ĐẠI HỌC QUỐC GIA HÀ NỘI
Trường Đại học Công nghệ

Báo cáo bài tập lớn
Machine learning:
Linear Regression and Logistic Regression

Môn học: Trí tuệ nhân tạo
Giảng viên: TS. Trần Hồng Việt

Nguyễn Cao Bảo Sơn	21021362
Nguyễn Khánh Sơn	21021363
Trần Công Sơn	21021364
Trần Đức Tài	21021365
Mai Văn Thái	21021366

Hà Nội - 2024

Mục lục

Danh mục hình ảnh

Danh mục phương trình

Danh mục công thức

Lời mở đầu

Trong lĩnh vực học máy, **Hồi quy tuyến tính** (Linear Regression) và **Hồi quy logistic** (Logistic Regression) là hai thuật toán phổ biến được sử dụng để giải quyết các bài toán học có giám sát. Cả hai mô hình đều có mục tiêu dự đoán giá trị đầu ra dựa trên một tập dữ liệu đầu vào, nhưng chúng khác nhau về bản chất của bài toán và cách thức thực hiện dự đoán.

Báo cáo sẽ đưa ra những khái niệm, cơ sở lý thuyết, nguyên tắc hoạt động, nhận xét mô hình và một vài ứng dụng Code trên ngôn ngữ Python của hai thuật toán.

Đường dẫn Github: sonchuckye/Machine-learning-Linear-and-Logistic-Regression: Machine learning: Linear and Logistic Regression (github.com)

I - Linear regression

1. Khái niệm

Linear regression hay còn gọi là phương pháp hồi quy tuyến tính là một loại thuật toán học máy có giám sát (***Supervised learning***) để mô tả mối quan hệ tuyến tính giữa biến mục tiêu và các biến độc lập bằng cách khớp phương trình tuyến tính với dữ liệu cho sẵn.

Tuyến tính (Linear) có thể được hiểu là thẳng, phẳng. Trong không gian hai chiều, hàm số được gọi là tuyến tính khi nó có đồ thị dạng đường thẳng. Trong không gian ba chiều, hàm số được gọi là tuyến tính khi nó có đồ thị dạng mặt phẳng. Trong không gian lớn hơn ba chiều, khái niệm siêu mặt phẳng (hyperplane) được sử dụng.

Linear regression có hai loại là Simple linear regression (hồi quy tuyến tính một biến, đầu vào có một thuộc tính) và Multiple linear regression (hồi quy tuyến tính đa biến, đầu vào có nhiều thuộc tính). Biến mục tiêu trong mô hình hồi quy tuyến tính thường có giá trị liên tục, chẳng hạn như GDP của một quốc gia hay giá của một ngôi nhà.

2. Simple linear regression

Mối tương quan giữa giá trị của biến mục tiêu và một biến độc lập có thể được ước lượng thông qua một đường thẳng phù hợp.

Phương trình 1. Hồi quy tuyến tính một biến độc lập

$$y = m * x + b$$

Trong đó:

- y: biến mục tiêu (label/target variable)
- x: biến độc lập (single feature)
- m: độ dốc, thể hiện độ thay đổi của y khi x thay đổi (slope, coeficients)
- b: đại diện cho giá trị dự đoán của biến mục tiêu khi giá trị của tất cả biến độc lập bằng 0 (intercept)

Với dữ liệu đầu vào đã được gán nhãn, mô hình sẽ tính toán các hệ số m và b sao cho hàm mất mát (***loss function/cost function)*** được tối ưu.

Hàm mất mát của linear regression có nhiều cách định nghĩa. Trong đó, nó có thể được định nghĩa từ sai số bình phương trung bình (MSE – Mean Square Error) hay sai số bình phương trung bình gốc (RMSE – Root Mean Square Error).

