**LINEAR REGRESSION**

1. **Giới thiệu**

Trước khi đi sâu vào chi tiết của Linear Regression, bạn có thể tự hỏi tại sao chúng ta lại xem xét thuật toán này. Nó không phải là một kỹ thuật từ số liệu thống kê?

Học máy, cụ thể hơn là lĩnh vực lập mô hình dự đoán, chủ yếu liêu quan đến việc giảm thiểu lỗi của mô hình hoặc đưa ra dự đoán chính xác nhất có thể, với chi phí có thể giải thích được. Trong học máy ứng dụng, chúng ta sẽ mượn, sử dụng lại các thuật toán từ nhiều lĩnh vực khác nhau, bao gồm cả số liệu thống kê và sử dụng chúng cho những mục đích này.

Do đó, Linear Regression được phát triển trong lĩnh vực thống kê và được nghiên cứu như một mô hình để hiểu mối quan hệ giữa các biến số đầu vào và đầu ra. Những đã được học máy mượn, nó vừa là thuật toán thống kê vừa là thuậ toán học máy.

1. **Linear Regression**

**Linear Regression** là một trong những thuật toán cơ bản và phổ biến nhất của **Supervised Learning (Học có giám sát),** trong đó **đầu ra dự đoán là liên tục.** Thuật toán này thích hợp để dự đoán các giá trị đầu ra là các đại lượng liên tục như doanh số hay giá cả thay vì cố gắng phân loại chúng thành các đại lượng rời rạc như màu sắc và chất liệu của quần áo, hay xác định đối tượng trong một bức ảnh là mèo hay chó,...

Ví dụ: một mô hình giả định mối quan hệ tuyến tính giữa các biến đầu vào (x) và biến đầu ra duy nhất (y). Cụ thể hơn, y có thể được tính từ sự kết hợp tuyến tính của các biến đầu vào (x).

Khi có một biến đầu vào duy nhất (x), phương pháp này gọi là **simple linear regression.** Khi có nhiều biến đầu vào, phương pháp này gọi là **multiple linear regression.**

Trong Linear Regression chúng ta sẽ gặp hai loại bài toán đó là Simple Linear Regression và Multiple Linear Regression. Simple Linear Regression chính là mối quan hệ giữa hai biến số liên tục trên trục hoành x và trên trục tung y. Phương trình Simple Linear Regression có dạng như phương trình đường thẳng y = ax + b với x là biến độc lập và y là biến phụ thuộc vào x. Đối với Multiple Linear Regression, bạn có thể hiểu một cách đơn giản là sẽ có nhiều biến độc lập x1, x2,….., xn và có nhiều hệ số a1, a2,…, an thay vì chỉ một biến x duy nhất.

1. **Cách thức hoạt động/ cơ sở tính toán**
2. **Cách thức hoạt động**

Về bản chất, một kỹ thuật Linear Regression đơn giản cố gắng vẽ một đồ thị đường giữa hai biến dữ liệu x và y. Là biến độc lập, x được vẽ theo trục hoành. Các biến độc lập còn được gọi là biến giải thích hoặc biến dự báo. Biến phụ thuộc, y được vẽ trên trục tung. Bạn có thể tham chiếu các giá trị y như các biến phản hồi hoặc các biến dự báo.

Để có cái nhìn tổng quan, hãy xem xét dạng đơn giản nhất của phương trình đồ thị đường giữa y và x:

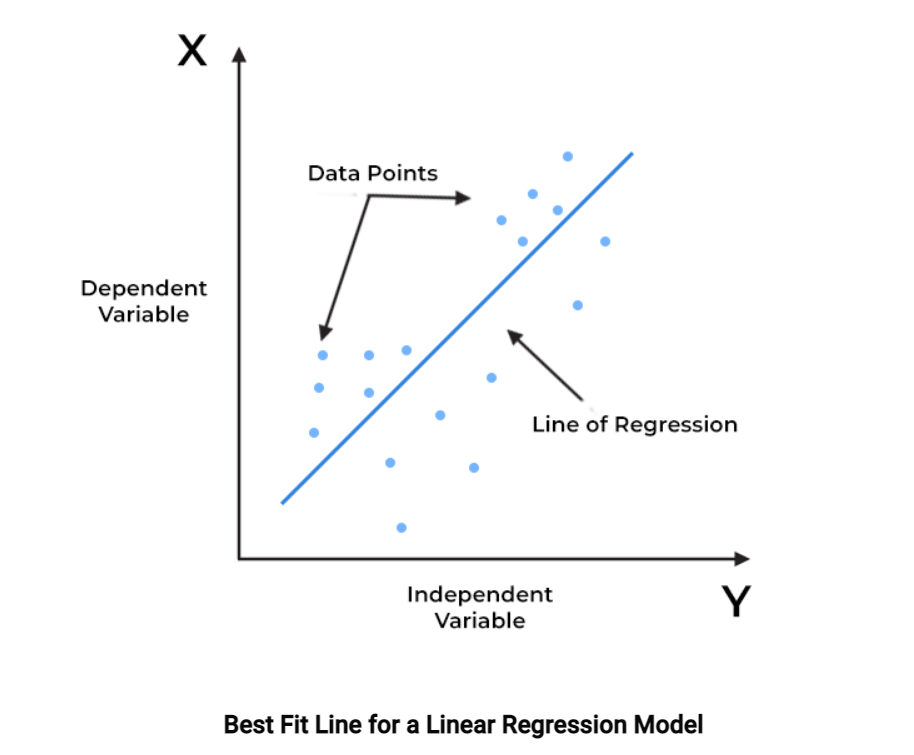
γ = β0 + β1\*x + ε

* **y** là giá trị dự đoán của biến phụ thuộc ( **y** ) đối với bất kỳ giá trị đã cho nào của biến độc lập ( **x** ).
* **β0** là phần chặn, giá trị dự đoán của **y** khi **x** bằng 0.
* **β1** là hệ số hồi quy – chúng ta kỳ vọng **y** sẽ thay đổi bao nhiêu khi **x** tăng.
* **x** là biến độc lập ( biến chúng ta mong đợi sẽ ảnh hưởng đến **y** ).
* **ε** là **sai** số của ước tính hoặc có bao nhiêu biến thể trong ước tính của chúng về hệ số hồi quy.

Linear Regression tìm đường phù hợp nhất thông qua dữ liệu của bạn bằng cách tìm kiếm hệ số hồi quy ( β1 ) giúp giảm thiểu tổng sai số ( ε ) của mô hình.

Ví dụ: Chẳng hạn giả sử rằng tập dữ liệu đầu vào cho (x,y) là (1,5), (2,8) và (3,11). Để xác định phương pháp hồi quy tuyến tính, bạn sẽ thực hiện các bước sau:

1. Vẽ một đường thẳng và đo lường mối tương quan giữa 1 và 5.
2. Tiếp tục thay đổi hướng của đường thẳng cho các giá trị mới (2, 8) và (3, 11) cho đến khi tất cả các giá trị đều phù hợp.
3. Xác định phương trình Linear Regression là y = 3\*x + 2.
4. Ngoại suy hoặc dự đoán y là 14 khi x là 4.



Trong hình trên: + Trục X là biến độc lập.

+ Trục Y là đầu ra / Biến phụ thuộc.

+ Line of Regression là đường phù hợp nhất cho mô hình.

Một đường được vẽ cho các điểm dữ liệu nhất định phù hợp với tất cả các vấn đề. Do đó, nó được gọi là ‘Best fit line’. Mục tiêu của thuật toán Linear Regression là tìm đường phù hợp nhất được thấy trong hình trên.

1. **Cơ sở tính toán**
2. **Mean Squared Error (MSE):**

Linear Regression thường sử dụng lỗi bình phương trung bình (MSE) để tính toán lỗi của mô hình. MSE được tính từ một tập hợp các giá trị x và y:

1. Tìm đường hồi quy.
2. Chèn các giá trị x của bạn vào phương trình Linear Regression để tìm các giá trị y mới ( y’ ).
3. Trừ giá trị Y mới từ giá trị ban đầu để nhận sai số.
4. Bình phương các sai số.
5. Cộng các sai số.
6. Tìm ý nghĩa.

Linear Regression khớp một đường với dữ liệu bằng cách tìm hệ số hồi quy dẫn đến MSE nhỏ nhất. MSE càng thấp, dự báo càng tốt.

MSE = \*

* n là số lượng mặt hàng
* Σ là ký hiệu tổng
* Thực tế là giá trị y ban đầu hoặc được quan sát
* Dự báo là giá trị y từ hồi quy

**Bài toán ví dụ:** Tìm MSE của tập giá trị sau: (43, 41), (44, 45), (45, 49), (46,47), (47, 44).

**Bước 1:** Tìm đường hồi quy. Sử dụng [online calculator](http://www.alcula.com/calculators/statistics/linear-regression/#gsc.tab=0) và nhận được đường hồi quy: y = 9.2 + 0.8\*x.

