Phương pháp số trong Khoa học dữ liệu

T.S. Nguyễn Thị Hoài Thương

Trường Đại học Khoa học tự nhiên TP.HCM
Khoa Toán Tin-học
Bộ môn Giải tích

ngththuong@hcmus.edu.vn

Ngày 13 tháng 7 năm 2022

Mục lục

Giới thiệu

- Hồi quy từ lâu đã trở thành một phần không thể thiếu trong việc phân tích dữ liệu. Nó liên quan đến việc tìm hiểu và phân tích mối quan hệ giữa các đối tượng nghiên cứu thể hiện qua biến mục tiêu và các biến độc lập (biến giải thích).
- Cụ thể chúng ta có các dạng biến sau:
  - **Biến định lượng** (quantitative/numerical variable) là biến biểu thị trực tiếp bằng con số. Ví dụ như tuổi, chiều cao, trọng lượng,...
  - **Biến định tính** (qualitative/categorical variable) hay biến phân loại là biến phản ánh tính chất hay loại hình, không biểu thị trực tiếp bằng con số. Ví dụ như giới tính, nghề nghiệp, tình trạng hôn nhân,...
  - **Biến nhị phân** (binary variable) là biến chỉ có hai giá trị, 2 biểu hiện không trùng nhau của một đơn vị, nếu đơn vị không có giá trị này thì phải chứa giá trị còn lại của biến thay phiên. Ví dụ: có hoặc không, sống hoặc chết, rời dịch vụ hoặc tiếp tục sử dụng dịch vụ,...

Giới thiệu

- Các loại biến hay loại dữ liệu của biến mục tiêu chính là cơ sở chọn lựa phương pháp hồi quy tương ứng
  - Với biến mục tiêu là *biến định lượng* thì phương pháp hồi quy mà chúng ta sẽ áp dụng đó chính là **phương pháp hồi quy tuyến tính**.
  - Với biến mục tiêu là *biến định tính hay biến nhị phân* thì phương pháp hồi quy chủ yếu và thường sử dụng là **phương pháp hồi quy logistic**.

Phân tích hồi quy tuyến tính - Đặt vấn đề

Ta xét bài toán sau:

- Giả sử chúng ta đã có số liệu thống kê về diện tích, số phòng ngủ, khoảng cách tới trung tâm và giá tiền từ N căn nhà lần lượt là

$$(a_1, b_1, c_1, d_1), \ldots, (a_N, b_N, c_N, d_N)$$

- Yêu cầu đặt ra là tìm một hàm dự đoán giá của một căn nhà sao cho nkhi có một ngôi nhà mới với các thông số về diện tích, số phòng ngủ và khoảng cách tới trung tâm, chúng ta có thể dự đoán được giá của căn nhà đó.

Chúng ta thấy rằng:

- Diện tích càng lớn thì giá nhà càng cao.
- Số lượng phòng ngủ càng nhiều thì giá nhà càng cao.
- Nhà càng xa trung tâm thì giá nhà càng giảm.

Phân tích hồi quy tuyến tính - Đặt vấn đề

Một hàm số đơn giản nhất có thể mô tả mới quan hệ giữa giá nhà và 3 đại lượng đầu vào (diện tích, số phòng ngủ, khoảng cách tới trung tâm) là

$$\begin{aligned} d \approx f(\mathbf{w}) = \widehat{d} \\ f(\mathbf{w}) = x_1 a + x_2 b + x_3 c + x_4 \end{aligned} \tag{1.1}$$

trong đó

- x_1, x_2, x_3, x_4 là các hằng số cần tìm.
- $\mathbf{w} = (a, b, c)$ là một vectơ hàng chứa các thông tin diện tích, số phòng ngủ, khoảng cách tới trung tâm của căn nhà.
- d là giá trị thực và $\widehat{d}$ là giá trị dự đoán.

$\rightsquigarrow$ Mối quan hệ $d \approx f(\mathbf{w})$ là mối quan hệ tuyến tính.

$\rightsquigarrow$ Bài toán đi tìm các hệ số tối ưu x_1, x_2, x_3, x_4 được gọi là **bài toán hồi quy tuyến tính** (linear regression).

Phân tích hồi quy tuyến tính - Đặt vấn đề

Lưu ý:

- d là giá trị thực (dựa trên số liệu thống kê).
- $\widehat{d}$ là giá trị mà mô hình hồi quy tuyến tính dự đoán được.

Nhìn chung, d và $\widehat{d}$ là hai giá trị khác nhau do có sai số mô hình. Tuy nhiên, chúng ta mong muốn rằng sự khác nhau này rất nhỏ.

Phân tích hồi quy tuyến tính - Phân tích toán học

Đặt

- $\mathbf{x} = (x_1, x_2, x_3, x_4)^T$ là một vectơ cột cần phải tối ưu.
- $\overline{\mathbf{w}} = (a, b, c, 1)$ là vectơ hàng chứa dữ liệu đầu vào.

Khi đó, phương trình (1.1) có thể được viết dưới dạng:

$$d \approx \overline{\mathbf{w}}\mathbf{x} = \widehat{d}.$$

Chúng ta mong muốn rằng sự sai khác giữa giá trị thực d và giá trị dự đoán $\widehat{d}$ là nhỏ nhất. Nói cách khác, chúng ta muốn giá trị sau càng nhỏ càng tốt

$$\frac{1}{2}(d - \widehat{d})^2 = \frac{1}{2}(d - \overline{\mathbf{w}}\mathbf{x})^2.$$

Lưu ý:

- Chúng ta thêm hệ số $1/2$ vào là để thuận tiện cho việc tính toán (khi đạo hàm thì số $1/2$ sẽ bị triệt tiêu).
- Chúng ta cần $(d - \widehat{d})^2$ vì $d - \widehat{d}$ có thể là số âm (việc nói $d - \widehat{d}$ nhỏ nhất sẽ không đúng vì khi $d - \widehat{d} = -\infty$ là rất nhỏ nhưng sự sai lệch rất lớn).

Phân tích hồi quy tuyến tính - Phân tích toán học

Với số liệu từ N căn nhà, ta mong muốn tổng sai số là nhỏ nhất. Tương đương với việc tìm **x** để hàm số sau đạt giá trị nhỏ nhất

$$L(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{N}(d_i - \overline{\mathbf{w}}_i\mathbf{x}) \tag{1.2}$$

trong đó: $\overline{\mathbf{w}}_i = (a_i, b_i, c_i, 1)$ và $L(\mathbf{x})$ được gọi là **hàm mất mát** (loss function) của bài toán hồi quy tuyến tính.

