

MINICURSO **INTRODUÇÃO AO MACHINE LEARNING**

15 de setembro de 2020

**Adriano Almeida
Felipe Carvalho
Felipe Menino**



Transmissão pelo YouTube:
<https://www.youtube.com/c/PGCAPHINPE>



Palestras

Minicursos

Hackathon



**COMPUTAÇÃO
APLICADA**



MATERIAL COMPLEMENTAR

Introdução ao Machine Learning

Adriano Almeida
Felipe Carvalho
Felipe Menino

Prefácio

Seja bem vinda(o) ao livro-texto do minicurso de **Introdução ao Machine Learning**. Criamos este material para compartilhar o pouco que sabemos e dividir nossas experiências. Neste material, você irá encontrar conteúdos sobre o conceito, técnicas e algumas dicas úteis sobre Machine Learning. Procuramos abordar os conceitos de forma didática, porque sabemos o quanto difícil é se inteirar de uma nova área, principalmente para as pessoas que não estão familiarizadas com o assunto. Este livro-texto não tem um público-alvo, escrevemos com o objetivo de atingir o máximo de pessoas em quaisquer áreas. O único pré-requisito para ler este livro-texto é ter curiosidade, porque não são as respostas que movem o mundo, e sim, as perguntas!

Introdução Ao Machine Learning
De alunos para alunos

Prefácio

A curiosidade é uma bengala

Introdução ao Machine Learning

DATAAT
Adriano Almeida, Felipe Carvalho e Felipe Menino

<https://dataat.github.io/introducao-ao-machine-learning/>

Prefácio

Seja bem vinda(o) ao livro-texto do minicurso de **Introdução ao Machine Learning**. Criamos este material para compartilhar o pouco que sabemos e dividir nossas experiências. Neste material, você irá encontrar conteúdos sobre o conceito, técnicas e algumas dicas úteis sobre Machine Learning. Procuramos abordar os conceitos de forma didática, porque sabemos o quanto difícil é se inteirar de uma nova área, principalmente para as pessoas que não estão familiarizadas com o assunto. Este livro-texto não tem um público-alvo, escrevemos com o objetivo de atingir o máximo de pessoas em quaisquer áreas. O único pré-requisito para ler este livro-texto é ter curiosidade, porque não são as respostas que movem o mundo, e sim, as perguntas!

Introdução Ao Machine Learning
De alunos para alunos

Prefácio

A curiosidade é uma bengala

Introdução ao Machine Learning

DATAAT
Adriano Almeida, Felipe Carvalho e Felipe Menino

<https://dataat.github.io/introducao-ao-machine-learning/>



DISCIPLINAS RELACIONADAS NA CAP

- CAP 394 - Introdução à data science (Dr. Rafael Santos e Dr. Gilberto Queiroz);
- CAP 359 - Princípios e aplicações de mineração de dados (Dr. Rafael Santos);
- CAP 351 - Neurocomputação (Dr. Marcos Quiles);
- CAP 354 - Inteligência artificial (Dr. Lamartine Guimarães);
- CAP 375 - Inteligência Computacional e Aplicações (Dr. Lamartine Guimarães);
- CAP 335 - Aprendizado Computacional e Reconhecimento de Padrões (Dr. Luciano Dutra)

Disponível em: <http://www.inpe.br/posgraduacao/cap/catalogo-disciplinas.php>



Workshop em
Computação
Aplicada

8-11 e 14-17 de setembro
Evento online

INTRODUÇÃO

<https://dataat.github.io/introducao-ao-machine-learning/introducao.html>

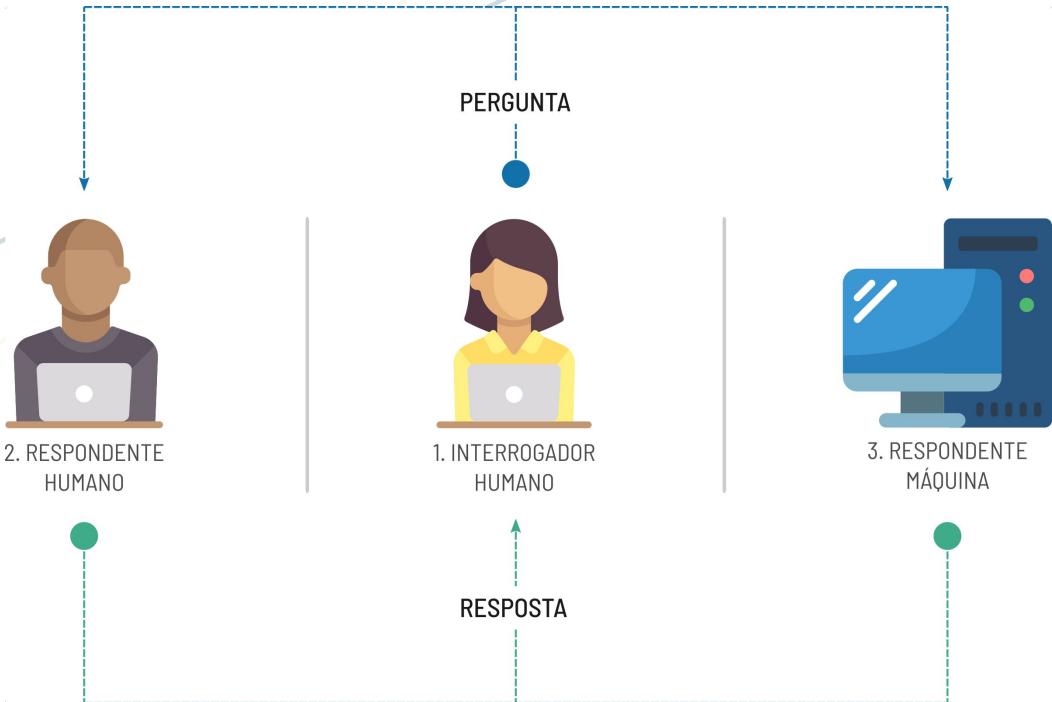
01

Transmissão pelo YouTube:

<https://www.youtube.com/c/PGCAPINPE>



“AS MÁQUINAS PODEM PENSAR?”



VOL. LIX. NO. 236.]

[October, 1950

M I N D A QUARTERLY REVIEW

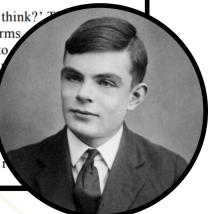
OF
PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

1. *The Imitation Game.*

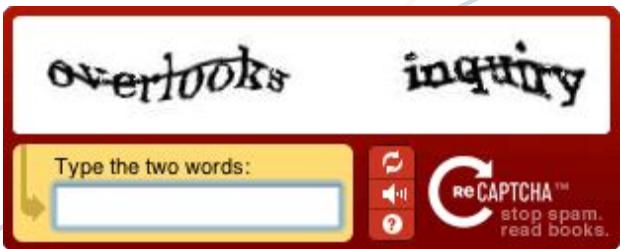
I PROPOSE to consider the question, ‘Can machines think?’ To begin with definitions of the meaning of the terms ‘think’. The definitions might be framed so as to cover the possible normal use of the words, but this attitude is dangerous, because it is difficult to decide exactly what the meaning of the words ‘machine’ and ‘think’ are. We must therefore examine how they are commonly used it is difficult to come to any conclusion that the meaning and the answer to the question ‘Can machines think?’ is to be sought in a statistical survey of public opinion. But this is absurd. Instead of attempting such a survey, let us replace the question by another, which is closely related and can be expressed in relatively unambiguous words.



(TURING, 1950)



"AS MÁQUINAS PODEM PENSAR?"



I'm not a robot



Privacy - Terms

Select all images with **crosswalks**
Click verify once there are none left.

VERIFY

CAPTCHA: Using Hard AI Problems for Security

Luis von Ahn¹, Manuel Blum¹, Nicholas J. Hopper¹, and John Langford²

¹ Computer Science Dept., Carnegie Mellon University, Pittsburgh PA 15213, USA

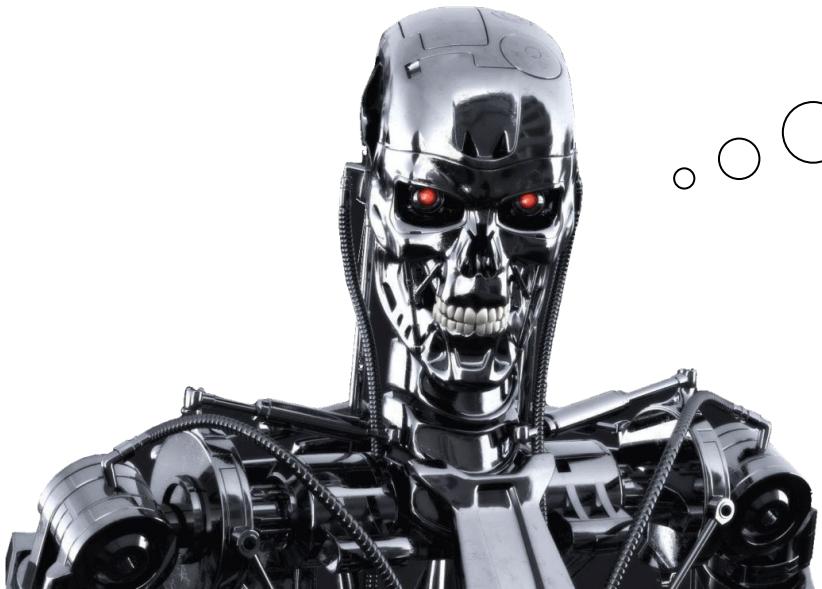
² IBM T.J. Watson Research Center, Yorktown Heights NY 10598, USA

Abstract. We introduce CAPTCHA, an automated test that humans can pass, but current computer programs can't pass: any program that has high success over a CAPTCHA can be used to solve an unsolved Artificial Intelligence (AI) problem. We provide several novel constructions of CAPTCHAS. Since CAPTCHAS have many applications in practical security, our approach introduces a new class of hard problems that can be exploited for security purposes. Much like research in cryptography has had a positive impact on algorithms for factoring and discrete log, we hope that the use of hard AI problems for security purposes allows us to advance the field of Artificial Intelligence. We introduce two families of AI problems that can be used to construct CAPTCHAS and we show that solutions to such problems can be used for steganographic communication. CAPTCHAS based on these AI problem families, then, imply a win-win situation: either the problems remain unsolved and there is a way to differentiate humans from computers, or the problems are solved and there is a way to communicate covertly on some channels.