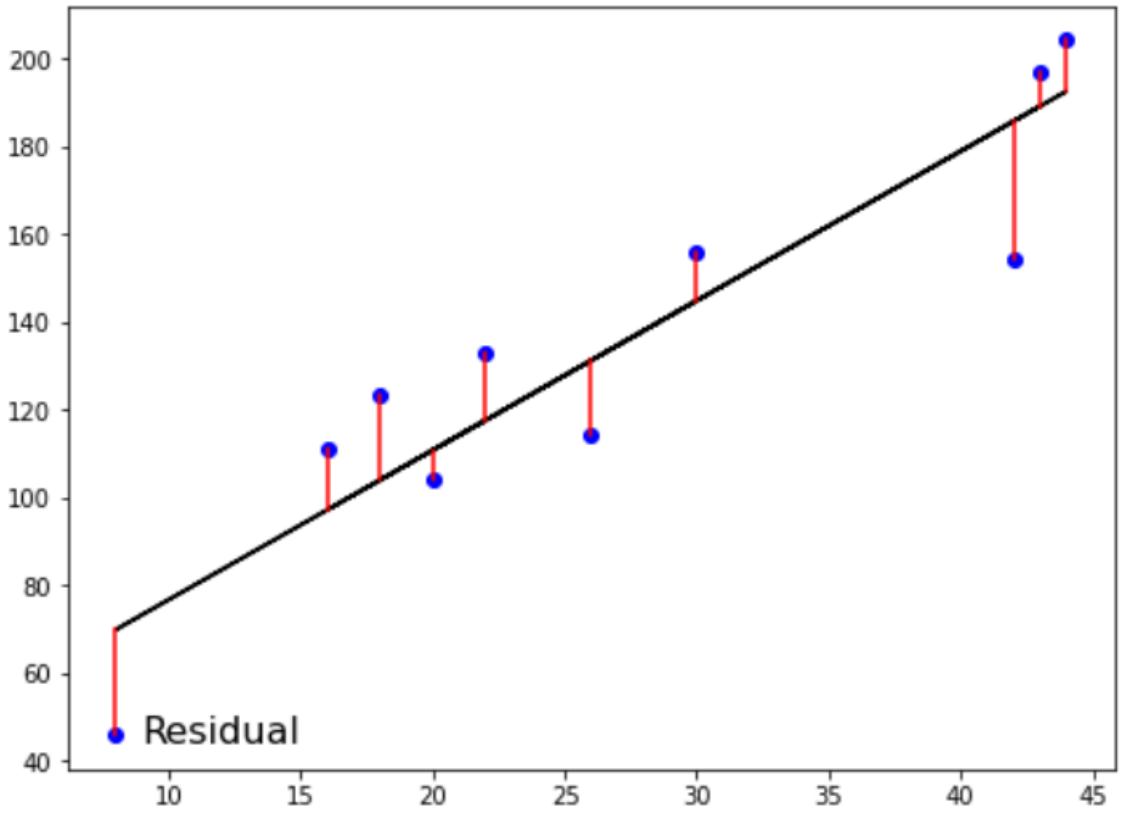
Công thức 1. Mean Square Error

$$MSE = \sum_{i=1}^{N} (\hat{y}_i - \mathrm{y}_i)^2$$

$\hat{y}_i - \mathrm{y}_i$ là sai lệch giữa giá trị dự đoán và giá trị thực của các điểm được lấy mẫu.

Công thức 2. Root Mean Square Error

$$RMSE = \sqrt{MSE}$$



Hình 1. Sai lệch giữa giá trị thực và giá trị dự đoán.

Sai lệch giữa giá trị thực và giá trị dự đoán gọi là phần dư (Residual). MSE là tổng bình phương phần dư các điểm dữ liệu được lấy mẫu.

Hàm mất mát có thể được tối ưu bằng phương pháp bình phương tối thiểu (Least Squares Method). Hệ số m và b có thể tính được từ hệ phương trình sau:

$$\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} x_i * m + N * b$$

$$\sum_{i=1}^{N} x_i * y_i = \sum_{i=1}^{N} {x_i}^2 * m + \sum_{i=1}^{N} x_i * b$$

3. Multiple linear regression

Khi dự đoán từ hai thuộc tính trở lên, mối quan hệ giữa đầu ra và các đại lượng đầu vào được mô tả bằng phương trình sau:

Phương trình 2. Hồi quy tuyến tính nhiều biến

$$y = m_1 * x_1 + m_2 * x_2 + \ldots\ldots + m_n * x_n + b$$

Viết lại các biến x dưới dạng vector hàng **X** và các hệ số m dưới dạng vector cột $\mathbf{M^T}$, ta có: y = $\mathbf{XM^T}$

Tương tự như hồi quy tuyến tính một biến, hồi quy tuyến tính đa biến cần tìm vector hệ số $\mathbf{M^T}$ để tối ưu hàm mất mát. Điểm tối ưu hệ số $\mathbf{M^T}$ cho bài toán Linear Regression có dạng:

$$M^T = (X^T X)^I X^T y$$

X^T: là chuyển vị của X

$(X^T X)^I$: là giả nghịch đảo của $X^T X$

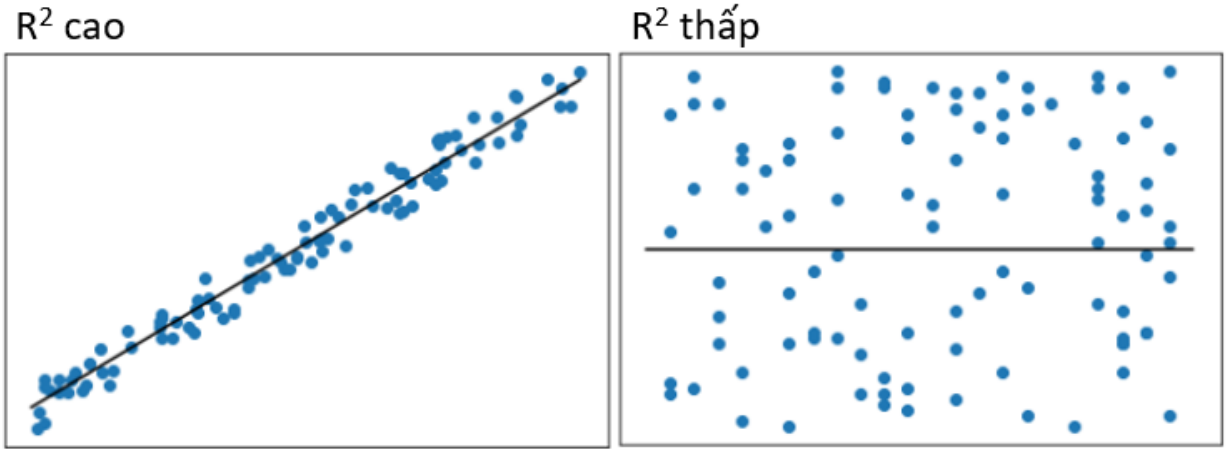
4. R-squared

R-squared cho ra định lượng về phương sai các giá trị mục tiêu, có thể nhận giá trị trong khoảng từ 0 đến 1. Nói cách khác, nó cho biết các dự đoán của mô hình khớp với thực tế như thế nào.

$$R^2 = 1 - \frac{RSS}{TSS}$$

Trong đó:

- RSS là tổng bình phương phần dư
- TSS = MSE



Hình 2. Đồ thị minh họa cho R-squared cao và thấp

5. Giả định về hồi quy tuyến tính

a. Hồi quy tuyến tính một biến

Độ tuyến tính: Các biến độc lập và phụ thuộc có một quan hệ tuyến tính với nhau. Điều này nghĩa là khi biến mục tiêu thay đổi sẽ tuân theo những thay đổi trong các biến độc lập theo kiểu tuyến tính. Điều này nghĩa là phải có một đường thẳng vẽ qua các điểm dữ liệu.

Tính độc lập: Các giám sát trong tập dữ liệu phải độc lập với nhau. Điều này nghĩa là các giá trị của biến mục tiêu tương ứng với biến độc lập dưới cùng một giám sát. Cũng tức là sai số ở mỗi điểm dữ liệu độc lập với nhau.

Đồng phương sai: Giả định này yêu cầu phương sai của sai số phải đồng nhất ở tất cả các giá trị của x. Điều này chỉ ra rằng số lượng của các biến độc lập không tác động đến phương sai của sai số.

Phân phối chuẩn của sai số: Phần dư phải được phân phối chuẩn (standard normal distribution). Điều này nghĩa là phần dư phải tuân theo một đường cong lên.

b. Hồi quy tuyến tính đa biến

Áp dụng cả 4 giả định của hồi quy tuyến tính một biến

Không có hiện tượng đa cộng tuyến: Có rất ít hoặc không có mối quan hệ giữa các biến độc lập. Đa cộng tuyến xảy ra khi hai hoặc nhiều biến có mối tương quan cao với nhau. Được phát hiện qua hai kỹ thuật: Ma trận tương quan và VIF(hệ số phóng đại phương sai)