Phân tích hồi quy tuyến tính - Nghiệm của bài toán hồi quy tuyến tính

Xét bài toán tối ưu sau

$$\min_{\mathbf{x}} L(\mathbf{x}) \tag{1.3}$$

trong đó $L(\mathbf{x})$ được định nghĩa như (1.2). Để tìm nghiệm $\mathbf{x}$ cho bài toán trên ta có thể làm như sau:

Cách 1: Giải phương trình đạo hàm của hàm mất mát để tìm nghiệm x

- Đặt $\mathbf{d} = (d_1, d_2, \ldots, d_N)$ là vectơ cột chứa thông tin giá trị của các căn nhà và $\overline{\mathbf{W}} = (\overline{\mathbf{w}}_1, \overline{\mathbf{w}}_2 \ldots, \overline{\mathbf{w}}_N)$ là ma trận chứa các dữ liệu đầu vào (chứa các thông tin diện tích, số phòng ngủ và khoảng cách tới trung tâm của các ngôi nhà).
- Khi đó ma trận mất mát $L(\mathbf{x})$ (1.2) được viết lại dưới dạng

$$L(\mathbf{x}) = \frac{1}{2}\|\mathbf{d} - \overline{\mathbf{W}}\mathbf{x}\|_2^2 \tag{1.4}$$

trong đó $\|z\|_2$ là chuẩn Euclidean.

Phân tích hồi quy tuyến tính - Nghiệm của bài toán hồi quy tuyến tính

- Đạo hàm của hàm mất mát $L(\mathbf{x})$, ta được

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{x}} &= \overline{\mathbf{W}}^T(\overline{\mathbf{W}}\mathbf{x} - \mathbf{d}) = 0 \\ \Leftrightarrow \overline{\mathbf{W}}^T\overline{\mathbf{W}}\mathbf{x} &= \overline{\mathbf{W}}^T\mathbf{d}. \end{aligned} \tag{1.5}$$

  - Nếu $\overline{\mathbf{W}}^T\overline{\mathbf{W}}$ khả nghịch thì phương trình (1.5) có nghiệm duy nhất là $\mathbf{x} = (\overline{\mathbf{W}}^T\overline{\mathbf{W}})^{-1}\overline{\mathbf{W}}^T\mathbf{d}$.
  - Nếu $\overline{\mathbf{W}}^T\overline{\mathbf{W}}$ không khả nghịch thì ta áp dụng thuật toán SVD để tìm ma trận giả nghịch đảo của $\overline{\mathbf{W}}^T\overline{\mathbf{W}}$. Khi đó, nghiệm của phương trình (1.5) có dạng $\mathbf{x} = (\overline{\mathbf{W}}^T\overline{\mathbf{W}})^{+}\overline{\mathbf{W}}^T\mathbf{d}$.

Cách 2: Dùng các thuật toán Gradient descent, Accelerated gradient descent, Stochastic gradient descent để tìm nghiệm $\mathbf{x}$ của bài toán (1.3).

Phân tích hồi quy tuyến tính - Bài tập

Bài tập 1: Cho bảng số liệu sau:

STT	Diện tích (m^2)	Số phòng ngủ	Khoảng cách tới TT	Giá(tỷ VNĐ)
1	40	1	30	1.1
2	60	2	32	1.55
3	53	2	30.1	1.68
4	71	2	35.7	1.75
5	80	2	24.5	5.5
6	56	2	27.6	2.3
7	75	2	27.6	3
8	79	2	27.6	3.5
9	56	2	29.7	2.4
10	60	2	29.7	2.9
11	72	2	29.7	3
12	95	3	29.7	4.2
13	47	1	19.3	1.5
14	91	2	18.1	2.2
15	68	1	21.4	1.5
16	69	2	17.5	3.15
17	82	2	25.1	3.4
18	60	2	26.5	2.245
19	68	2	26.5	2.4

Phân tích hồi quy tuyến tính - Bài tập

Dựa vào bảng số liệu trên, hãy dự đoán giá của một căn nhà có diện tích là 79 m^2, 2 phòng ngủ, khoảng cách tới trung tâm là 26.5 km bằng cách:

a) Giải phương trình đạo hàm mất mát.

b) Dùng các thuật toán Gradient descent, Accelerated gradient descent, Stochastic gradient descent.

c) Sử dụng thư viện scikit-learn.

Biết giá trị thực tế của căn nhà trên là 2.5 tỷ VNĐ, hãy so sánh các kết quả trên với nhau.

Phân tích hồi quy tuyến tính - Bài tập

Bài tập 2: Cho bảng số liệu chiều cao và cân nặng của 14 người như sau:

STT	Chiều cao (cm)	Cân nặng (kg)
1	147	49
2	150	50
3	153	51
4	155	52
5	158	54
6	160	56
7	163	58
8	168	60
9	170	72
10	173	63
11	175	64
12	178	66
13	180	67
14	183	68

Phân tích hồi quy tuyến tính - Bài tập

Bài toán đặt ra là từ bảng số liệu trên, hãy dự đoán cân nặng của một người có chiều cao là 165 cm bằng cách:

a) Giải phương trình đạo hàm mất mát.

b) Dùng các thuật toán Gradient descent, Accelerated gradient descent, Stochastic gradient descent.

c) Sử dụng thư viện scikit-learn.

Biết cân nặng thực tế của người đó trên là 59 kg, hãy so sánh các kết quả trên với nhau.