(VON AHN et al. 2003)



REVOLUÇÃO DAS MÁQUINAS?!



Eu posso pensar!
então...

Transmissão pelo YouTube:

<https://www.youtube.com/c/PGCAPINPE>



REVOLUÇÃO DAS MÁQUINAS?!



Eu posso pensar!
então...

Transmissão pelo YouTube:

<https://www.youtube.com/c/PGCAPINPE>



REVOLUÇÃO DAS MÁQUINAS?!



amazon

Google

facebook

Spotify®

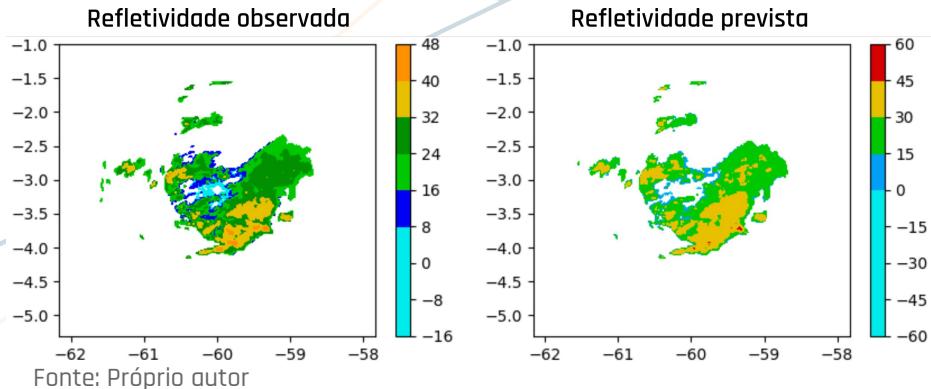
YouTube

Transmissão pelo YouTube:

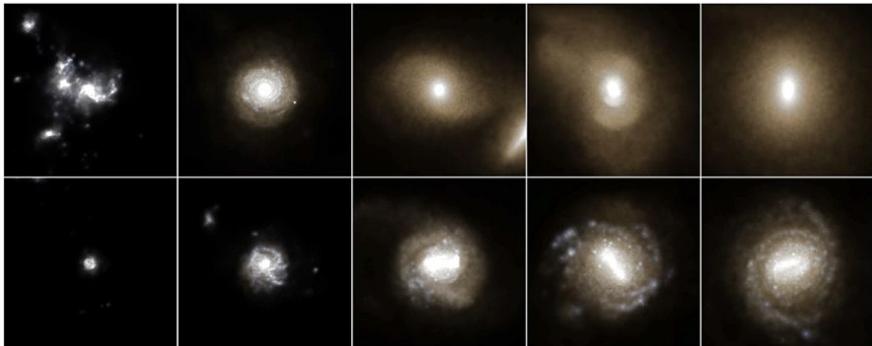
<https://www.youtube.com/c/PGCAPINPE>



REVOLUÇÃO DAS MÁQUINAS?!



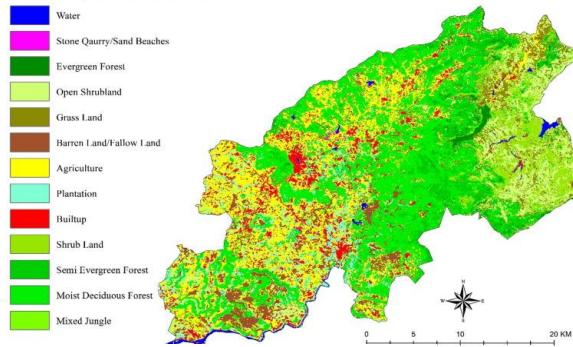
Fonte: Próprio autor



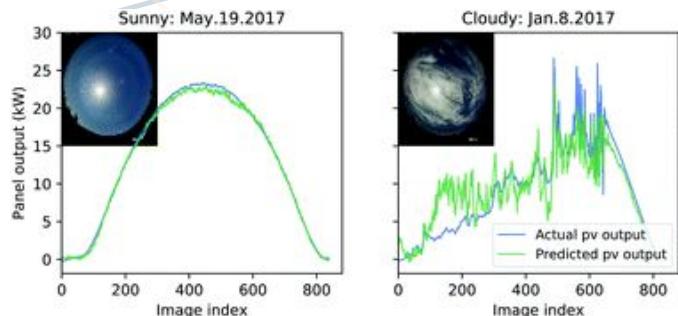
Fonte: <https://blog.galaxyzoo.org/2015/09/21/new-images-for-galaxy-zoo-part-2-illustris/>

Multi-Temporal Land use 2013

- Water
- Stone Quarry/Sand Beaches
- Evergreen Forest
- Open Shrubland
- Grass Land
- Bare Land/Fallow Land
- Agriculture
- Plantation
- Builtup
- Shrub Land
- Semi Evergreen Forest
- Moist Deciduous Forest
- Mixed Jungle

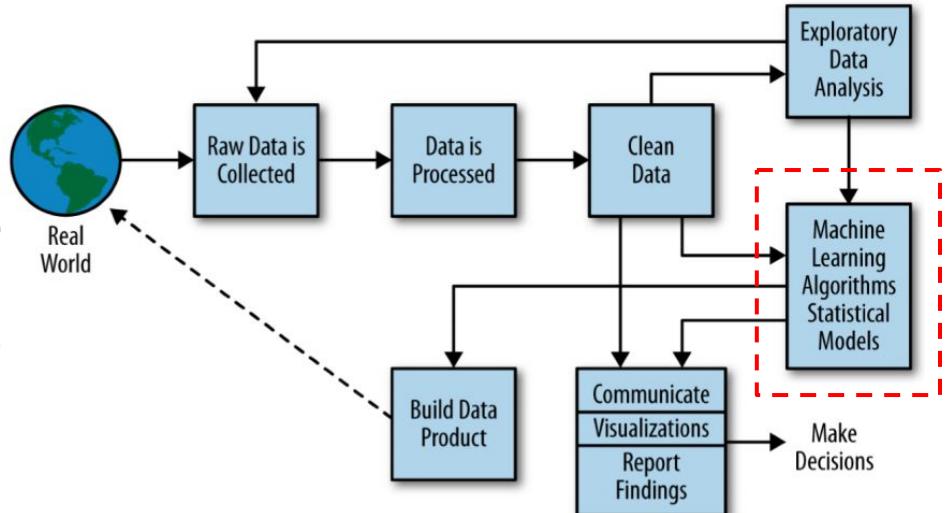


Fonte: (Kantakumar, 2015)



Fonte: Sun, Szucs e Brandt, 2018

APRENDIZADO DE MÁQUINA



Fonte: Schutt e O'Neil, 2013.

“Um campo de estudo que oferece aos computadores a capacidade de aprender sem serem explicitamente programados”

(SAMUEL, 1959)

APRENDIZADO DE MÁQUINA

*"Um computador aprende com a experiência **E** a respeito de alguma classe de tarefas **T** e desempenho medido por **P**, se seu desempenho nas tarefas em **T**, conforme medido por **P**, melhora com a experiência **E**"*

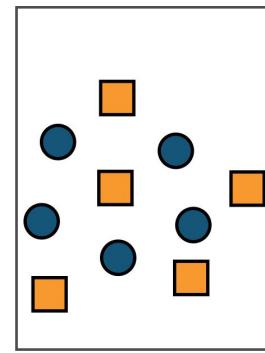
(MITCHELL, 1997)

MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:
<https://www.youtube.com/c/PGCAPINPE>



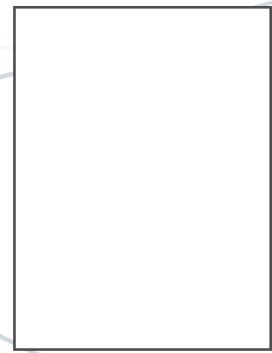
Separador de quadrados e círculos



Todos os dados



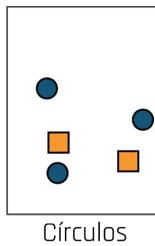
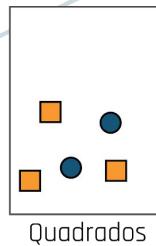
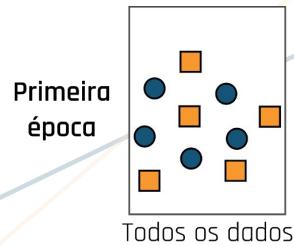
Quadrados



Círculos

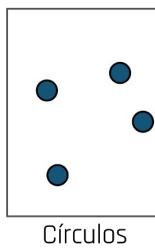
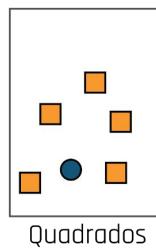


APRENDIZADO DE MÁQUINA



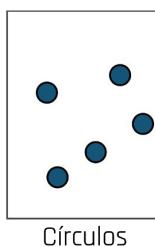
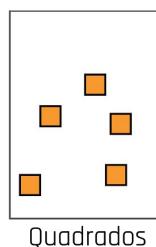
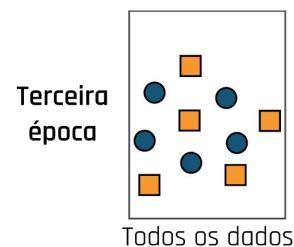
	Quadrados	Círculos
Quadrados	3	2
Círculos	2	3

