Algoritmos Supervisionados – Regressão Logística

ROCKETSEAT

Plataforma completa de aprendizado contínuo em programação.

#BoostingPeople

rocketseat.com.br

Algoritmos Supervisionados

Regressão Logística

O objetivo deste módulo é apresentar o algoritmo de Regressão Logística, usado amplamente para classificação binária e multiclasse que entrega, além da classificação, a probabilidade da instância pertencer a classe de interesse. Como de costume, faremos um projeto prático completo, desde o EDA, otimização de hiperparâmetros até a entrega do modelo através de uma API em Container Docker.

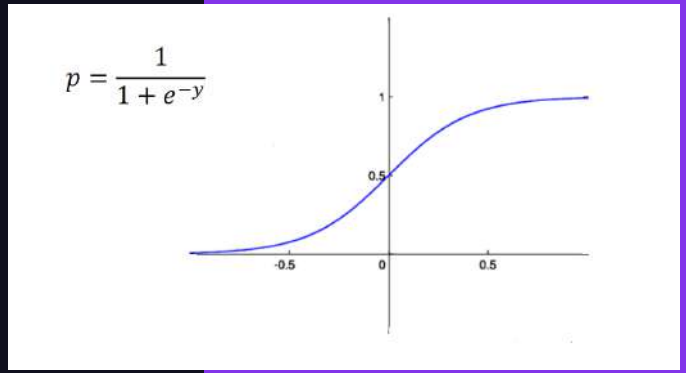
Agenda

- O que é o algoritmo de Regressão Logística?
- Métricas de Classificação
- Projeto – Regressão Logística

O que é o algoritmo de Regressão Logística?

A regressão logística é um algoritmo de aprendizado supervisionado utilizado principalmente para problemas de classificação binária, embora possa ser estendido para problemas de classificação multiclasse. Apesar do nome "regressão", a regressão logística é usada para classificação, não para predição de valores contínuos.

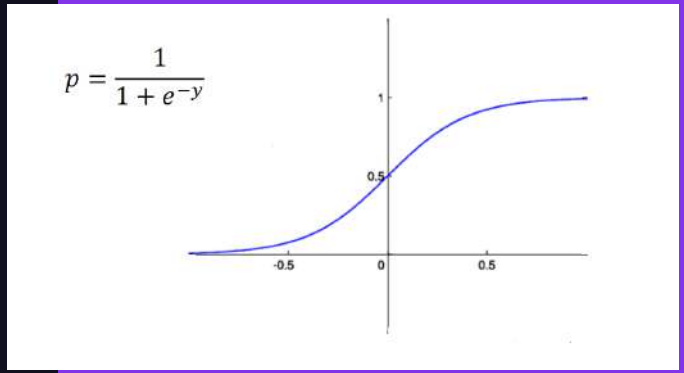
O algoritmo funciona modelando a probabilidade de que uma determinada amostra pertença a uma classe específica.



O que é o algoritmo de Regressão Logística?

$$p = \frac{1}{1+e^{-y}}$$

Função Sigmoide



Onde p representa a probabilidade de uma instância pertencer a uma classe analisada ou evento (geralmente positivo)

E y representa um número real dado pela combinação linear dos atributos utilizados na predição, derivado da regressão linear.

O que é o algoritmo de Regressão Logística?

Para realizar uma classificação a partir de uma probabilidade se define um limiar de decisão (threshold), onde registros cuja probabilidade ultrapasse esse limiar são classificados como sendo da classe 1, caso contrário serão considerados da classe 0. Desta forma, a regressão logística não só é capaz de classificar instâncias mas também de informar a certeza/incerteza associada com a classificação, através do valor da probabilidade calculada.

O que é o algoritmo de Regressão Logística?

$$y = f(x) = b_0 + b_1 x_1 + \ldots + b_n x_n$$

$$p = \frac{1}{1+e^{-(b_0+b_1 x_1+\ldots+b_n x_n)}}$$

$$e^{b_0+b_1 x_1+\ldots+b_n x_n} = \frac{p}{1-p}$$

$$b_0 + b_1 x_1 + \ldots + b_n x_n = ln(\frac{p}{1-p})$$

Função Logit

$$p = \frac{1}{1+e^{-y}}$$

Função Sigmoide

O que é o algoritmo de Regressão Logística?

$$odds(p) = \frac{p}{1-p}$$

Onde $odds(p)$ é a $chance(p)$ de ocorrência de um evento, que é dada pela probabilidade do evento ocorrer sobre a probabilidade do evento não ocorrer.

Note que é a partir da definição de chance que deriva a capacidade da regressão logística em calcular a probabilidade e incerteza associada a uma instância classificada, por meio da função logit.

O que é o algoritmo de Regressão Logística?

Suponha que tenhamos uma base onde queremos classificar (prever) se um aluno será aprovado no teste final com base na informação se ele dormiu bem na noite anterior e obtivemos os seguintes coeficientes na regressão linear:

$$y = f(x) = b_0 + b_1 x_1 + \ldots + b_n x_n$$

- b_0 ou a = 0,12 (y-intercept)
- b_1 ou b = 0,45 (dormiu bem na noite anterior – 0 ou 1)

O que é o algoritmo de Regressão Logística?

Definir a probabilidade de um aluno passar no teste final, considerando que ele dormiu bem na noite anterior:

$$p = \frac{1}{1 - e^{-y}} \text{ onde } y = 0.12 + (0.45 * 1) = 0.57$$

$$p = \frac{1}{1 - e^{-(0,57)}} = 0,63876318 \; ou \; 63,87\%$$

$$chance(p) = \frac{p}{1 - p} = \frac{0,6387}{1 - 0,6387} = 1,768 \; pra \; 1$$

O que é o algoritmo de Regressão Logística?