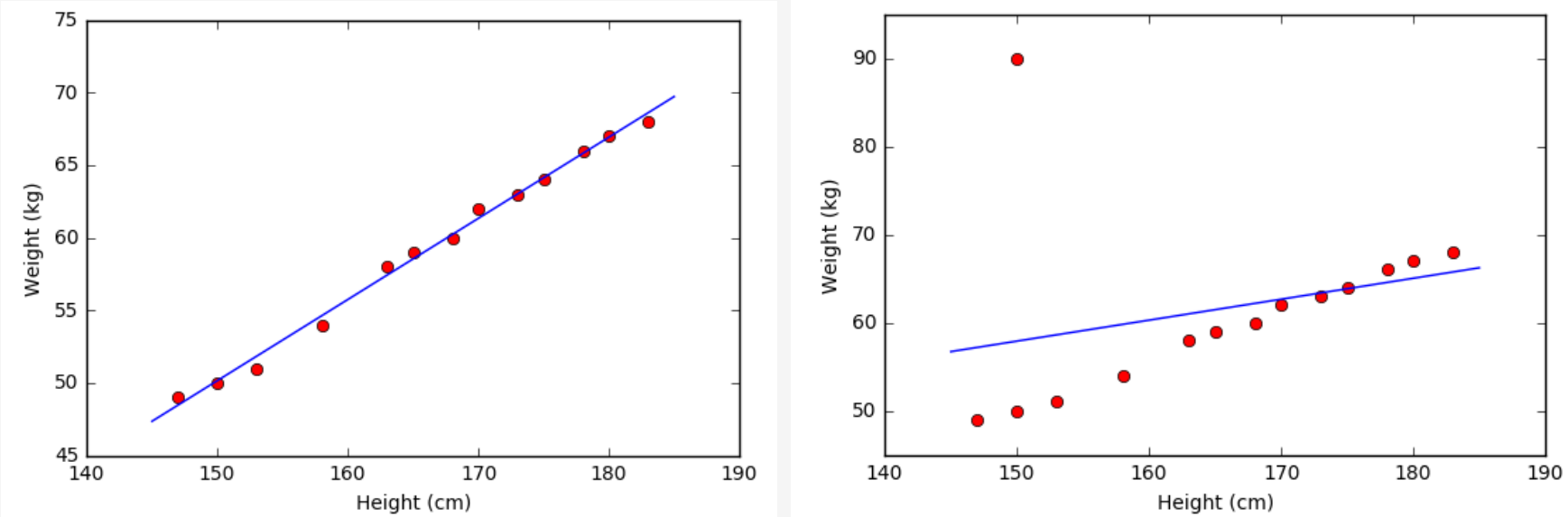
Phân tích hồi quy tuyến tính - Ưu và nhược điểm

Ưu điểm:

- Áp dụng cho các bài toán đơn giản. Nó chỉ nghiên cứu mối quan hệ tuyến tính giữa một biến độc lập và biến phụ thuộc.
- Áp dụng cho biến định lượng.

Nhược điểm:

- Nó rất nhạy cảm với nhiễu. Trong bài tập về mối quan hệ giữa chiều cao và cân nặng bên trên, nếu chỉ có một cặp dữ liệu nhiễu (150cm, 90 kg) thì kết quả sẽ sai khác đi rất nhiều.



Hình: Mối liên hệ giữa chiều cao và cân nặng với số liệu không có điểm nhiễu (bên trái) và với số liệu có điểm nhiễu (bên phải).

Vì vậy, trước khi thực hiện phân tích hồi quy tuyến tính, chúng ta cần phải loại bỏ các điểm nhiễu. Bước này được gọi là tiền xử lí.

- Nó không biểu diễn được các mô hình phức tạp.

Phân tích hồi quy logistic - Giới thiệu

- Mô hình hồi quy Binary logistic (gọi đơn giản là hồi quy logistic) là một trong những mô hình phổ biến dùng trong nghiên cứu nhằm ước lượng xác suất của một sự việc sẽ xảy ra.
- Trong cuộc sống có rất nhiều hiện tượng tự nhiên xảy ra ở đủ các lĩnh vực kinh tế-xã hội, môi trường,... mà chúng ta ứng dụng mô hình hồi quy logistic để dự đoán như:
  - Đơn hàng này có được chấp nhận hay không?
  - Người mua hàng này có thích hay không?
  - Chỉ số môi trường ở đây có sạch hay không?

Phân tích hồi quy logistic - Nhắc lại một số kiến thức

Quy tắc dây chuyền (chain rule)

Nếu $z = f(y)$ và $y = g(x)$ (hay $z = f(g(x))$) thì $\frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx}$.

Ví dụ: Cho $z = y^2$ và $y = 2x + 1$. Tính $\frac{dz}{dx}$?

Áp dụng quy tắc chain rule, ta có:

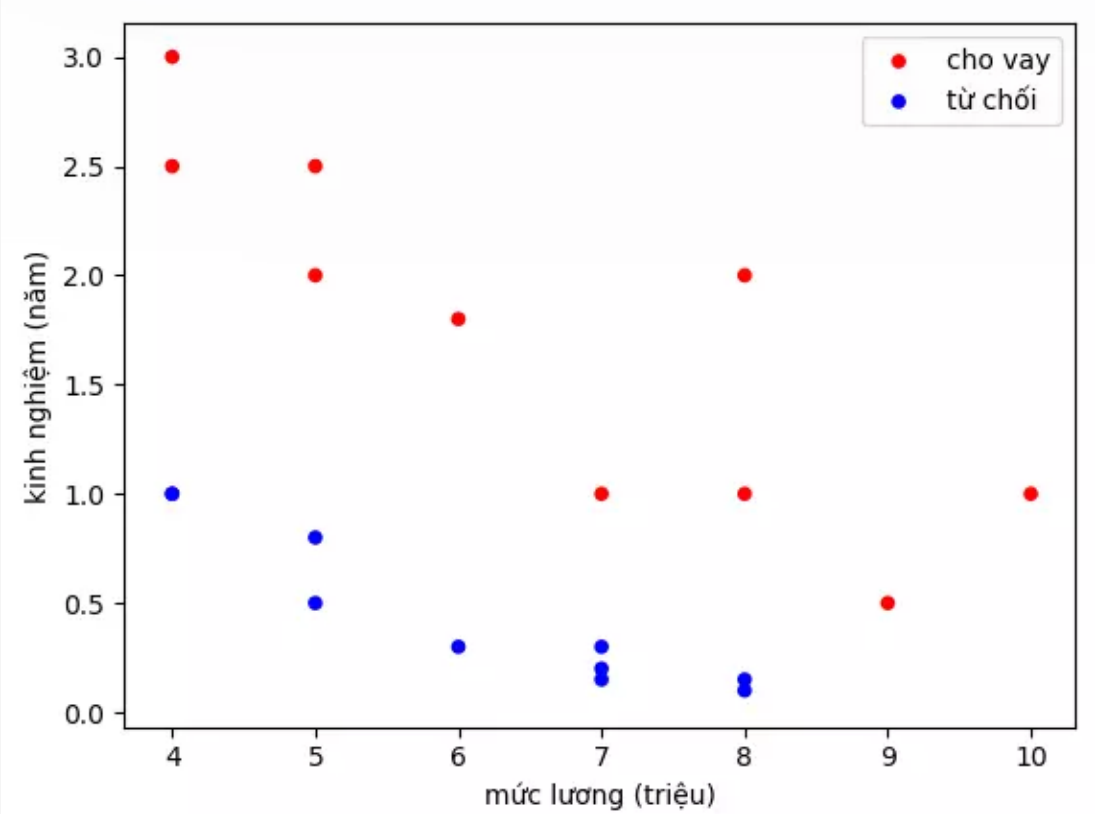
$$\frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx} = 2y \times 2 = 4y = 4(2x + 1).$$