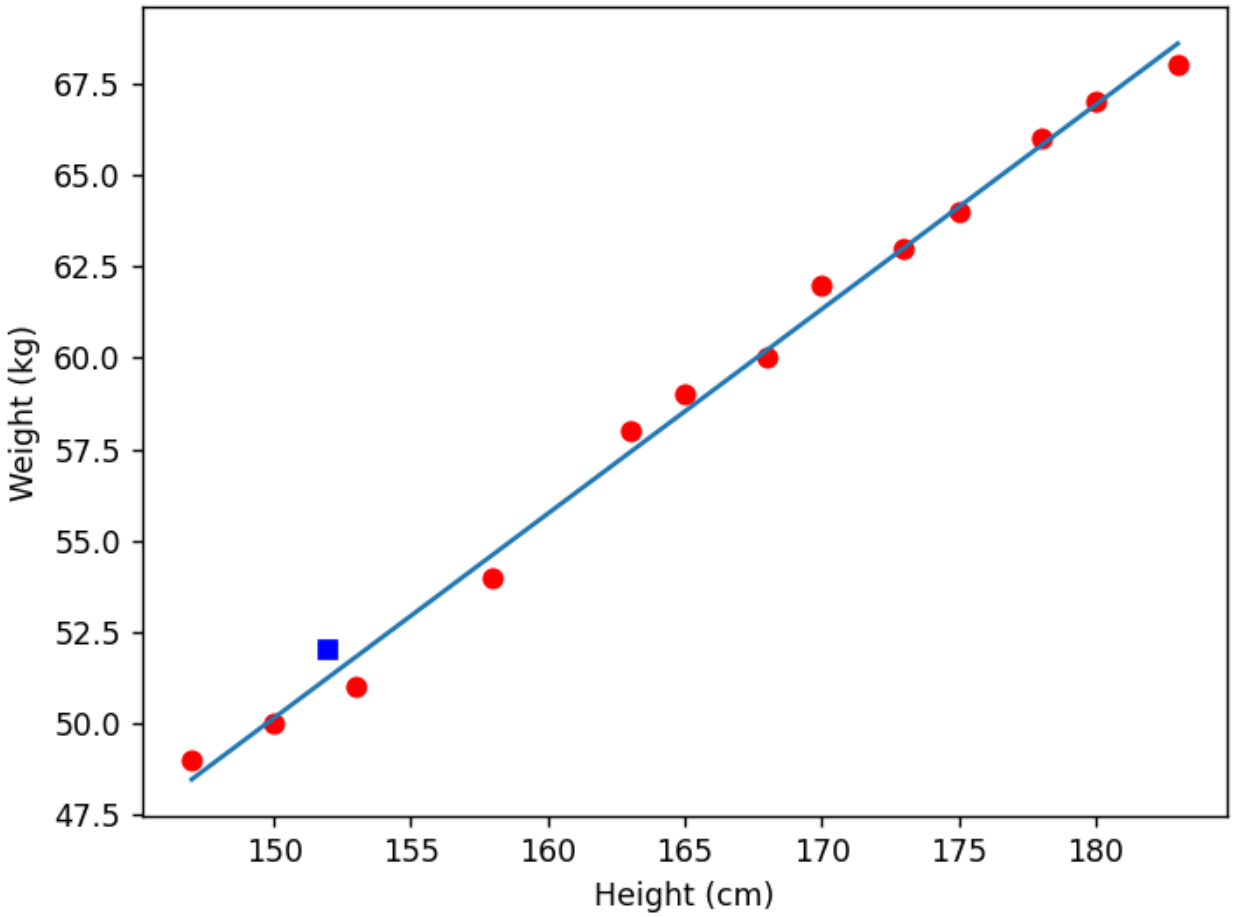
Tính cộng: Mô hình giả định rằng tác động của những thay đổi trong biến dự đoán đối biến phản hồi là nhất quán bất kể giá trị của các biến khác. Nghĩa là không có tương tác giữa các biến trong các tác động của chúng lên biến mục tiêu.

Lựa chọn đặc trưng (biến độc lập): Phải lựa chọn cẩn thận các biến độc lập mà ta sẽ đưa vào mô hình. Trong đó có cả các biến không liên quan hoặc dư thừa, chúng có thể dẫn đến hiện tượng chính xác quá mức hoặc phức tạp hóa mô hình.

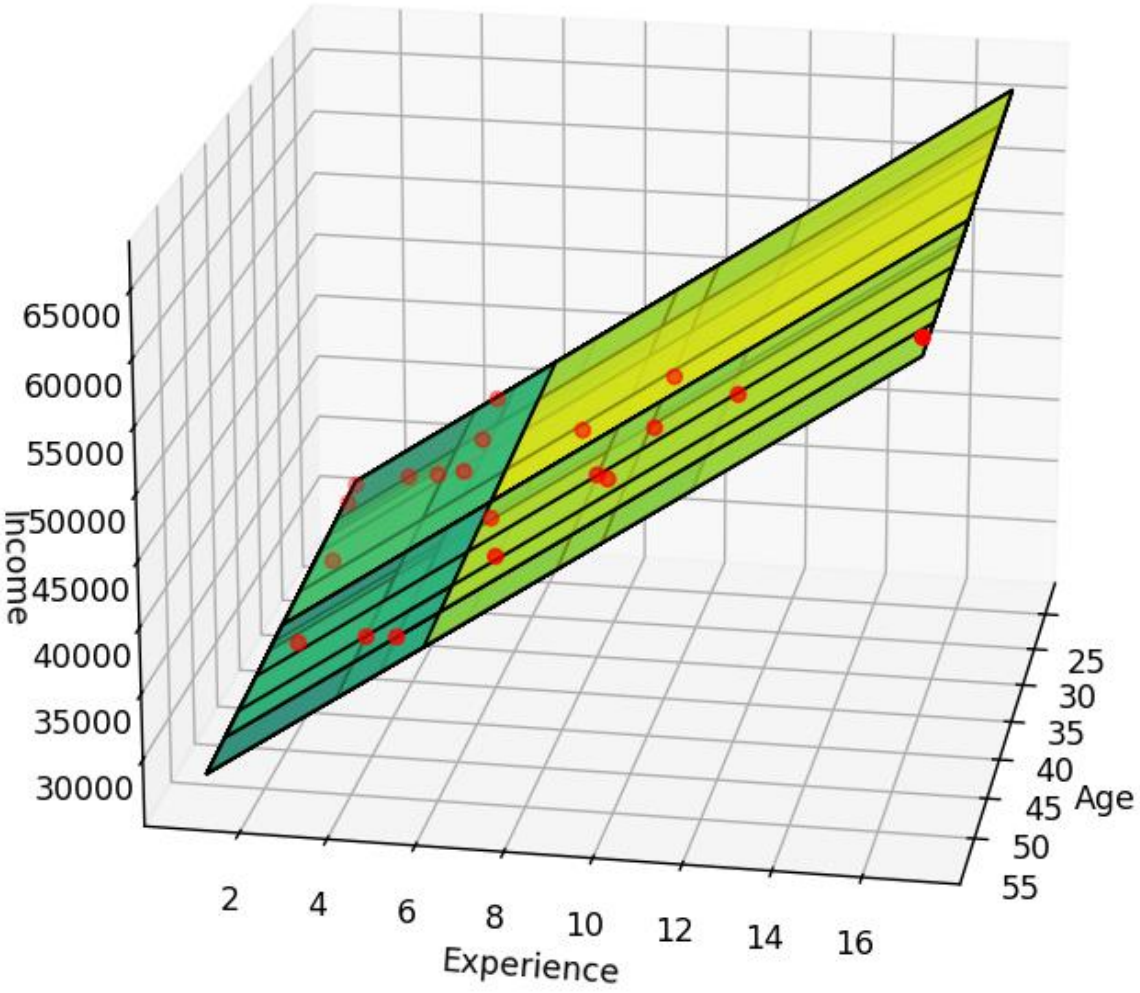
Quá chính xác: Xảy ra khi mô hình quá khớp với dữ liệu huấn luyện, sai số và các biến động ngẫu nhiên không thể hiện mối quan hệ thực sự giữa các biến.

