

# Introdução ao aprendizado de máquina

## Aula 4 - Problemas de classificação e regressão logística

- Aplicações
- Função logística
- Entropia cruzada
- Acurácia, precisão e recall
- Classes desequilibradas
- Múltiplas classes



# Alguns exemplos

# Alguns exemplos de problemas de classificação:



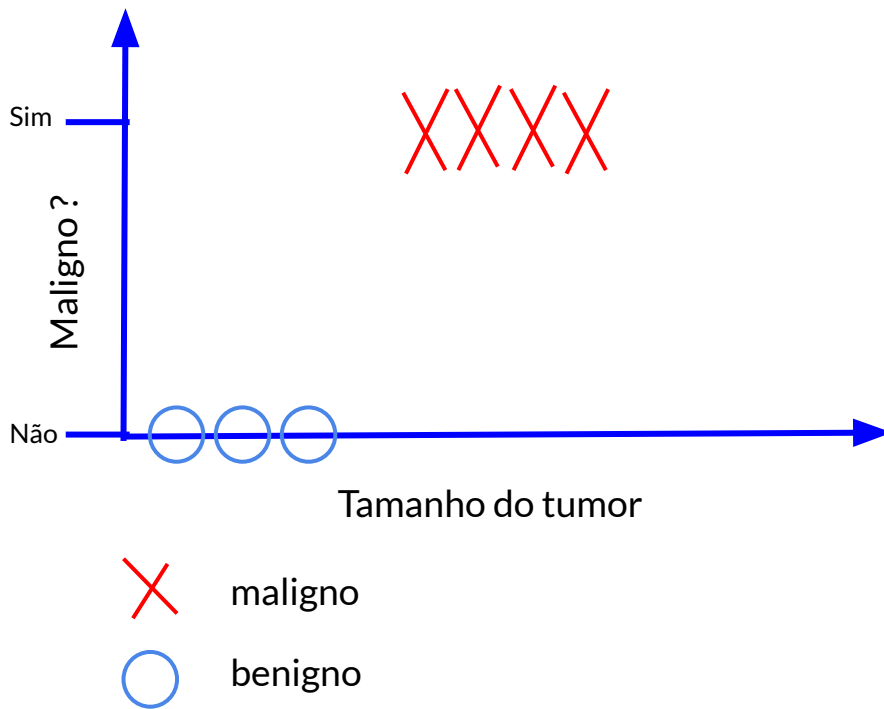
1. Filtros de spam.
2. Reconhecimento de dígitos.
3. O tumor é maligno?
4. Em que candidato uma pessoa votará.
5. A compra é fraudulenta?
6. A receita da empresa será maior que a receita prevista?

# Tamanho do tumor: problemas com a regressão linear

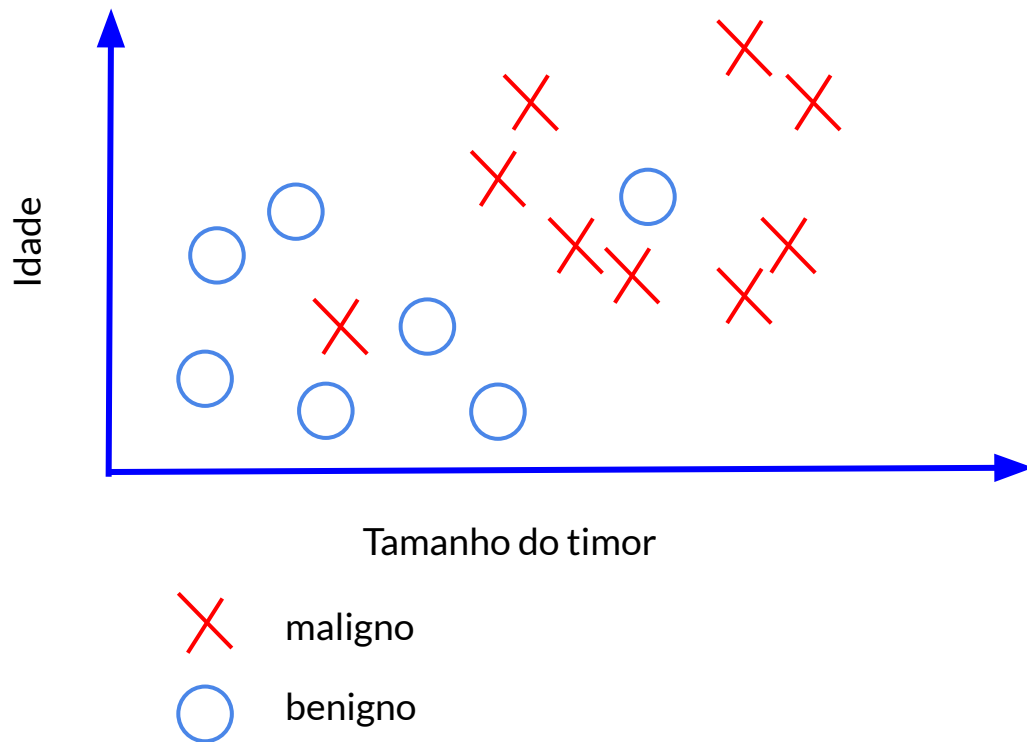
Exemplo: o tumor é maligno?

- O que acontece com uma regressão linear nesse exemplo?
- O que ocorre se observarmos um tumor muito maior?

$$\text{previsão} = \begin{cases} 1 & \text{se } \hat{y} \geq 0.5 \\ 0 & \text{se } \hat{y} < 0.5 \end{cases}$$



# Exemplo com várias variáveis





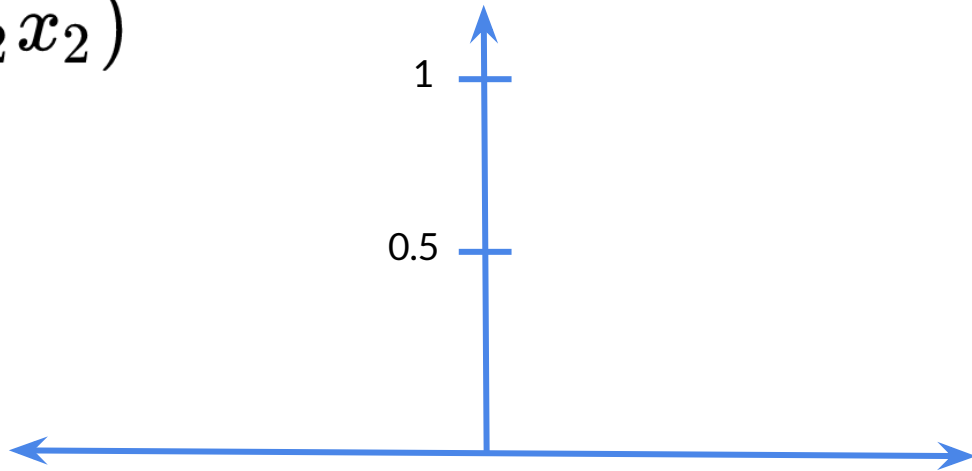
# Função Logística

# Função logística ou sigmoid

$$h_{\theta} = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

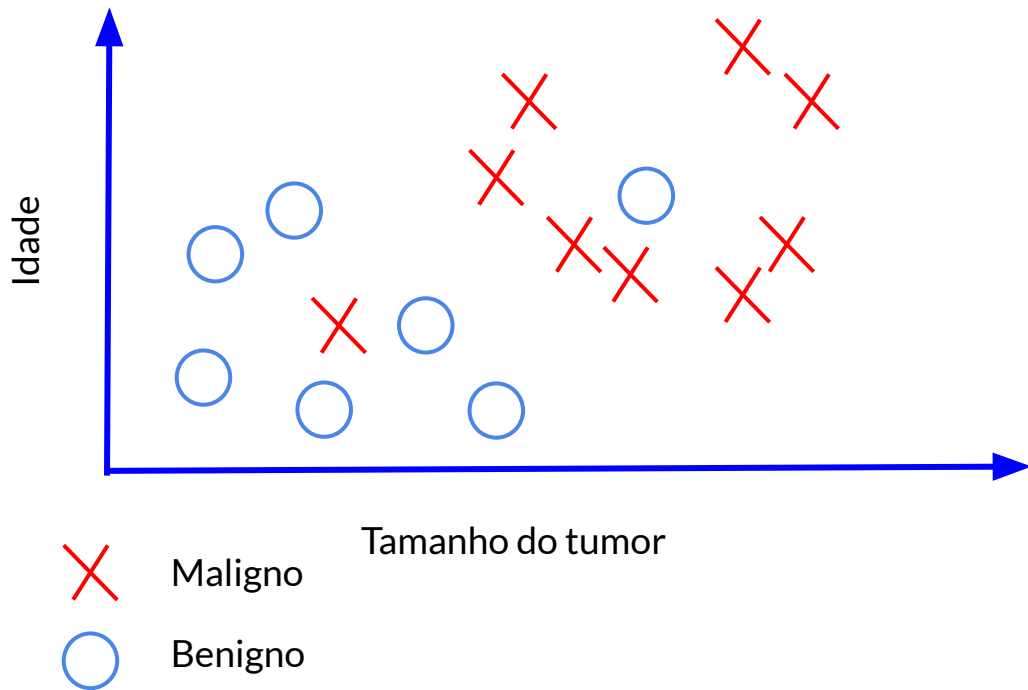
$$h_{\theta} = \Omega(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

Interpretação:  
previsão é a probabilidade de  $y = 1$



# Intuição da fronteira de decisão

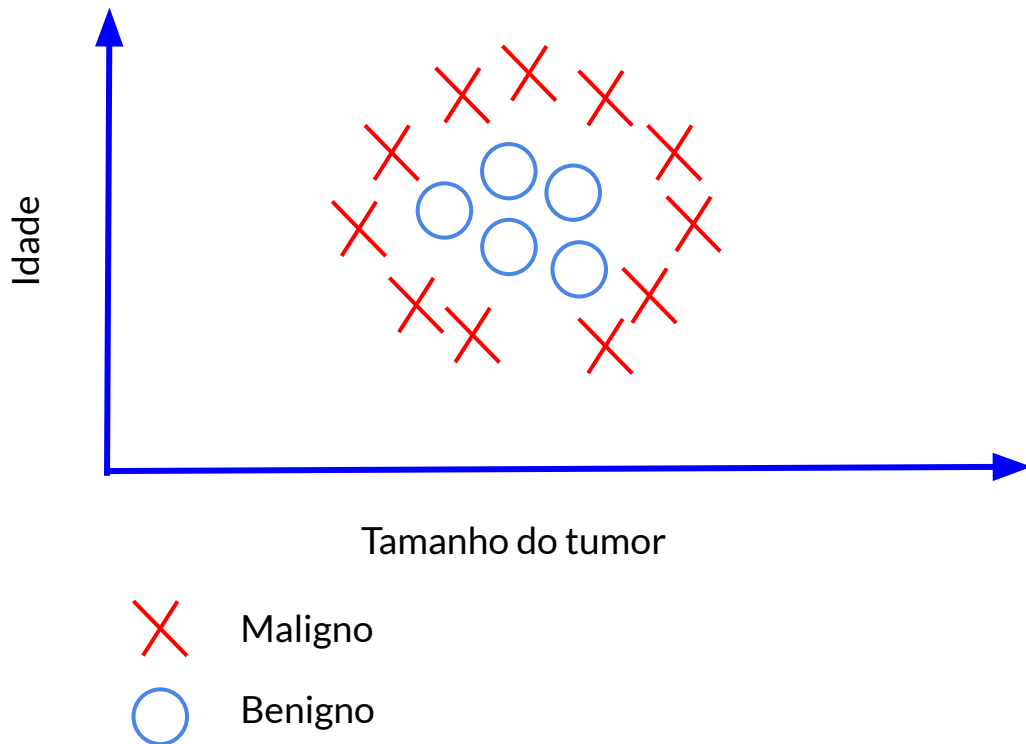
$$\hat{y} = \Omega(-5 + 2 * \text{tamanho\_tumor} + 1 * \text{idade})$$



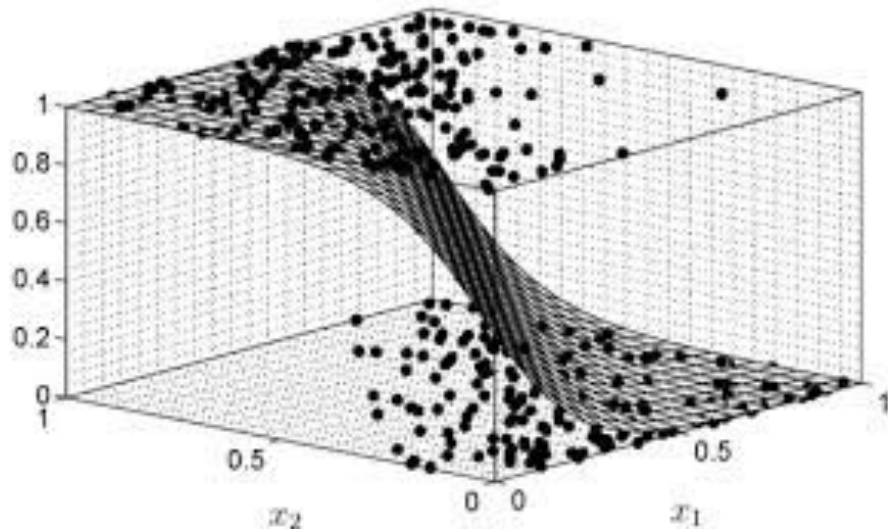


# Fronteira de decisão não linear

$$\hat{y} = \Omega(\theta_0 + \theta_1 \text{tamanho} + \theta_2 \text{idade} + \theta_3 \text{tamanho}^2 + \theta_4 \text{idade}^2)$$



# Vizualizando a regressão logística em 3D





# Função custo

# Função entropia cruzada

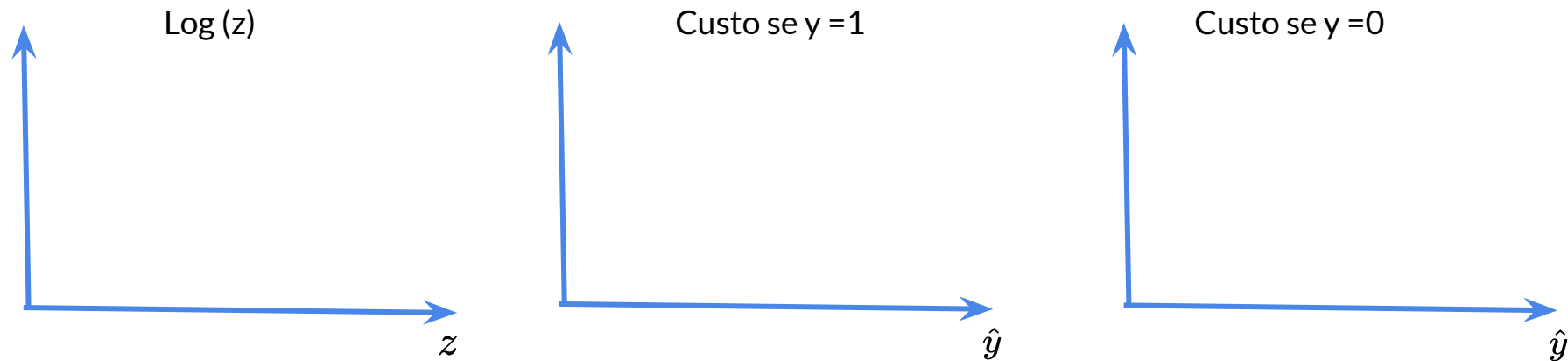


$$\hat{y} = \frac{1}{1 + e^{-(\theta_0 + \theta_1 * x)}}$$

$$Custo(y_i, \hat{y}_i) = \begin{cases} -\log(\hat{y}_i) & \text{se } y_i = 1 \\ -\log(1 - \hat{y}_i) & \text{se } y_i = 0 \end{cases}$$

# Intuição da entropia cruzada

$$J(\theta) = - \sum_{i=1}^n y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)$$



Essa função custo gera um problema de otimização convexa




# Métrica de performance

# Acurácia



		Previsão maligno?	
		Não	Sim
Era maligno?	Não	40 VN	10 FP
	Sim	20 FN	30 VP


# Problemas com acurácia



		Previsão fraude?	
		Não	Sim
Era fraude?	Não	995 VN	0 FP
	Sim	5 FN	0 VP

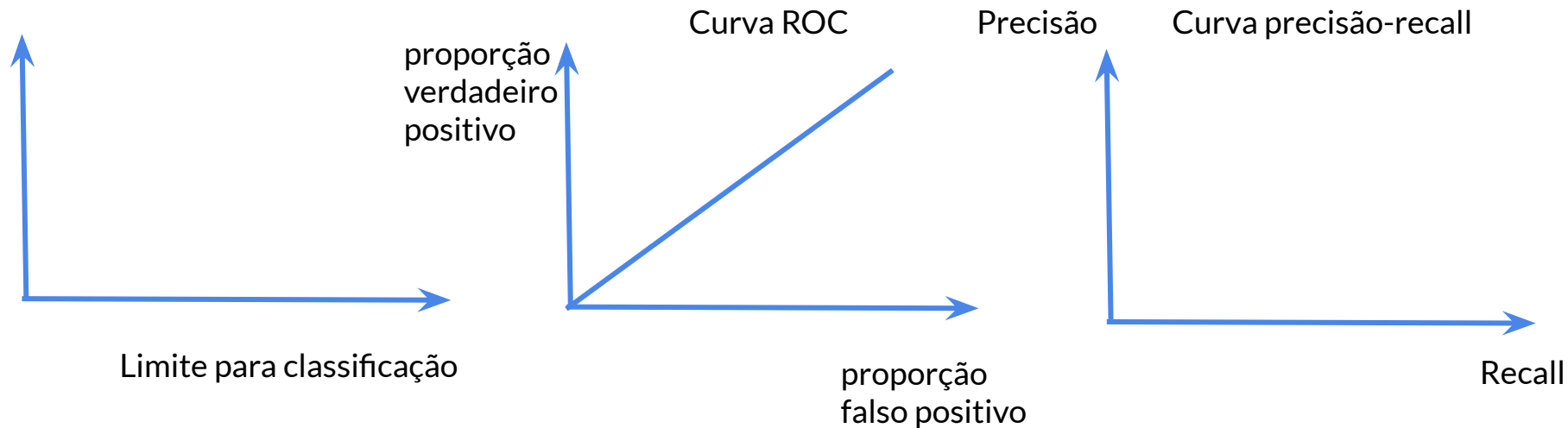


# Precisão, Revocação (recall) e F-Score



		Previsão fraude?	
		Não	Sim
Era fraude?	Não	995 VN	0 FP
	Sim	5 FN	0 VP

# ROC e curva de precisão-recall





# Classes desequilibradas

# Classes desequilibradas

---



Lidar com classes desequilibradas é um desafio de vários algoritmos de ML

Temos 3 abordagens diferentes para lidar com esse problema;

1. Usar uma métrica que leve em consideração precisão e recall
2. Modificar a função custo para punir mais erros na classe menos comum
3. Bootstrap uma nova amostragem baseada na base dados original de tal forma que as classes fiquem equilibradas e depois ajustar as probabilidades previstas para obter previsões realistas.



# Múltiplas classes

# Classificação com múltiplas classes



- 3 partidos políticos: Trabalhista, Liberal e Conservador. Queremos prever em quem um indivíduo votará.
- Podemos treinar 3 modelos diferentes e depois modificar a probabilidade de votar em cada partido.

$$\begin{aligned} P_{\text{modificada}}(\text{Trabalhista}) &= \frac{P(\text{Trabalhista})}{\sum_i P(\text{partido } i)} \\ &= \frac{P(\text{Trabalhista})}{P(\text{Trabalhista}) + P(\text{Liberal}) + P(\text{Conservador})} \end{aligned}$$

# Classificação com múltiplas classes