Phân tích hồi quy logistic - Đặt vấn đề

- Ngân hàng bạn đang làm đang có chương trình cho vay ưu đãi cho các đối tượng mua chung cư. Tuy nhiên, gần đây có nhiều chung cư hấp hẫn (giá tốt, vị trí đẹp,...) nên lượng hồ sơ người nộp cho chương trình ưu đãi tăng lên nhiều.
- Bình thường bạn có thể duyệt 10-20 hồ sơ một ngày để quyết định hồ sơ có được cho vay hay không. Tuy nhiên gần đây bạn nhận được 1000-2000 hồ sơ mỗi ngày.
- Bạn không thể xử lý hết hồ sơ và bạn cần có một giải pháp để có thể dự đoán hồ sơ mới là có nên cho vay hay không. Sau khi phân tích, bạn nhận thấy có 2 yếu tố quyết định đến việc hồ sơ có được chấp nhận hay không. Đó là mức lương và kinh nghiệm làm việc.

Phân tích hồi quy logistic - Đặt vấn đề

- Dưới đây là đồ thị biểu diễn những dữ liệu bạn có từ trước đến nay

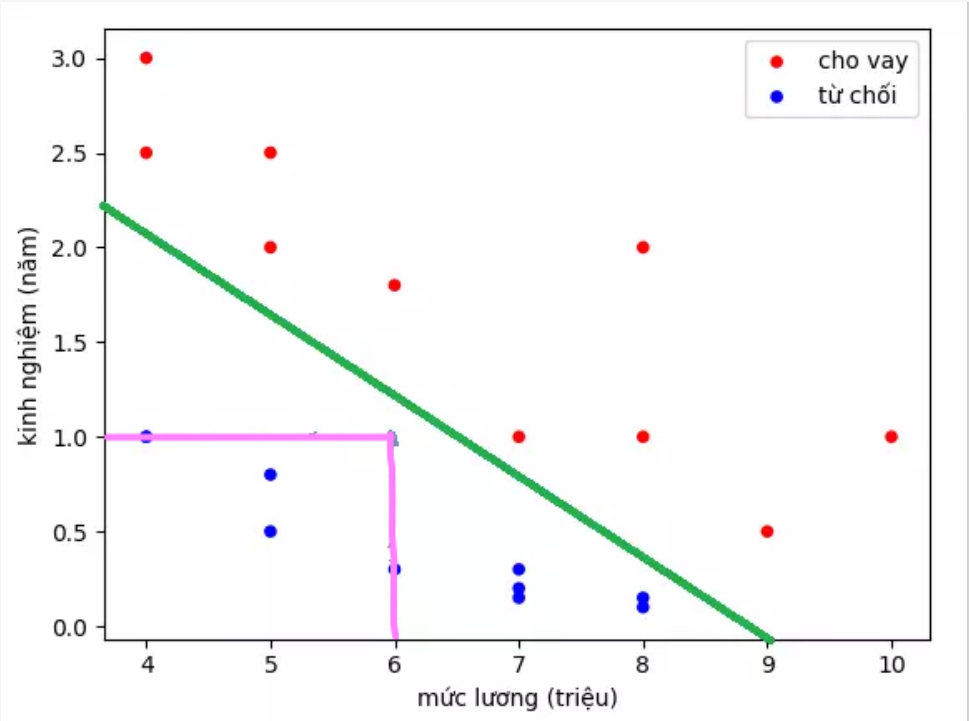


Hình: Đồ thị giữa mức lương, số năm kinh nghiệm và kết quả cho vay.

Về mặt logic, bây giờ chúng ta cần tìm đường thẳng phân chia giữa các điểm cho vay và từ chối. Sau đó, đưa ra quyết định cho 1 điểm mới dựa vào đường thẳng đó.

Phân tích hồi quy logistic - Đặt vấn đề

- Ví dụ như đường màu xanh là đường phân chia. Từ đây, chúng ta có thể dự đoán được rằng hồ sơ của người có mức lương 6 triệu và 1 năm kinh nghiệm là không chấp nhận.

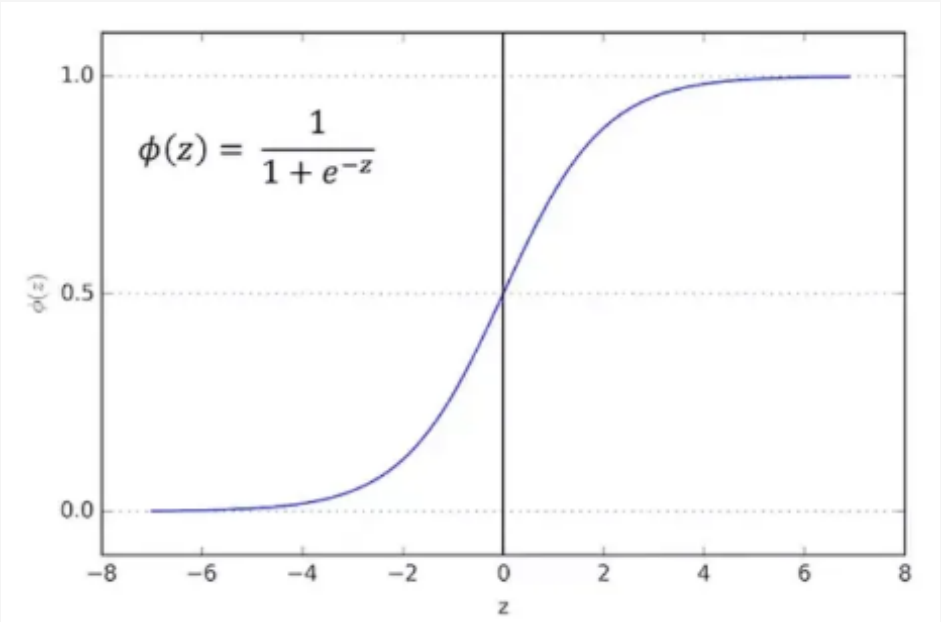


- Tuy nhiên, do ngân hàng đang gặp khó khăn nên hạn chế cho vay, ngân hàng yêu cầu hồ sơ đạt trên 80% mới cho vay. Bây giờ không chỉ dừng lại ở việc quyết định cho vay hay không, mà còn phải tìm xác suất hồ sơ đó cho vay là bao nhiêu.

