Regressão Logística

April 29, 2025

0.1 Regressão Logística

A regressão logística é usada para classificação binária, onde usamos a função sigmoide , que recebe entradas como variáveis independentes e produz um valor de probabilidade entre 0 e 1.

Por exemplo, temos duas classes: Classe 0 e Classe 1.

Se o valor da função logística para uma entrada for maior que 0,5 (valor limite), então ela pertence à Classe 1; caso contrário, ela pertence à Classe 0.

Ela é chamada de regressão porque é a extensão da regressão linear , mas é usada principalmente para problemas de classificação.

- A regressão logística prevê a saída de uma variável dependente categórica. Portanto, o resultado deve ser um valor categórico ou discreto.
- Pode ser Sim ou Não, 0 ou 1, verdadeiro ou falso, etc., mas em vez de fornecer o valor exato como 0 e 1, ele fornece os valores probabilísticos que estão entre 0 e 1.
- Na regressão logística, em vez de ajustar uma linha de regressão, ajustamos uma função logística em forma de "S", que prevê dois valores máximos (0 ou 1).

0.2 Função Sigmóide

A função sigmoide é uma função matemática usada para mapear os valores previstos para probabilidades.

Ela mapeia qualquer valor real em outro valor dentro de um intervalo de 0 e 1. O valor da regressão logística deve estar entre 0 e 1, que não pode ultrapassar esse limite, então ela forma uma curva como a forma "S".

A curva em forma de S é chamada de função sigmoide ou função logística.

Na regressão logística, usamos o conceito de valor limite, que define a probabilidade de 0 ou 1.

Por exemplo, valores acima do valor limite tendem a 1, e um valor abaixo do valor limite tende a 0.

0.3 Tipos de Regressão Logística

A Regressão Logística pode ser classificada em três categorias:

1. **Binomial**: Na regressão logística binomial, pode haver apenas dois tipos possíveis de variáveis dependentes, como 0 ou 1, aprovado ou reprovado, etc.
2. **Multinomial**: Na regressão logística multinomial, pode haver 3 ou mais tipos possíveis não ordenados da variável dependente, como "gato", "cachorro" ou "ovelha"

3. **Ordinal**: Na regressão logística ordinal, pode haver 3 ou mais tipos ordenados possíveis de variáveis dependentes, como “baixo”, “médio” ou “alto”.

0.4 A Regressão Logística e seu funcionamento

O modelo de regressão logística transforma a saída de valor contínuo da função de regressão linear em saída de valor categórico usando uma função sigmoide, que mapeia qualquer conjunto de variáveis independentes de valor real de entrada em um valor entre 0 e 1.

Essa função é conhecida como função logística.

Sejam os recursos de entrada independentes:

$$\begin{pmatrix} x_{11} & \cdots & x_{1m} \\ x_{21} & \cdots & x_{2m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix}$$

e a variável dependente é Y tendo apenas valor binário, ou seja, 0 ou 1.

$$Y = \begin{cases} 0 \text{ se classe 1} \\ 1 \text{ se classe 2} \end{cases}$$

$$z = \left(\sum_{i=1}^{n} w_i x_i \right) + b$$

[1]:
```
from IPython.display import Image
```

Onde x_i é a i-ésima observação de X, $w_i = [w_1, w_2, w_3 \ldots, w_m]$ são os pesos ou coeficiente, e b é o termo de viés, também conhecido como intercepto. Isso pode ser representado simplesmente como o produto escalar do peso e do viés.

$$z = w \cdot X + b$$

tudo o que discutimos acima é a regressão linear .

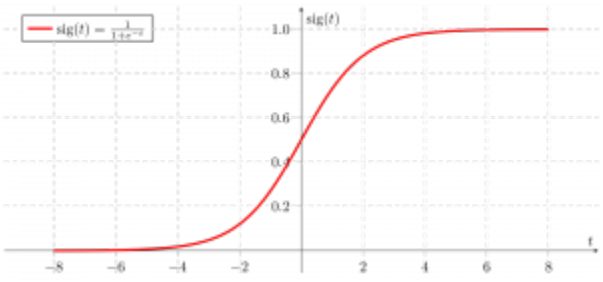
0.5 Função Sigmóide

Agora usamos a função sigmoide onde a entrada será z e encontramos a probabilidade entre 0 e 1, ou seja, y previsto.

