# Regressão Logística

## 1. Introdução

A regressão logística é um modelo matemático usado para determinar a probabilidade de um evento ocorrer apresentando a relação entre recursos e calculando a probabilidade de um determinado resultado. O termo regressão deve-se ao processo de cálculo da regressão logística, pois a saída do modelo será um valor contínuo sobre uma curva de probabilidades e a partir de um **limiar de corte** chamado de *cutoff* ou *threshold* se determina uma saída na forma de classificação. Desta maneira, por conhecermos a variável dependente, a regressão logística se encaixa como um tipo de aprendizado supervisionado.

Ela pode ser expressa pela relação matemática:

$$Y = \left( \frac{1}{1 + e^{-(a+bx)}} \right)$$

Vale destacar que o trecho dentro dos parênteses:

$$a + bx$$

remete a regressão linear, que é utilizada para encontrar um padrão em dados contínuos, porém é a expressão utilizada na relação matemática da regressão logística para formar a curva “S”, em que os dados são classificados.

Com requisito para aplicar a regressão logística, a variável dependente (alvo) obrigatoriamente precisa ser categórica. Por sua vez, as variáveis explicativas (características) devem ser independentes entre si para evitar multicolinearidade.

## 2. Função Logística

A regressão logística pode ser expressa como:

$$log \left( \frac{p(x)}{1 - p(x)} \right) = \beta + \beta_1 x$$

Em que, o lado esquerdo é chamado de função logit ou logit-odds e,

$$\left(\frac{p(x)}{1-p(x)}\right)$$

é chamado de odds.

As probabilidades significam a razão entre a probabilidade de sucesso e a probabilidade de fracasso. Portanto, na regressão logística, a combinação linear de entradas é mapeada para o log(odds), em que a saída é igual a 1.

Se fizermos a inversão da função acima, obtemos:

$$P(x) = \frac{e^{\beta_0} + \beta_1 x}{1 + e^{\beta_0} + \beta_1 x}$$

Esta relação matemática é conhecida como a função sigmoide e fornece uma curva em forma de S atribuindo um valor de probabilidade variando de:

$$0 < p < 1$$

A curva sigmoide ou função logística é a função matemática da regressão logística que representa a transição de estados entre 0 e 1. A partir desta curva, podem ser observadas as probabilidades envolvidas de cada ponto composto na curva S.

De acordo com o valor de corte estabelecido para a curva S, a probabilidade é transformada em um valor absoluto no qual permite classificar nosso resultado. Por exemplo, se o valor de corte para a curva S for 0.5, probabilidades menores do que 0.5 geram o valor absoluto 0 em que a função não é ativada e probabilidades maiores ou iguais a 0.5 geram o valor absoluto 1, ativando a função do nosso classificador.

Abaixo, através da Figura 1 - Função Sigmoid, é possível visualizar a função sigmoid considerando o exemplo compartilhado.

