Báo cáo thực tập ngày 19/06/2024

Nguyễn Tiến Đạt - Thực tập sinh AI

1 HỒI QUY LOGISTIC - LOGISCTIC REGRESSION

Nếu hồi quy tuyến tính (linear regression) cho giá trị đầu ra là số thực thì hồi quy logistic (logistic regression) cho đầu ra là giá trị nhị phân (0 hoặc 1).

1.1 Bài toán

Lương	Thời gian làm việc	Khả năng cho vay
10	1	1
9	0.5	1
5	2	1
...	...	...
6	0.3	0
7	0.15	0
...	...	...

Xét một ví dụ về sự liên quan giữa Lương và Thời gian làm việc của khoảng 20 người. Bài toán đặt ra là xây dựng một mô hình đánh giá khả năng cho vay của một người dựa trên Lương và Thời gian làm việc. Nhìn chung Lương cao hoặc Thời gian làm việc lâu năm thì khả năng cho vay càng cao. Tuy nhiên sẽ có những lúc ngân hàng cần lọc ra những những hồ sơ chắc chắn trên một ngưỡng % nào đó mới được vay. Vậy nên có những bài toán phân loại người ta sẽ quan tâm đến xác suất hơn là dự đoán chính xác giá trị $0/1$.

1.2 Hàm sigmoid

Xác suất của một sự kiện trong khoảng [0,1]. Công việc bây giờ là phải tìm ra xác suất của hồ sơ nào cho vay dựa trên hai thông số Lương và Thời gian làm việc. Hay giá trị của hàm cần cần một vài tính chất quan trọng:

- Là các hàm số liên tục nhận giá trị thực, bị chặn trong khoảng [0,1].
- Thể hiện khả năng xác suất càng cao thì đầu ra càng gần 1 và ngược lại, xác suất càng thấp thì đầu ra càng gần 0.
- Có khả năng đạo hàm để thuận lợi trong việc tối ưu.

Hàm thoả mãn được ba tính chất trên được sử dụng nhiều nhất là hàm *sigmoid*.

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

Thêm nữa:
$\lim_{z\to-oo}\phi(z) = 0$

$\lim_{z\to oo}\phi(z) = 1$

$$\phi'(z) = \frac{e^{-z}}{(1+e^{-z})^2} = \frac{1}{1+e^{-z}}\cdot\frac{e^{-z}}{1+e^{-z}} = \phi(z)(1-\phi(z))$$

1.3 Model

Với dòng thứ i trong bảng dữ liệu, gọi $x_1^{(i)}$ là lương và $x_2^{(i)}$ là thời gian làm việc của hồ sơ thứ i.

$p(x^{(i)}=1) = \hat{y_i}$ là xác suất mà model dự đoán hồ sơ thứ i được cho vay.

$p(x^{(i)}=0) = 1-\hat{y_i}$ là xác suất mà model dự đoán hồ sơ thứ i không được cho vay.

$\Rightarrow p\left(x^{(i)}=1\right) + p\left(x^{(i)}=0\right) = 1$

Hàm sigmoid: $\phi(z) = \dfrac{1}{1+e^{-z}}$

Công thức của logistic regression là:

$$\widehat{y_i} = \phi\left(w_0 + w_1x_1^{(i)} + w_2x_2^{(i)}\right) = \frac{1}{1+e^{-\left(w_0+w_1x_1^{(i)}+w_2x_2^{(i)}\right)}}$$

1.4 Loss function

Ta vẫn dùng hàm mất mát để đánh giá độ tốt của model với $\hat{y}$ (giá trị dự đoán của mô hình) càng gần y (giá trị thực) càng tốt.

- Nếu hồ sơ thứ i được cho vay với $y_i = 1$ thì $\hat{y_i}$ càng gần 1 càng tốt.
- Nếu hồ sơ thứ i không được cho vay với $y_i = 0$ thì $\hat{y_i}$ càng gần 0 càng tốt.

Với mỗi điểm $(x^{(i)}, y_i)$ ta có công thức của loss function (binary_crossentropy) là:

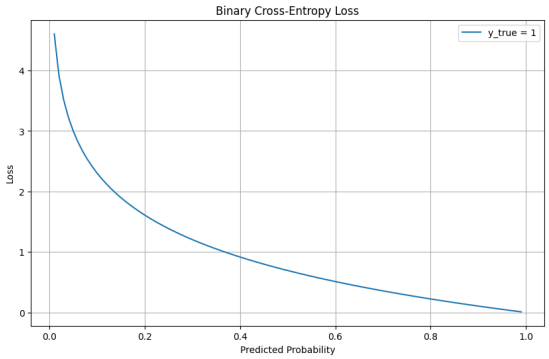
$$L = -\left(y_i * log\left(\widehat{y_i}\right) + (1-y_i) * log\left(1-\widehat{y_i}\right)\right)$$

