Classificação usando Regressão Logística

Onde estamos e para onde vamos

Nas aulas anteriores, aprendemos sobre **Regressão Linear**, onde a variável alvo de saída y podia assumir um **número infinito de valores possíveis**.

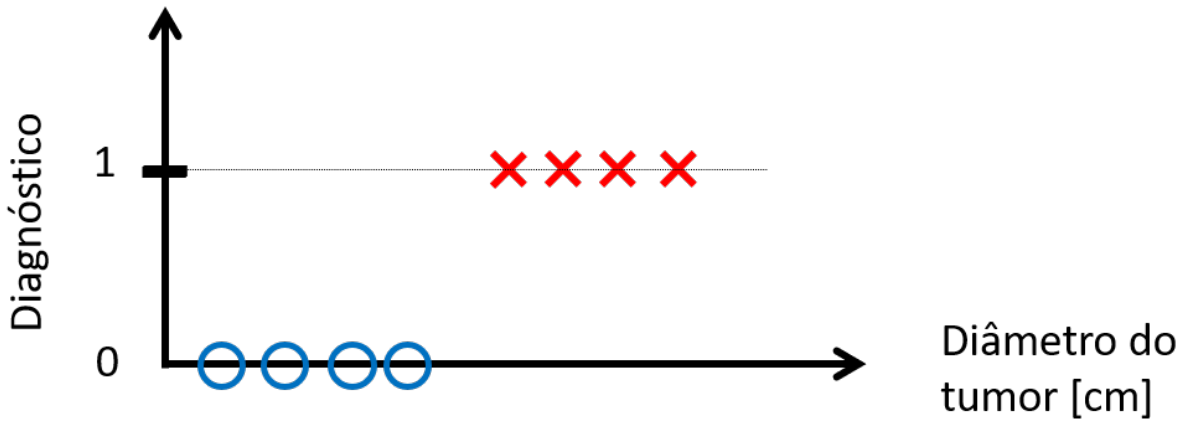
Nesta aula e nas próximas, aprenderemos sobre **Classificação**, onde a variável de saída y pode assumir apenas um pequeno conjunto de valores possíveis, denominados "classes".

Relembrando o que são problema de classificação

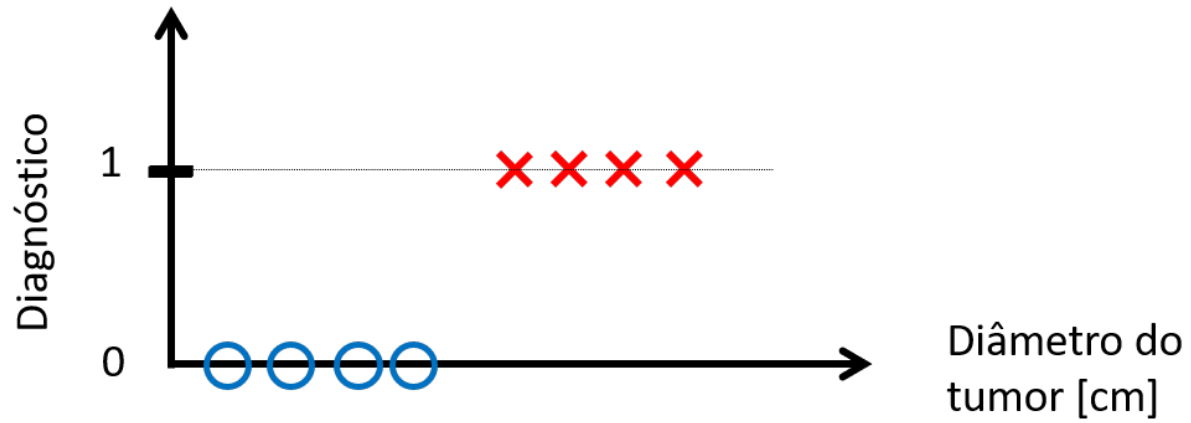
Pergunta	Resposta "y"
Este e-mail é SPAM?	Sim (1) ou Não (0)
Essa transação é fraudulenta?	Sim (1) ou Não (0)
Esse tumor é maligno?	Sim (1) ou Não (0)

- Nos problemas de classificação acima, y pode assumir apenas 1 dentre 2 valores possíveis.
- Problemas desse tipo são denominados **problemas de classificação binária** (apenas 2 possíveis classes/categorias)
- Geralmente, utiliza-se *1 = Sim = True* (classe positiva) e *0 = Não = False* (classe negativa)
- Porém, trata-se apenas de uma convenção. Poderia ser o contrário.