Phân tích hồi quy logistic - Hàm sigmoid

Giờ chúng ta phải tìm xác suất cho vay của một hồ sơ, đương nhiên là giá trị của hàm cần trong đoạn $[0, 1]$. Hàm mà luôn có giá trị trong đoạn $[0, 1]$, liên tục mà lại dễ sử dụng thì đó làm hàm sigmoid với công thức như sau:

$$\phi(z) = \frac{1}{1 + e^{-z}} \tag{1.6}$$



Hình: Đồ thị hàm sigmoid.

Phân tích hồi quy logistic - Hàm sigmoid

Từ đồ thị trên ta thấy rằng:

- Giá trị của hàm sigmoid sẽ tiệm cận đến 1 khi z tiến đến $+\infty$.
- Giá trị của hàm sigmoid sẽ tiệm cận đến 0 khi z tiến đến $-\infty$.
- Giá trị của hàm sigmoid sẽ bằng 0.5 khi $z = 0$.

Nhờ vào đặc tính này mà hàm sigmoid được sử dụng nhiều trong lĩnh vực trí tuệ nhân tạo với vai trò là hàm kích hoạt.

Tính chất:

- Là một hàm số liên tục và nhận giá trị trong khoảng $(0, 1)$.
- Vì là hàm liên tục nên hàm sigmoid sẽ có đạo hàm tại mọi điểm (để áp dụng gradient descent).

Phân tích hồi quy logistic - Thiết lập mô hình

Với hồ sơ thứ i mà ta đang xét, ta gọi:

- $x_1^{(i)}$ là lương và $x_2^{(i)}$ là thời gian làm việc của người nộp hồ sơ vay.
- $p(y_i = 1) = \widehat{y}_i$ là xác suất mà chúng ta dự đoán hồ sơ được cho vay.
- $p(y_i = 0) = 1 - \widehat{y}_i$ là xác suất mà chúng ta dự đoán hồ sơ không được cho vay.

Với **phân tích hồi quy tuyến tính**, ta có hàm dự đoán sau

$$\widehat{y}_i = w_0 + w_1 x_1^{(i)} + w_2 x_2^{(2)}$$

trong đó: w_0, w_1, w_2 là các hằng số cần tìm.

Với **phân tích hồi quy logistic**, ta có hàm dự đoán như sau:

$$\widehat{y}_i = \phi\left(w_0 + w_1 x_1^{(i)} + w_2 x_2^{(2)}\right) = \frac{1}{1 + e^{-\left(w_0 + w_1 x_1^{(i)} + w_2 x_2^{(2)}\right)}}.$$

Lưu ý: Ở đây ta dùng hàm sigmoid để chuyển giá trị $w_0 + w_1 x_1^{(i)} + w_2 x_2^{(2)}$ thành xác suất mà chúng ta dự đoán hồ sơ thứ i có được cho vay hay không?

Phân tích hồi quy logistic - Hàm mất mát (Loss function)

Từ mô hình bài toán trên ta có nhận xét như sau:

- Nếu hồ sơ thứ i là cho vay, tức là $y_i = 1$ thì ta cũng muốn $\widehat{y}_i$ càng gần 1 càng tốt. Hay nói cách khác là mô hình dự đoán xác suất người thứ i được vay vốn càng cao càng tốt.
- Nếu hồ sơ thứ i không được vay, tức là $y_i = 0$ thì ta cũng muốn $\widehat{y}_i$ càng gần 0 càng tốt. Hay nói cách khác là mô hình dự đoán xác suất người thứ i được vay vốn càng thấp càng tốt.

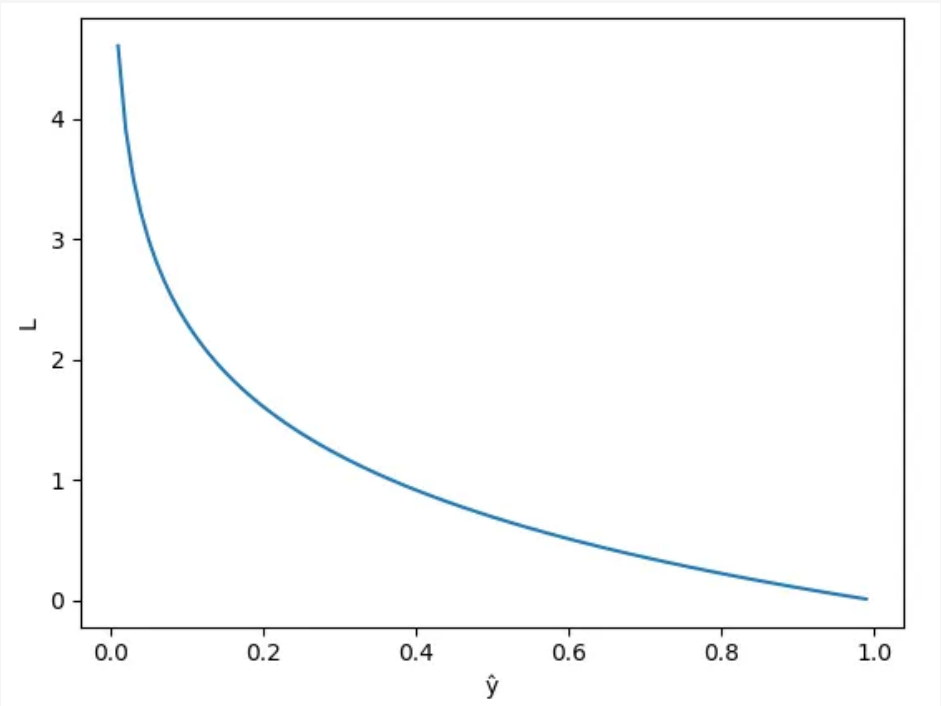
Với mỗi điểm $(x^{(i)}, y_i)$, ta xét hàm số sau:

$$J_i = -(y_i \ln(\widehat{y}_i) + (1 - y_i) \ln(1 - \widehat{y}_i)). \tag{1.7}$$