6. Linear Regression với thư viện scikit-learn

Thư viện scikit-learn của Python xây dựng mô hình hồi quy tuyến tính với các lý thuyết được nêu ra phía trên. Trong đường dẫn Github, Linear Regression được dùng để dự đoán cân nặng dựa trên chiều cao và thu nhập dựa trên tuổi và kinh nghiệm làm việc. Độ chính xác đánh giá qua R-squared lên đến 97%.

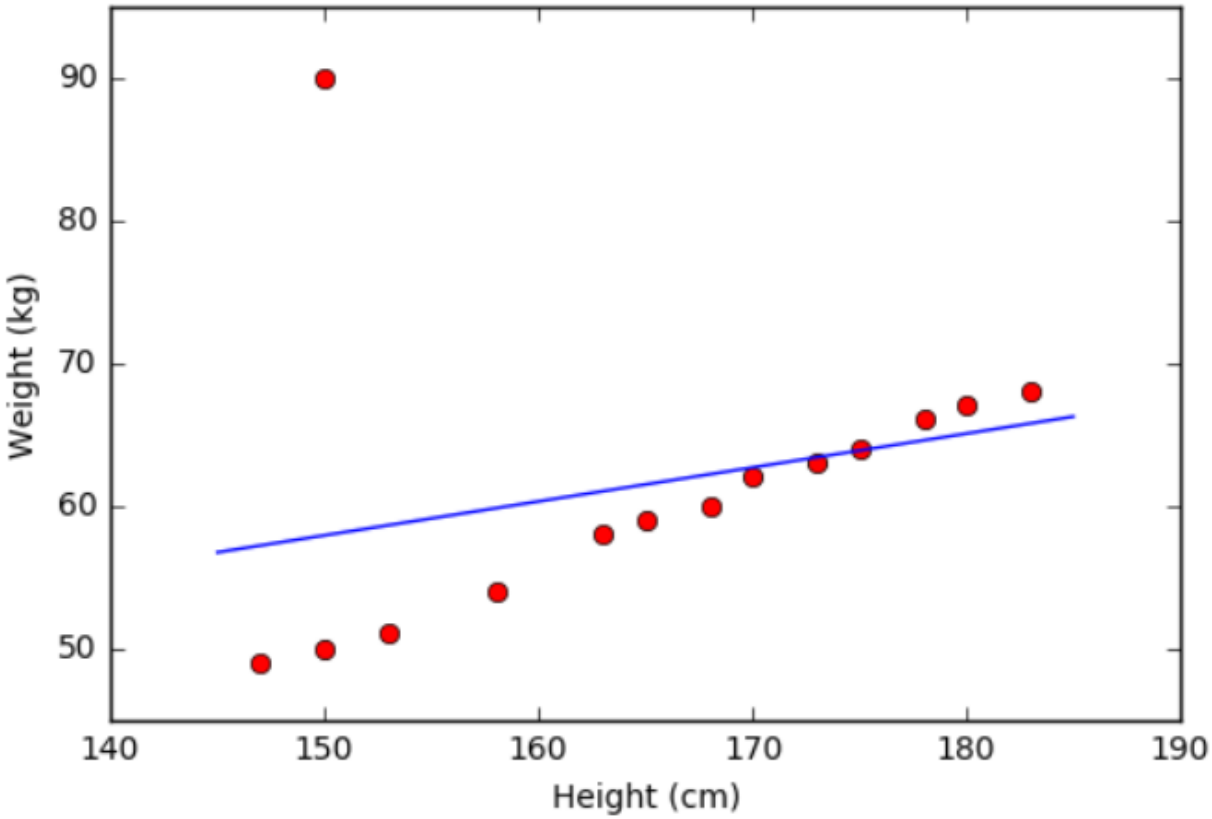


Hình 3. Hồi quy tuyến tính dự đoán cân nặng dựa trên chiều cao



Hình 4. Hồi quy tuyến tính dự đoán thu nhập dựa trên tuổi và kinh nghiệm làm việc

