Definição & Exemplos

Na regressão linear, utilizamos variáveis explicativas para tentar prever uma variável resposta, ambas tendo valores contínuos. Já a regressão logística é aplicada quando queremos atribuir um rótulo a partir de valores, sendo esses contínuos ou não. Por exemplo, em um hospital poderíamos ter pacientes com o rótulo “doente” ou “não doente” baseando-se em atributos como temperatura corporal e pressão sanguínea.

- variável resposta y - rótulo previsto;
- variável(is) explicativa(s) X - atributo(s) previsor(es).

Em outras palavras, na regressão logística, calcularemos a probabilidade de um indivíduo pertencer a uma classe ou não a partir de um ou mais valores apresentados, de forma a realizarmos uma classificação binária.

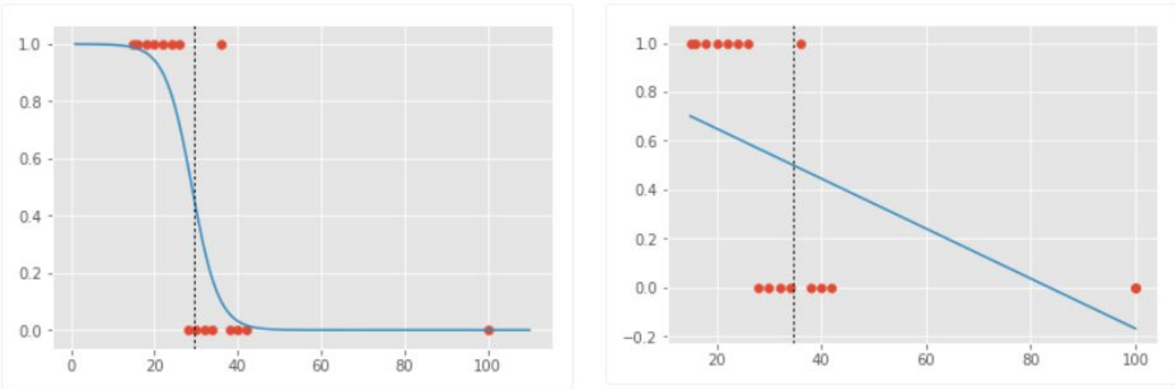

Figura 1[1]: Tentativa de classificação por regressão logística (esquerda) e regressão linear (direita).

Para estimar os valores na regressão, utilizamos a seguinte equação:

$$y = b_0 + b_1 * x_1 \rightarrow y = a * x + b$$

onde y é o atributo que queremos prever, b_0 é uma constante angular, b_1 é um coeficiente e x_1 é a variável explicativa em questão. Perceba que na regressão logística a nossa reta sofre um “achatamento” para que ela retorne apenas valores entre 0 e 1. Dessa forma, para que tenhamos como resultado um valor que represente uma probabilidade, o valor de y será passado para outra função, a qual é chamada de função logística ou sigmóide.

Suponhamos que haja uma base de crédito com os seguintes atributos previsores:

- Histórico de pagamento = Bom (0)
- Dívida = Média (0)
- Garantias = Nenhuma (1)
- Renda anual => 30 (2)

Vamos elencar coeficientes b_n para cada atributo. Nesse caso, temos que:

[1] Fonte: https://matheusfacure.github.io/2017/02/25/regr-log/

- b_0 = -0,12
- b_1 = -0,71
- b_2 = 0,24
- b_3 = Garantias = -0,54
- b_4 = Renda anual = 1,07

Calculando y, obtemos:

$$y \;=\; -0,12 \;+\; (b_1 * Histórico\ de\ pagamento) \;+\; (b_2 *\ Dívida) \;+\; (b_3 * Garantias) \;+\; (b_4 * Renda\ anual)$$
$$y \;=\; -0,12 \;+\; (-0,71 * 0) \;+\; (0,24 * 0) \;+\; (-0,54 * 1) \;+\; (1,07 * 2)$$
$$y \;=\; 1,48$$

Agora, usaremos y como entrada para a função sigmoid. Logo,

$$f(y) \;=\; \frac{1}{1+e^{-y}}$$
$$f(1,48) \;=\; \frac{1}{1+e^{-1,48}}$$
$$f(1,48) \;=\; 0,8145$$

Portanto, o usuário com esse perfil teria 81% de probabilidade de pagar determinado empréstimo.

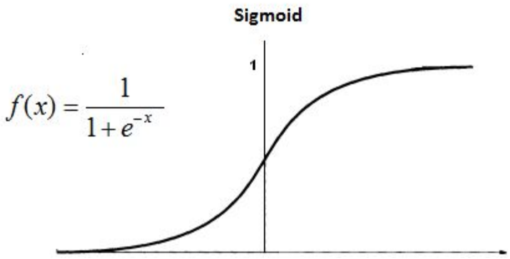

Figura 2[2]: Função sigmoid ou logística

[2] A função sigmoid é utilizada em redes neurais em problemas de classificação binária.

Exemplo gráfico & Cálculo do erro

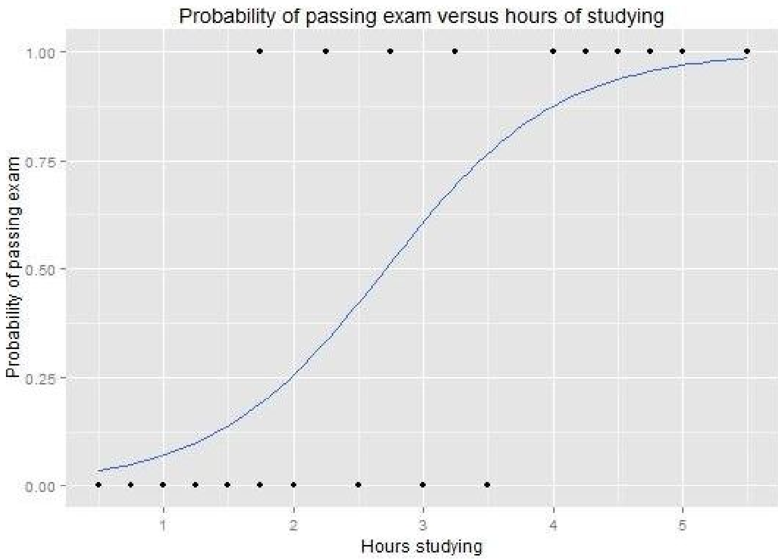

Figura 3: Probabilidade de passar no exame a partir das horas de estudo.

Pelo gráfico, podemos perceber que quanto mais uma pessoa estuda para um exame, ela tem uma probabilidade maior de passar.

Pelo fato de estarmos trabalhando em um problema de classificação, a função MSE (Mean Squared Error) não será mais útil, pois os resíduos não serão mais medidos em termos de distância. Tendo isso em vista, utilizaremos a função de custo chamada Binary Cross Entropy.

$$L = -\sum_{i} y_i log\tilde{y}_i + (1 - y_i)\log{(1 - \tilde{y}_i)}$$