Phân tích hồi quy logistic - Hàm mất mát (Loss function)

Ta thấy rằng:

- Nếu $y_i = 1$ thì $J_i = -\ln(\widehat{y}_i)$

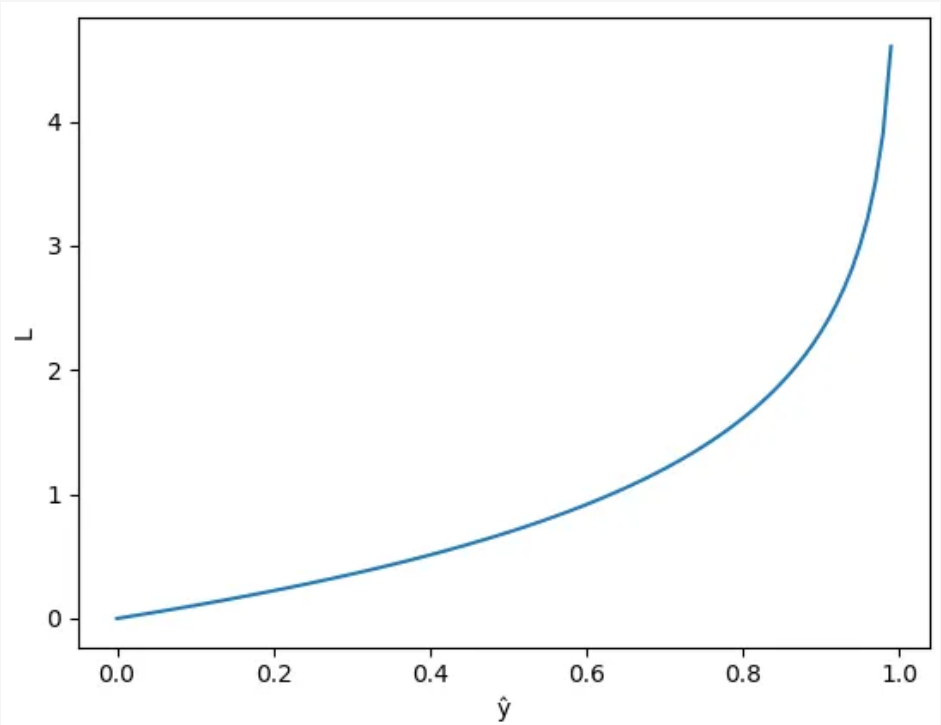


Hình: Đồ thị hàm J_i trong trường hợp $y_i = 1$.

- Khi mô hình dự đoán $\widehat{y}_i$ gần 1, tức là giá trị dự đoán gần với giá trị thật y_i, thì giá trị của J_i nhỏ và xấp xỉ bằng 0.
- Khi mô hình dự đoán $\widehat{y}_i$ gần 0, tức là giá trị dự đoán ngược lại với giá trị thật y_i, thì giá trị của J_i rất lớn.

Phân tích hồi quy logistic - Hàm mất mát (Loss function)

- Nếu y_i thì $J_i = -\ln(1 - \widehat{y}_i)$



Hình: Đồ thị hàm J_i trong trường hợp $y_i = 0$.

- Khi mô hình dự đoán $\widehat{y}_i$ gần 0, tức là giá trị dự đoán gần với giá trị thật y_i, thì giá trị của J_i nhỏ và xấp xỉ bằng 0.
- Khi mô hình dự đoán $\widehat{y}_i$ gần 1, tức là giá trị dự đoán ngược lại với giá trị thật y_i, thì giá trị của J_i rất lớn.

Phân tích hồi quy logistic - Hàm mất mát (Loss function)

- Như vậy, khi giá trị dự đoán $\widehat{y_i}$ càng gần giá trị thật y_i thì L càng nhỏ. Do đó, mô hình bài toán (1.7) trở thành bài toán tìm giá trị nhỏ nhất của J_i.
- Với số liệu từ N bộ hồ sơ , ta mong muốn rằng tổng sai số là nhỏ nhất. Tương đương với việc tìm $\mathbf{w} = (w_0, w_1, w_2)$ để hàm sau đạt giá trị nhỏ nhất

$$L(\mathbf{w}) = -\sum_{i=1}^{N}(y_i \ln(\widehat{y_i}) + (1 - y_i)\ln(1 - \widehat{y_i})) \tag{1.8}$$

trong đó: $\widehat{y_i} = \dfrac{1}{1 + e^{-\left(w_0 + w_1 x_1^{(i)} + w_2 x_2^{(i)}\right)}}$ và hàm $L(\mathbf{w})$ được gọi là **hàm mất mát** (loss function) của bài toán hồi quy logistic.

Phân tích hồi quy logistic - Nghiệm của bài toán hồi quy logistic

Xét bài toán tối ưu sau:

$$\min_{\mathbf{w}} L(\mathbf{w})$$

trong đó $L(\mathbf{w})$ được định nghĩa như (1.8). Để tìm cực tiểu $\mathbf{w}$ cho bài toán trên ta áp dụng thuật toán Gradient descent (hoặc Accelerated Gradient descent).

Áp dụng thuật toán Gradient descent:

- **Bước 1:** Tìm vectơ gradient $\nabla L(\mathbf{w})$. Ta có:

$$\nabla L(\mathbf{w}) = \sum_{i=1}^{N} \nabla J_i \tag{1.9}$$

trong đó

$$\begin{aligned} J_i &= -(y_i \ln(\widehat{y}_i) + (1 - y_i) \ln(1 - \widehat{y}_i)) \\ \widehat{y}_i &= \frac{1}{1 + e^{-\left(w_0 + w_1 x_1^{(i)} + w_2 x_2^{(i)}\right)}} \end{aligned} \tag{1.10}$$

Phân tích hồi quy logistic - Nghiệm của bài toán hồi quy logistic

Từ (1.10), ta có

$$\begin{aligned}
\frac{dJ_i}{d\widehat{y}_i} &= -\left(\frac{y_i}{\widehat{y}_i} - \frac{1-y_i}{1-\widehat{y}_i}\right) \\
\frac{d\widehat{y}_i}{dw_0} &= \widehat{y}_i(1-\widehat{y}_i) \\
\frac{d\widehat{y}_i}{dw_1} &= x_1^{(i)}\widehat{y}_i(1-\widehat{y}_i) \\
\frac{d\widehat{y}_i}{dw_1} &= x_2^{(i)}\widehat{y}_i(1-\widehat{y}_i)
\end{aligned}$$

Do đó, khi áp dụng quy tắc dây chuyền (chain rule), ta được

$$\begin{aligned}
\frac{dJ_i}{dw_0} &= \frac{dJ_i}{d\widehat{y}_i}\frac{d\widehat{y}_i}{dw_0} = \widehat{y}_i - y_i \\
\frac{dJ_i}{dw_1} &= \frac{dJ_i}{d\widehat{y}_i}\frac{d\widehat{y}_i}{dw_1} = x_1^{(i)}(\widehat{y}_i - y_i) \\
\frac{dJ_i}{dw_2} &= \frac{dJ_i}{d\widehat{y}_i}\frac{d\widehat{y}_i}{dw_2} = x_2^{(i)}(\widehat{y}_i - y_i)
\end{aligned} \tag{1.11}$$

Phân tích hồi quy logistic - Nghiệm của bài toán hồi quy logistic

Từ (1.9) và (1.11), ta có

$$\frac{dL}{dw_0} = \sum_{i=1}^{N} (\widehat{y}_i - y_i)$$

$$\frac{dL}{dw_1} = \sum_{i=1}^{N} x_1^{(i)} (\widehat{y}_i - y_i)$$

$$\frac{dL}{dw_2} = \sum_{i=1}^{N} x_2^{(i)} (\widehat{y}_i - y_i)$$

Đặt

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(N)} \\ x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(N)} \end{pmatrix}, \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix}, \widehat{\mathbf{y}} = \begin{pmatrix} \widehat{y}_1 \\ \widehat{y}_2 \\ \dots \\ \widehat{y}_N \end{pmatrix}$$

Khi đó

$$\nabla L(\mathbf{w}) = \mathbf{X}(\widehat{\mathbf{y}} - \mathbf{y}).$$

Phân tích hồi quy logistic - Nghiệm của bài toán hồi quy logistic

- **Bước 2:** Giả sử ta có giá trị ban đầu $\mathbf{w}_0$, áp dụng thuật toán Gradient descent, ta có

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla L(\mathbf{w}_t)$$

  trong đó η là learning rate.

Sau khi thực hiện thuật toán Gradient descent ta sẽ tìm được w_0, w_1, w_2. Với mỗi hồ sơ mới, ta sẽ biết được lương $x_1^{(i)}$ và thời gian làm việc $x_2^{(i)}$ của người nộp hồ sơ vay. Khi đó, ta áp dụng công thức (1.10) để tính phần trăm cho vay $\widehat{y}_i$.

- Nếu $\widehat{y}_i$ có giá trị lớn hơn hoặc bằng ngưỡng cho vay mà ngân hàng quy định thì bạn sẽ dự đoán được hồ sơ đó sẽ được vay.
- Ngược lại thì không cho vay.

Phân tích hồi quy logistic - Quan hệ giữa phần trăm và đường phân cách

Giả sử bạn lấy mốc ở chính giữa là 50%, tức là hồ sơ mới dự đoán $\widehat{y_i} \geq 0.5$ thì cho vay. Ngược lại, nếu nhỏ hơn 0.5 thì không cho vay. Khi đó

$$\begin{aligned}
\widehat{y_i} \geq 0.5 &\Leftrightarrow \frac{1}{1+e^{-\left(w_0+w_1x_1^{(i)}+w_2x_2^{(i)}\right)}} \geq 0.5 \\
&\Leftrightarrow e^{-\left(w_0+w_1x_1^{(i)}+w_2x_2^{(i)}\right)} \leq 1 = e^0 \\
&\Leftrightarrow w_0 + w_1x_1^{(i)} + w_2x_2^{(i)} \geq 0
\end{aligned}$$

Tương tự

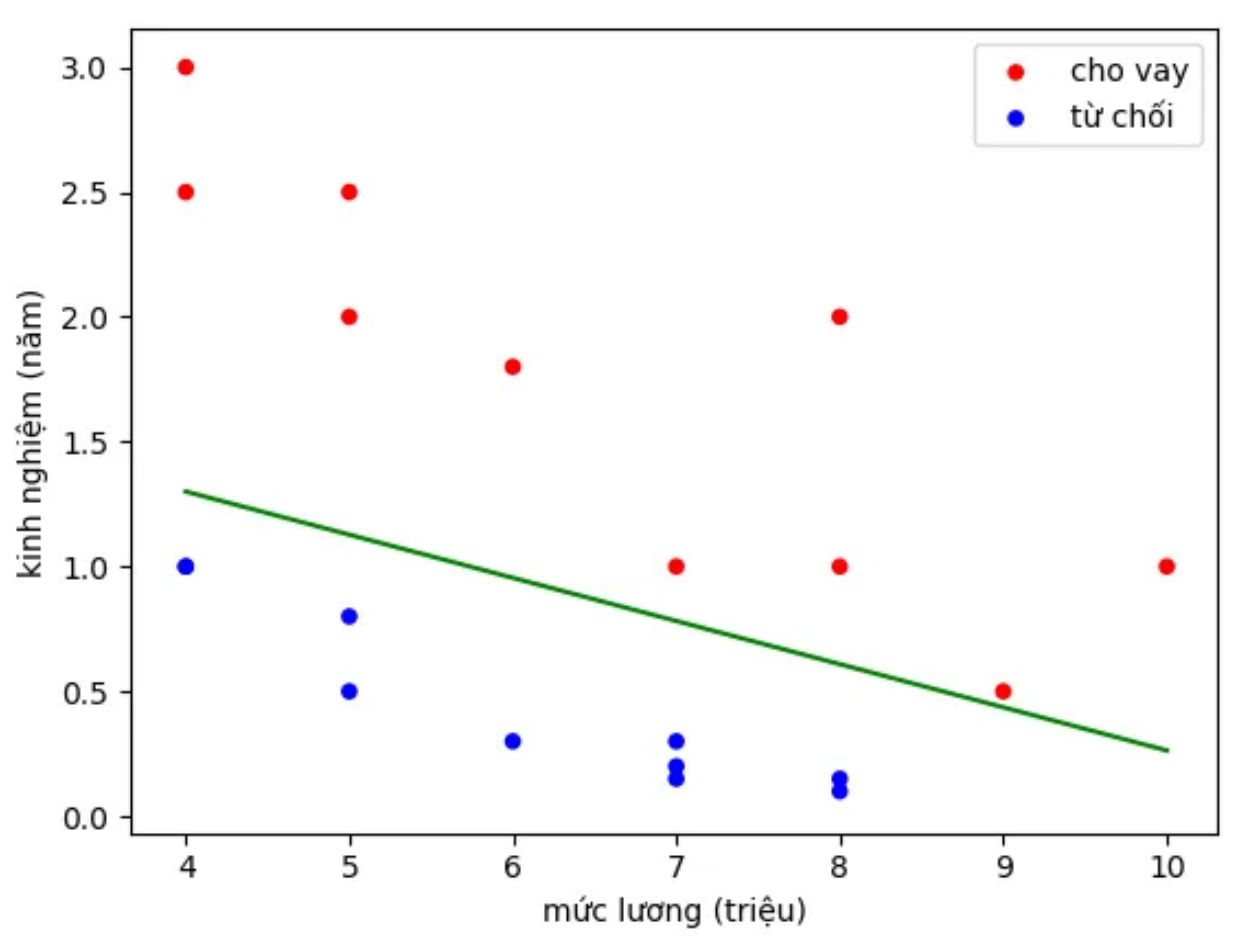
$$\widehat{y_i} < 0.5 \Leftrightarrow w_0 + w_1x_1^{(i)} + w_2x_2^{(i)} < 0$$