Acurácia = 33,33%



	Quadrados	Círculos
Quadrados	5	0
Círculos	1	4

Acurácia = 90%



	Quadrados	Círculos
Quadrados	5	0
Círculos	0	5

Acurácia = 100%



TIPOS DE APRENDIZADO



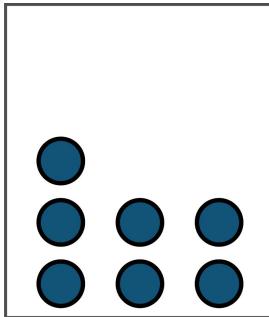
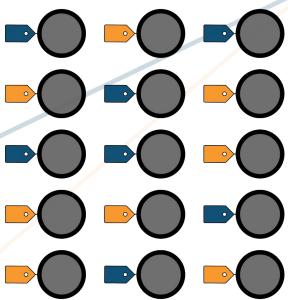
APRENDIZADO SUPERVISIONADO

MC2 - INTRODUÇÃO AO MACHINE LEARNING

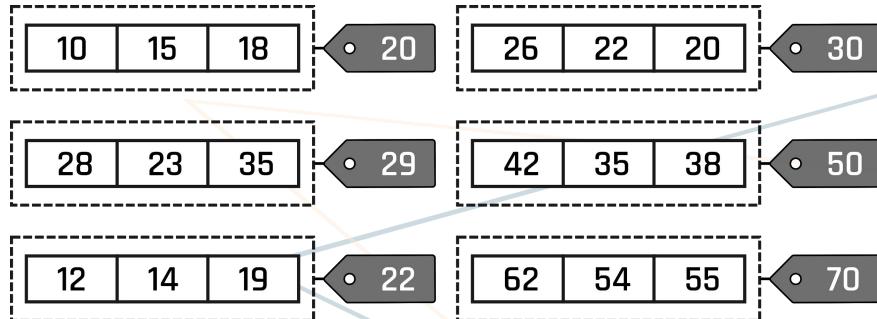
Transmissão pelo YouTube:
<https://www.youtube.com/c/PGCAPINPE>



Classificação



Regressão



Dados disponíveis

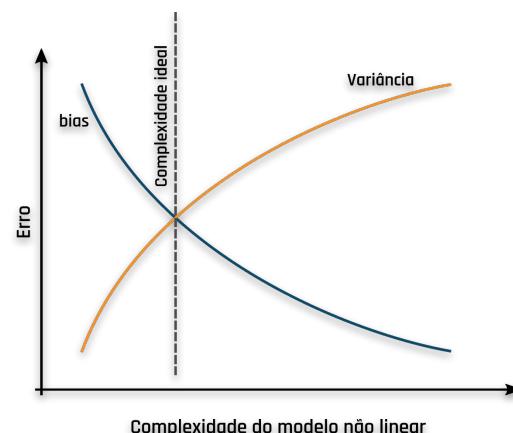
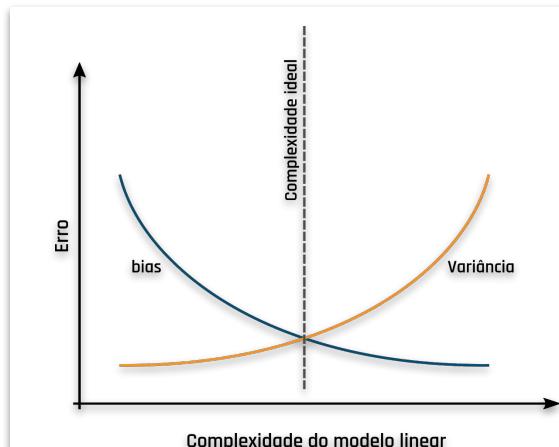
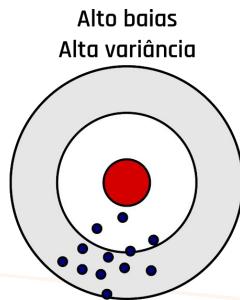
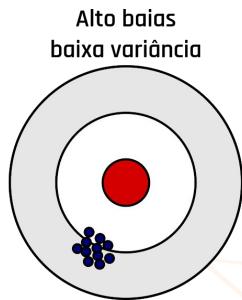
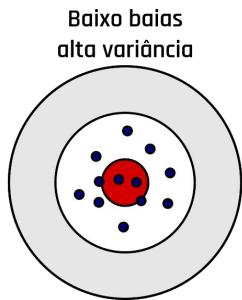
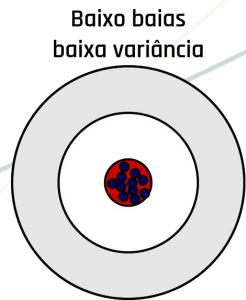


APRENDIZADO SUPERVISIONADO

MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:

<https://www.youtube.com/c/PGCAPINPE>



APRENDIZADO SUPERVISIONADO

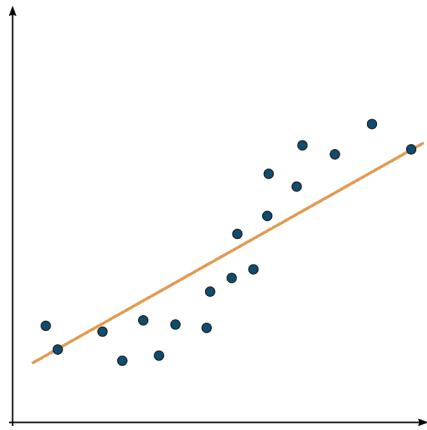
MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:

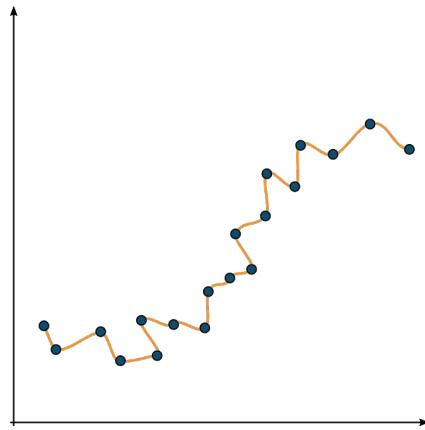
<https://www.youtube.com/c/PGCAPINPE>



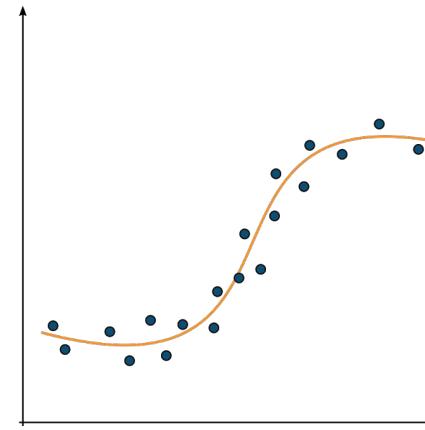
Modelo com Underfitting



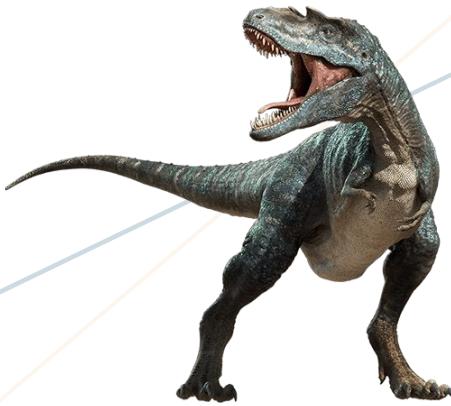
Modelo com Overfitting



Modelo ideal

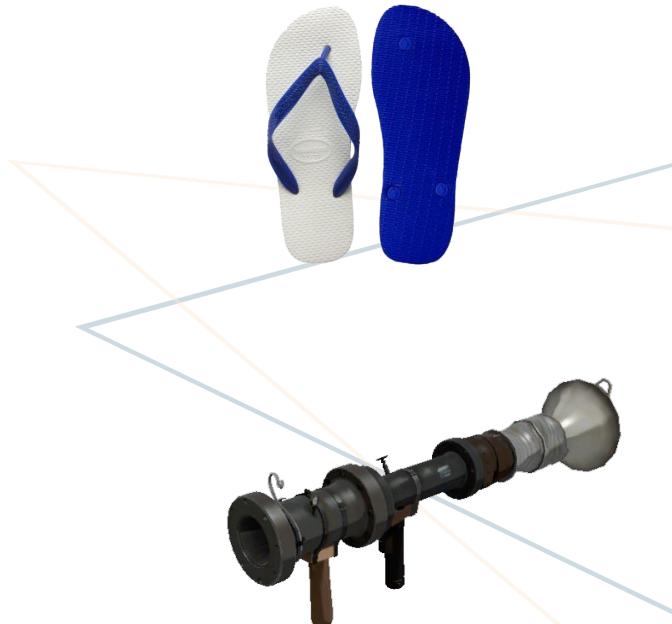


APRENDIZADO SUPERVISIONADO



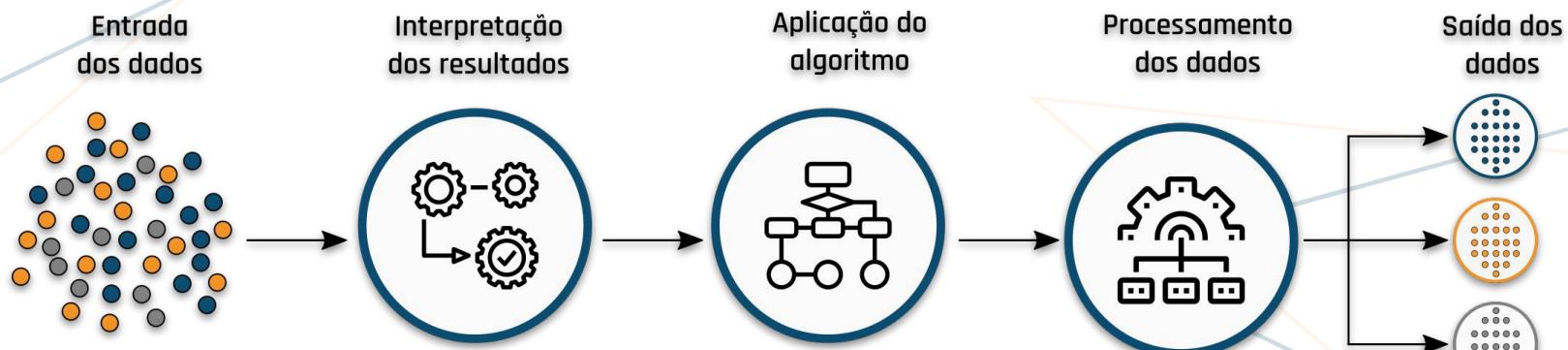
MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:
<https://www.youtube.com/c/PGCAPINPE>





