**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI**

**PHÂN HIỆU TẠI THÀNH PHỐ HỒ CHÍ MINH**



**BÁO CÁO BÀI TẬP LỚN**

**CHUYÊN ĐỀ CÔNG NGHỆ PHẦN MỀM**

Giảng viên hướng dẫn: Trần Phong Nhã

Sinh viên thực hiện: Nguyễn Tuấn Đạt - 6051071029

Nguyễn Gia Ngọc - 6051071077

Trần Minh Nhân - 6051071080

Võ Thành Nhân - 6051071081

Huỳnh Phước Trường - 6051071134

Lớp: CQ.60.CNTT

TP. Hồ Chí Minh, năm 2022

LỜI CẢM ƠN

Lời đầu tiên chúng em xin gửi lời cảm ơn chân thành đến quý thầy, cô giáo trong **Bộ môn Công nghệ thông tin – Phân hiệu Trường Đại học Giao thông vận tải.** Những người đã truyền dạy, đã trang bị cho chúng em kho tàng kiến thức về bầu trời công nghệ thông tin rộng lớn. Ở đây, chúng em không chỉ học được kiến thức về sách vở mà còn học được các bài học, kỹ năng sống. Đặc biệt, chúng em xin gửi lời cảm ơn chân thành và sâu sắc đến thầy Trần Phong Nhã, người đã giúp chúng em trong quá trình làm bài tập lớn, người đã bỏ thời gian quý báu, thậm chí là thời gian nghỉ ngơi để hướng dẫn, để giúp chúng em có thêm kiến thức và kinh nghiệm về môn chuyên đề công nghệ phần mềm. Em thật chẳng biết dùng lời nào để diễn tả được công lao của thầy. “**CẢM ƠN THẦY**. **CẢM ƠN THẦY VỀ TẤT CẢ**”.

Trong quá trình học tập và tìm hiểu chúng em đã nổ lực rất nhiều với mong muốn hoàn thành bài tập một cách tốt nhất, nhưng sai lầm là không thể không mắc phải. Em mong thầy, cô bộ môn có thể thông cảm và cho em những ý kiến, đóng góp để chúng em có thể hoàn thành tốt bài tập này.

Sau cùng, em xin kính chúc Thầy lời chúc sức khoẻ, luôn hạnh phúc và thành công hơn nữa trong công việc cũng như trong cuộc sống.

Chúng em xin chân thành cảm ơn!

NHẬN XÉT CỦA GIẢNG VIÊN

*Tp. Hồ Chí Minh, ngày … tháng … năm …*

Giảng viên hướng dẫn

**MỤC LỤC**

[Phần 1 SƠ LƯỢC THUẬT TOÁN HỒI QUY 6](#_Toc119357732)

[1.1 Khái niệm 6](#_Toc119357733)

[1.2 Các thuật toán hồi quy phổ biến nhất trong Machine Learning 6](#_Toc119357734)

[1.2.1 Linear Regression 6](#_Toc119357735)

[1.2.2 Logistic Regression 14](#_Toc119357736)

[1.2.3 Stepwise Regression 22](#_Toc119357737)

[1.2.4 Ordinary Least Squares Regression (OLSR) 22](#_Toc119357738)

[1.2.5 Multivariate Adaptive Regression Splines (MARS) 23](#_Toc119357739)

[1.2.6 Locally Estimated Scatterplot Smoothing (LOESS) 25](#_Toc119357740)

[Phần 2 DEMO 30](#_Toc119357741)

[2.1 Linear Regression 30](#_Toc119357742)

[2.2 Logistic Regression 31](#_Toc119357743)

[2.3 Stepwise Regression 32](#_Toc119357744)

[2.4 Ordinary Least Squares Regression (OLSR) 33](#_Toc119357745)

[2.5 Multivariate Adaptive Regression Splines (MARS) 34](#_Toc119357746)

[2.6 Locally Estimated Scatterplot Smoothing (LOESS) 35](#_Toc119357747)

**MỤC LỤC HÌNH ẢNH**

[Hình 1 Đường hồi quy tuyến tính 8](#_Toc119357748)

[Hình 2 Phương trình hồi quy logistic 17](#_Toc119357749)

[Hình 3 Hàm hồi quy logistic 19](#_Toc119357750)

[Hình 4 Đồ thị Satter 23](#_Toc119357751)

[Hình 5 Biểu đồ phân tích hồi quy LOWESS 27](#_Toc119357752)

[Hình 6 LOWESS Approximation of the Sine Wave 28](#_Toc119357753)

[Hình 7 LOWESS Approximation of the Sine Wave 28](#_Toc119357754)

[Hình 8 LOWESS Approximation of the Sine Wave 29](#_Toc119357755)

[Hình 9 Linear Regression 30](#_Toc119357756)

[Hình 10 Logistic Regression 31](#_Toc119357757)

[Hình 11 Stepwise Regression 32](#_Toc119357758)

[Hình 12 Ordinary Least Squares Regression (OLSR) 33](#_Toc119357759)

[Hình 13 Multivariate Adaptive Regression Splines (MARS) 34](#_Toc119357760)

[Hình 14 Locally Estimated Scatterplot Smoothing (LOESS) 35](#_Toc119357761)

1. SƠ LƯỢC THUẬT TOÁN HỒI QUY
   1. Khái niệm

Thuật toán hồi quy có liên quan mật thiết với việc mô hình hóa mối quan hệ giữa các biến, mà ta có thể tinh chỉnh bằng cách sử dụng một thước đo tần suất các lỗi xuất hiện trong dự đoán được thực hiện bởi thuật toán machine learning.

Những phương pháp này là một đặc điểm của thống kê.Ngoài ra, đã được chọn tham gia machine learning thống kê. Điều này có thể gây nhầm lẫn bởi vì chúng ta có thể sử dụng hồi quy để chỉ lớp vấn đề và lớp thuật toán.

* 1. Các thuật toán hồi quy phổ biến nhất trong Machine Learning
* Linear Regression
* Logistic Regression
* Stepwise Regression
* Ordinary Least Squares Regression (OLSR)
* Multivariate Adaptive Regression Splines (MARS)
* Locally Estimated Scatterplot Smoothing (LOESS)
  + 1. Linear Regression

Thuật toán Linear Regression (Hồi Quy Tuyến Tính) thuộc nhóm Supervised learning (Học có giám sát ). Hồi quy tuyến tính là một phương pháp rất đơn giản nhưng đã được chứng minh được tính hữu ích cho một số lượng lớn các tình huống. Trong bài viết này, bạn sẽ khám phá ra chính xác cách thức tuyến tính làm việc như thế nào. Trong việc phân tích dữ liệu, bạn sẽ tiếp xúc với thuật ngữ "Regression" (Hồi quy ) rất thường xuyên. Trước khi đi sâu vào Hồi quy tuyến tính, hãy tìm hiểu khái niệm Hồi quy trước đã. Hồi quy chính là một phương pháp thống kê để thiết lập mối quan hệ giữa một biến phụ thuộc và một nhóm tập hợp các biến độc lập. Ví dụ:

Tuổi = 5 + Chiều cao \* 10 + Trọng lượng \* 13

Ở đây chính ta đang thiết lập mối quan hệ giữa Chiều cao & Trọng lượng của một người với Tuổi của anh/cô ta. Đây là một ví dụ rất cơ bản của Hồi quy.

* + - 1. Hồi quy tuyến tính giản đơn

Introduction

"Hồi quy tuyến tính" là một phương pháp thống kê để hồi quy dữ liệu với biến phụ thuộc có giá trị liên tục trong khi các biến độc lập có thể có một trong hai giá trị liên tục hoặc là giá trị phân loại. Nói cách khác "Hồi quy tuyến tính" là một phương pháp để dự đoán biến phụ thuộc (Y) dựa trên giá trị của biến độc lập (X). Nó có thể được sử dụng cho các trường hợp chúng ta muốn dự đoán một số lượng liên tục. Ví dụ, dự đoán giao thông ở một cửa hàng bán lẻ, dự đoán thời gian người dùng dừng lại một trang nào đó hoặc số trang đã truy cập vào một website nào đó v.v...