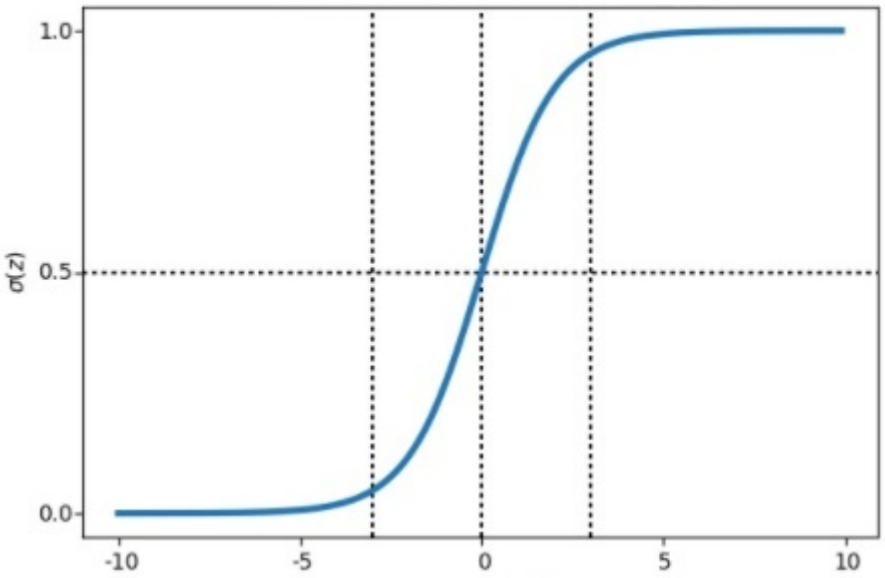


**Figura 1** – Função Sigmoid

Geometricamente, a probabilidade envolvida pela regressão logística indica o quão distante o exemplo está do hiperplano. O hiperplano é um separador linear. Separadores lineares, além de serem representados por hiperplanos, também podem ser representados por retas ou planos.

Abaixo, através da Figura 2 – Hiperplano (Separador Linear), é possível verificar um exemplo em que temos uma superfície de decisão de uma regressão logística com dois atributos onde podemos visualizar o hiperplano estabelecendo a separação linear.

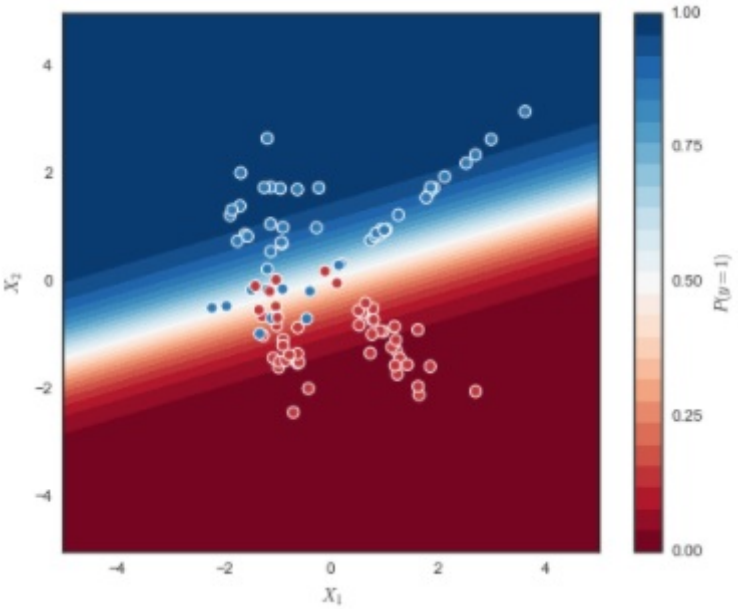


**Figura 2** – Hiperplano (Separador Linear)

### Tipos da Regressão Logística

Existem três tipos básicos de regressão logística, são elas:

- **Regressão logística binomial:** É quando a variável dependente pode assumir apenas dois valores como resultado. Por exemplo: verdadeiro ou falso; 0 ou 1; positivo ou negativo.
- **Regressão logística multinomial:** É quando a variável dependente pode assumir três ou mais valores, não tendo uma relação de ordem entre os mesmos. Por exemplo, se quiséssemos classificar a previsão do tempo, poderíamos classificar entre ensolarado, nublado ou chuvoso com bases na análise de padrões das variáveis independentes. Outro exemplo poderia ser identificar qual tipo de prato que os clientes de um determinado restaurante preferem, assim, poderíamos classificar os pratos entre carne, vegano ou vegetariano.
- **Regressão logística ordinal:** Semelhante a regressão multinomial, podendo haver três ou mais valores. No entanto, neste tipo há uma ordem que as medições seguem. Um bom exemplo poderia ser classificar um hotel dentro de uma escala de 1 a 5 estrelas.

## 3. Campos de Aplicação

Existem diversos campos de aplicação onde a regressão logística pode ser utilizada, abaixo compartilho alguns exemplos:

- **Comércio eletrônico:** A regressão logística é utilizada neste campo para validar a eficácia de campanhas publicitárias e promocionais. Desta maneira, com base nos resultados obtidos pela previsão da regressão logística, as empresas podem otimizar estratégias e atingir metas de negócio com redução de despesas e perdas.