**Bước 2:** Tìm các giá trị y’ mới:

* 9.2 + 0.8\*43 = 43.6
* 9.2 + 0.8\*44 = 44.4
* 9.2 + 0.8\*45 = 45.2
* 9.2 + 0.8\*46 = 46
* 9.2 + 0.8\*47 = 46.8

**Bước 3:** Tìm lỗi (y – y’):

* 41 – 43.6 = -2.6
* 45 – 44.4 = 0.6
* 49 – 45.2 = 3.8
* 47 – 46 = 1
* 44 – 46.8 = -2.8

**Bước 4:** Square the Errors:

* -2.62 = 6.76
* 0.62 = 0.36
* 3.82 = 14.44
* 12 = 1
* -2.82 = 7.84

Ảnh có chứa bàn

Mô tả được tạo tự động

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Height(X) | Weight(Y) | Estimated(Y’) | Error(Y-Y’) | Error Squared |
| 43 | 41 | 43.6 | -2.6 | 6.76 |
| 44 | 45 | 44.4 | 0.6 | 0.36 |
| 45 | 49 | 45.2 | 3.8 | 14.44 |
| 46 | 47 | 46 | 1 | 1 |
| 47 | 44 | 46.8 | -2.8 | 7.84 |
|  | | | | | | |
| Regression line: y = 9.2 + 0.8x | | | | | |

Bảng kết quả sau khi tính

1. **Mean Square Error (MAE)**

MAE biểu thị sự khác biệt giữa giá trị ban đầu và giá trị dự đoán được trích xuất bằng cách lấy trung bình mức chênh lệch tuyệt đối trên tập dữ liệu.

Các bước thực hiện MAE:

* Tìm tất cả các lỗi tuyệt đối, yi – y.
* Tinhs tổng của tất cả.
* Chia cho số n lỗi. Ví dụ có 10 phép đo, hãy chia cho 10.

MAE =

1. **R Squared (R2)**

R2 biểu thị hệ số về mức độ phù hợp của các giá trị so với các giá trị ban đầu. Giá trị từ 0 đến 1 được hiểu là tỷ lệ phần trăm. Giá trị càng cao thì mô hình càng tốt.

R2 = 1 -

1. **Các tham số/ Siêu tham số của Linear Regression**

Cú pháp

Sklearn.linear\_model.LinearRegression

*class*sklearn.linear\_model.**LinearRegression**(*\**, *fit\_intercept=True*, *copy\_X=True*, *n\_jobs=None*, *positive=False*)

1. **Siêu tham số**

* **Fit\_intercept : bool, default=True**

Có nên tính toán hệ số chặn cho mô hình này không. Nếu được đặt thành False sẽ không có phần chặn nào được sử dụng trong phép tính (tức là dữ liệu dự kiến sẽ được căn giữa).

* **Copy\_X : bool, default=True**

Nếu True, X sẽ được sao chép; nếu không, nó có thể bị ghi đè.

* **n\_jobs : int, default=False**

Số lượng công việc được sử dụng tính toán. Điều này sẽ chỉ cung cấp khả năng tăng tốc trong trường hợp có vấn đề đủ lớn, đó là nếu trước tiên n\_targets > 1 và thứ hai là X thưa thớt hoặc nếu dương được đặt thành True. False có nghĩa là 1 trừ khi trong ngữ cảnh joblib.parallel\_backend -1 có nghĩa là sử dụng tất cả các bộ xử lý.

1. **Tham số**

* **coef\_ : array of shape (n\_features, ) or (n\_targets, n\_features)**

Các hệ số ước lượng cho bài toán Linear Regression. Nếu nhiều mục tiêu được chuyển trong quá trình điều chỉnh (y 2D), thì đây là mảng 2D có hình dạng (n\_targets, n\_features), trong khi nếu chỉ một mục tiêu được chuyển, thì đây là mảng 1D có độ dài n\_features.

* **Rank\_ : int**

Hạng của ma trận x. Chỉ khả dụng khi x dày đặc.

* **Singular\_ : array of shape (min(X, y),)**

Các giá trị riêng của x. Chỉ khả dụng khi x dày đặc.

* **Intercept\_ : float or array of shape (n\_targets,)**

Thuật ngữ độc lập trong mô hình tuyến tính. Đặt thành 0,0 nếu fit\_intercept = False.

* **n\_features\_in\_ : int**

Số lượng các tính năng nhìn thấy trong quá trình phù hợp.

* **Feature\_names\_in\_ : ndarray của hình dạng (n\_features\_in\_,)**

Tên củacác tính năng nhìn thấy trong quá trình phù hợp. Chỉ được xác định khi x có tên tính năng là tất cả các chuỗi.