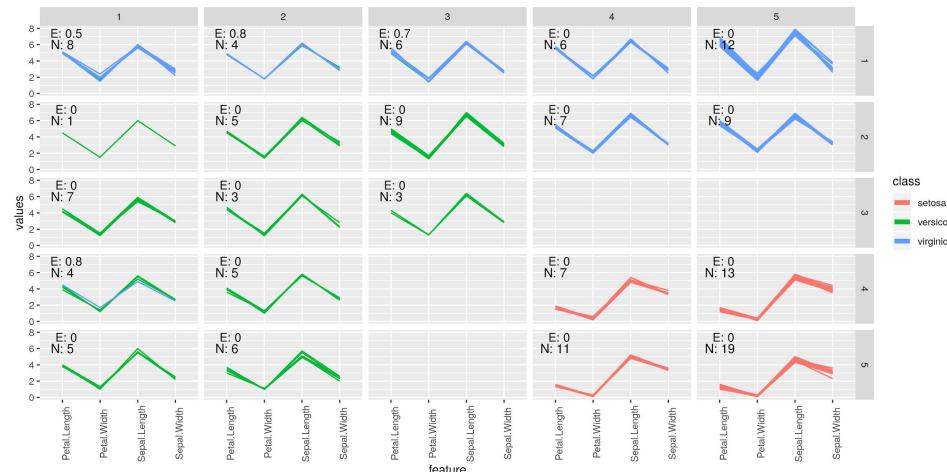
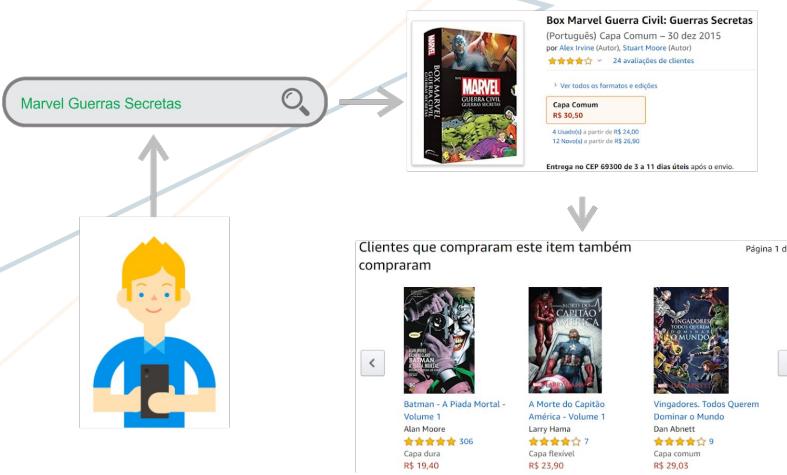
APRENDIZADO NÃO SUPERVISIONADO



Transmissão pelo YouTube:
<https://www.youtube.com/c/PGCAPINPE>



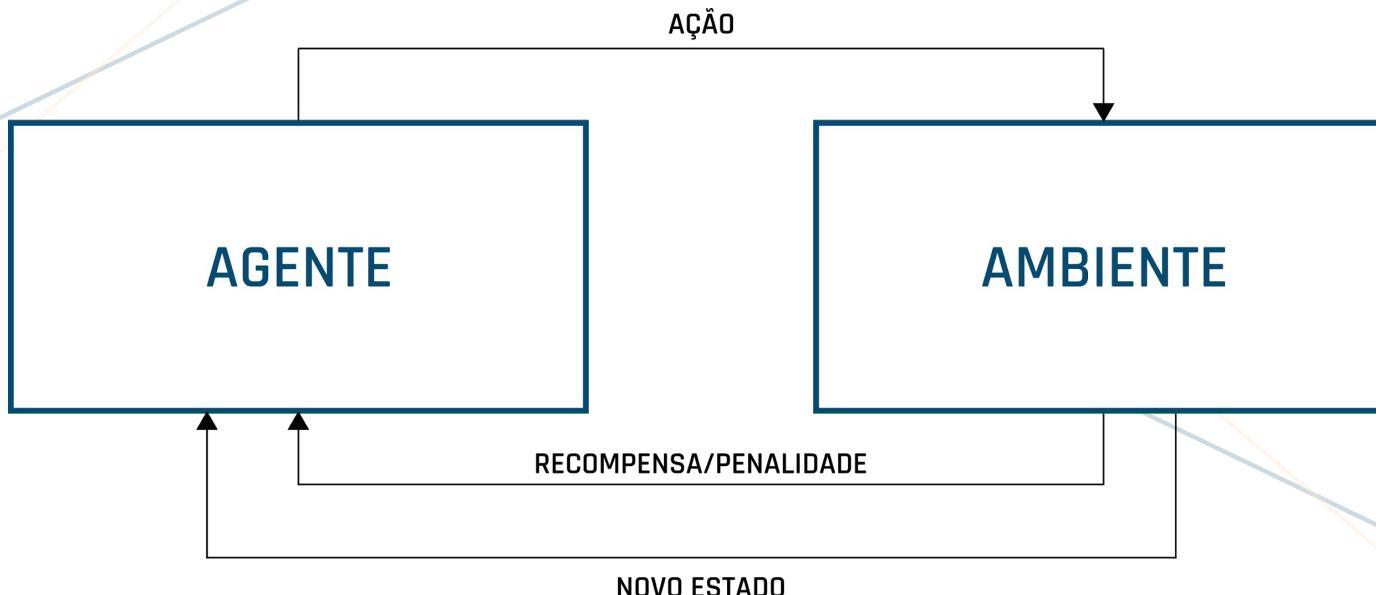
APRENDIZADO NÃO SUPERVISIONADO



(DE SOUZA, 2019)



APRENDIZADO POR REFORÇO



APRENDIZADO POR REFORÇO



MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:
<https://www.youtube.com/c/PGCAPINPE>





Workshop em
Computação
Aplicada

8-11 e 14-17 de setembro
Evento online

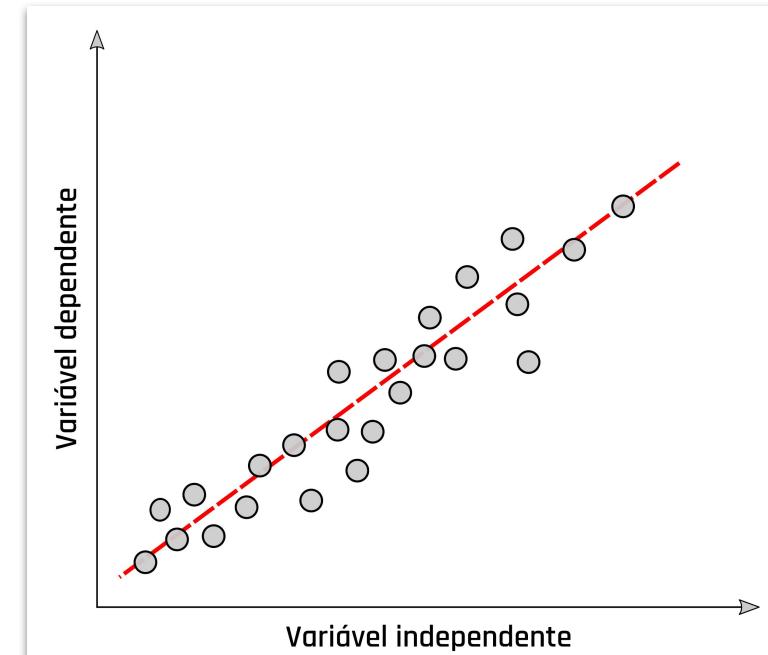
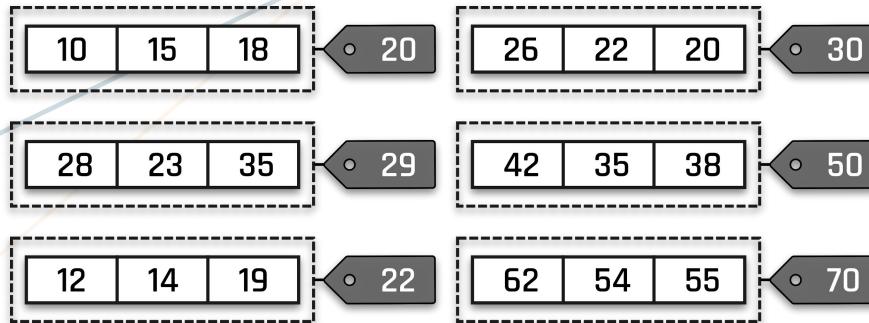
REGRESSÃO