- **Saúde:** Pode ser aplicada na saúde para prever a possibilidade de um determinado paciente desenvolver uma doença ou não. Por exemplo, doenças cardíacas podem ser previstas com base em variáveis como peso, altura, sexo, idade e outros fatores genéticos.
- **Setor financeiro:** Fortemente utilizada em linhas de crédito. Como por exemplo, para identificar a inadimplência de um cliente, a emissão ou não de um cartão crédito e identificar a probabilidade de um cliente pagar suas despesas. Ainda, a regressão logística pode ser utilizada para saber as características de um produto financeiro e se o mesmo pode ser ofertado para um determinado cliente.
- **Teste de produto:** A aplicação neste campo é comumente utilizada para testar o sucesso ou falha de produtos. Com base nas características (variáveis) do protótipo pode ser identificado se o mesmo funcionará de maneira satisfatória em cenários de uso do produto, como também avaliar a qualidade de funcionalidades do mesmo.
- **Política:** Pode ser aplicada para prever o resultado eleitoral. É um tipo classificação oferecido pela regressão logística que considera padrões de eleições passadas com variáveis atuais do candidato como partido, idade, sexo, propostas sociais e ambientais, local ou endereço residencial e outras para criar a previsão de voto.
- ***Marketing*:** A regressão logística pode ser utilizada para identificar o potencial de interesse do cliente em uma nova linha de produtos, taxa de conversão de venda, aderência de assinaturas ou cancelamento da mesma. Além disso, a regressão logística ajuda a maximizar o retorno sobre o investimento (ROI) em campanhas de *marketing*, um benefício para os resultados financeiros no longo prazo para as empresas.

## 4. Aplicação Prática

A implementação da Regressão Logística é feita utilizando o *Python* e a biblioteca scikit-learn, seguindo o exemplo abaixo para a importação da biblioteca para o modelo:

```
from sklearn.linear_model import LogisticRegression
```

Carregada a biblioteca para o ambiente onde está sendo desenvolvido o código, os próximos passos funcionam de forma análoga para os demais modelos de *Machine Learning*, sendo estes passos **instanciar o modelo**, **treinar o modelo** (*fit*) e **gerar as novas predições** (*predict*):

```
# Instanciar o Modelo
model = LogisticRegression()

# Fit do Modelo
model.fit(X_train, # o X são os atributos (features) do dataset de treino
          y_train) # o y é a variável resposta (target) procurada do dataset
```

```
de treino

# Predict do Modelo
y_pred = model.predict(X_test) # dados de teste para gerar novas predições
```

## Materiais Complementares

- Let's Code channel at Youtube; O que é classificação e regressão em machine learning?
- LARA, Evandro de Avila e. Polytomous ordinal *logistic regression*: Assessing the potential of Clonostachys rosea in biocontrol of Botrytis cinerea. 2012. 52 f. Dissertação (Mestrado em Estatística Aplicada e Biometria) - Universidade Federal de Viçosa, Viçosa, 2012.
- BATISTA, António Sarmento. Regressão Logística: Uma introdução ao modelo estatístico-Exemplo de aplicação ao Revolving Credit. Vida Econômica Editorial, 2015.
- FÁVERO, Luiz Paulo et al. Análise de dados: modelagem multivariada para tomada de decisões. 2009.
- GONZALEZ, Leandro de Azevedo. FÁVERO, Luiz Paulo et al. Análise de dados: modelagem multivariada para tomada de decisões. 2009. 2018. Monografia (FÁVERO, Luiz Paulo et al. Análise de dados: modelagem multivariada para tomada de decisões. 2009.) - Universidade Federal do Maranhão, [S. l.], 2018.

## Referências

James, Gareth, et al. An Introduction to Statistical Learning: With Applications in R. Alemanha, Springer New York, 2013.

Tópico anterior

Próximo Tópico