Chuẩn bị

Để bắt đầu với Hồi quy tuyến tính, chúng ta hãy đi lướt qua một số khái niệm toán học về thống kê.

* Tương quan (r) - Giải thích mối quan hệ giữa hai biến, giá trị có thể chạy từ -1 đến +1
* Phương sai (σ2) - Đánh giá độ phân tán trong dữ liệu của bạn
* Độ lệch chuẩn (σ) - Đánh giá độ phân tán trong dữ liệu của bạn (căn bậc hai của phương sai)
* Phân phối chuẩn
* Sai số (lỗi) - {giá trị thực tế - giá trị dự đoán}

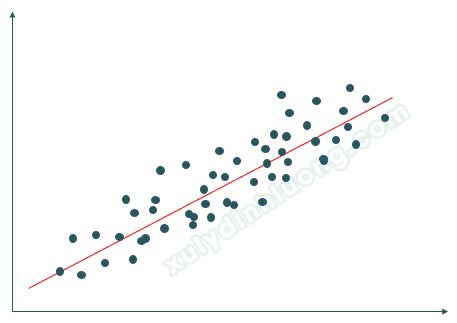
Giả định

Không một kích thước nào phù hợp cho tất cả, điều này cũng đúng đối với Hồi quy tuyến tính. Để thoả mãn hồi quy tuyến tính, dữ liệu nên thoả mãn một vài giả định quan trọng. Nếu dữ liệu của bạn không làm theo các giả định, kết quả của bạn có thể sai cũng như gây hiểu nhầm.

* Tuyến tính & Thêm vào : Nên có một mối quan hệ tuyến tính giữa biến độc lập và biến không độc lập và ảnh hưởng của sự thay đổi trong giá trị của các biến độc lập nên ảnh hưởng thêm vào tới các biến phụ thuộc.
* Tính bình thường của phân bổ các lỗi : Sự phân bổ sai khác giữa các giá trị thực và giá trị dự đoán (sai số) nên được phân bổ một cách bình thường.
* Sự tương đồng: Phương sai của các lỗi nên là một giá trị không đổi so với
* Thời gian
* Dự đoán
* Giá trị của các biến độc lập
* Sự độc lập về thống kê của các lỗi: Các sai số (dư) không nên có bất kỳ mối tương quan nào giữa chúng. Ví dụ: Trong trường hợp dữ liệu theo chuỗi thời gian, không nên có sự tương quan giữa các sai số liên tiếp nhau.

Đường hồi quy tuyến tính

Trong khi sử dụng hồi quy tuyến tính, mục tiêu của chúng ta là để làm sao một đường thẳng có thể tạo được sự phân bố gần nhất với hầu hết các điểm. Do đó làm giảm khoảng cách (sai số) của các điểm dữ liệu cho đến đường đó.



Hình 1 Đường hồi quy tuyến tính

Ví dụ, ở các điểm ở hình trên (trái) biểu diễn các điểm dữ liệu khác nhau và đường thẳng (bên phải) đại diện cho một đường gần đúng có thể giải thích mối quan hệ giữa các trục x & y. Thông qua, hồi quy tuyến tính chúng ta cố gắng tìm ra một đường như vậy. Ví dụ, nếu chúng ta có một biến phụ thuộc Y và một biến độc lập X - mối quan hệ giữa X và Y có thể được biểu diễn dưới dạng phương trình sau:

Y = Β0 + Β1\*X

Ở đây

* Y = Biến phụ thuộc
* X = biến độc lập
* Β0 = Hằng số
* Β1 = Hệ số mối quan hệ giữa X và Y

Một vài tính chất của hồi quy tuyến tính

* Đường hồi quy luôn luôn đi qua trung bình của biến độc lập (x) cũng như trung bình của biến phụ thuộc(y)
* Đường hồi qui tối thiểu hóa tổng của "Diện tích các sai số". Đó là lý do tại sao phương pháp hồi quy tuyến tính được gọi là "Ordinary Least Square (OLS)"
* Β1 giải thích sự thay đổi trong Y với sự thay đổi X bằng một đơn vị. Nói cách khác, nếu chúng ta tăng giá trị của X bởi một đơn vị thì nó sẽ là sự thay đổi giá trị của Y

Tìm đường hồi quy tuyến tính

Sử dụng công cụ thống kê ví dụ như Excel, R, SAS ... bạn sẽ trực tiếp tìm hằng số (B0 và B1) như là kết quả của hàm hồi quy tuyến tính. Như lý thuyết ở trên, nó hoạt động trên khái niệm OLS và cố gắng giảm bớt diện tích sai số, các công cụ này sử dụng các gói phần mềm tính các hằng số này.

* + - 1. Hồi quy nhiều tuyến tính

Nhiều hồi quy tuyến tính cố gắng mô hình hóa mối quan hệ giữa hai hoặc nhiều đặc điểm và phản hồi bằng cách điều chỉnh phương trình tuyến tính với dữ liệu quan sát được.

Rõ ràng, nó không là gì khác ngoài một phần mở rộng của hồi quy tuyến tính đơn giản.

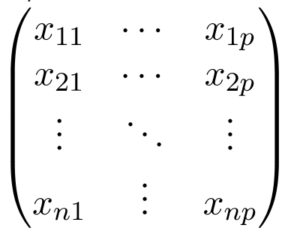
Hãy xem xét một tập dữ liệu với p đặt điểm(hoặc biến độc lập) và một phản hồi (hoặc biến phụ thuộc).

Ngoài ra, tập dữ liệu chứa n hàng quan sát.

Chúng ta xác định:

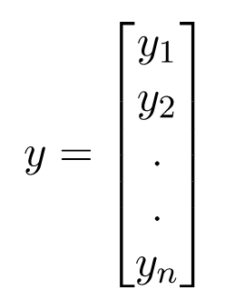
X (ma trận đặc trưng) = ma trận cỡ n X p trong đó x\_ {ij} biểu thị các giá trị của đặc điểm thứ j cho lần quan sát thứ i.

Vì thế



Và

y (vectơ phản hồi) = vectơ có kích thước n trong đó y\_ {i} biểu thị giá trị của phản ứng cho lần quan sát thứ i.

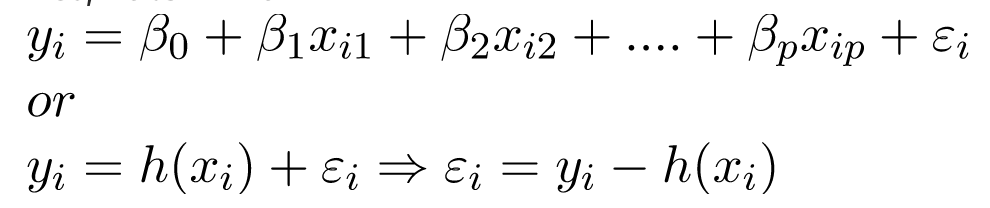


Đường hồi quy cho p đối tượng được biểu diễn như sau:

https://cafedev.vn/wp-content/uploads/2020/12/Screen-Shot-2020-12-18-at-1.39.34-PM-300x24.png

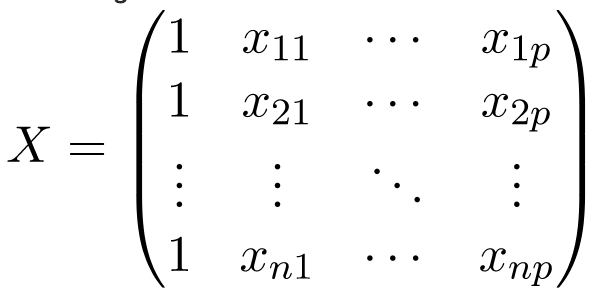
Trong đó h (x\_i) là giá trị phản hồi dự đoán cho lần quan sát thứ i và b\_0, b\_1,…, b\_p là các hệ số hồi quy.