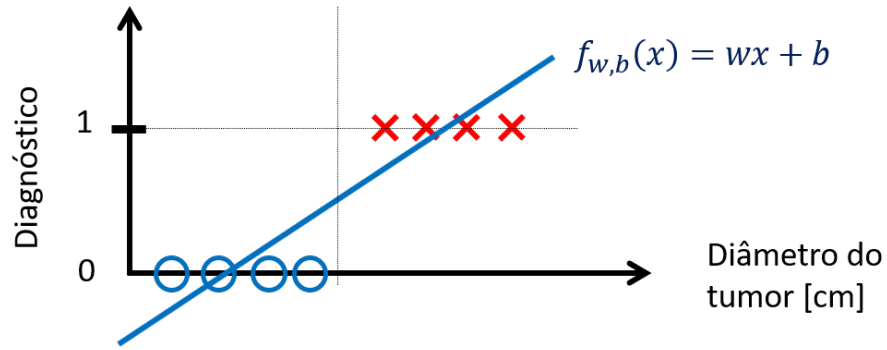
Por que Regressão Linear não é adequada para problemas de classificação?



Por que Regressão Linear não é adequada para problemas de classificação?



Usando Regressão Linear, poderíamos chegar em:



Se $f_{w,b}(x) < 0.5$, então $\hat{y} = 0$

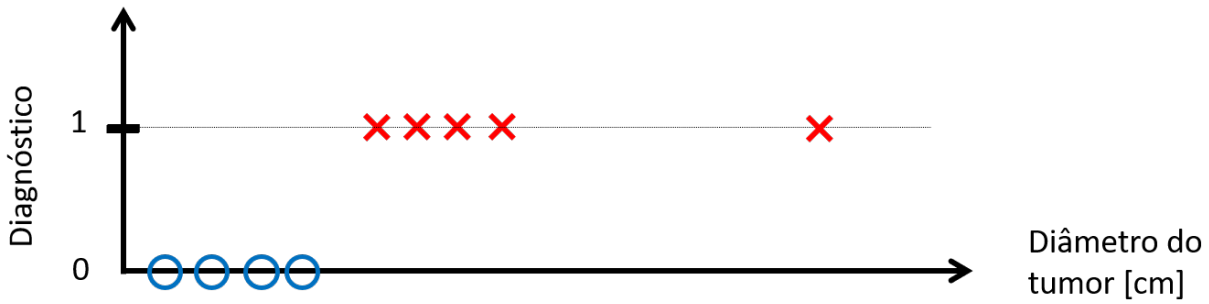
Se $f_{w,b}(x) \geq 0.5$, então $\hat{y} = 1$

Pergunta:

Parece razoável?

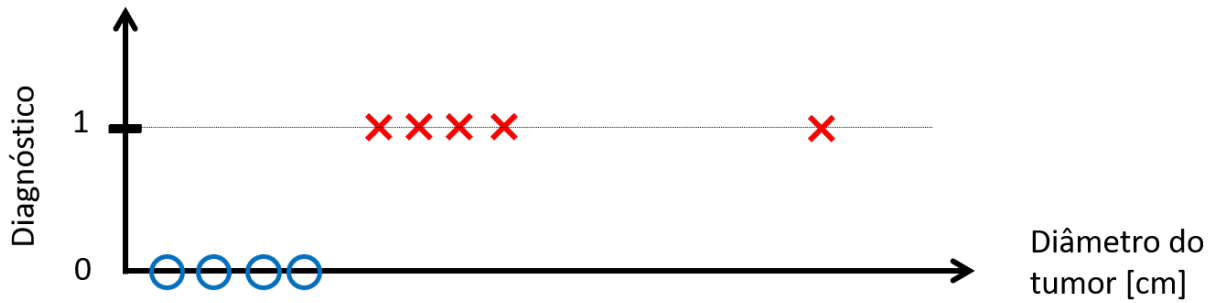
Por que Regressão Linear não é adequada para problemas de classificação?