<https://dataat.github.io/introducao-ao-machine-learning/regressao.html>

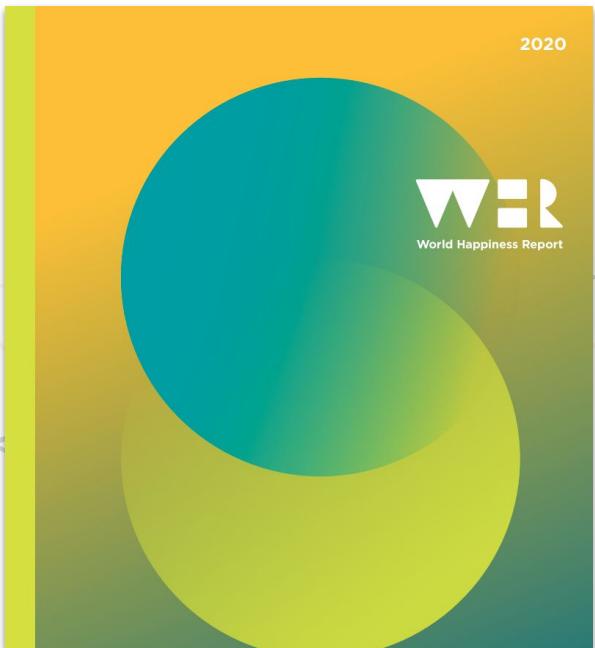
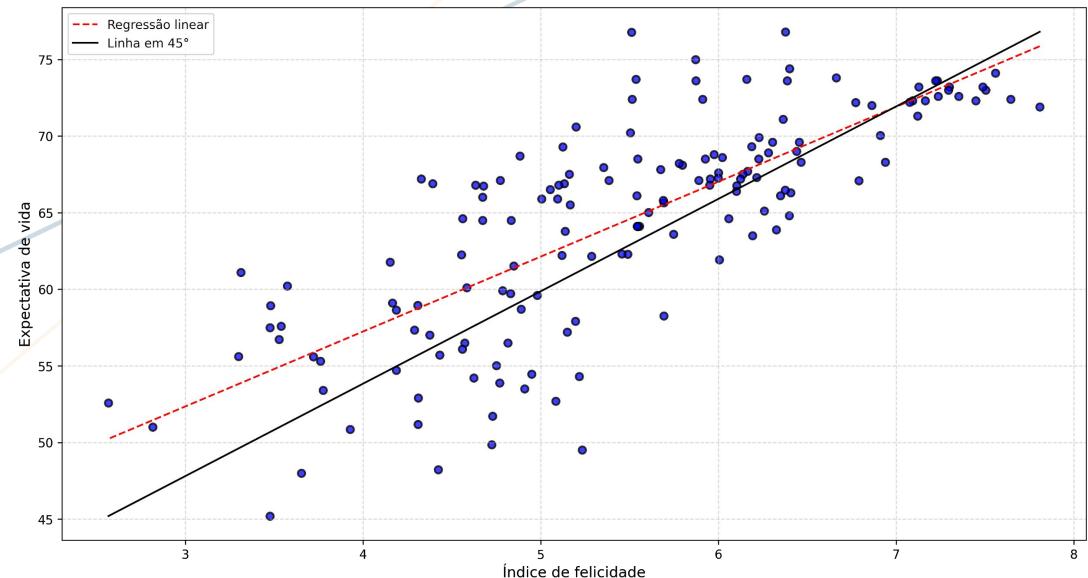
02



VARIÁVEIS INDEPENDENTE E DEPENDENTE



VARIÁVEIS INDEPENDENTE E DEPENDENTE

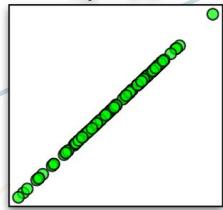


(HELLIWELL et al. 2020)

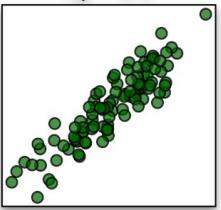


CORRELAÇÕES ENTRE VARIÁVEIS

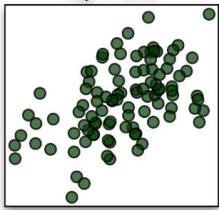
Correlação positiva perfeita
 $r_{xy} = 1.0$



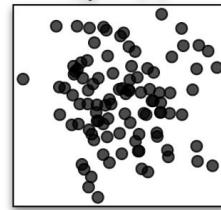
Correlação positiva alta
 $r_{xy} = 0.9$



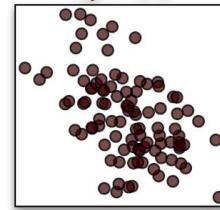
Correlação positiva baixa
 $r_{xy} = 0.5$



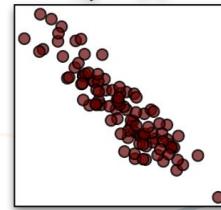
Sem correlação
 $r_{xy} = 0.0$



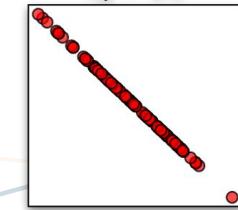
Correlação negativa baixa
 $r_{xy} = -0.5$



Correlação negativa alta
 $r_{xy} = -0.9$



Correlação negativa perfeita
 $r_{xy} = -1.0$



$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$



DANGER
ZONE



2.1 Regressão Linear

Exemplo disponível

A regressão linear é um dos métodos mais intuitivos e utilizados para essa finalidade. Esses métodos são divididos em dois grupos, a regressão linear simples (RLS) e regressão linear múltipla (RLM). A RLS tem como objetivo estabelecer uma relação entre duas variáveis através de uma função, que pode ser definida por:

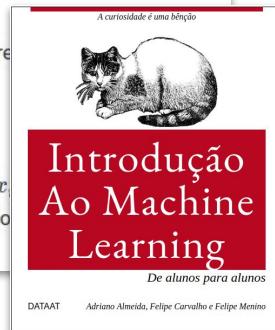
$$y_i = \alpha + \beta x_i \quad (2.2)$$

Onde y_i é a variável alvo, α e βx_i são coeficientes calculados pela regressão, que representam o intercepto no eixo y e inclinação da reta, respectivamente.

A RLM é semelhante semelhante à RLS, porém possui múltiplas variáveis preditivas definida por:

$$y_i = \alpha + \beta x_{i1} + \beta x_{i2} + \dots + \beta x_{in}$$

Onde y_i é a variável alvo, α continua sendo o coeficiente de intercepto e βx é o coeficiente angular da p -ésima variável. Ambos os métodos podem ainda serem somados para minimizar o erro.



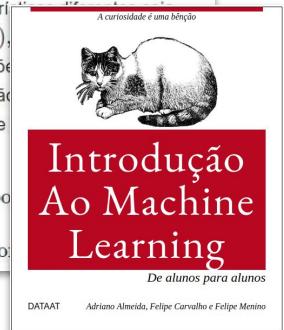
2.2 Máquina de vetores de suporte

Exemplo disponível

A máquina de vetores de suporte (SVM - sigla do inglês, support vector machine) é um modelo de aprendizado de máquina supervisionado concebido a partir de um conceito inicialmente proposto por Vapnik and Chervonenkis (1963). A SVM podem ser utilizada tanto para tarefas de classificação, quanto para tarefas de regressão, sendo uma ótima alternativa aos modelos de redes neurais artificiais profundas que tem custo computacional muito superior em dados com muitas dimensões. Outra vantagem na utilização dos modelos baseados em SVM é que eles não são sensíveis aos outliers, ou seja, valores extremos não causam ruído no treinamento.

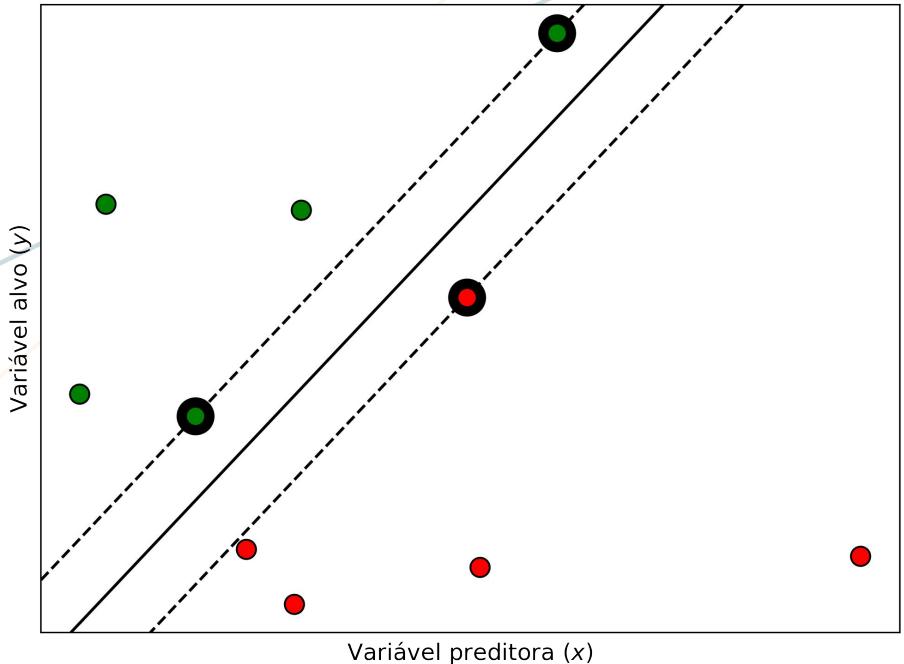
O funcionamento básico das SVM consiste em ajustar a equação de uma reta, denominada hiperplano de tal forma que a distância entre ela e os pontos com características maximizada. Um conjunto de n pontos é definido como $(\vec{x}_1, y_1), (\vec{x}_2, y_2), \dots, (\vec{x}_n, y_n)$, onde \vec{x}_i são as variáveis independentes representadas por um vetor de d -dimensões e y_i são as variáveis dependentes. A distância maximizada entre o hiperplano e as fronteiras são chamadas de margens e os pontos que estão no limite dessa margem são os vetores de decisão. Os componentes podem ser modelados da seguinte forma:

$$\vec{w} \cdot \vec{x} - b = \begin{cases} -1, & \text{primeiro vetor de suporte} \\ 0, & \text{hiperplano} \\ 1, & \text{segundo vetor de suporte} \end{cases}$$