Ngoài ra, chúng ta có thể viết:

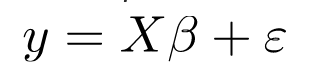


Trong đó e\_i đại diện cho sai số còn lại trong lần quan sát thứ i.

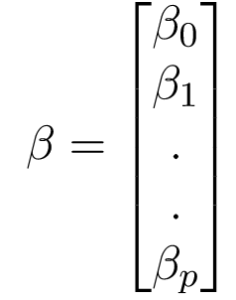
Chúng ta có thể tổng quát hóa mô hình tuyến tính của mình hơn một chút bằng cách biểu diễn ma trận đặc trưng X dưới dạng:



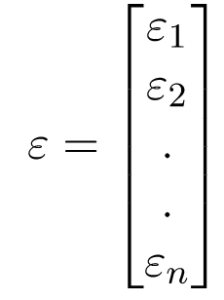
Vì vậy, bây giờ, mô hình tuyến tính có thể được biểu diễn dưới dạng ma trận như sau:



Ở đâu



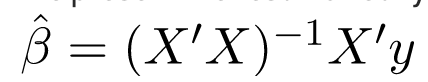
Và



Bây giờ, chúng ta xác định ước tính của b, tức là b ’bằng phương pháp Bình phương nhỏ nhất.

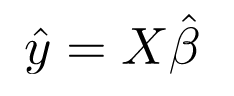
Như đã được giải thích, phương pháp Least Squares có xu hướng xác định b ’mà tổng sai số dư được giảm thiểu.

Chúng ta trình bày kết quả trực tiếp tại đây:



Trong đó biểu diễn chuyển vị của ma trận trong khi -1 biểu thị nghịch đảo của ma trận.

Khi biết các ước lượng bình phương nhỏ nhất, b ’, mô hình hồi quy tuyến tính bội số hiện có thể được ước tính là:



Trong đó y ’là vectơ phản hồi ước tính.

* + - 1. Các ứng dụng:
* Đường xu hướng: Đường xu hướng biểu thị sự thay đổi của một số dữ liệu định lượng theo thời gian (như GDP, giá dầu, v.v.). Các xu hướng này thường tuân theo một mối quan hệ tuyến tính. Do đó, hồi quy tuyến tính có thể được áp dụng để dự đoán các giá trị trong tương lai. Tuy nhiên, phương pháp này thiếu giá trị khoa học trong trường hợp các thay đổi tiềm ẩn khác có thể ảnh hưởng đến dữ liệu.
* Kinh tế học: Hồi quy tuyến tính là công cụ thực nghiệm chủ yếu trong kinh tế học. Ví dụ: nó được sử dụng để dự đoán chi tiêu tiêu dùng, chi tiêu đầu tư cố định, đầu tư hàng tồn kho, mua hàng xuất khẩu của một quốc gia, chi tiêu cho nhập khẩu, nhu cầu nắm giữ tài sản lưu động, nhu cầu lao động và cung lao động.
* Tài chính: Mô hình tài sản giá vốn sử dụng hồi quy tuyến tính để phân tích và định lượng rủi ro hệ thống của một khoản đầu tư.
* Sinh học: Hồi quy tuyến tính được sử dụng để mô hình hóa mối quan hệ nhân quả giữa các thông số trong hệ thống sinh học.
  + 1. Logistic Regression
       1. Khái niệm:

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.

* + - 1. Tại sao hồi quy logistic lại quan trọng?

Hồi quy logistic là một kỹ thuật quan trọng trong lĩnh vực trí tuệ nhân tạo và máy học (AI/ML). Mô hình ML là các chương trình phần mềm có thể được đào tạo để thực hiện các tác vụ xử lý dữ liệu phức tạp mà không cần sự can thiệp của con người. Mô hình ML được xây dựng bằng hồi quy logistic có thể giúp các tổ chức thu được thông tin chuyên sâu hữu ích từ dữ liệu kinh doanh của mình. Họ có thể sử dụng những thông tin chuyên sâu này để phân tích dự đoán nhằm giảm chi phí hoạt động, tăng độ hiệu quả và đổi chỉnh quy mô nhanh hơn. Ví dụ: doanh nghiệp có thể khám phá các mẫu hình cải thiện khả năng giữ chân nhân viên hoặc tạo ra thiết kế sản phẩm mang về nhiều lợi nhuận hơn.

Dưới đây là một số lợi ích của việc sử dụng hồi quy logistic so với các kỹ thuật ML khác.

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.

* + - 1. Hồi quy logistic có những ứng dụng nào?

Hồi quy logistic có một số ứng dụng thực tế trong nhiều ngành công nghiệp khác nhau.

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.

* + - 1. Phân tích hồi quy hoạt động như thế nào?

Hồi quy logistic là một trong số ít các kỹ thuật phân tích hồi quy khác nhau thường được các nhà khoa học dữ liệu sử dụng trong máy học (ML). Để hiểu rõ về hồi quy logistic, trước tiên chúng ta phải hiểu kỹ thuật phân tích hồi quy cơ bản. Dưới đây là một ví dụ về phân tích hồi quy tuyến tính để cho thấy cách thức hoạt động của phân tích hồi quy.

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)

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.

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

Dự đoán các giá trị không xác định

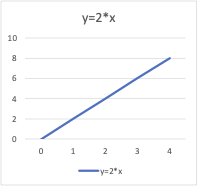
Đối với các giá trị không xác định, phần mềm sẽ áp dụng phương trình để dự đoán. Nếu bạn biết rằng tháng 7 sẽ có sáu ngày mưa, phần mềm sẽ ước tính giá trị doanh số trong tháng 7 là 14.

* + - 1. Mô hình hồi quy logistic hoạt động như thế nào?

Để hiểu rõ về mô hình hồi quy logistic, trước tiên chúng ta phải hiểu các phương trình và biến.

Phương trình

Trong toán học, phương trình cho ta mối quan hệ giữa hai biến: x và y. Bạn có thể sử dụng các phương trình hoặc hàm này để vẽ đồ thị theo trục x và trục y bằng cách nhập các giá trị khác nhau của x và y. Ví dụ: nếu bạn vẽ đồ thị cho hàm y = 2\*x, bạn sẽ có một đường thẳng như hình dưới đây. Do đó hàm này còn được gọi là hàm tuyến tính.



Hình 2 Phương trình hồi quy logistic

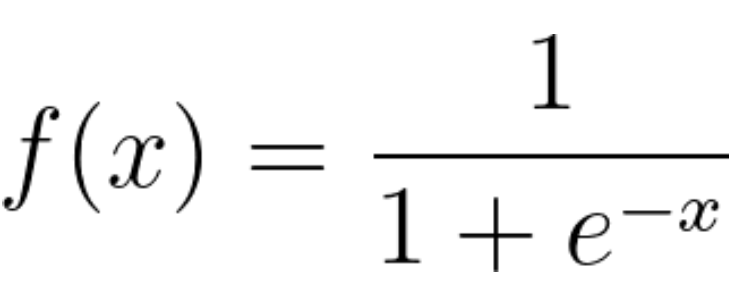
Biến

Trong thống kê, biến là các yếu tố dữ liệu hoặc thuộc tính có giá trị khác nhau. Bất kỳ phân tích nào cũng có một số biến nhất định là biến độc lập hoặc biến giải thích. Những thuộc tính này là nguyên nhân của một kết quả. Các biến khác là biến phụ thuộc hoặc biến đáp ứng; giá trị của chúng phụ thuộc vào các biến độc lập. Nhìn chung, hồi quy logistic khám phá cách các biến độc lập ảnh hưởng đến một biến phụ thuộc bằng cách xem xét các giá trị dữ liệu lịch sử của cả hai biến.