Adicionando apenas uma amostra de paciente com tumor maligno...

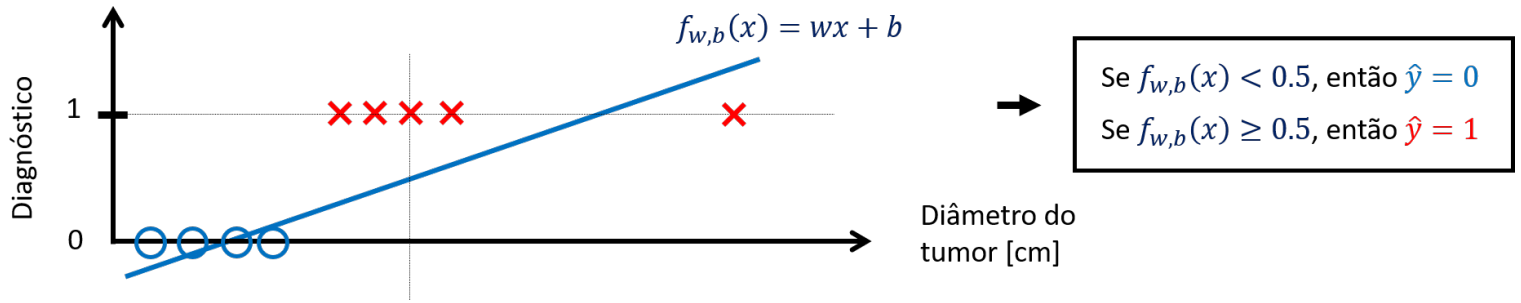


Por que Regressão Linear não é adequada para problemas de classificação?

Adicionando apenas uma amostra de paciente com tumor maligno...



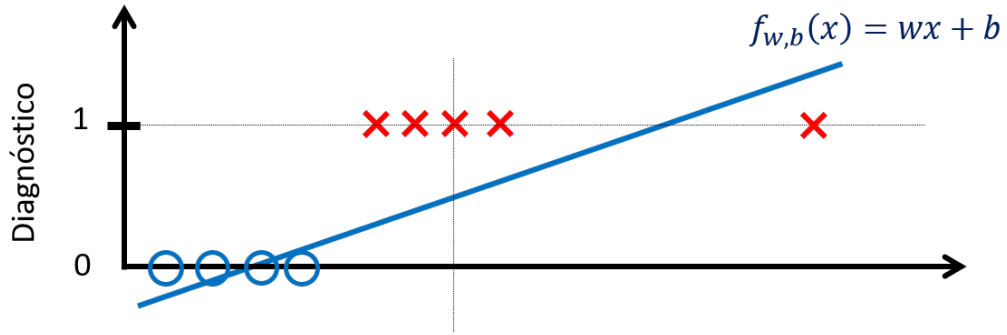
Usando Regressão Linear, poderíamos chegar em:



Observação:

Note que a adição dessa nova amostra não deveria ter mudado a classificação, mas mudou já que o modelo foi treinando via Regressão Linear.

Por que Regressão Linear não é adequada para problemas de classificação?