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

[2]:
```
Image(filename='sigmoid-function-300x138.png')
```

[2]:



Conforme mostrado acima, a função sigmoide converte os dados da variável contínua em probabilidade , ou seja, entre 0 e 1.

- $\sigma(z)$ tende à 1 quando $z \to \infty$
- $\sigma(z)$ tende à 0 quando $z \to -\infty$
- $\sigma(z)$ é sempre limitado entre 0 e 1

onde a probabilidade de ser uma classe pode ser medida como:

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

0.6 Equação de regressão logística

$\frac{p(x)}{1-p(x)} = e^z \log\left[\frac{p(x)}{1-p(x)}\right] = z \log\left[\frac{p(x)}{1-p(x)}\right] = w{\cdot}X{+}b$Aplicando exponencial em ambos os lados$\frac{p(x)}{1-p(x)} = e^{\iota}$

então a equação final de regressão logística será:

$$p(X; b, w) = \frac{e^{w \cdot X + b}}{1 + e^{w \cdot X + b}} = \frac{1}{1 + e^{w \cdot X + b}}$$

0.7 Função de verossimilhança para regressão logística

As probabilidades previstas serão:

para $y = 1$As probabilidades previstas serão: $p(X; b, w) = p(x)$

para $y = 0$ As probabilidades previstas serão: $1 - p(X; b, w) = 1 - p(x)$

$$L(b, w) = \prod_{i=1}^{n} p(x_i)^{y_i} \left(1 - p(x_i)\right)^{1-y_i}$$

Aplicando logaritmo natural em ambos os lados

$\log\left(L(b, w)\right) = \sum_{i=1}^{n} y_i \log p(x_i) + (1 - y_i) \log\left(1 - p(x_i)\right)$ $\log\left(L(b, w)\right) = \sum_{i=1}^{n} y_i \log p(x_i) + \log\left(1 - p(x_i)\right) - y_i \log\left(1 - p($

0.8 Gradiente da função de log-verossimilhança

Para encontrar as estimativas de máxima verossimilhança, diferenciamos em relação à w,

$$\frac{\partial J(L(b,w))}{\partial w_j} = -\sum_{i=n}^{n} \frac{1}{1+e^{w\cdot x_i+b}} e^{w\cdot x_i+b} x_{ij} + \sum_{i=1}^{n} y_i x_{ij} \frac{\partial J(L(b,w))}{\partial w_j} = -\sum_{i=n}^{n} p(x_i;b,w) x_{ij} + \sum_{i=1}^{n} y_i x_{ij} \frac{\partial J(L(b,w))}{\partial w_j} = \sum_{i=}^{n}$$

0.9 Implementação de código para regressão logística

0.9.1 Regressão logística binomial:

A variável alvo pode ter apenas 2 tipos possíveis: "0" ou "1", que pode representar "vitória" vs "perda", "aprovado" vs "reprovado", "morto" vs "vivo", etc., neste caso, funções sigmoides são usadas, o que já foi discutido acima.

Importando bibliotecas necessárias com base no requisito do modelo. Este código Python mostra como usar o conjunto de dados de câncer de mama para implementar um modelo de Regressão Logística para classificação.

[8]:
```
# import the necessary libraries
from sklearn.datasets import load_breast_cancer
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

# load the breast cancer dataset
X, y = load_breast_cancer(return_X_y=True)

# split the train and test dataset
X_train, X_test,\
    y_train, y_test = train_test_split(X, y,
                                       test_size=0.20,
                                       random_state=23)
# LogisticRegression
clf = LogisticRegression(max_iter=2500, random_state=0)
clf.fit(X_train, y_train)

# Prediction
y_pred = clf.predict(X_test)

acc = accuracy_score(y_test, y_pred)
print("Logistic Regression model accuracy (in %):", acc*100)
```

```
Logistic Regression model accuracy (in %): 96.49122807017544
```

0.10 Regressão Logística Multinomial

A variável alvo pode ter 3 ou mais tipos possíveis que não são ordenados (ou seja, os tipos não têm significância quantitativa), como "doença A" vs "doença B" vs "doença C".

Neste caso, a função softmax é usada no lugar da função sigmoide. A função softmax para classes K será:

$$(softmax)(z_i)\frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}}$$

Aqui, K representa o número de elementos no vetor z, e i, j itera sobre todos os elementos no vetor.