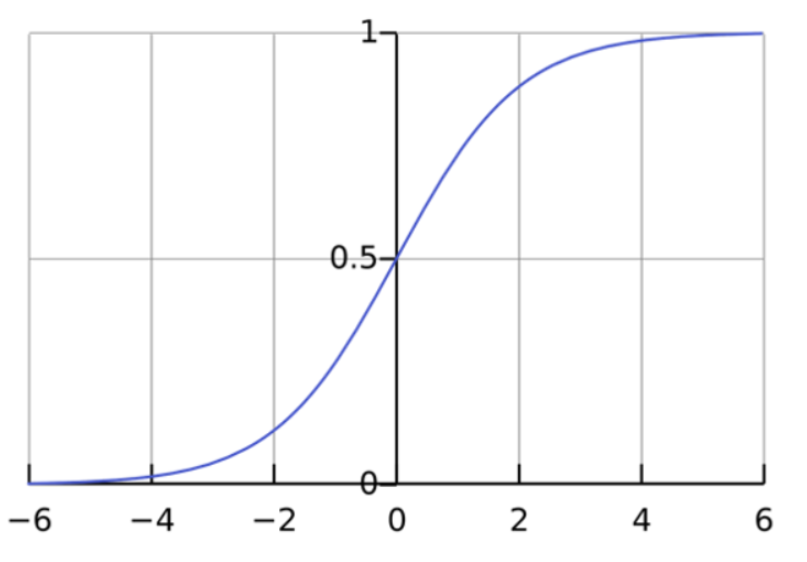
Trong ví dụ ở trên của chúng tôi, x được gọi là biến độc lập, biến dự đoán hoặc biến giải thích vì nó có một giá trị đã xác định. Y được gọi là biến phụ thuộc, biến kết quả hoặc biến đáp ứng vì giá trị của nó không xác định.

* + - 1. 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.



Hình 3 Hàm hồi quy logistic

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.

* + - 1. 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

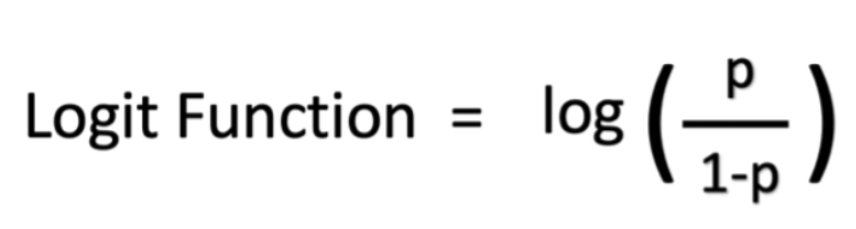
y=f(β0+β1x1+β2x2+…βnxn)

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.

* + - 1. 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:



* + - 1. 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.

* + - 1. Làm thế nào để so sánh hồi quy logistic với các kỹ thuật ML khác?

Hai kỹ thuật phân tích dữ liệu phổ biến là phân tích hồi quy tuyến tính và deep learning.

Phân tích hồi quy tuyến tính

Như đã giải thích ở trên, hồi quy tuyến tính lập mô hình mối quan hệ giữa các biến phụ thuộc và độc lập bằng tổ hợp tuyến tính. Phương trình hồi quy tuyến tính là

y= β0X0+β1X1+β2X2+…βnXn+ ε, trong đó β1 đến βn và ε là các hệ số hồi quy.

Hồi quy logistic so với hồi quy tuyến tính

Hồi quy tuyến tính dự đoán một biến phụ thuộc liên tục bằng một tập hợp các biến độc lập cho trước. Một biến liên tục có thể có một phạm vi giá trị, chẳng hạn như giá cả hoặc độ tuổi. Do đó, hồi quy tuyến tính có thể dự đoán giá trị thực của biến phụ thuộc. Kỹ thuật này có thể trả lời các câu hỏi như "Giá gạo sau 10 năm nữa sẽ là bao nhiêu?"

Không giống như hồi quy tuyến tính, hồi quy logistic là một thuật toán phân loại. Kỹ thuật này không thể dự đoán giá trị thực sự cho dữ liệu liên tục. Kỹ thuật này có thể trả lời các câu hỏi như "Liệu giá gạo trong 10 năm nữa có tăng 50% hay không?"

Deep learning

Deep learning sử dụng mạng nơ-ron hoặc các thành phần phần mềm mô phỏng bộ não con người để phân tích thông tin. Các phép toán deep learning dựa trên khái niệm toán học của vectơ.

Hồi quy logistic so với deep learning

Hồi quy logistic ít phức tạp và có cường độ điện toán ít hơn so với deep learning. Quan trọng hơn là nhà phát triển không thể điều tra hoặc sửa đổi các phép toán deep learning do tính chất phức tạp và dựa trên máy móc của chúng. Mặt khác, các phép toán hồi quy logistic lại minh bạch và dễ khắc phục sự cố hơn.

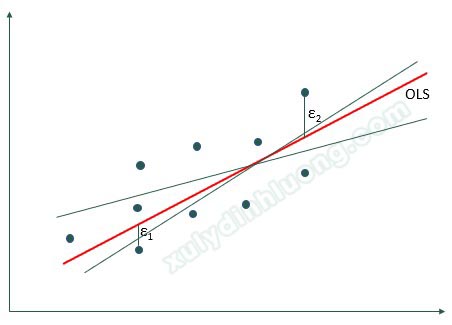
* + 1. Stepwise Regression

Trong bất kỳ hiện tượng nào, sẽ có một số yếu tố đóng vai trò lớn hơn trong việc xác định một kết quả. Nói một cách dễ hiểu, hồi quy từng bước là một quá trình giúp xác định yếu tố nào quan trọng và yếu tố nào không. Một số biến nhất định có giá trị p khá cao và không góp phần có ý nghĩa vào độ chính xác của dự đoán của chúng tôi. Từ đó, chỉ các yếu tố quan trọng được giữ lại để đảm bảo rằng mô hình tuyến tính thực hiện dự đoán của nó dựa trên các yếu tố có thể giúp nó tạo ra kết quả chính xác nhất.

Trong bài viết này, nhóm em sẽ trình bày cách sử dụng hồi quy từng bước sử dụng cách tiếp cận loại bỏ ngược. Đây là nơi ban đầu tất cả các biến được bao gồm và trong mỗi bước, biến không có ý nghĩa thống kê nhất bị loại bỏ. Nói cách khác, biến “vô dụng” nhất được kick. Điều này được lặp lại cho đến khi tất cả các biến còn lại có ý nghĩa thống kê.

* + 1. Ordinary Least Squares Regression (OLSR)

Một trong các phương pháp ước lượng hồi quy tuyến tính phổ biến là **bình phương nhỏ nhất OLS** (Ordinary Least Squares). Với tổng thể, sai số (error) ký hiệu là e, còn trong mẫu nghiên cứu sai số lúc này được gọi là phần dư (residual) và được ký hiệu là ε. Biến thiên phần dư được tính bằng tổng bình phương tất cả các phần dư cộng lại. Nguyên tắc của phương pháp **hồi quy OLS** là làm cho biến thiên phần dư này trong phép hồi quy là nhỏ nhất. Khi biểu diễn trên mặt phẳng Oxy, đường hồi quy OLS là một đường thẳng đi qua đám đông các điểm dữ liệu mà ở đó, khoảng cách từ các điểm dữ liệu (trị tuyệt đối của ε) đến đường hồi quy là ngắn nhất.



Hình 4 Đồ thị Satter

Từ đồ thị scatter biểu diễn mối quan hệ giữa các biến độc lập và biến phụ thuộc, các điểm dữ liệu sẽ nằm phân tán nhưng có xu hướng chung tạo thành dạng một đường thẳng. Chúng ta có thể có rất nhiều đường đường thẳng hồi quy đi qua đám đông các điểm dữ liệu này chứ không phải chỉ một đường duy nhất, vấn đề là ta phải chọn ra đường thẳng nào mô tả sát nhất xu hướng dữ liệu. **Bình phương nhỏ nhất OLS** sẽ tìm ra đường thẳng đó dựa trên nguyên tắc cực tiểu hóa khoảng cách từ các điểm dữ liệu đến đường thẳng. Trong hình ở trên đường màu đỏ là đường hồi quy OLS.

* + 1. Multivariate Adaptive Regression Splines (MARS)

**Multivariate Adaptive Regression Splines**, hoặc MARS, là một thuật toán cho các bài toán hồi quy phi tuyến tính phức tạp.

Thuật toán liên quan đến việc tìm một tập hợp các hàm tuyến tính đơn giản mà tổng hợp lại dẫn đến hiệu suất dự đoán tốt nhất. Theo cách này, MARS là một dạng tập hợp các hàm tuyến tính đơn giản và có thể đạt được hiệu suất tốt trong các bài toán hồi quy đầy thách thức với nhiều biến đầu vào và các mối quan hệ phi tuyến tính phức tạp.

Thuật toán MARS liên quan đến việc khám phá một tập hợp các hàm tuyến tính từng mảnh đơn giản đặc trưng cho dữ liệu và sử dụng chúng tổng hợp để đưa ra dự đoán. Theo một nghĩa nào đó, mô hình là một tập hợp các hàm tuyến tính.

Một hàm tuyến tính mảnh là một hàm bao gồm các hàm nhỏ hơn. Trong trường hợp này, nó là một hàm có kết quả đầu ra là 0 hoặc giá trị đầu vào trực tiếp.

“Hàm đúng” của một biến đầu vào liên quan đến việc chọn một giá trị cụ thể cho biến đó và xuất giá trị 0 cho tất cả các giá trị dưới giá trị và xuất giá trị nguyên trạng cho tất cả các giá trị trên giá trị đã chọn.

**f(x) = x if x > value, else 0**

Hoặc ngược lại, "hàm trái" có thể được sử dụng trong đó các giá trị nhỏ hơn giá trị đã chọn được xuất trực tiếp và các giá trị lớn hơn giá trị đã chọn sẽ xuất ra giá trị không.

**f(x) = x if x < value, else 0**

Đây được gọi là hàm bản lề, trong đó giá trị được chọn hoặc điểm phân tách là "nút" của hàm. Nó cũng được gọi là một hàm tuyến tính chỉnh lưu trong mạng nơ-ron.

Thuật toán MARS tạo ra nhiều hàm này, được gọi là hàm cơ sở cho một hoặc nhiều biến đầu vào.

Sau đó, một mô hình hồi quy tuyến tính được học từ đầu ra của mỗi hàm cơ sở này với biến mục tiêu. Điều này có nghĩa là đầu ra của mỗi hàm cơ sở được tính theo hệ số. Dự đoán được thực hiện bằng cách tính tổng kết quả đầu ra có trọng số của tất cả các hàm cơ bản trong mô hình.

Chìa khóa của thuật toán MARS là cách các hàm cơ sở được chọn. Điều này bao gồm hai bước: giai đoạn phát triển hoặc thế hệ được gọi là giai đoạn thuận và giai đoạn cắt tỉa hoặc tinh chế được gọi là giai đoạn lùi.

* Giai đoạn chuyển tiếp: Tạo các hàm cơ sở ứng viên cho mô hình.
* Giai đoạn lùi lại: Xóa các chức năng cơ bản khỏi mô hình.

Giai đoạn chuyển tiếp bao gồm việc tạo ra các hàm cơ sở và thêm vào mô hình. Giống như một cây quyết định, mỗi giá trị cho mỗi biến đầu vào trong tập dữ liệu huấn luyện được coi là một ứng cử viên cho một hàm cơ sở.

Các hàm luôn được thêm vào từng cặp, đối với phiên bản bên trái và bên phải của hàm tuyến tính từng đoạn của cùng một điểm phân tách. Một cặp hàm đã tạo chỉ được thêm vào mô hình nếu nó làm giảm lỗi của mô hình tổng thể.

Giai đoạn lùi lại liên quan đến việc chọn các chức năng để xóa khỏi mô hình, từng chức năng một. Một chức năng chỉ bị xóa khỏi mô hình nếu nó không dẫn đến tác động đến hiệu suất (trung tính) hoặc tăng hiệu suất dự đoán.

Sự thay đổi về hiệu suất của mô hình trong giai đoạn lùi được đánh giá bằng cách sử dụng xác nhận chéo của tập dữ liệu đào tạo, được gọi là xác nhận chéo tổng quát hoặc GCV. Do đó, có thể ước tính ảnh hưởng của từng mô hình tuyến tính từng đoạn đối với hiệu suất của mô hình.

Số lượng chức năng được sử dụng bởi mô hình được xác định tự động, vì quá trình cắt tỉa sẽ tạm dừng khi không thể thực hiện thêm cải tiến nào.

Hai siêu tham số chính cần xem xét là tổng số hàm ứng viên cần tạo, thường được đặt thành một số rất lớn và mức độ của các hàm cần tạo.

Mức độ là số lượng biến đầu vào được xem xét bởi mỗi hàm tuyến tính từng phần. Theo mặc định, giá trị này được đặt thành một, nhưng có thể được đặt thành giá trị lớn hơn để cho phép các tương tác phức tạp giữa các biến đầu vào được mô hình nắm bắt. Mức độ thường được giữ ở mức nhỏ để hạn chế độ phức tạp tính toán của mô hình (bộ nhớ và thời gian thực thi).

Một lợi ích của thuật toán MARS là nó chỉ sử dụng các biến đầu vào để nâng cao hiệu suất của mô hình. Giống như các thuật toán tổng hợp rừng ngẫu nhiên và đóng bao, MARS đạt được một kiểu lựa chọn tính năng tự động.

* + 1. Locally Estimated Scatterplot Smoothing (LOESS)

**Locally Weighted Scatterplot Smoothing** có trọng số cục bộ nằm trong nhóm các thuật toán hồi quy dưới ô của Học tập có giám sát. Điều này có nghĩa là bạn cần một tập hợp dữ liệu được gắn nhãn với một biến mục tiêu số để đào tạo mô hình của mình.

Vì **LOWESS** là một kỹ thuật so khớp phi tham số, chúng ta không cần phải giả định rằng dữ liệu tuân theo bất kỳ phân phối cụ thể nào. Tuy nhiên, đồng thời, phù hợp không tham số có nghĩa là cuối cùng, chúng ta sẽ không có một phương trình tổng thể để bạn sử dụng để dự đoán giá trị của các điểm dữ liệu mới. Tuy nhiên, đừng lo lắng, vì cung cấp một giải pháp cho vấn đề này trong phần Python ở phần sau của câu chuyện này.

Với những ưu nhược điểm trên, **LOWESS** thường được sử dụng để thực hiện các phân tích sau:

* Phân tích hồi quy trong đó hồi quy tuyến tính đơn giản không thể tạo ra một đường phù hợp do các thuộc tính dữ liệu không tuân theo mối quan hệ tuyến tính. Lưu ý, một trường hợp đặc biệt của hồi quy nhiều tuyến tính được gọi là hồi quy đa thức cũng có thể được sử dụng trong trường hợp này.
* Khám phá và phân tích dữ liệu trong khoa học xã hội, chẳng hạn như bầu cử và hành vi bỏ phiếu.
* Việc khớp một đường với một biểu đồ phân tán hoặc biểu đồ thời gian trong đó các giá trị dữ liệu bị nhiễu, các điểm dữ liệu thưa thớt hoặc các mối liên hệ yếu kém cản trở khả năng của bạn để nhìn thấy một đường phù hợp nhất.

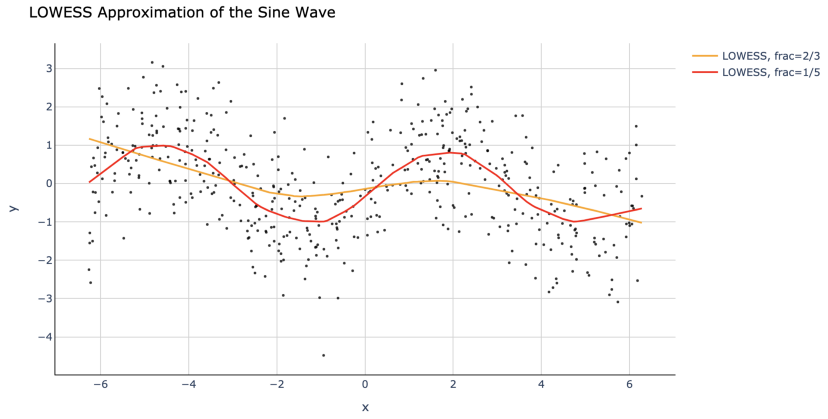
**Làm thế nào để thuật toán Locally Weighted Scatterplot Smoothing có trọng số cục bộ hoạt động?**

LOWESS chia nhỏ dữ liệu để thực hiện hồi quy tuyến tính trên các phần cục bộ của dữ liệu.

Cho biết rằng một bổ sung quan trọng cho **LOWESS** so với OLS là nó áp dụng trọng số như ta có thể đã đoán từ tên của thuật toán – **LOWESS**.

Hàm trọng số cung cấp trọng số lớn nhất cho các điểm dữ liệu gần điểm ước tính nhất và trọng số nhỏ nhất cho các điểm dữ liệu cách xa nhất. Nó dựa trên ý tưởng rằng các điểm gần nhau trong không gian biến giải thích có nhiều khả năng liên quan đến nhau theo cách đơn giản hơn là các điểm xa nhau hơn.

Chúng tôi sẽ sử dụng một ví dụ để minh họa cách hoạt động của LOWESS. Chúng ta hãy bắt đầu bằng cách tạo biểu đồ phân tán trong đó các điểm dữ liệu tuân theo một mô hình sóng sin, nhưng chúng có một số nhiễu ngẫu nhiên được thêm vào, làm cho mô hình ít rõ ràng hơn.

Sau đó, chúng ta thực hiện phân tích hồi quy LOWESS một vài lần bằng cách sử dụng các siêu tham số khác nhau và thêm các đường cong LOWESS vào biểu đồ:

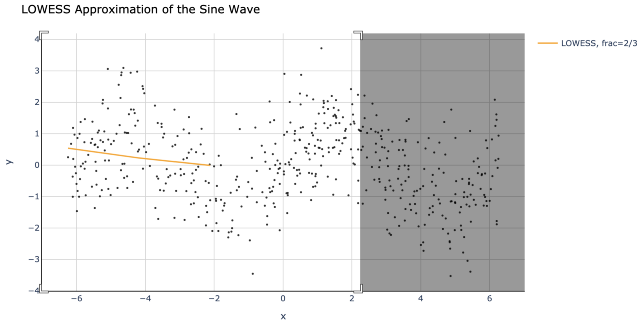
Hình 5 Biểu đồ phân tích hồi quy LOWESS

Như chúng ta có thể thấy, có rất nhiều nhiễu trong dữ liệu, với mối quan hệ giữa x và y là phi tuyến tính (thực tế là tuần hoàn). Rõ ràng, hồi quy tuyến tính đơn giản sẽ không cho chúng ta một kết quả có ý nghĩa ở đây vì nó chỉ có thể tạo ra một đường thẳng (không giống như hồi quy **LOWESS** hoặc đa thức).

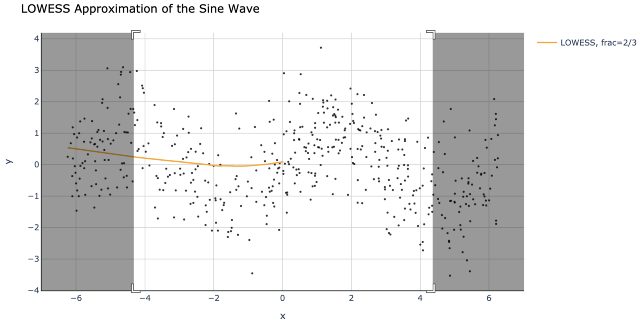
Điều đầu tiên mà **LOWESS** làm là chọn một tập hợp con của dữ liệu. Chúng ta có thể kiểm soát kích thước của các tập hợp con bằng cách chỉ định siêu tham số 'phân số'. Hai đường mà chúng ta vẽ ở trên có giá trị siêu tham số là 2/3 và 1/5.

Hãy lấy đường màu cam làm ví dụ, với giá trị siêu tham số mặc định là 2/3. Điều này có nghĩa là thuật toán **LOWESS** lấy một tập hợp con của hai phần ba toàn bộ dữ liệu và thực hiện hồi quy bình phương tuyến tính nhỏ nhất có trọng số trong khoảng thời gian của dữ liệu đó trong khi di chuyển từng điểm một và lặp lại quá trình tương tự cho đến khi nó trải qua mọi điểm duy nhất.

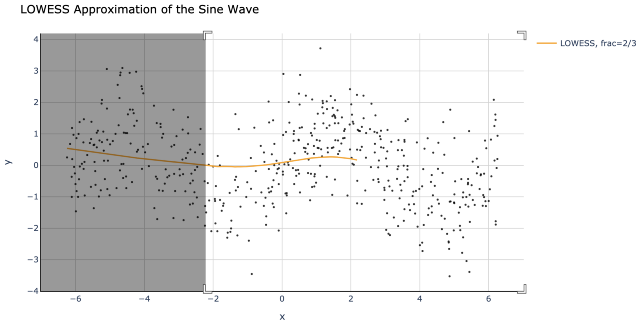
Bắt đầu từ bên trái và đi qua phải từng điểm một, chúng ta có thể thấy đường màu cam vẫn thẳng như thế nào cho đến khi x = -2.1. Điều này là do toàn bộ dữ liệu nằm trong khoảng từ -2π đến + 2π (-6,28 đến +6,28) với 2/3 cửa sổ ban đầu nằm trong khoảng từ -6,28 đến +2,1. Do đó, chỉ khi thuật toán vượt ra ngoài -2,1 (điểm giữa từ -6,28 đến +2,1), nó bắt đầu gặp các điểm dữ liệu mới trong cửa sổ của nó, làm thay đổi độ dốc của đường cong.