Nhược điểm của thuật toán này đó là mối tương quan giữa dữ liệu đầu vào và đầu ra không tuyến tính thì sai số dự đoán có thể lớn. Khi đó, các mô hình phù hợp hơn như hồi quy phi tuyến có thể được sử dụng. Một nhược điểm khác là mô hình dễ bị ảnh hưởng bởi nhiễu. Nhiễu có thể làm sai lệch mô hình dự đoán đi đáng kể. Để giải quyết vấn đề này, có thể sử dụng một số thuật toán lọc trong bước tiền xử lý dữ liệu như Moving Average, Median Filter,… Với Median Filter, giá trị nhiễu sẽ được thay thế bằng trung vị những giá trị lân cận.



Hình 5. Minh họa ảnh hưởng của dữ liệu nhiễu với Linear Regression

II – Logistic Regression

1. Khái niệm

Logistic regression (Hồi quy logistic) là một mô hình học máy có giám sát được sử dụng để phân loại nhị phân (Binary Classification), trong đó sử dụng hàm sigmoid, lấy đầu vào dưới dạng các biến độc lập và tạo ra giá trị xác suất trong khoảng từ 0 đến 1, dự đoán khả năng biến mục tiêu nhận giá trị là 1.

Đầu ra (biến mục tiêu) được dự đoán của logistic regression thường được viết chung dưới dạng hàm Sigmoid:

$$f(x) = \Theta(w^T x)$$

Đây là một hàm số liên tục nhận giá trị thực, bị chặn trong khoảng (0, 1), đại diện cho xác suất giá trị dự đoán là 0 hoặc 1. Nếu coi điểm có tung độ là 0.5 làm điểm phân chia thì những điểm càng xa về phía bên trái sẽ có xác suất cao nhận giá trị 0. Ngược lại, càng xa điểm này về phía bên phải có xác suất cao nhận giá trị 1.

$$f(s) = \frac{1}{1+e^{-s}} \triangleq \sigma(s)$$

Đây là hàm Sigmoid được sử dụng nhiều nhất trong hồi quy Logistic vì nó bị chặn trong khoảng (0, 1) và giới hạn khi tiến tới âm vô cùng là 0, giới hạn khi tiến tới dương vô cùng là 1. Khi đó, xác xuất để biến mục tiêu y nhận giá trị là 1 và 0 là như sau:

$$P(y = 1) = \sigma(z)$$
$$P(y = 0) = 1 - \sigma(z)$$

Ta có tỷ lệ (Odds):

$$\frac{P(x)}{1-P(x)}=e^{z}$$

Phương trình 3. Hồi quy Logistic

$$P(X;b,w) = \frac{1}{1+e^{-w^{T}X+b}}$$

Trong đó:

- P(X; b, w) là xác suất biến để điểm dữ liệu X rơi vào class 1
- X là vector cột các giá trị biến độc lập (feature)
- w^{T} là vector hàng hệ số (coefficient)
- b là đại diện cho giá trị dự đoán của biến mục tiêu khi giá trị của tất cả biến độc lập bằng 0 (intercept)

Mô hình hồi quy tuyến tính tìm vector hệ số w và b để tối ưu hàm mất mát dựa trên các điểm dữ liệu được lấy mẫu. Quy trình tối ưu hàm mất mát ở đây là bài toán tìm giá trị nhỏ nhất cho phương trình:

$$J(w) = -\log P(y|X;w,b) = -\sum_{i=1}^{N}(y_i \log z_i + (1-y_i)\log(1-z_i))$$

2. Các loại hồi quy logistic

Logistic Regression cũng có thể được sử dụng cho đầu vào một biến hoặc nhiều biến. Hồi quy logistic có ba loại chính: nhị phân, đa phân loại và thứ bậc.

Hồi quy logistic nhị phân: Xử lý hai giá trị có thể xảy ra, tức là: có hoặc không. Chỉ có hai kết quả có thể xảy ra. Khái niệm này thường được biểu diễn dưới dạng 0 hoặc 1 trong mã hóa. Ví dụ bao gồm:

- Liệu có cho vay cho một khách hàng ngân hàng không (kết quả là có hoặc không).
- Đánh giá nguy cơ ung thư (kết quả là cao hoặc thấp).
- Đội bóng có thắng trận vào ngày mai không (kết quả là có hoặc không).

Hồi quy logistic đa thức: Là một mô hình có nhiều lớp mà một mục có thể được phân loại vào. Có một tập hợp các lớp đã được xác định trước với ba hoặc nhiều hơn trước khi chạy mô hình. Ví dụ bao gồm:

- Phân loại văn bản theo ngôn ngữ chúng đến từ.
- Dự đoán liệu một học sinh sẽ đi học đại học, trường nghề hay vào lực lượng lao động.
- Mèo của bạn có thích thức ăn ướt, thức ăn khô hay thức ăn của con người không?

Hồi quy logistic thứ tự: Hồi quy logistic thông thường cũng là một mô hình trong đó có nhiều lớp mà một mục có thể được phân loại thành; tuy nhiên, trong trường hợp này cần phải có thứ tự các lớp. Các lớp học không cần phải tương xứng. Khoảng cách giữa mỗi lớp có thể khác nhau. Ví dụ bao gồm:

- Xếp hạng nhà hàng trên thang điểm từ 0 đến 5 sao.
- Dự đoán kết quả trên bục giảng của một sự kiện Olympic.
- Đánh giá sự lựa chọn của ứng viên, cụ thể là ở những nơi áp dụng cách bầu chọn có sắp xếp theo thứ tự.

3. Giả định về hồi quy logistic

Quan sát độc lập (Independent observations):

- Giả định: Mỗi quan sát là độc lập với các quan sát khác, nghĩa là không có sự tương quan giữa bất kỳ biến đầu vào nào.
- Ví dụ: Khi nghiên cứu về khả năng mắc bệnh tiểu đường, các quan sát về từng cá nhân là độc lập, không có sự tương tác hoặc ảnh hưởng lẫn nhau giữa các cá nhân này.

Biến mục tiêu nhị phân (Binary dependent variables):

- Giả định: Biến mục tiêu phải là nhị phân hoặc phân đôi, nghĩa là nó chỉ có thể nhận hai giá trị. Đối với các biến có hơn hai giá trị, hàm SoftMax sẽ được sử dụng.
- Ví dụ: Trong mô hình dự đoán bệnh tiểu đường, biến mục tiêu có thể là mắc bệnh tiểu đường (1) hoặc không mắc bệnh tiểu đường (0).

Mối quan hệ tuyến tính giữa các biến độc lập và log-odds (Linearity relationship between independent variables and log odds):

- Giả định: Mối quan hệ giữa các biến độc lập và log-odds của biến mục tiêu nên là tuyến tính.
- Ví dụ: Nếu mô hình dự đoán bệnh tiểu đường sử dụng BMI là biến độc lập, thì log-odds của việc mắc bệnh tiểu đường nên có mối quan hệ tuyến tính với chỉ số BMI.

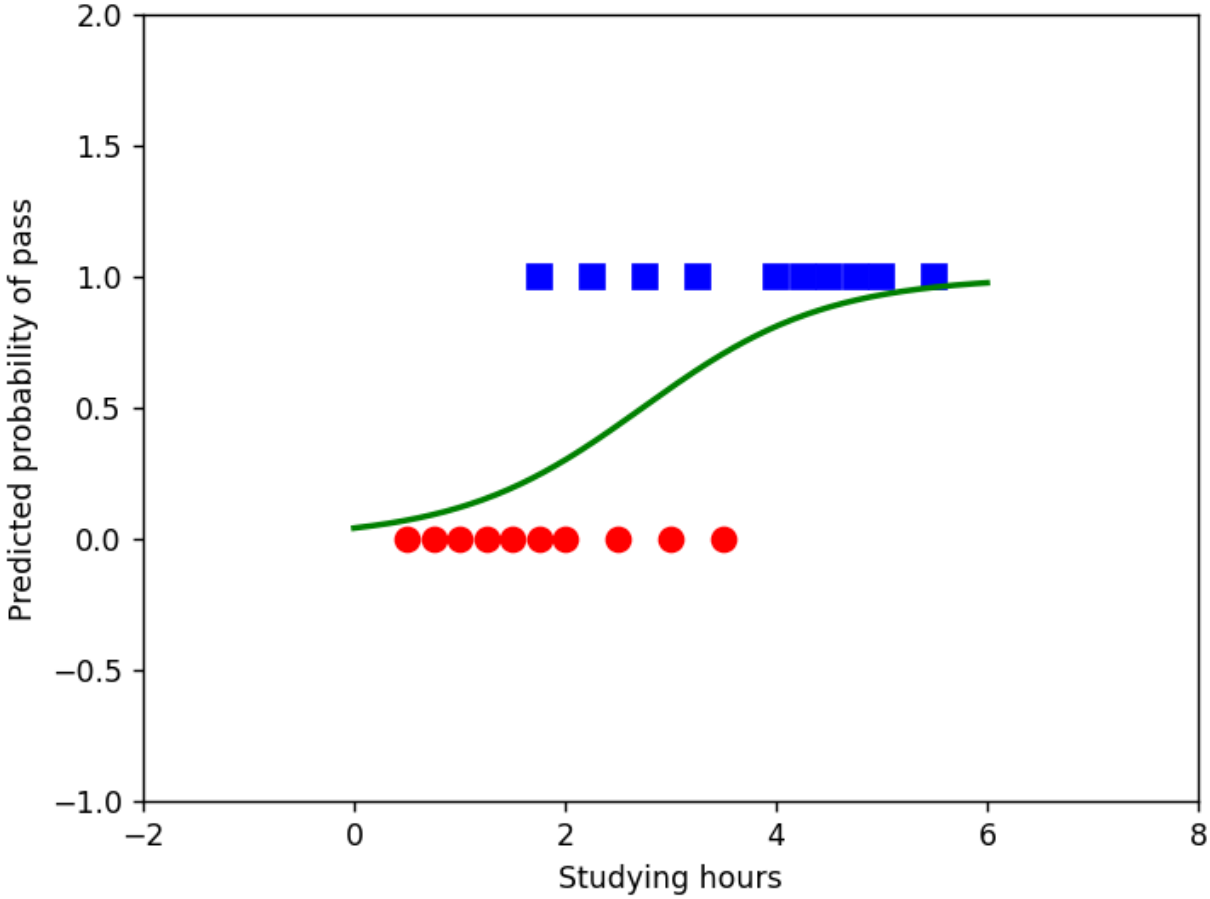
Không có giá trị ngoại lai (No outliers):