O treinamento de um modelo de regressão logística envolve encontrar a função sigmoide que se ajuste aos dados de treinamento de forma mais precisa possível. Em outras palavras, buscamos determinar os coeficientes $(b0, b1, \ldots, bn)$ que reduzam ao máximo os erros de predição, resultando no melhor desempenho do modelo. Para alcançar esse objetivo, empregamos uma função de erro ou custo denominada Entropia Cruzada Binária, cuja definição matemática é a seguinte:

$$H_i = -(y_i \cdot ln(p) + (1 - y_i) \cdot ln(1 - p))$$

Métricas de Classificação

Área sobre a Curva ROC (AUC)

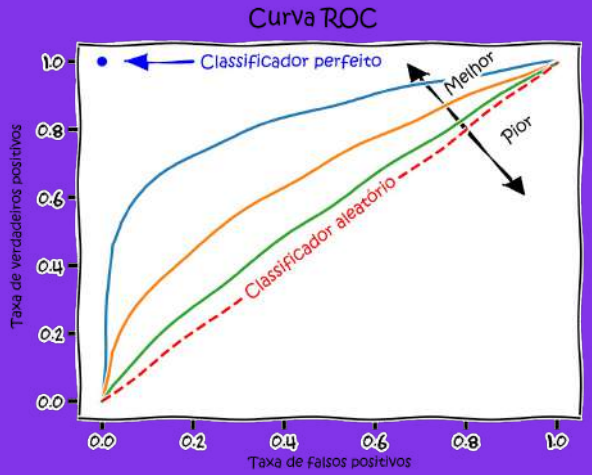
A Curva ROC (Receiver Operating Caracteristics) é uma representação gráfica da taxa de verdadeiros positivos (Recall) em função da taxa de falsos positivos (1 - Especificidade) em diferentes limiares de classificação.

Ela mostra o desempenho do modelo em várias configurações de limiares, permitindo que você visualize como o modelo equilibra a taxa de verdadeiros positivos e a taxa de falsos positivos.

A AUC é a área sob a Curva ROC (Area Under Curve) e varia de 0 a 1, onde:

- AUC = 0.5: O modelo é tão bom quanto um classificador aleatório.
- AUC < 0.5: O modelo é pior que um classificador aleatório.
- AUC > 0.5: O modelo é melhor que um classificador aleatório.

Quanto maior o valor de AUC, melhor é a capacidade discriminativa do modelo. Um AUC próximo de 1 indica um modelo excelente, capaz de distinguir perfeitamente entre as classes.



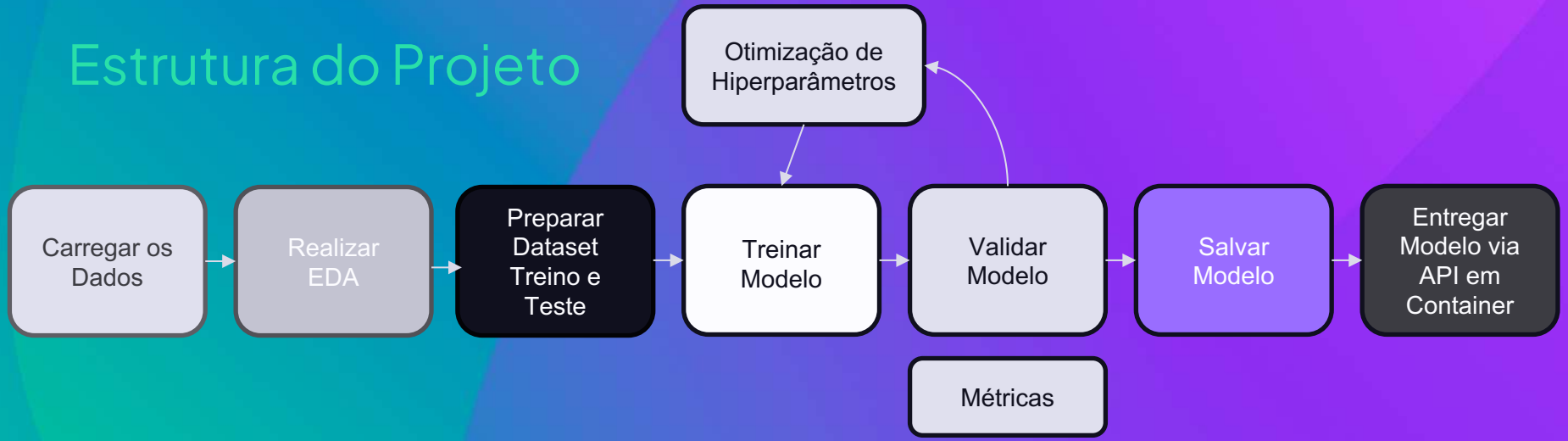
Projeto – Regressão Logística

Uma cooperativa agrícola recebe frutas de diversos produtores e precisa identificar, através de um processo manual de controle de qualidade, quais destas frutas tem condições de serem distribuídas para os consumidores.

Com o objetivo de otimizar este processo de controle de qualidade, a cooperativa gostaria de automatizar esta decisão sobre distribuir ou não um determinado item, com base nas características de cada fruta, que é enviada pelo produtor rural.

Desta forma, iremos desenvolver um classificador binário através do algoritmo de Regressão Logística, para predizer a qualidade da fruta (boa ou ruim), com base nas características da mesma.

Estrutura do Projeto



Code Time ...

rocketseat.com.br