<https://dataat.github.io/introducao-ao-machine-learning/regressao.html>

MÁQUINA DE VETORES DE SUPORTE

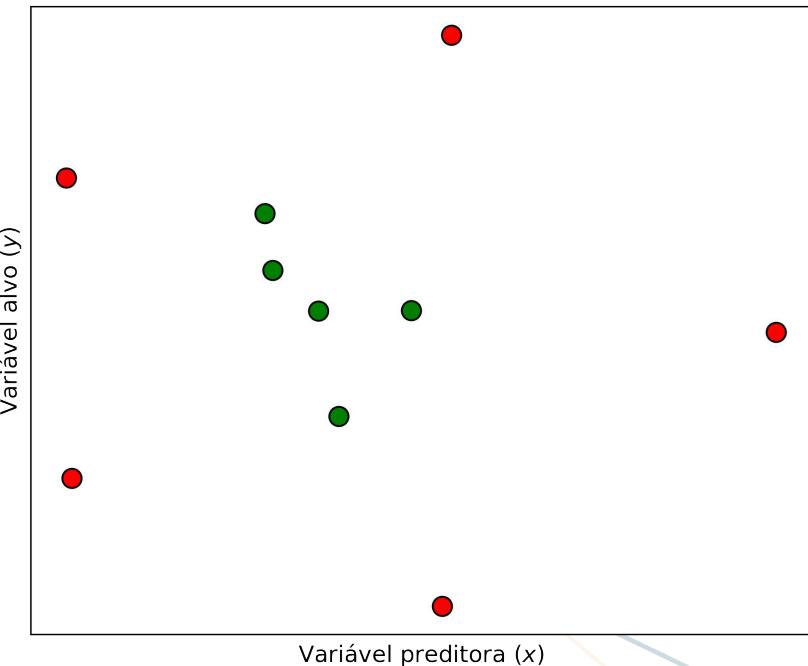
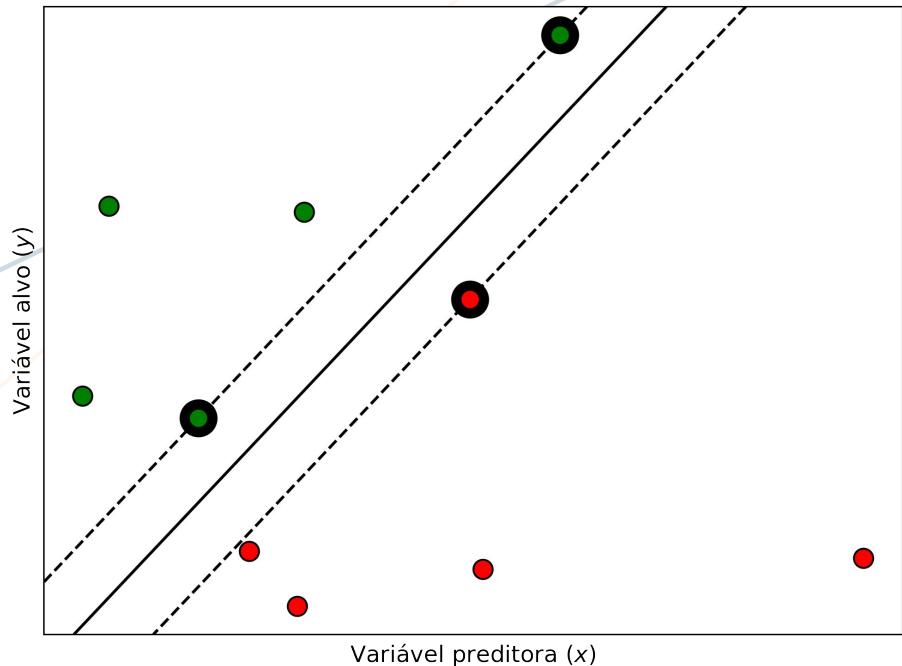


$$(\vec{x}_1, y_1), (\vec{x}_2, y_2), \dots, (\vec{x}_n, y_n)$$

$$\vec{w} \cdot \vec{x} - b = \begin{cases} -1, & \text{primeiro vetor de suporte} \\ 0, & \text{hiperplano} \\ 1, & \text{segundo vetor de suporte} \end{cases}$$

(VAPNIK E CHERVONENKIS, 1963)

MÁQUINA DE VETORES DE SUPORTE



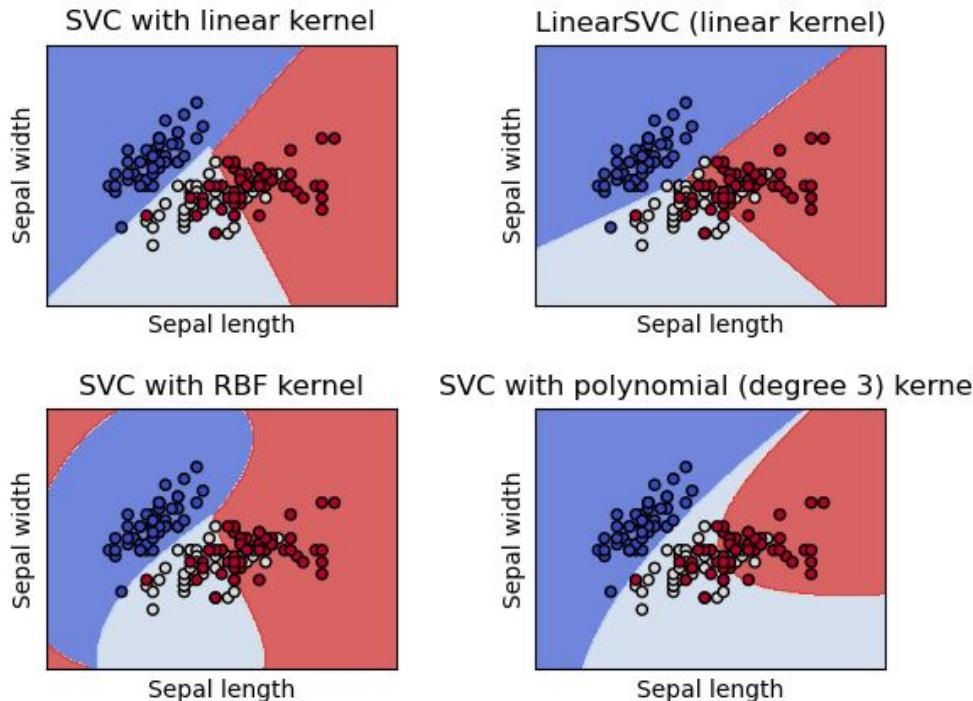
MÁQUINA DE VETORES DE SUPORTE

KERNELS

MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:

<https://www.youtube.com/c/PGCAPINPE>



<https://scikit-learn.org/stable/modules/svm.html>

(BOSER, GUYON, AND VAPNIK, 1992)

MÁQUINA DE VETORES DE SUPORTE

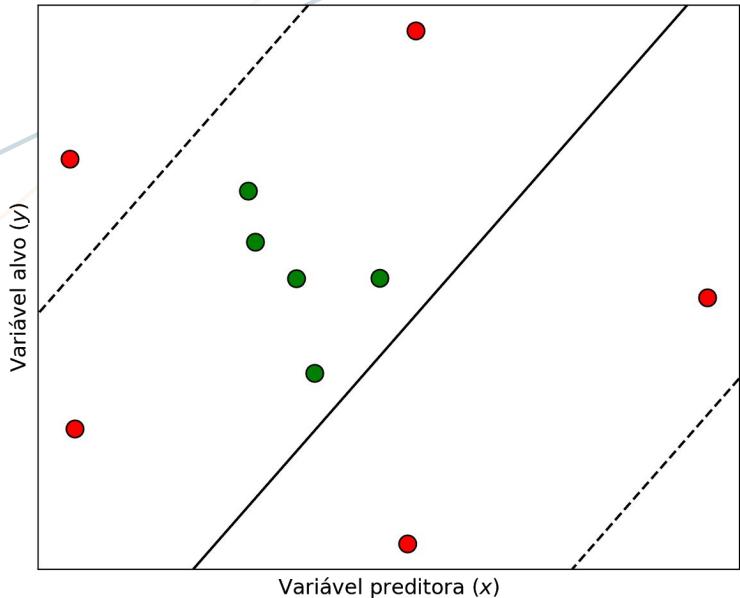
MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:

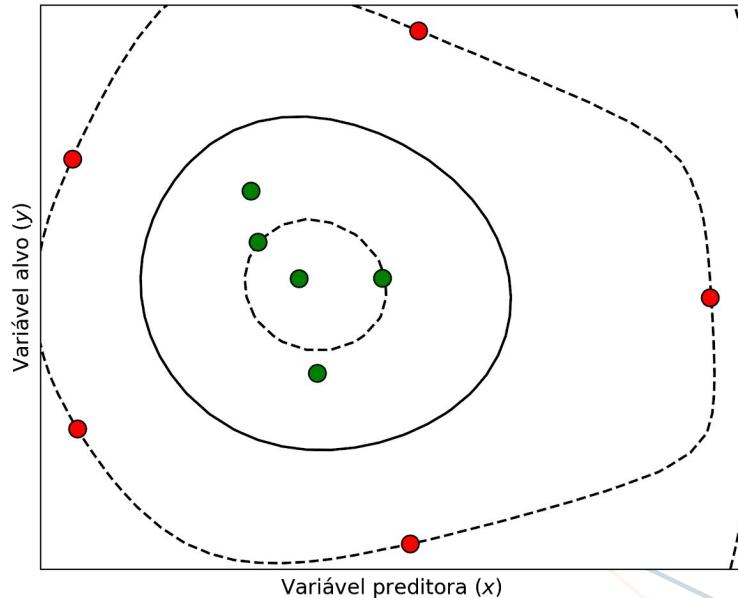
<https://www.youtube.com/c/PGCAPINPE>



Kernel Linear



Kernel não linear



MÁQUINA DE VETORES DE SUPORTE

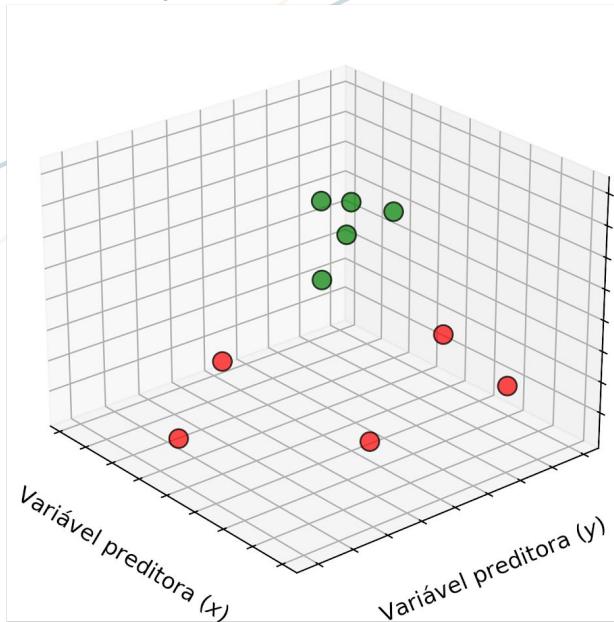
MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:

<https://www.youtube.com/c/PGCAPINPE>

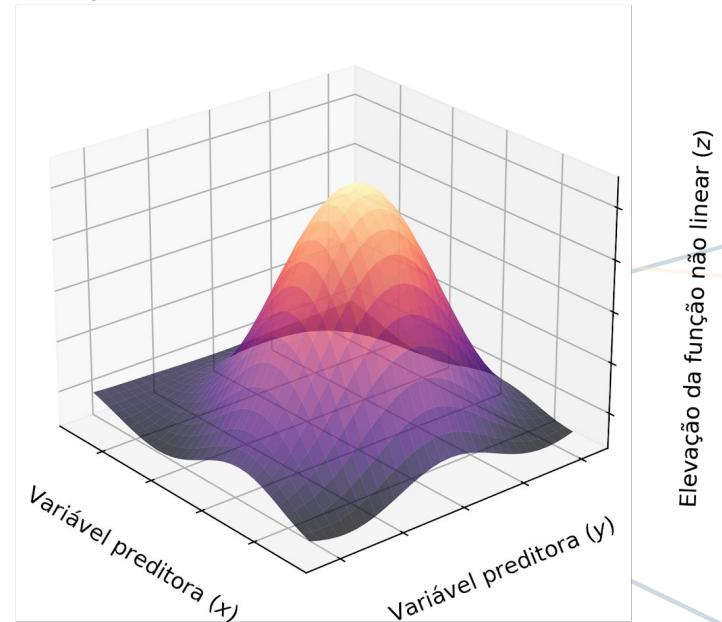


Distribuição tridimensional dos dados



Elevação da função não linear (z)

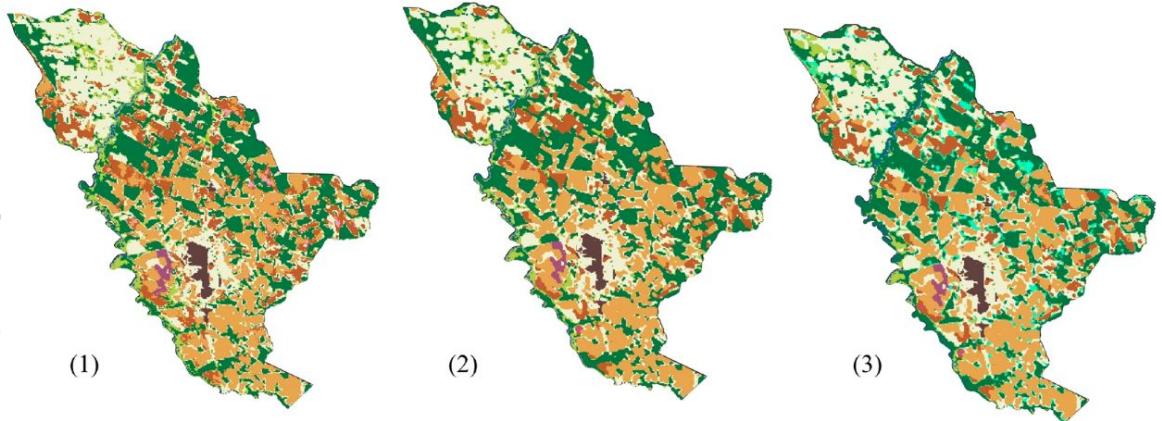
Função não linear do modelo com SVM



Elevação da função não linear (z)



MÁQUINA DE VETORES DE SUPORTE



1. Savanna	4. Pasture	7. Soy_Fallow	10. Sugarcane	13. Secondary Vegetation
2. Fallow_Cotton	5. Soy_Corn	8. Soy_Millet	11. Urban Area	
3. Forest	6. Soy_Cotton	9. Soy_Sunflower	12. Water	

www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN
DATA DESCRIPTOR

Land use and cover maps for Mato Grosso State in Brazil from 2001 to 2017

Rolf Simoes^{1,*}, Michelle C. A. Picoli¹, Gilberto Camara^{1,2}, Adeline Maciel¹, Lorena Santos¹, Pedro R. Andrade¹, Alber Sanchez², Karine Ferreira³ & Alexandre Carvalho³

This paper presents a dataset of yearly land use and land cover classification maps for Mato Grosso State, Brazil, from 2001 to 2017. Mato Grosso is one of the world's fast moving agricultural frontiers. To ensure multi-year compatibility, the study uses MODIS sensor-ready products and an innovative method to obtain annual land cover classes to classify satellite imagery time series. The dataset provides information about crop and pasture expansion, mineral extraction, as well as spatially explicit estimates of increases in agricultural productivity and trade-offs between crop and pasture expansion. Therefore, the dataset provides new and relevant information to understand the impact of environmental policies on the expansion of tropical agriculture in Brazil. Using such results, researchers can make informed assessments of the interplay between production and protection within Amazon, Cerrado, and Pantanal biomes.

Background & Summary
Brazil is one of the top agricultural producers and exporters, being the largest extent of tropical rainforest and home to an estimated 15% to 20% of the world's biodiversity. Such unique position leads to the need for balancing agriculture production and environmental protection. Without substantial investments in productivity and areas under protection, the expansion of agricultural production in Brazil can be a significant factor in environmental degradation. It is vital to understand the impact of environmental policies on the expansion of tropical agriculture in Brazil.

In Nationally Determined Contribution (NDC) to the United Nations Framework Convention on Climate Change (UNFCCC), Brazil committed to achieve zero net emissions by 2030 in the Amazon rainforest by 2030. The forest emission balance will be achieved by restoring and reforesting 12 million hectares. Brazil's NDC also makes a firm commitment to promote low-carbon agriculture and to increase biofuel use for transportation. Overall, achieving the emission reduction goals Brazil set in its NDC will highly depend on how the country meets the targets associated with the land use sector.

Since the election of the current government in late 2018, there has been a tension between the desire for more agricultural exports and rural protectionism required to strengthen cattle breeding. While the export sector supports the country's pledges to the Paris Agreement, most cattle ranchers and smallholders do not want to commit to environmental protection policies.¹ Since the traditional rural sector is one of the primary supporters of current government, there are increasing concerns about whether Brazil will be committed to achieve its NDC. Comparing the environmental impact of different agricultural sectors is therefore important for all stakeholders involved in the debate.

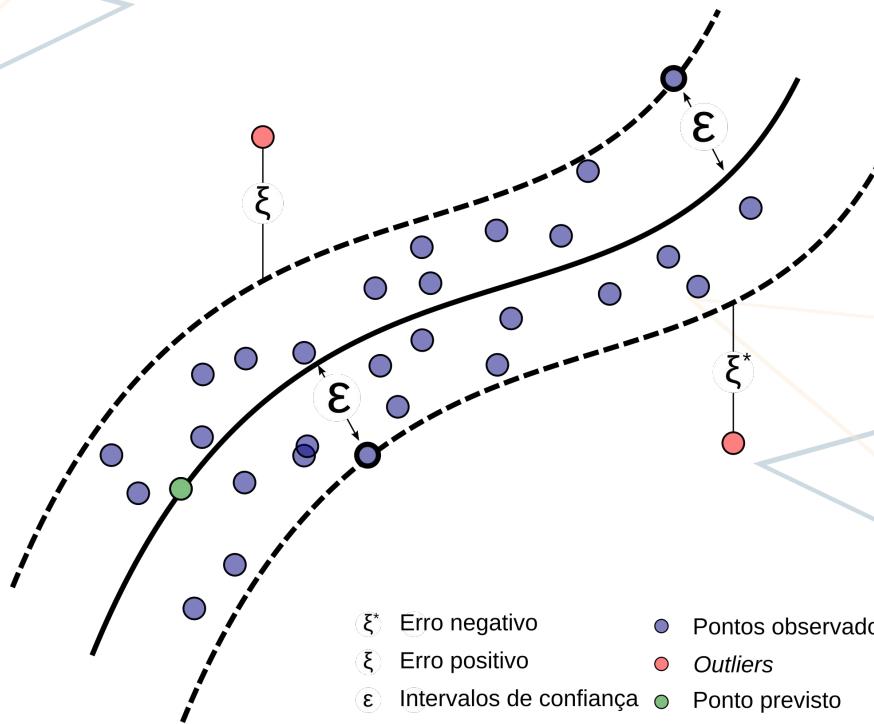
The legal basis for land policies in Brazil is the Forest Code. When created in 1965, it established a proportion of rural properties that must be permanently maintained as forest (legal reserve). It also prohibited clearing vegetation in sensitive areas such as steep slopes and along riverbanks and streams. In 2012, Congress approved a revision of the Forest Code. It stipulates that landowners, in the Legal Amazon, must conserve 80% of their property in forest areas, 35% in cerrado areas, 20% in general fields. To monitor compliance with the new Forest Code, Brazil has been successfully using wall-to-wall satellite-based monitoring². Using satellite observations