Hàm loss function trên toàn bộ dữ liệu:

$$J = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i * log\left(\widehat{y_i}\right) + (1-y_i) * log\left(1-\widehat{y_i}\right)\right)$$

1.4.1 Đánh giá

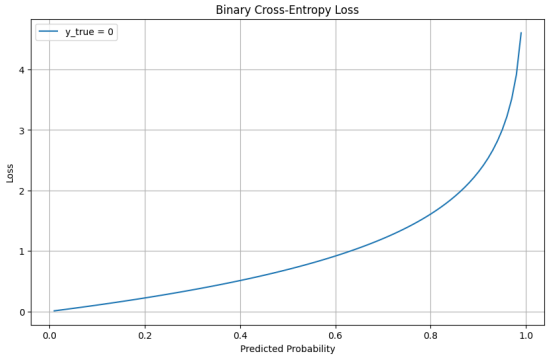
- Nếu $y_i = 1 => L = -log(\hat{y_i})$



Nhận xét:

- Loss "giảm" dần từ 0 về 1.
- Khi model dự đoán $\hat{y_i}$ gần 1, gần với giá trị thực y_i thì Loss nhỏ, xấp xỉ 0.
- Khi model dự đoán $\hat{y_i}$ gần 0, giá trị dự đoán ngược lại với giá trị thực y_i thì Loss rất lớn.

- Nếu $y_i = 0 => L = -log(1-\hat{y_i})$



Nhận xét:

- Loss tăng dần từ 0 lên 1.
- Khi model dự đoán $\hat{y_i}$ gần 0, gần với giá trị thực y_i thì Loss nhỏ, xấp xỉ 0.
- Khi model dự đoán $\hat{y_u}$ gần 1, giá trị dự đoán ngược lại với giá trị thực y_i thì Loss rất lớn.

Vậy loss function càng nhỏ thì model dự đoán càng gần với giá trị thực.
=> Tối ưu model Logistic Regression bằng cách tìm giá trị nhỏ nhất của loss function.

1.5 Chain rule

Quy tắc dây chuyền (chain rule) là một công thức biểu thị đạo hàm của thành phần của hai hàm khả vi f và g theo đạo hàm của f và g. Cụ thể:
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: nếu $z = (2x+1)^2 \Rightarrow z = f(g(x))$
với $f(x) = x^2,\ g(x) = 2x+1$

Áp dụng chain rule ta có:
$\frac{dz}{dx} = \frac{dz}{dy} * \frac{dy}{dx} = \frac{d(2x+1)^2}{d(2x+1)} * \frac{d(2x+1)}{d(x)}$
$= 2 * (2x+1) * 2 = 4 * (2x+1)$

Áp dụng chain rule để tính hàm sigmoid: $\phi(z) = \frac{1}{1+e^{-z}}$

$\Rightarrow \quad \phi(z) \quad = \quad f(g(z)) \quad \text{với} \quad f(z) \quad = \quad \frac{1}{1+z}, \quad g(z) \quad = \quad e^{-z}$

$\Rightarrow \frac{d(\phi(z))}{dz} = \frac{d\left(\frac{1}{1+e^{-z}}\right)}{d\left(e^{-z}\right)} * \frac{d\left(e^{-z}\right)}{dz} = \frac{-1}{\left(1+e^{-z}\right)^2} * \left(-e^{-z}\right)$

$= \frac{e^{-z}}{\left(1+e^{-z}\right)^2} = \frac{e^{-z}}{1+e^{-z}} * \frac{1}{1+e^{-z}} = \phi(z)(1-\phi(z))$

1.6 Áp dụng Gradient Descent

Ta tính đạo hàm của loss function với w để có thể áp dụng gradient descent cho bài toán tối ưu loss function.

Với mỗi điểm $\left(x^{(i)}, y_i\right)$ ta có:

$$\begin{cases} L &= -\left(y_i * log\left(\widehat{y_i}\right) + (1-y_i) * log\left(1-\widehat{y_i}\right)\right) \\ \widehat{y_i} &= \phi\left(w_o + w_1 x_1^{(i)} + w_2 x_2^{(i)}\right) = \phi(z) \\ z &= w_o + w_1 x_1^{(i)} + w_2 x_2^{(i)} \end{cases}$$

với $\widehat{y_i}$ là giá trị dự đoán của model và y_i là giá trị thực của dữ liệu.

Áp dụng chain rule ta có:
$\frac{dL}{dw_0} = \frac{dL}{d\widehat{y_i}} * \frac{d\widehat{y_i}}{dz} * \frac{dz}{dw_0}$

$\frac{dL}{d\widehat{y_i}} = -\frac{d(y_i * log(\widehat{y_i}) + (1-y_i) * log(1-\widehat{y_i}))}{d\widehat{y_i}} = -\left(\frac{y_i}{\widehat{y_i}} - \frac{1-y_i}{1-\widehat{y_i}}\right) = \frac{\widehat{y_i} - y_i}{\widehat{y_i} * (1-\widehat{y_i})}$
$\frac{d\widehat{y_i}}{dw_0} = \frac{d\left(\phi\left(w_o + w_1 x_1^{(i)} + w_2 x_2^{(i)}\right)\right)}{dw_0} = \widehat{y_i} * (1-\widehat{y_i})$

$\Rightarrow \frac{dL}{dw_0} = \frac{\widehat{y_i} - y_i}{\widehat{y_i} * (1-\widehat{y_i})} * \widehat{y_i} * (1-\widehat{y_i}) = \widehat{y_i} - y_i$

$\frac{d\widehat{y_i}}{dw_1} = x_1^{(i)} * (\widehat{y_i} - y_i) \Rightarrow \frac{dL}{dw_1} = x_1^{(i)} * (\widehat{y_i} - y_i)$

$\frac{d\widehat{y_i}}{dw_2} = x_2^{(i)} * (\widehat{y_i} - y_i) \Rightarrow \frac{dL}{dw_2} = x_2^{(i)} * (\widehat{y_i} - y_i)$

Vậy đạo hàm của loss function trên toàn bộ điểm dữ liệu là:
$\frac{dL}{dw_0} = \frac{1}{N} * (\widehat{y_i} - y_i)$

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

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

1.7 Biểu diễn bài toán dưới dạng ma trận

$$X = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} \\ ... & ... & ... \\ 1 & x_1^{(n)} & x_2^{(1)} \end{bmatrix},\ y = \begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{bmatrix},\ w = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}$$

$\hat{y} = \phi(X * W)$

$J = -\frac{1}{N} * sum(y \otimes log(\hat{y}) + (1-y) \otimes log(1-\hat{y}))$

$\frac{dJ}{dw} = \frac{1}{N} * X^T * (\hat{y} - y)$

Sau khi thực hiện thuật toán gradient descent ta sẽ tìm được các trọng số w_0, w_1, w_2. Với mỗi hồ sơ mới $x^{(k)}$ ta sẽ tính được xác suất cho vay $\widehat{y_k} = \phi\left(w_o + w_1 x_1^{(k)} + w_2 x_2^{(k)}\right)$ rồi cho sánh với ngưỡng cho vay t của công ty. Nếu $\hat{y_k} \geq t$ thì cho vay, ngược lại không cho vay.

1.8 Một vài tính chất của Logistic Regression

1. Logistic Regression được sử dụng nhiều cho bài toán Classification.

Mặc dù có tên là "regression" tuy nhiên mô hình logistic regression được sử dụng nhiều trong các bài toán phân loại (classification). Sau khi tìm được mô hình, việc xác định class y cho một điểm dữ liệu x được xác định bằng việc tính xác suất.

$$P(y=1|x;w); P(y=0|x,w)$$

Nếu biểu thức trái lớn hơn, ta kết luận điểm dữ liệu thuộc class 1 và ngược lại.

2. Đường ranh giới tạo bởi logistic regression là một siêu phẳng

Giả sử các điểm dữ liệu thuộc class 1 có xác suất đầu ra lớn hơn 0.5:

$$\begin{aligned} &\widehat{y_i} \geqslant 0.5 \\ \Leftrightarrow &\frac{1}{1+e^{-\left(w_o + w_1 x_1^{(i)} + w_2 x_2^{(i)}\right)}} \geqslant 0.5 \\ \Leftrightarrow &2 \geqslant 1 + e^{-\left(w_o + w_1 x_1^{(i)} + w_2 x_2^{(i)}\right)} \\ \Leftrightarrow &e^{-\left(w_o + w_1 x_1^{(i)} + w_2 x_2^{(i)}\right)} \leqslant 1 = e^0 \\ \Leftrightarrow &w_o + w_1 x_1^{(i)} + w_2 x_2^{(i)} \geqslant 0 \end{aligned}$$

Tương tự: $\widehat{y_i} < 0.5 \Leftrightarrow w_o + w_1 x_1^{(i)} + w_2 x_2^{(i)} < 0$

Tập hợp các điểm thuộc class 1 tạo thành *nửa không gian* $\hat{y_i} > 0$, tập hợp các điểm thuộc lớp 0 tạo thành nửa không gian còn lại. Ranh giới giữa 2 lớp đó là siêu phẳng $w_0 + w_1 * x_1 + w_2 * x_2 = 0$.