1. **Tại sao sử dụng Linear Regression**

Linear Regression sử dụng để đưa ra dựa đoán cho một số đầu vào. Nhưng ngoài việc đưa ra dự đoán, Linear Regression còn thực hiện nhiều việc khác bao gồm:

1. Linear Regression linh hoạt hơn và có khả năng ứng dụng rộng rãi.

* Linear Regression cho phép bạn hiểu được mối quan hệ giữa các biến. Bằng cách sử dụng cách phép đo thống kê như R-Squared/ R-Squared đã điều chỉnh, Linear Regression có thể cho biết mô hình giải thích được bao nhiêu phần trăm tổng biến thiên trong dữ liệu.
* Linear Regression cho biết yếu tố dự đoán nào trong mô hình có ý nghĩa thống kê và yếu tố nào không. Ví dụ cung cấp cho mô hình Linear Regression 50 tính năng, bạn có thể tìm ra tính năng nào là yếu tố dự báo tốt cho biến mục tiêu và tính năng nào không.
* Linear Regression có thể đưa ra khoảng tin cậy cho từng hệ số hồi quy mà nó ước lượng.

1. Linear Regression ít **Black box** và dễ giao tiếp. Một mô hình đơn giản có nghĩa là việc truyền đạt cách thức hoạt động của chính mô hình đó và cách diễn giải kết quả của một mô hình sẽ dễ dàng hơn.
2. Linear Regression sẽ giúp hiểu rõ hơn về luận thống kê tổng thể.
3. **Khi nào sử dụng Linear Regression**

Linear Regression cung cấp một tính toán khoa học để xác định và dự đoán kết quả trong tương lai. Khả năng tìm dự đoán và đánh giá chúng có thể giúp mang lại lợi ích cho nhiều doanh nghiệp và cá nhân, chẳng hạn như các hoạt động được tối ưu hóa và tài liệu nghiên cứu chi tiết.

**Ví dụ:** Bạn có thể thu nhập dữ liệu để giúp bạn tối ưu hóa các hoạt động tiếp thị hoặc sản xuất của mình bằng cách sử dụng quy trình này để phân tích mối quan hệ giữa các yếu tối góp phần khác nhau. Tương tự như vậy, bạn cũng có thể có được các tài liệu nghiên cứu dữ liệu chi tiết về mối quan hệ giữa các yếu tố quan trọng và đưa chúng vào các bài thuyết trình của các bên liên quan, kế hoạch cải tiến hoặc tài liệu nghiên cứu điển hình.

Những ngành nghề sử dụng Linear Regression thường xuyên:

* **Đội kinh doanh**. Ví dụ: Bạn có thể sử dụng nó để kiểm tra các mô hình mua hàng hoặc tương tác của người tiêu dùng để dự đoán khi nào sản phẩm hoặc dịch vụ của bạn có thể đạt được nhu cầu cao hơn.
* **Nhà phân tích thể thao**. Ví dụ: Có thể tham khảo quy trình để xác định xem một đội có thể thể hiện tốt như thế nào trong trận đấu tiếp theo hoặc cách các đội chuyên nghiệp trong cùng một môn thể thao xếp hạng với nhau. Hoặc có thể sử dụng để phân tích từng người chơi thể thao và thu nhập thông tin chi tiết về cách chơi cá nhân của họ.
* **Nhà môi giới**. Ví dụ: Mức độ ô nhiễm tác động có thể có có đối với nhiệt độ hoặc những thứ đơn giản hơn như lượng nước mà cây nhận được ảnh hưởng như thế nào đến sự phát triển của chúng. Giúp dự đoán các điều kiện môi trường trong tương lai để thông báo cho các chuyên gia về tính bền vững về những hoạt động hiện tại có thể cần điều chỉnh.
* **Nhà phân tích tài chính**. Ví dụ: Các mô hình định giá tài sản, vốn giúp các chuyên gia tài chính xác định mối quan hệ giữa lợi nhuận tài sản dự kiến và rủi ro thị trường liên quan.Việc sử dụng mô hình Linear Regression để đánh giá các dự án và kết quả tài chính giúp các chuyên gia này hỗ trợ khả năng sinh lời của công ty họ.