Se $f_{w,b}(x) < 0.5$, então $\hat{y} = 0$

Se $f_{w,b}(x) \geq 0.5$, então $\hat{y} = 1$

Observações:

- Note que a Regressão Linear fez com que algumas amostras de tumor maligno fossem classificadas como benignos.
- Isso acontece porque a Regressão Linear não é uma técnica adequada para problemas de classificação
- Para problemas de classificação, temos a **Regressão Logística** que, apesar de ter o termo "regressão" no seu nome, ela não serve para regressão, mas sim para classificação.
- A **Regressão Logística** é um dos algoritmos para classificação mais utilizados, e começaremos por ele.
- A Regressão Logística baseia-se na **função sigmoide**, também chamada de função logística. Essa função sempre fornece valores entre 0 e 1.

Pergunta:

Qual alternativa abaixo representa um problema de classificação?

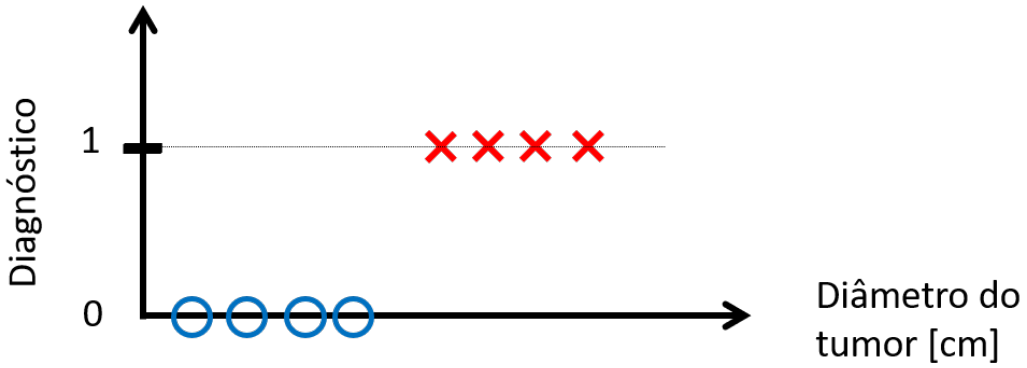
A) Estimar o peso de uma baleia com base em seu comprimento

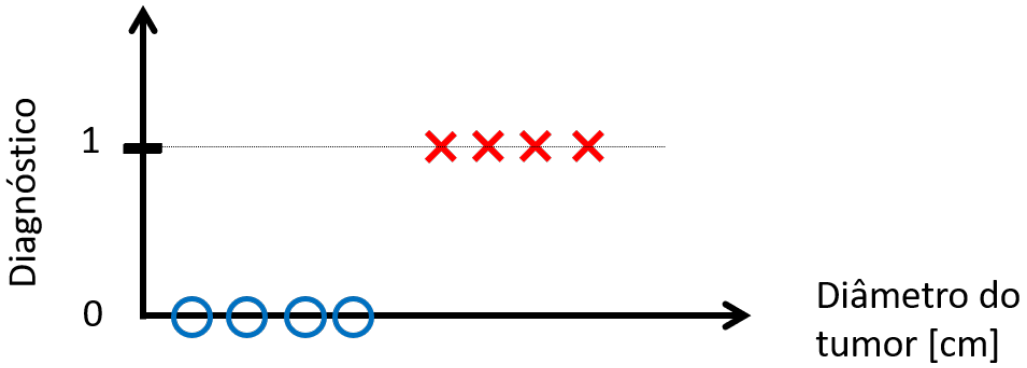
B) Decidir se um animal é uma baleia ou não.