Hình 6 LOWESS Approximation of the Sine Wave



Hình 7 LOWESS Approximation of the Sine Wave

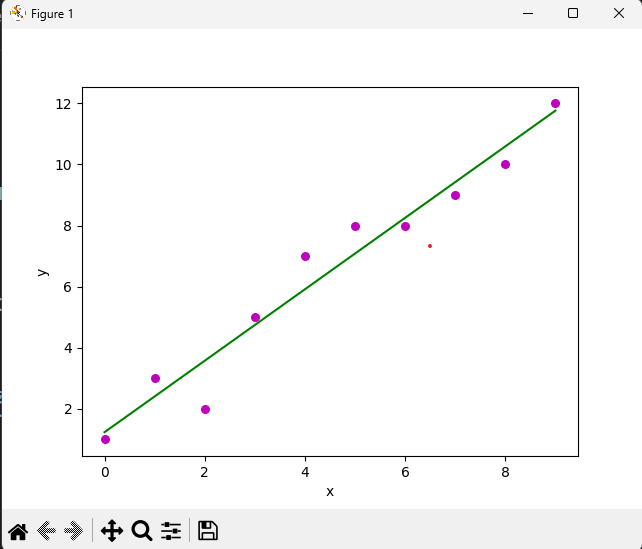


Hình 8 LOWESS Approximation of the Sine Wave

Nếu bây giờ chúng ta nhìn vào đường màu đỏ (frac = 1/5), chúng ta có thể thấy việc thu hẹp cửa sổ xuống 1/5 dữ liệu khiến thuật toán nhạy cảm hơn với các xu hướng cục bộ như thế nào. Hình ảnh gif nổi bật ở đầu câu chuyện này minh họa cách đường màu đỏ này (frac = 1/5) di chuyển trên dữ liệu thực hiện hồi quy tuyến tính có trọng số. Đây là quá trình tương tự như được mô tả cho đường màu cam (frac = 2/3) ngoại trừ, cửa sổ mà thuật toán nhìn thấy hẹp hơn.

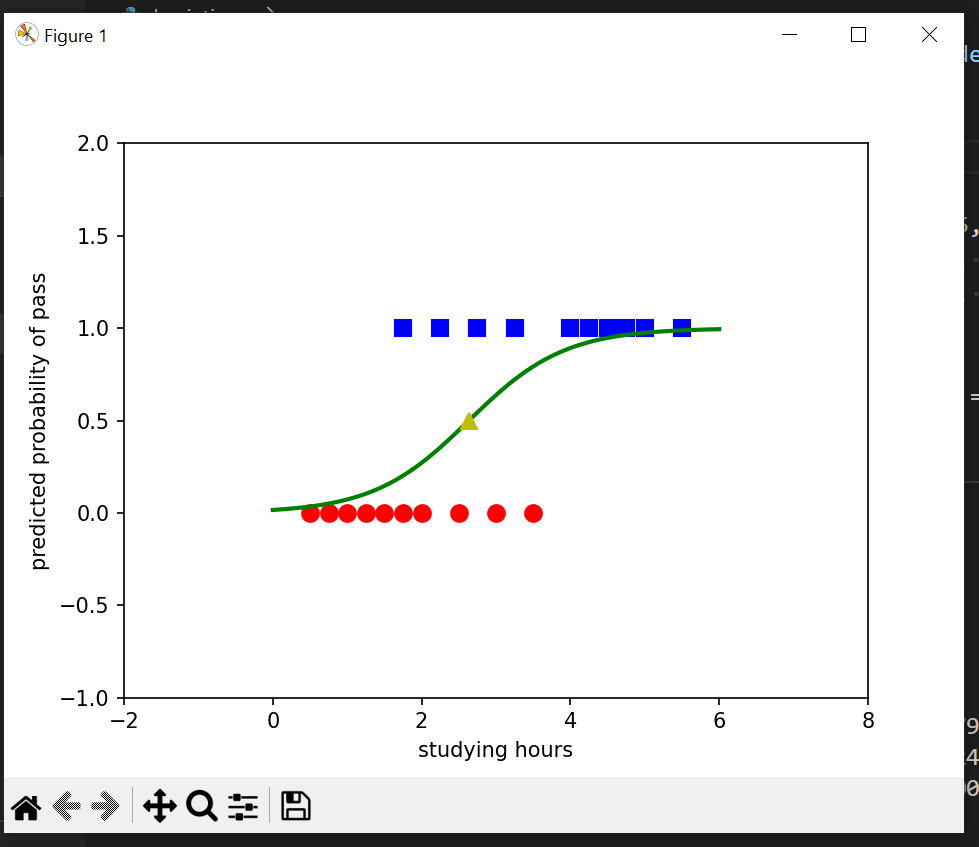
Mặc dù trong ví dụ này, việc làm cho cửa sổ nhỏ hơn đã giúp chúng tôi có được sự xấp xỉ tốt hơn của sóng sin, nhưng không phải lúc nào bạn cũng mong muốn làm cho một siêu thông số 'phân số' nhỏ hơn. Điều này phần lớn phụ thuộc vào dữ liệu bạn đang phân tích, vì việc làm cho một cửa sổ nhỏ hơn sẽ có nguy cơ trang bị quá mức. Bạn sẽ thấy điều này trong ví dụ tiếp theo của tôi, nơi 2/3 hoạt động tốt hơn nhiều so với 1/5.