1. **Ưu điểm, hạn chế và tối ưu**
2. **Ưu điểm**

* **Dễ dàng thực hiện:** Mô hình Linear Regression dễ thực hiện về mặt tính toán vì nó không đòi hỏi nhiều chi phí kỹ thuật, kể cả trước khi khởi chạy mô hình cũng như trong quá trình bảo trì mô hình.
* **Khả năng diễn giải:** Không giống như các mô hình **deep learning** khác (neural networks), Linear Regression tính tương đối đơn giản. Kết quả là thuật toán này vượt trội so với các mô hình black-box thiếu sót trong việc chứng minh biến đầu vào nào khiến biến đầu ra thay đổi.
* **Khả năng mở rộng:** Linear Regression không nặng về mặt tính toán và do đó, rất phù hợp trong các trường hợp cần mở rộng quy mô. Ví dụ: mô hình có thể mở rộng quy mô tốt liên quan đến khối lượng dữ liệu tăng lên (big data).
* **Tối ưu cho cài đặt trực tuyến:** Tính dễ tính toán của các thuật toán này cho phép chúng được sử dụng trong cài đặt trực tuyến. Mô hình có thể được đào tạo và đào tạo lại với mỗi ví dụ mới để tự dự đoán trong thời gian thực, không giống như **neural networks** hoặc **SVM** nặng tính toán và yêu cầu nhiều tài nguyên máy tính cũng như thời gian đợi đáng kể để đào tạo lại tập dữ liệu mới. Tất cả những yếu tố này làm cho các mô hình tính toán chuyên sâu như vậy trở nên đắt đỏ và không phù hợp với các ứng dụng thời gian thực.

Các tính năng trên làm nổi bật lý do tại sao Linear Regression là một mô hình phổ biến để giải quyết các vấn đề về Machine learning trong thực tế.

1. **Hạn chế**

Hạn chế đầu tiên của Linear Regression là nó rất **nhạy cảm với nhiễu** (sensitive to noise). Vì vậy, trước khi thực hiện Linear Regression, các nhiễu (outlier) cần phải được loại bỏ. Bước này được gọi là tiền xử lý (pre-processing).

Hạn chế thứ hai của Linear Regression là nó **không biễu diễn được các mô hình phức tạp**. Mặc dù chúng ta thấy rằng phương pháp này có thể được áp dụng nếu quan hệ giữa outcome và input không nhất thiết phải là tuyến tính, nhưng mối quan hệ này vẫn đơn giản nhiều so với các mô hình thực tế.

1. **Tối ưu**

Linear Regression là một mô hình đơn giản, lời giải cho phương trình đạo hàm bằng 0 cũng khá đơn giản. Trong hầu hết các trường hợp, chúng ta không thể giải được phương trình đạo hàm bằng 0.

Nhưng có một điều chúng ta nên nhớ, **còn tính được đạo hàm là còn có hy vọng.**

1. **Thực nghiệm**

**Source Code:** [**https://colab.research.google.com/drive/1dsozLwdpCyYXYTQ\_UZk0OcIMTlLgRacP?usp=sharing**](https://colab.research.google.com/drive/1dsozLwdpCyYXYTQ_UZk0OcIMTlLgRacP?usp=sharing)

1. **Bộ dữ liệu sử dụng**
2. **Cách thực hiện**
3. **Kết quả thu được**
4. **Tài liệu tham khảo**

<https://www.indeed.com/career-advice/career-development/when-to-use-linear-regression#:~:text=Understanding%20linear%20regression%20is%20important,operations%20and%20detailed%20research%20materials>.

<https://towardsdatascience.com/3-reasons-why-you-should-use-linear-regression-models-instead-of-neural-networks-16820319d644>

<https://www.javatpoint.com/linear-regression-in-machine-learning>

<https://mlu-explain.github.io/linear-regression/>

<https://machinelearningmastery.com/linear-regression-for-machine-learning/>

<https://www.geeksforgeeks.org/ml-linear-regression/>

quyền","status":true,"is\_default":true,"role\_type\_id":1,"permission":[{"id":1,"name":"View Data","nameVi":"Xem Dữ Liệu","description":"view","status":true,"code":"view\_data"},{"id":2,"name":"Send Messages","nameVi":"Gửi tin nhắn","description":"send messages","status":true,"code":"send\_messages"}]},"user":{"id":1,"fullname":"Owner","gender":true,"birthday":null,"select\_username":"email","username":"owner","email":"owner@gmail.com","phone":"0932221911","address":"123 NTT","is\_active":true,"activeDay":"2023-02-19 15:18:44","codeconfirm":null,"code\_otp\_confirm":null,"avatar":null,"connection\_image":null,"job":null,"introduce\_yourself":null,"language":"vi","ward\_id":null,"create\_by":null,"role\_id":1,"organization\_id":1,"connection\_count":null,"connection\_count\_default":null,"reason\_register":null,"is\_removed":false}}

Dashboard