Như vậy

$$w_0 + w_1x_1^{(i)} + w_2x_2^{(i)} = 0$$

chính là đường phân cách giữa điểm cho vay và từ chối.

Phân tích hồi quy logistic - Quan hệ giữa phần trăm và đường phân cách

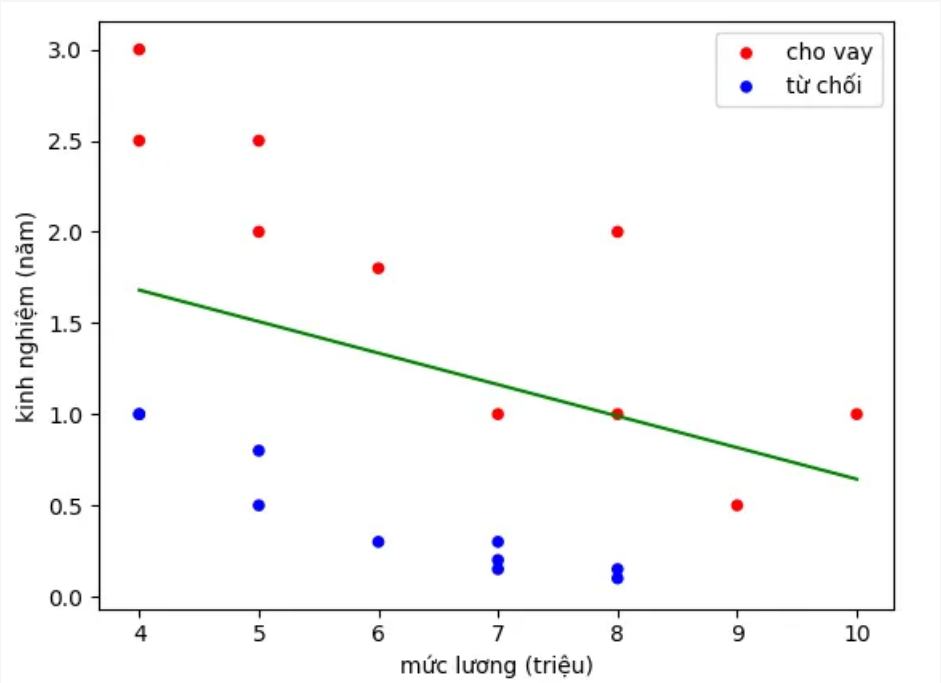


Hình: Đường phân cách với xác suất trên 50%.

Phân tích hồi quy logistic - Quan hệ giữa phần trăm và đường phân cách

Trong trường hợp tổng quát, bạn lấy xác suất lớn hơn s $(0 < s < 1)$ thì mới cho vay. Khi đó

$$\widehat{y_i} > s \Leftrightarrow w_0 + w_1 x_1^{(i)} + w_2 x_2^{(i)} > -\ln\left(\frac{1}{s} - 1\right).$$



Hình: Đường phân cách với xác suất $s = 0.8$.

Phân tích hồi quy logistic - Quan hệ giữa phần trăm và đường phân cách

Với $s = 0.8$, ta thấy rằng đường phân cách gần với điểm màu đỏ hơn so với $s = 0.5$. Thậm chí một số hồ sơ cũ được cho vay nhưng nếu giờ nộp lại cũng từ chối. Đồng nghĩa với việc khi ngân hàng thắt chặt việc cho vay lại thì một số hồ sơ sẽ bị loại.

Phân tích hồi quy logistic - Bài tập

Bài tập 1: Cho bảng số liệu sau:

	Lương	Thời gian làm việc	Cho vay
0	10	1.0	1
1	5	2.0	1
2	6	1.8	1
3	7	1.0	1
4	8	2.0	1
5	9	0.5	1
6	4	3.0	1
7	5	2.5	1
8	8	1.0	1

	Lương	Thời gian làm việc	Cho vay
9	4	2.50	1
10	8	0.10	0
11	7	0.15	0
12	4	1.00	0
13	5	0.80	0
14	7	0.30	0
15	4	1.00	0
16	5	0.50	0
17	6	0.30	0
18	7	0.20	0
19	8	0.15	0

Phân tích hồi quy logistic - Bài tập

a) Từ bảng số liệu trên, áp dụng thuật toán Gradient descent để viết hàm tính xác suất cho vay của một hồ sơ bất kì.

b) Giả sử ngân hàng yêu cầu hồ sơ đạt 80% mới cho vay, hãy vẽ đường phân cách giữa hồ sơ cho vay và không cho vay. Từ đó xác định xem một người có mức lương là 9 triệu và kinh nghiệm làm việc là 0.5 năm thì có được vay hay không?

Bài tập 2: Tương tự như **Bài tập 1** nhưng với thuật toán Accelerated Gradient descent.