1. DEMO
   1. Linear Regression



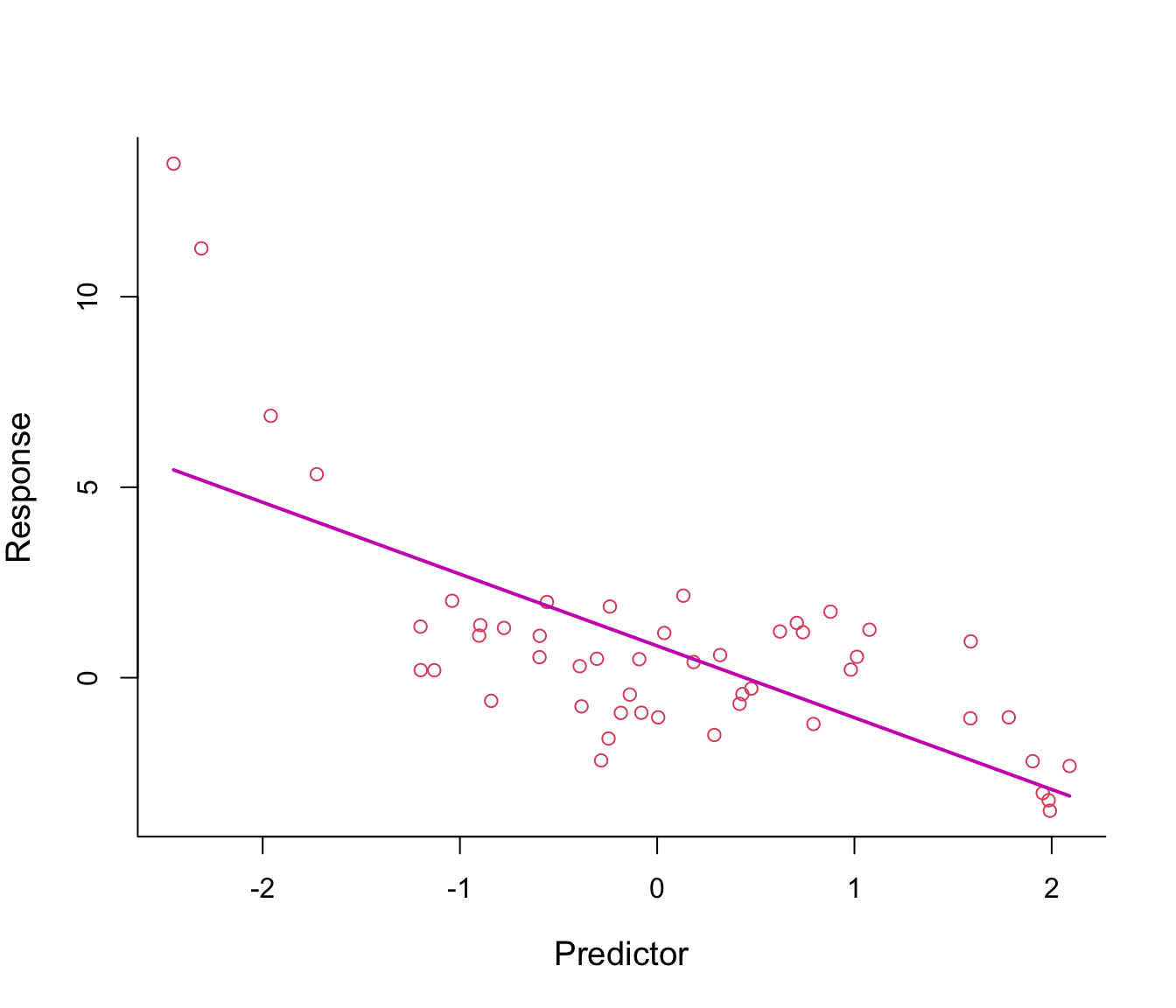
Hình 9 Linear Regression

* 1. Logistic Regression



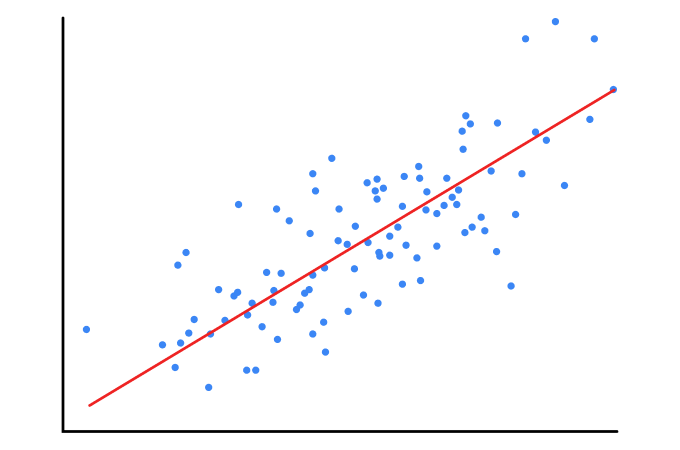
Hình 10 Logistic Regression

* 1. Stepwise Regression



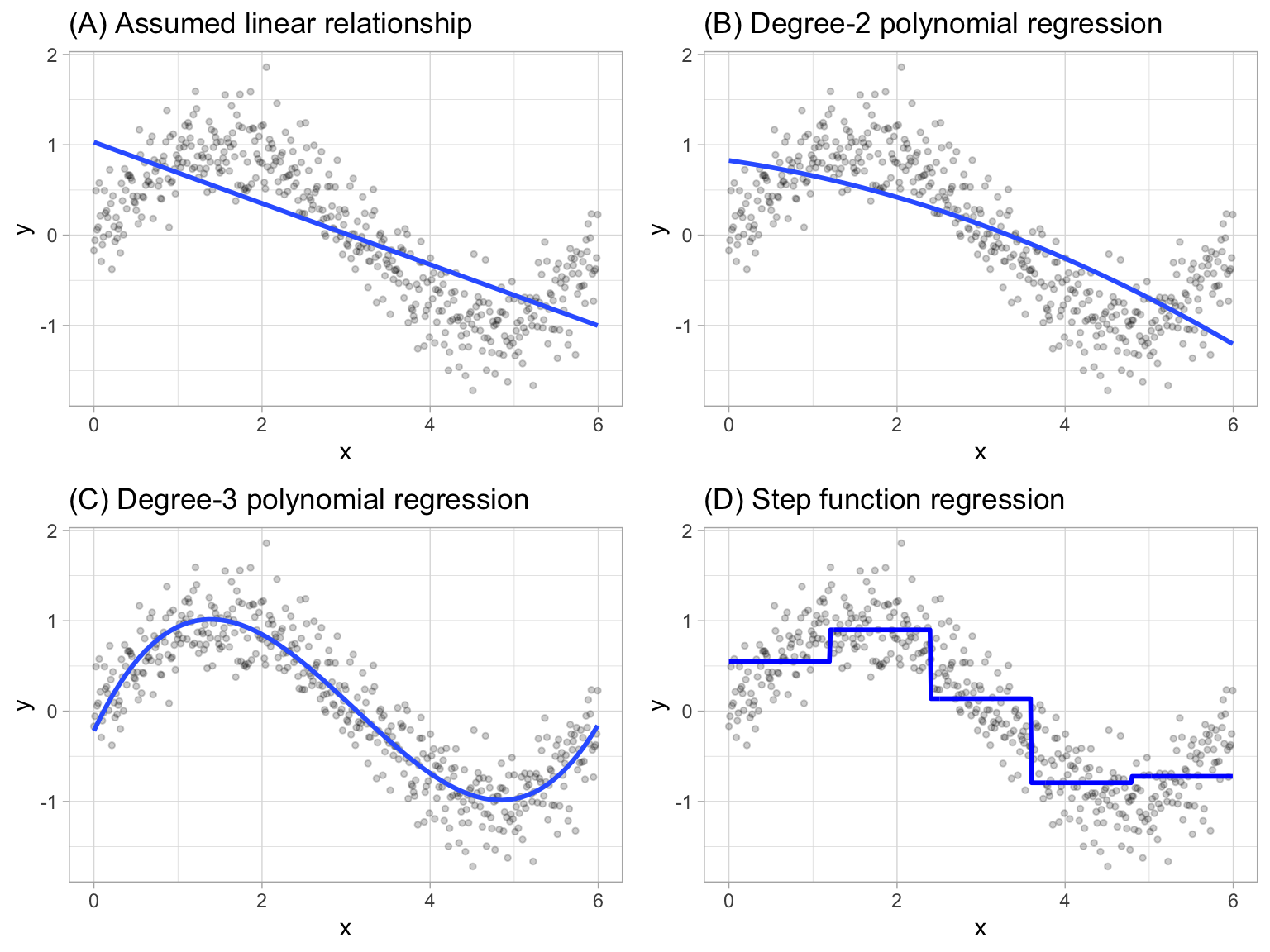
Hình 11 Stepwise Regression

* 1. Ordinary Least Squares Regression (OLSR)



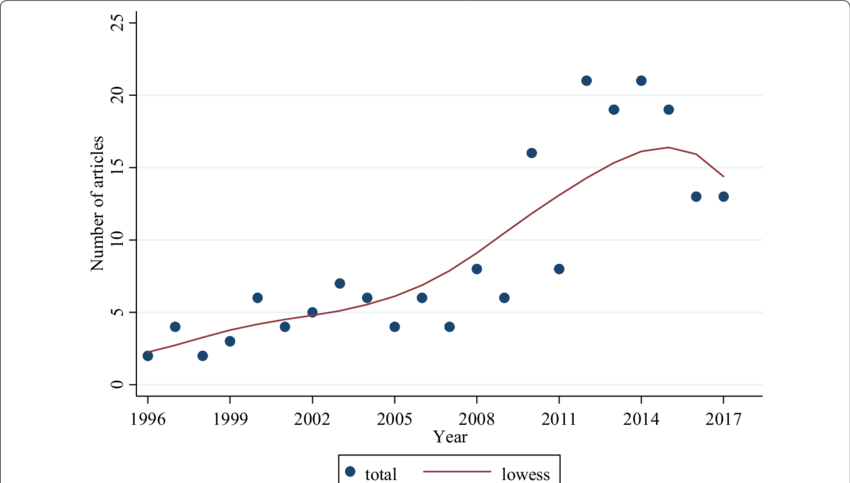
Hình 12 Ordinary Least Squares Regression (OLSR)

* 1. Multivariate Adaptive Regression Splines (MARS)



Hình 13 Multivariate Adaptive Regression Splines (MARS)

* 1. Locally Estimated Scatterplot Smoothing (LOESS)



Hình 14 Locally Estimated Scatterplot Smoothing (LOESS)