Regressão Logística

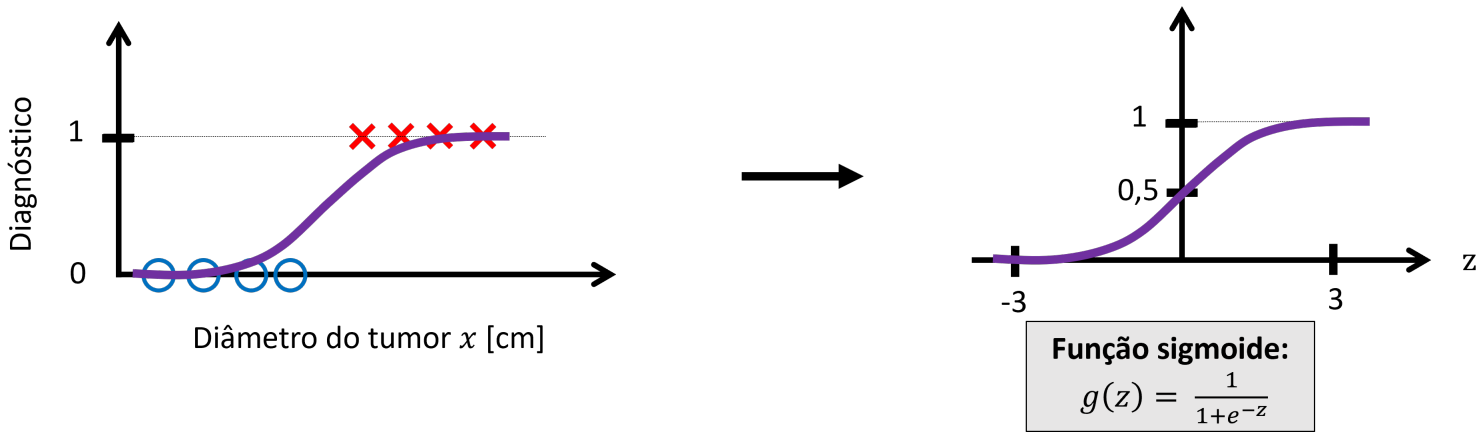
Regressão Logística

- A **Regressão Logística** é um dos algoritmos de classificação mais simples e utilizados.
- Devido a sua simplicidade, costuma ser um bom método para realização de um teste inicial sobre os seus dados.


Diagnóstico
1
0
Diâmetro do
tumor [cm]

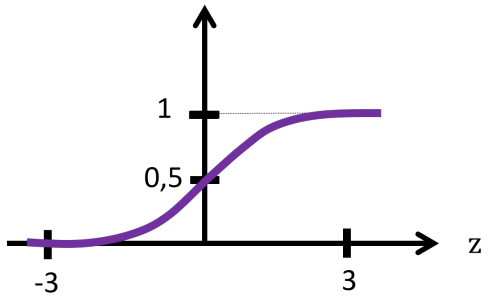


Usando Regressão Logística, podemos chegar em algo do tipo:



- Note que a saída está sempre entre 0 e 1 (por exemplo, 0.7)
- A Regressão Logística baseia-se na **função sigmoide** $g(z)$, também chamada de função logística. Essa função sempre fornece valores entre 0 e 1, ou seja, $0 < g(z) < 1$.

Analisando com detalhes a função sigmoide



Função sigmoide:
$$g(z) = \frac{1}{1+e^{-z}}$$

Quando z é elevado, por exemplo, $z = 100$, tem-se

$$g(z) \approx \frac{1}{1}$$

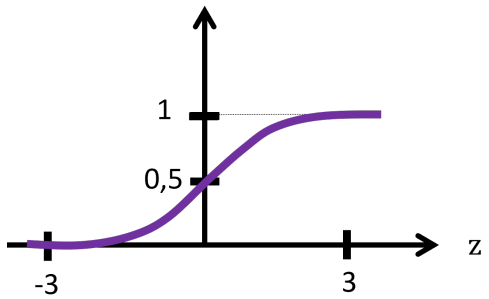
De forma análoga, quando z é bem negativo, por exemplo, $z = -100$, tem-se

$$g(z) \approx \frac{1}{1+\infty} = 0$$

Quando, $z = 0$, tem-se

$$g(z) = \frac{1}{1+1} = 0.5$$

Definindo o modelo de Regressão Logística



Função sigmoide:
$g(z) = \frac{1}{1+e^{-z}}$

Passo 1:

Definimos z como um modelo linear do tipo

$$z = \vec{w} \cdot \vec{x} + b$$

Passo 2:

Passamos este z pela função sigmoide

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

Resultado

O resultado desse passo-a-passo é o modelo de Regressão Logística

$$f_{\vec{w},b}(\vec{x}) = g(\vec{w} \cdot \vec{x} + b) = \frac{1}{1 + e^{-(\vec{w} \cdot \vec{x} + b)}}$$

Comparação Regressão Linear com Regressão Logística

Na Regressão Linear, usada em problemas de Regressão, tínhamos

$$f_{\vec{w},b}(\vec{x}) = \vec{w} \cdot \vec{x} + b$$

Note que $f_{\vec{w},b}(\vec{x})$ recebia $\vec{x}$ e fornecia valores entre $-\infty$ e $+\infty$

Agora, na Regressão Logística, usada em problemas de Classificação, temos

$$f_{\vec{w},b}(\vec{x}) = \frac{1}{1 + e^{-(\vec{w} \cdot \vec{x} + b)}}$$

Note que $f_{\vec{w},b}(\vec{x})$ agora recebe $\vec{x}$ e fornece valores entre 0 e 1.

Pergunta

- Mas se trata-se de um problema de classificação, onde estão as classes?
- $f_{\vec{w},b}(\vec{x})$ não deveria fornecer OU 0 OU 1?

Comparação Regressão Linear com Regressão Logística

Na Regressão Linear, usada em problemas de Regressão, tínhamos

$$f_{\vec{w},b}(\vec{x}) = \vec{w} \cdot \vec{x} + b$$

Note que $f_{\vec{w},b}(\vec{x})$ recebia $\vec{x}$ e fornecia valores entre $-\infty$ e $+\infty$

Agora, na Regressão Logística, usada em problemas de Classificação, temos

$$f_{\vec{w},b}(\vec{x}) = \frac{1}{1 + e^{-(\vec{w} \cdot \vec{x} + b)}}$$

Note que $f_{\vec{w},b}(\vec{x})$ agora recebe $\vec{x}$ e fornece valores entre 0 e 1.

Pergunta

- Mas se trata-se de um problema de classificação, onde estão as classes?
- $f_{\vec{w},b}(\vec{x})$ não deveria fornecer OU 0 OU 1?

Resposta

Para resolvermos esse problema, basta interpretarmos $f_{\vec{w},b}(\vec{x})$ como sendo a **probabilidade** da classe ser 1.

Exemplo

Seja

$$f_{\vec{w},b}(\vec{x}) = \frac{1}{1 + e^{-(\vec{w} \cdot \vec{x} + b)}}$$

a probabilidade da classe ser 1

Exemplo:

x: tamanho do tumor
y: 1 se o tumor for maligno, ou 0 se o turmo for benigno

$$f_{\vec{w},b}(\vec{x}) = 0.7 \quad \rightarrow \quad y = 1 \text{ com 70 \% de chance}$$

$$f_{\vec{w},b}(\vec{x}) = 0.4 \quad \rightarrow \quad y = 0 \text{ com 60 \% de chance}$$

Notação formal:

$$f_{\vec{w},b}(\vec{x}) = P(y = 1 | \vec{x}; \vec{w}, b)$$

Pergunta:

Lembre-se que a função sigmoide é dada por $g(z) = \frac{1}{1+e^{-z}}$. Se z é um número bastante negativo, então:

A) $g(z)$ é próximo de -1

B) $g(z)$ é próximo de 0

De olho no código!

De olho no código!

Vamos agora ver como implementar a função Sigmoide em Python.

Acesse o Python Notebook usando o QR code ou o link abaixo:

https://colab.research.google.com/github/xaximpvp2/master/blob/main/codigo_aula9_funcao_sigmoide_e_classificacao.ipynb

Atividade de aula

Parte 1

Rode todo o código. Certifique-se de que você o compreendeu.

Parte 2

1. Explique, com as suas próprias palavras, como a função Sigmoide é utilizada na construção do Método de Regressão Logística.
2. Para os dados que estão no código, descubra, por tentativa e erro, quais valores para w e b resultaram num bom modelo de Regressão Logística?