Então a probabilidade para a classe c será dada por:

$$P\left(Y = c|\overrightarrow{X} = x\right) = \frac{e^{w_c \cdot x + b_c}}{\sum_{k=1}^{K} e^{w_k \cdot x + b_w}}$$

Na Regressão Logística Multinomial, a variável de saída pode ter mais de duas saídas discretas possíveis . Considere o Conjunto de Dados Digit.

[13]:
```
from sklearn.model_selection import train_test_split
from sklearn import datasets, linear_model, metrics

# load the digit dataset
digits = datasets.load_digits()

# defining feature matrix(X) and response vector(y)
X = digits.data
y = digits.target

# splitting X and y into training and testing sets
X_train, X_test,\
    y_train, y_test = train_test_split(X, y,
                                       test_size=0.4,
                                       random_state=1)

# create logistic regression object
reg = linear_model.LogisticRegression(max_iter=4000)

# train the model using the training sets
reg.fit(X_train, y_train)

# making predictions on the testing set
y_pred = reg.predict(X_test)

# comparing actual response values (y_test)
# with predicted response values (y_pred)
print("Logistic Regression model accuracy(in %):",
      metrics.accuracy_score(y_test, y_pred)*100)
```

```
Logistic Regression model accuracy(in %): 96.52294853963839
```

0.10.1 Avaliando o modelo de regressão logística

Podemos avaliar o modelo de regressão logística usando as seguintes métricas:

Acurácia: A precisão fornece a proporção de instâncias classificadas corretamente.

$$Acurcia = \frac{Verdadeiros\ Positivos + Verdadeiros\ Negativos}{Total}$$

Precisão: A precisão se concentra na exatidão das previsões positivas.

$$Preciso = \frac{Verdadeiros\ Positivos}{Verdadeiros\ Positivos + Falsos\ Positivos}$$

Recall (recuperação) (Sensibilidade ou Taxa de Verdadeiros Positivos): O recall mede a proporção de instâncias positivas previstas corretamente entre todas as instâncias positivas reais.

$$Recuperao = \frac{Verdadeiros\ Positivos}{Verdadeiros\ Positivos + Falsos\ Negativos}$$

Pontuação F1: A pontuação F1 é a média harmônica de precisão e recuperação.

$$F1\ Score = \frac{Preciso * Recuperao}{Preciso + Recuperao}$$

Área sob a curva característica de operação do receptor (AUC-ROC): A curva ROC desenha a taxa de verdadeiro positivo em relação à taxa de falso positivo em vários limites. A AUC-ROC mede a área sob essa curva, fornecendo uma medida agregada do desempenho de um modelo em diferentes limites de classificação.

Área sob a curva de precisão-recall (AUC-PR): Semelhante à AUC-ROC, a AUC-PR mede a área sob a curva de precisão-recall, fornecendo um resumo do desempenho de um modelo em diferentes compensações entre precisão e recall. Compensação de precisão-recall na definição de limiar de regressão logística

0.11 Diferenças entre regressão linear e logística

A diferença entre regressão linear e regressão logística é que a saída da regressão linear é o valor contínuo que pode ser qualquer coisa, enquanto a regressão logística prevê a probabilidade de uma instância pertencer a uma determinada classe ou não.

Regressão Linear	Regressão Logística
A regressão linear é usada para prever a variável dependente contínua usando um determinado conjunto de variáveis independentes.	A regressão logística é usada para prever a variável dependente categórica usando um determinado conjunto de variáveis independentes.
A regressão linear é usada para resolver problemas de regressão.	É usado para resolver problemas de classificação.
Prevemos o valor de variáveis contínuas	Prevemos valores de variáveis categóricas

Regressão Linear	Regressão Logística
encontramos a melhor linha de ajuste.	encontramos a Curva S.
O método de estimativa dos mínimos quadrados é usado para estimar a precisão.	O método de estimativa de máxima verossimilhança é usado para estimativa de precisão.
A saída deve ser um valor contínuo, como preço, idade, etc.	A saída deve ser um valor categórico, como 0 ou 1, Sim ou não, etc.
Exigia uma relação linear entre variáveis dependentes e independentes.	Não é necessária uma relação linear.
Pode haver colinearidade entre as variáveis independentes.	Deve haver pouca ou nenhuma colinearidade entre variáveis independentes.

[ ]: