

Introdução a Aprendizagem de máquina para Bioinformática

Aula 3

Msc. Amanda Araújo Serrão de Andrade

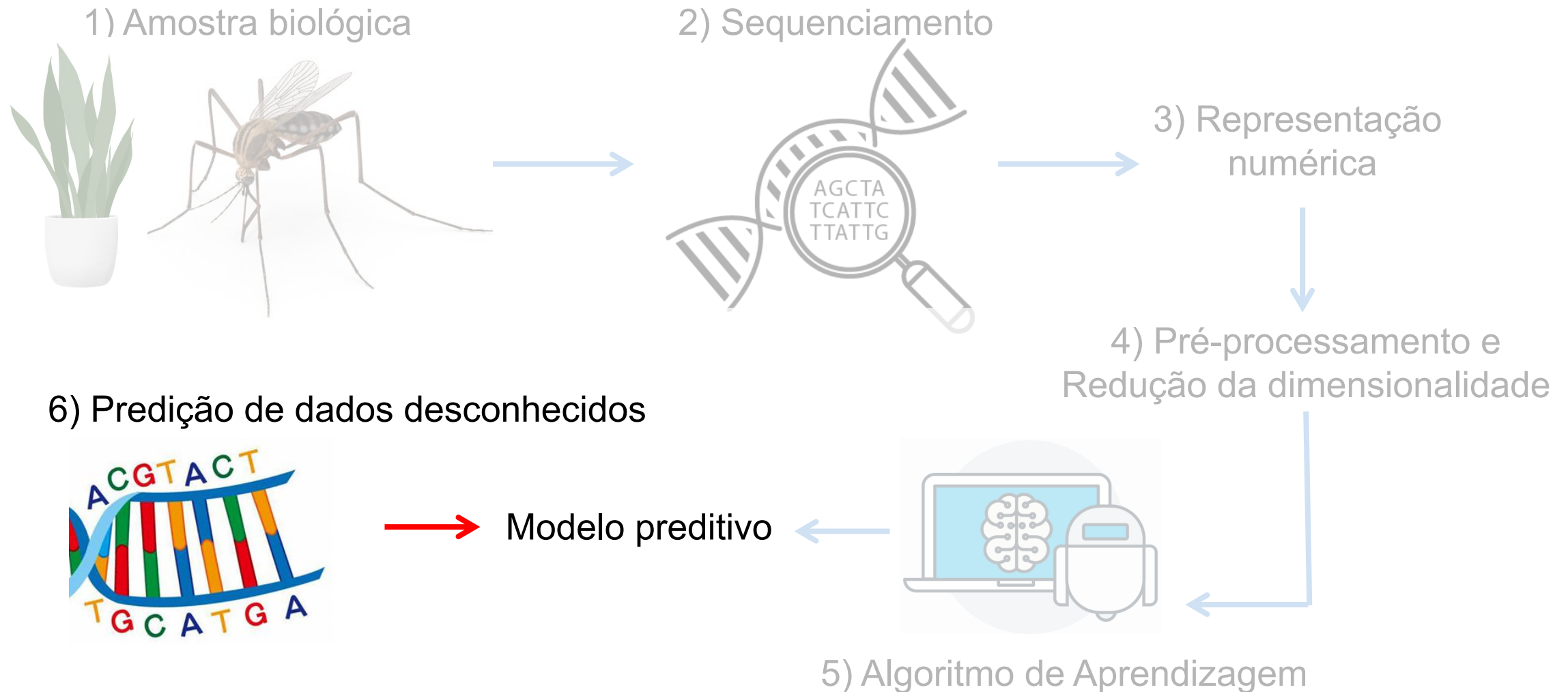
Biomédica, Mestre em Modelagem Computacional e Doutoranda em Genética

Orientadora: Dra. Ana Tereza Ribeiro de Vasconcelos

Sumário da aula

- Como saber se meu modelo está funcionando?
- Principais métricas de avaliação
- Análise dos resultados da classificação (prática)
- Comparação de performance entre diferentes algoritmos
- Qual linguagem de programação utilizar?

Exemplo prático: Como identificar se um vírus desconhecido infecta plantas ou mosquitos?



Como saber se o modelo está funcionando?

- O modelo que funciona aprendeu a partir do conjunto de dados de treinamento sendo capaz de prever as classes de dados desconhecidos.

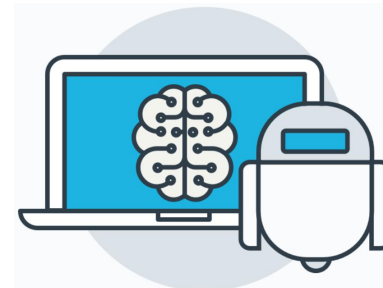
Matriz de características

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

80%

Dados de treino

Modelo preditivo



Como saber se o modelo está funcionando?

- O modelo que funciona aprendeu a partir do conjunto de dados de treinamento sendo capaz de prever as classes de dados desconhecidos.

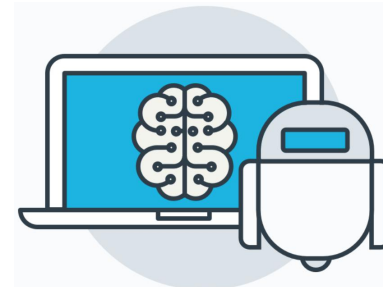
Matriz de características

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

Dados de teste

20%

Modelo preditivo



- Classificação de dados desconhecidos
- Métricas de desempenho

6) Predição de dados desconhecidos

A matriz de confusão

- Importante saída da predição de um modelo de classificação, pode ser binário ou ter mais classes.
- Melhor avaliada a partir da predição de dados desconhecidos.
- Definição de uma classe positiva e outra classe negativa.
- Prática: vírus de mosquito é a classe positiva.

		Classe prevista	
		Vírus de mosquito	Vírus de plantas
Classe real	Vírus de mosquito	2	1
	Vírus de plantas	0	5

6) Predição de dados desconhecidos

A matriz de confusão

- Importante saída da predição de um modelo de classificação, pode ser binário ou ter mais classes.
- Melhor avaliada a partir da predição de dados desconhecidos.
- Definição de uma classe positiva e outra classe negativa.
- Prática: vírus de mosquito é a classe positiva.

		Classe prevista	
		Vírus de mosquito	Vírus de plantas
Classe real	Vírus de mosquito	Verdadeiro positivo (TP)	Falso negativo (FN)
	Vírus de plantas	Falso positivo (FP)	Verdadeiro negativo (TN)

6) Predição de dados desconhecidos

A matriz de confusão

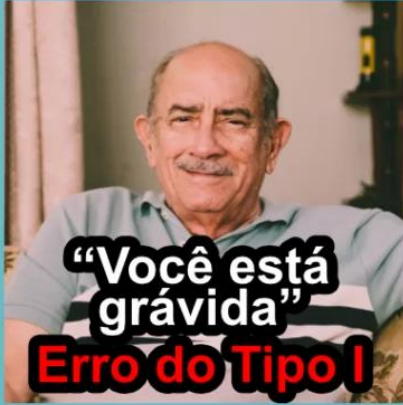
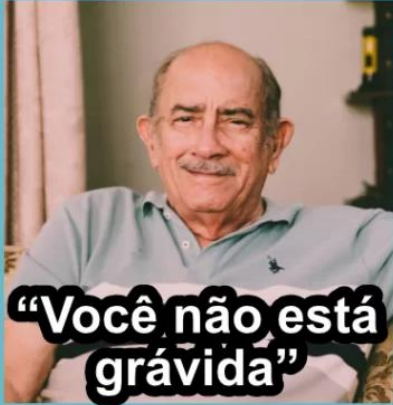


		Classe prevista	
		Grávida	Não grávida
Classe real	Não grávida		
	Grávida		

Fig 1. Exemplos dos erros encontrados em uma matriz de confusão. Fonte: <https://psicometriaonline.com.br>

Principais métricas para a avaliação dos resultados

- As métricas são obtidas a partir da análise matemática da matriz de confusão.
-
- As principais métricas são: Acurácia, Sensibilidade, Especificidade, curva ROC e área embaixo da curva.
- Apresentam vantagens e desvantagens, a depender do algoritmo utilizado.
- Outras métricas tem sido utilizadas, como por exemplo, o coeficiente de Matthews.

6) Predição de dados desconhecidos

Acurácia

- Taxa de acertos do modelo, seja verdadeiro positivo ou verdadeiro negativo.
- Não é uma métrica indicada em casos de desbalanceamento dos dados.
- Tem caído em desuso na Bioinformática.
- Existe a acurácia balanceada.

$$\text{Acurácia} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} = \frac{\text{Predições corretas}}{\text{Todas as predições}}$$

Sensibilidade

- Sensibilidade (taxa de verdadeiros positivos): avalia a capacidade do método de detectar com sucesso resultados classificados como positivos.
- Também conhecida como revocação ou Recall.

$$\text{Sensibilidade} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

6) Predição de dados desconhecidos

Especificidade

- Especificidade (taxa de verdadeiros negativos): avalia a capacidade do método de detectar resultados negativos.

$$\text{Especificidade} = \frac{\text{TN}}{\text{FP} + \text{VN}}$$

6) Predição de dados desconhecidos

Receiver Operating Characteristics (ROC) curve

- A curva ROC é um gráfico que permite avaliar um classificador binário.
- Sensibilidade vs Especificidade;

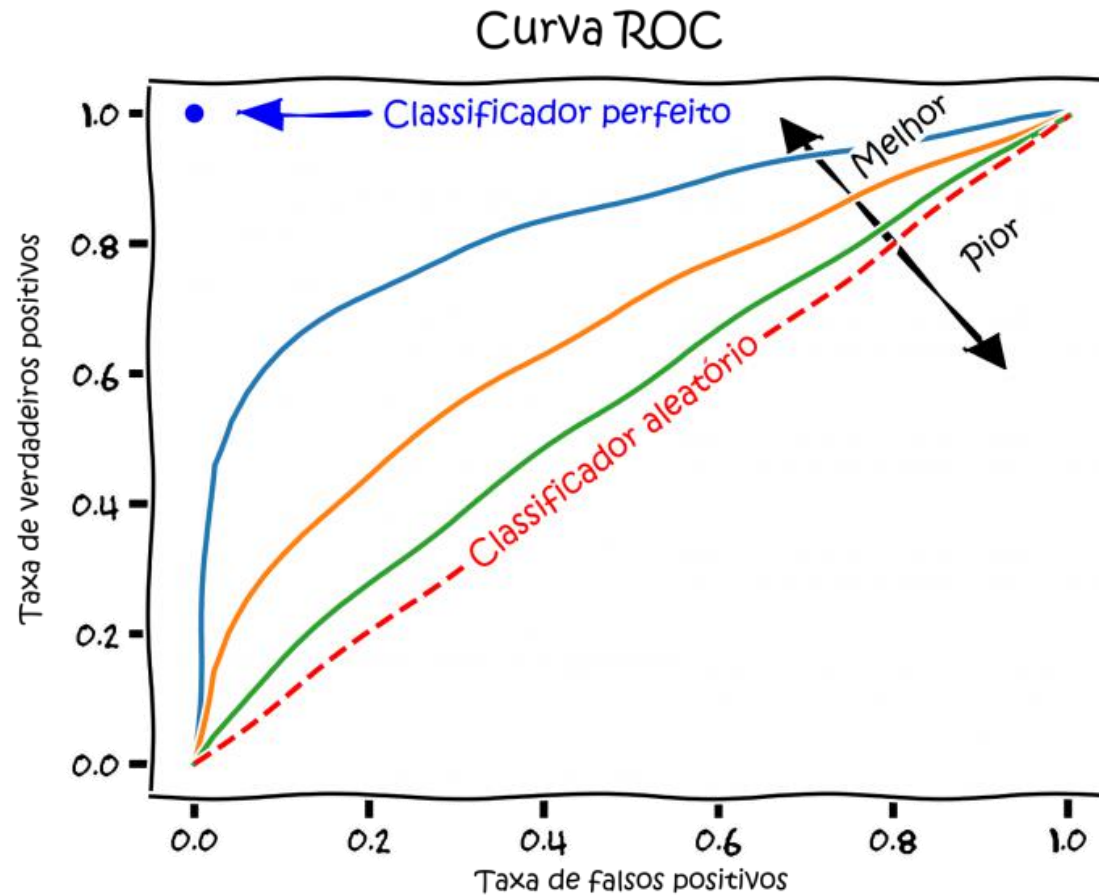


Fig 2. Ilustração de uma curva ROC. Fonte: adaptado e traduzido de MartinThoma (CC0 1.0 domínio público).

6) Predição de dados desconhecidos

Area Under the Curve (AUC)

- AUC calcula a área da forma bidimensional formada abaixo da curva;
- Varia de 0 a 1; Quão correta estão as predições do modelo.

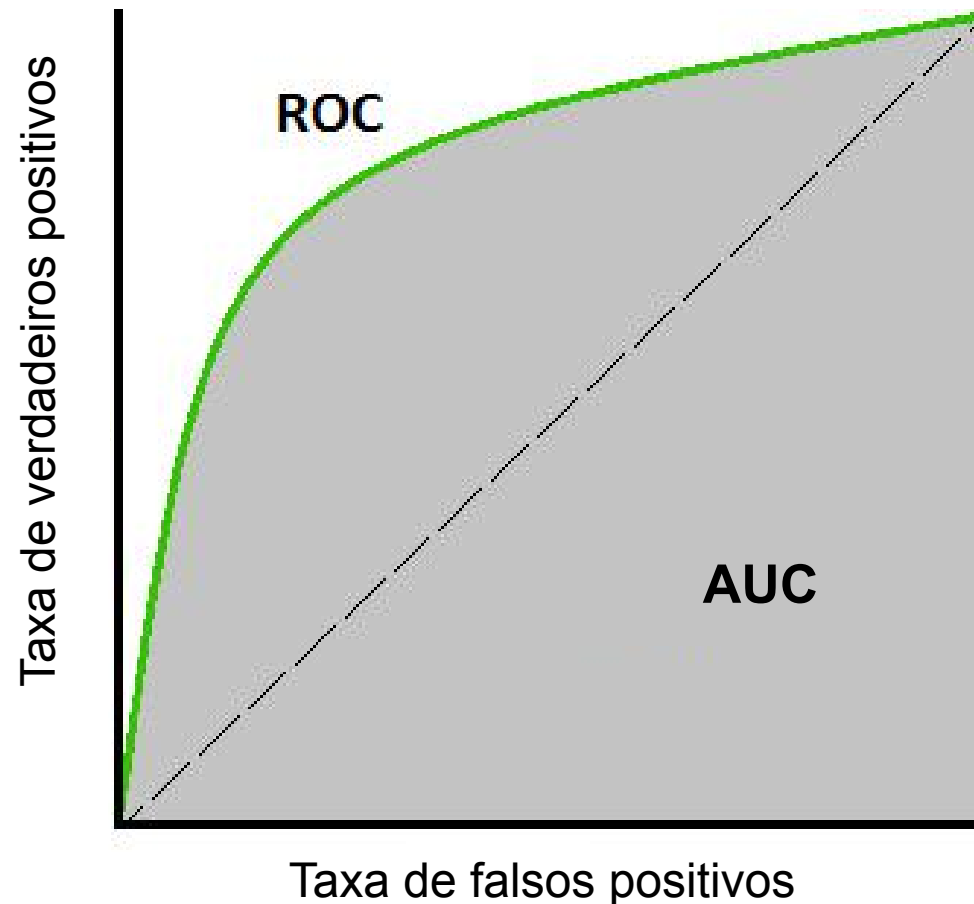
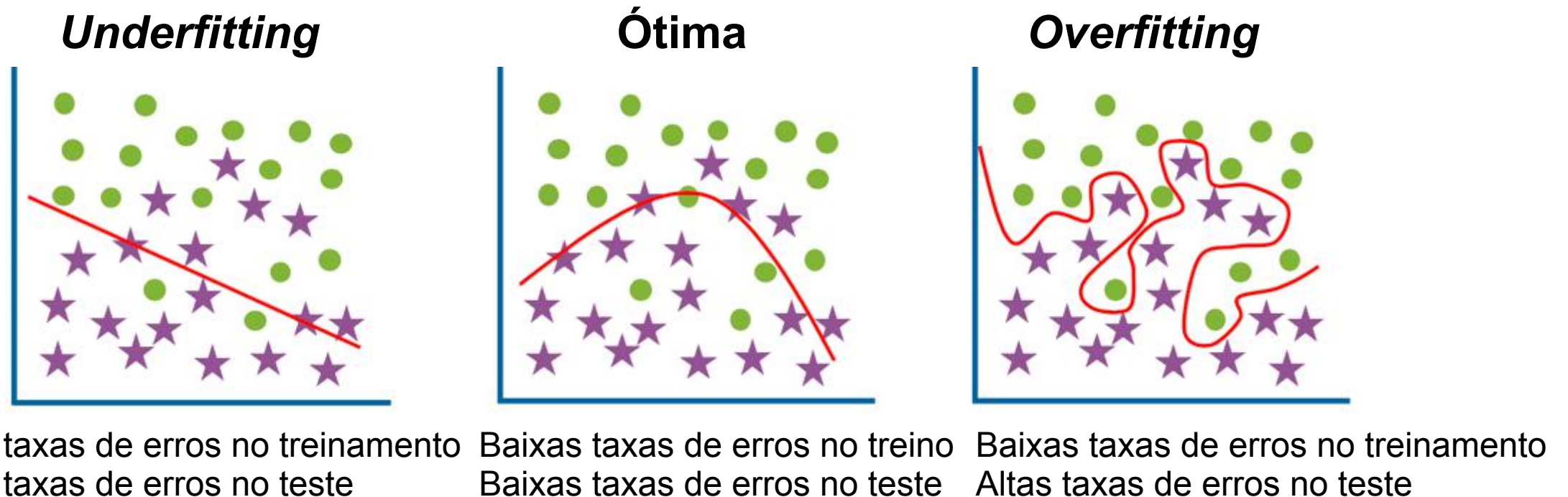


Fig 3. Obtenção do valor de AUC. Fonte: adaptado e traduzido de towardsdatascience.com

Underfitting e Overfitting

- Underfitting: o modelo não consegue identificar padrões discriminatórios.
- Overfitting: o modelo “decora” os dados de treino. Não tem generalização.



Qual linguagem de programação utilizar?



Linguagem: R
Pacote: Caret



Linguagem: Python
Pacote: Scikit learn



Dúvidas?

aandradebio@gmail.com

- As dúvidas mais específicas devem vir acompanhadas com código e um conjunto de dados de teste para agilizar a resposta



Obrigada!!

aandradebio@gmail.com

atr@Incc.br