hay hàm logit trong toán học làm phương trình giữa *x* và *y*. Hàm logit ánh xạ *y* làm hàm sigmoid của *x*.

Nếu vẽ phương trình hồi quy logistic này, bạn sẽ có một đường cong hình chữ S như hình dưới đây.



Như bạn có thể thấy, hàm logit chỉ trả về các giá trị giữa 0 và 1 cho biến phụ thuộc, dù giá trị của biến độc lập là gì. Đây là cách hồi quy logistic ước tính giá trị của biến phụ thuộc. Phương pháp hồi quy logistic cũng lập mô hình phương trình giữa nhiều biến độc lập và một biến phụ thuộc.

Phân tích hồi quy logistic với nhiều biến độc lập

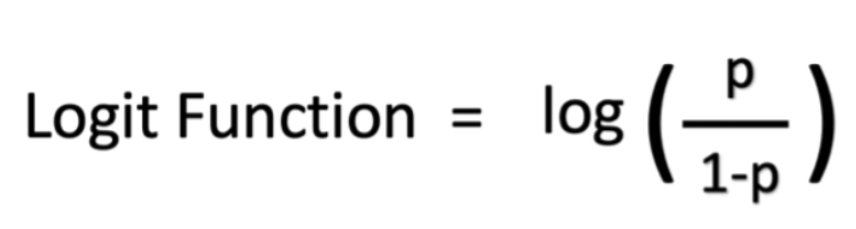
Trong nhiều trường hợp, nhiều biến giải thích ảnh hưởng đến giá trị của biến phụ thuộc. Để lập mô hình các tập dữ liệu đầu vào như vậy, công thức hồi quy logistic phải giả định mối quan hệ tuyến tính giữa các biến độc lập khác nhau. Bạn có thể sửa đổi hàm sigmoid và tính toán biến đầu ra cuối cùng như sau

*y* = *f*(β0 + β1*x*1 + β2*x*2+… βn*x*n)

Ký hiệu β đại diện cho hệ số hồi quy. Mô hình logit có thể đảo ngược tính toán các giá trị hệ số này khi bạn cho nó một tập dữ liệu thực nghiệm đủ lớn có các giá trị đã xác định của cả hai biến phụ thuộc và biến độc lập. Log của tỷ số odds

Mô hình logit cũng có thể xác định tỷ số thành công trên thất bại hay log của tỷ số odds. Ví dụ: nếu bạn đang chơi poker với bạn bè và thắng bốn ván trên mười ván, tỷ số chiến thắng của bạn là bốn phần sáu, hoặc 4/6, và đó là tỷ số thành công trên thất bại của bạn. Mặt khác, xác suất thắng là 4/10.

Về mặt toán học, tỷ số odds về mặt xác suất của bạn là *p*/(1 - *p*) và log của tỷ số odds là log (*p*/(1 - *p*)). Bạn có thể biểu diễn hàm logistic bằng log của tỷ số odds như hình dưới đây:



Phân tích hồi quy logistic có những loại nào?

Có ba cách tiếp cận phân tích hồi quy logistic dựa trên kết quả của biến phụ thuộc.

Hồi quy logistic nhị phân

Hồi quy logistic nhị phân phù hợp với các vấn đề phân lớp nhị phân chỉ có hai kết quả có thể xảy ra. Biến phụ thuộc chỉ có thể có hai giá trị, chẳng hạn như có và không hoặc 0 và 1.

Dù hàm logistic tính toán một phạm vi giá trị giữa 0 và 1, mô hình hồi quy nhị phân vẫn sẽ làm tròn kết quả đến các giá trị gần nhất. Nói chung, kết quả dưới 0,5 sẽ được làm tròn thành 0 và kết quả trên 0,5 sẽ được làm tròn thành 1, do đó hàm logistic trả về một kết quả nhị phân.

Hồi quy logistic đa thức

Hồi quy đa thức có thể phân tích các vấn đề có một số kết quả có thể xảy ra, miễn là số kết quả hữu hạn. Ví dụ: kỹ thuật này có thể dự đoán xem giá nhà sẽ tăng 25%, 50%, 75% hay 100% dựa trên dữ liệu dân số, nhưng sẽ không thể dự đoán được giá trị chính xác của một ngôi nhà.

Hồi quy logistic đa thức hoạt động bằng cách ánh xạ các giá trị kết quả cho các giá trị khác nhau giữa 0 và 1. Hàm logistic có thể trả về một khoảng dữ liệu liên tục như 0,1, 0,11, 0,12, v.v., do đó hồi quy đa thức cũng nhóm đầu ra đến các giá trị gần nhất có thể có.

Hồi quy logistic thứ tự

Hồi quy logistic thứ tự, hay mô hình logit có thứ tự, là một loại hồi quy đa thức đặc biệt cho các vấn đề trong đó các số đại diện cho các bậc chứ không phải là giá trị thực tế. Ví dụ: bạn sẽ sử dụng hồi quy thứ tự để dự đoán đáp án cho câu hỏi khảo sát yêu cầu khách hàng đánh giá dịch vụ của bạn ở mức kém, ổn, tốt hoặc xuất sắc dựa trên một giá trị số, chẳng hạn như số lượng mặt hàng họ mua từ bạn trong năm.

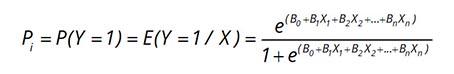
Hồi quy logistic nhị phân ( Binary logistic regression )

* Đặc trưng của hồi quy nhị phân là biến phụ thuộc chỉ có hai giá trị: 0 và 1.

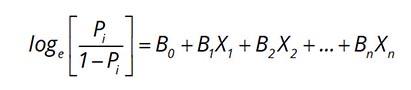
* Trên thực tế, có rất nhiều hiện tượng tự nhiên, hiện tượng kinh tế, xã hội,… mà chúng ta cần dự đoán khả năng xảy ra của nó như chiến dịch quảng cáo có được chấp nhận hay không, người vay có trả được nợ hay không, công ty có phá sản hay không, khách hàng có mua hay không,… Những biến nghiên cứu có hai biểu hiện như vậy được mã hóa thành hai giá trị 0 và 1.

2. Phương trình hồi quy Binary Logistic

Hồi quy Binary Logistic, chúng ta sẽ ước lượng xác suất xảy ra sự kiện Y (probability) khi biết giá trị X. Biến phụ thuộc Y có hai giá trị 0 và 1, với 0 là không xảy ra sự kiện và 1 là xảy ra sự kiện. Từ đặc điểm này, chúng ta có thể đánh giá được khả năng xảy ra sự kiện (Y = 1) nếu xác suất dự đoán lớn hơn 0.5, ngược lại, khả năng không xảy ra sự kiện (Y = 0) nếu xác suất dự đoán nhỏ hơn 0.5. Ta có hàm xác suất như sau:



Trong đó Pi = P(Y = 1) = E(Y = 1/ X) là xác suất xảy ra sự kiện. Thực hiện các phép chuyển đổi toán học, chúng ta thu được phương trình hồi quy Binary Logistic như sau:



Trong đó:

* *Pi:* xác suất xảy ra sự kiện (Y = 1)
* *1-Pi:* xác suất không xảy ra sự kiện (Y = 0)
* *B0:* hằng số hồi quy
* *B1, B2, … Bn:* hệ số hồi quy