- Giả định: Không nên có các giá trị ngoại lai trong tập dữ liệu.
- Ví dụ: Trong tập dữ liệu nghiên cứu bệnh tiểu đường, một giá trị ngoại lai có thể là một người có chỉ số BMI cao bất thường mà không tương ứng với khả năng mắc bệnh tiểu đường. Những giá trị như vậy có thể ảnh hưởng đến tính chính xác của mô hình.

4. Logistic Regression với thư viện scikit-learn

Thư viện scikit-learn của Python xây dựng mô hình hồi quy logistic với các lý thuyết được nêu ra phía trên. Trong đường dẫn Github, mô hình hồi quy tuyến tính được sử dụng để dự đoán khả năng đậu/trượt dựa trên thời gian học tập và mức độ hài lòng của khách hàng đối với các dịch vụ trên máy bay.

Các dự đoán thu về kết quả với độ chính xác khoảng 80%.



Hình 6. Hồi quy Logistic dự đoán khả năng đậu dựa trên thời gian học

Kết luận

Linear Regression là thuật toán đơn giản dự đoán giá trị biến mục tiêu dựa trên mối tương quan giữa các điểm dữ liệu và nhãn được lấy mẫu. Logistic Regression là mô hình học máy có giám sát sử dụng trong bài toán phân loại. Cả hai thuật toán đều được ứng dụng trong nhiều lĩnh vực như dự đoán thị trường, thời tiết.

Sự khác biệt chính của 2 thuật toán này như sau:

Linear Regression	Logistic Regression
Đầu ra phải có giá trị liên tục, chẳng hạn như giá cả, tuổi, v.v.	Đầu ra phải là giá trị phân loại như 0 hoặc 1, Có hoặc không, v.v.
Nó đòi hỏi mối quan hệ tuyến tính giữa các biến phụ thuộc và độc lập.	Nó không yêu cầu mối quan hệ tuyến tính.
Có thể có sự va chạm giữa các biến độc lập.	Không nên có sự va chạm giữa các biến độc lập.

Đường dẫn Github: [sonchuckye/Machine-learning-Linear-and-Logistic-Regression: Machine learning: Linear and Logistic Regression (github.com)]

Phân công công việc

Nguyễn Cao Bảo Sơn	Viết báo cáo Logistic Regression
Nguyễn Khánh Sơn	Chạy code, sửa báo cáo, làm slide
Trần Công Sơn	Viết báo cáo Logistic Regression
Trần Đức Tài	Viết báo cáo Linear Regression
Mai Văn Thái	Viết báo cáo Linear Regression

Tham Khảo

[1] Banerjee, Sriparna and Chaudhuri, Sheli and Mehra, Raghav and Misra, Arundhati, "A comprehensive review of Median Filter, Hybrid Median Filter and their proposed variants," 2020.

[2] Barron, Emmanuel and Greco, John, "Linear Regression," pp. 219-251, 2024.

[3] Bishop, Christopher M, "Pattern recognition and Machine Learning," *Springer*, 2006.

[4] Cox, David R, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 215-242, 1958.

[5] Cramer, Jan Salomon, "The origins of logistic regression," *Econometrics eJournal*, 2002.

[6] Duda, Richard and Hart, Peter and G.Stork, David, "Pattern Classification," *Wiley Interscience*, 2001.

[7] Kivedal, Bjørnar Karlsen, "Simple Linear Regression," pp. 59-93, 2024.

[8] Smith, S, Digital signal processing: a practical guide for engineers and scientists, Newnes, 2003.

[9] Maindonald, John and Braun, W. and Andrews, Jeffrey, "Multiple Linear Regression," pp. 144-207, 2024.