(SIMOES ET AL., 2020)

REGRESSÃO COM MÁQUINA DE VETORES DE SUPORTE

MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:
<https://www.youtube.com/c/PGCAPINPE>



(DRUKER ET AL., 1997)

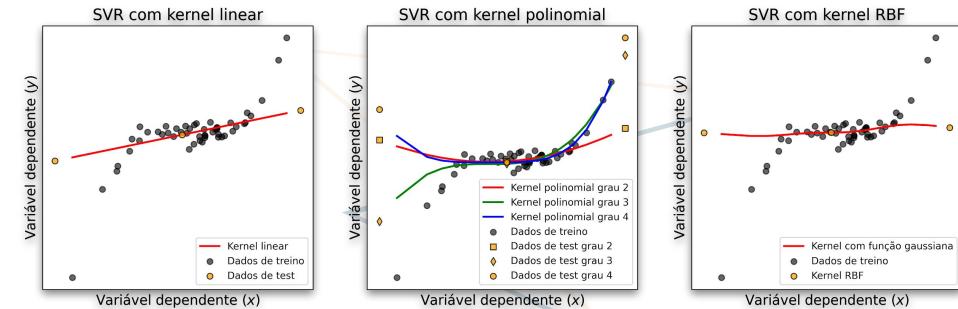


EXEMPLO EM PYTHON

```
● ● ●
from sklearn.svm import SVR

model_linear = SVR(kernel='linear').fit(x_train, y_train)
model_polyg2 = SVR(kernel='poly', degree=2).fit(x_train, y_train)
model_polyg3 = SVR(kernel='poly', degree=3).fit(x_train, y_train)
model_polyg4 = SVR(kernel='poly', degree=4).fit(x_train, y_train)
model_rbf = SVR(kernel='rbf').fit(x_train, y_train)

y_test_linear = model_linear.predict(x_test)
y_test_polyg2 = model_polyg2.predict(x_test)
y_test_polyg3 = model_polyg3.predict(x_test)
y_test_polyg4 = model_polyg4.predict(x_test)
y_test_rbf = model_rbf.predict(x_test)
```



HORA DE PRATICAR!

MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:

<https://www.youtube.com/c/PGCAPINPE>



Workshop em
Computação
Apliada
8-11 e 14-17 de setembro
Evento online

MC2 - INTRODUÇÃO AO MACHINE LEARNING
Transmissão pelo YouTube:
<https://www.youtube.com/c/PGCAPINPE>

Exemplo de aplicação do algoritmo de regressão linear

Instrutores: Adriano Almeida, Felipe Carvalho e Felipe Menino

Realização: Dia 15/09

Descrição: Este notebook tem como propósito apresentar a aplicação do algoritmo de regressão linear no conjunto de dados *World Happiness Report 2020* utilizando a biblioteca *scikit-learn* na linguagem de programação *Python*.

Descrição do conjunto de dados ^



O conjunto de dados utilizado neste notebook é o *World Happiness Report 2020* publicado por Helliwell et al. (2020). O primeiro *World Happiness* foi apresentado em 2012 na Assembleia Geral das Nações Unidas. Este é o oitavo relatório, e assim como os anteriores, busca fornecer informações sobre o estado de felicidade e bem-estar da população em diversos países ao redor do mundo. Este conjunto de dados contém informações de 153 países, provenientes principalmente dos atributos descritos a seguir:

<https://www.kaggle.com/lordadriano/intro-ml-python-linear-regression-worcap2020>



Workshop em
Computação
Apliada
8-11 e 14-17 de setembro
Evento online

MC2 - INTRODUÇÃO AO MACHINE LEARNING
Transmissão pelo YouTube:
<https://www.youtube.com/c/PGCAPINPE>

Exemplo de aplicação da regressão com máquinas de vetores de suporte

Instrutores: Adriano Almeida, Felipe Carvalho e Felipe Menino

Realização: Dia 15/09

Descrição: Este notebook tem como propósito apresentar a aplicação da regressão com máquinas de suporte de vetores no conjunto de dados *Boston House Prices* utilizando a biblioteca *scikit-learn* na linguagem de programação *Python*.

Descrição do conjunto de dados ^



O conjunto de dados utilizado neste notebook é o *Boston House Prices*, originalmente publicado por Harrison e Rubinfeld (1978). Estes dados foram coletados de diferentes áreas da capital de Massachusetts, Boston, pelo U.S Census Service em 1970 e possuem informações associadas aos preços dos imóveis dessas regiões. O *Boston House Prices* é amplamente utilizado como exemplo em exercícios de aprendizado de máquina, em especial para tarefas de regressão. Este conjunto de dados possui 506 registros e 13 atributos descritos a seguir:

<https://www.kaggle.com/lordadriano/intro-ml-python-svr-worcap2020>



Workshop em
Computação
Aplicada

8-11 e 14-17 de setembro
Evento online

CLASSIFICAÇÃO

<https://dataat.github.io/introducao-ao-machine-learning/classificacao.html>

03



Aprendizado de Máquina



CLASSIFICAÇÃO



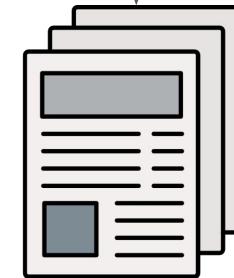
- SPAMs causam danos diariamente (Blanzieri e Bryl, 2008)
 - Roubo de cartão de crédito
 - Vendas falsas
 - Compras inexistentes
- Fazer a identificação de SPAMs pode ajudar a reduzir esses problemas
- **Questão:** Como ?

CLASSIFICAÇÃO



Exemplos de SPAM

Analista



Conjunto de regras para
identificação de SPAM

MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:
<https://www.youtube.com/c/PGCAPINPE>



Implementação

CLASSIFICAÇÃO

MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:

<https://www.youtube.com/c/PGCAPINPE>



Teste com email

Atualização de conta Caixa de entrada ×

Supor**e** DO INPE (email:**suportedoinpe@iinpe.com.br**)
para mim ▾ 17:14 (há 2 minutos)

Olá Usuáorio,

Identificamos que sua senha venceu e precisa ser inserida novamente no sistema

Clique aqui para atualizar

Obrigado

Equipe de Suporte do INPE

SPAM!

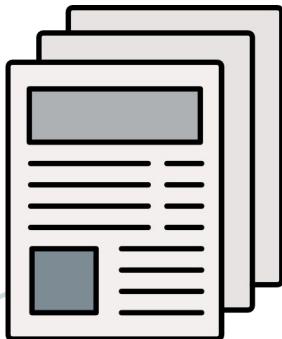
The screenshot shows an email from 'Supor'e DO INPE' with a subject 'Atualização de conta'. The message body contains several lines of text, all of which are highlighted with red dashed boxes. The text includes a greeting, a statement about the password expiring, and a link to update it. The word 'SPAM!' is written in large letters at the bottom right of the email window.

CLASSIFICAÇÃO

MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:

<https://www.youtube.com/c/PGCAPINPE>



Teste com email

MUITAS OFERTAS AGORA Caixa de entrada ×

Equipe Bela casa (email: belacasa@bellacasa.com) 18:05 (há 5 minutos)

para mim ▾

Fala usuário!

Estamos de casa nova! Para comemorar a mudança de nosso novo site estamos com muitas ofertas. Para conferir, clique aqui.

!!!PARA APROVEITAR AS OFERTAS, ATUALIZEI AQUI SEU CARTÃO DE CRÉDITO!!!!

Esperamos você lá

Equipe da LOJA BELA CASA

NÃO É SPAM!

The screenshot shows an email inbox with one message from 'Equipe Bela casa'. The subject is 'MUITAS OFERTAS AGORA'. The message body contains promotional text about new website offers and a link to 'clique aqui'. It ends with 'Esperamos você lá' and 'Equipe da LOJA BELA CASA'. The email is timestamped at 18:05 (5 minutes ago). A large text 'NÃO É SPAM!' is overlaid on the bottom right of the inbox. On the left side of the slide, there is a vertical list of three bullet points: 'Erros ortográficos', 'Pedido de senhas', and 'Email suspeito'. There is also a small icon of a document with a list of items.

CLASSIFICAÇÃO

MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:
<https://www.youtube.com/c/PGCAPINPE>

As técnicas de classificação são utilizadas para a identificação do rótulo de determinadas observações com base em características e informações previamente conhecidas (Lantz, 2013)

CLASSIFICAÇÃO

MC2 - INTRODUÇÃO AO MACHINE LEARNING

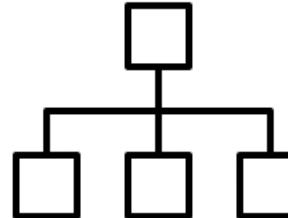
Transmissão pelo YouTube:

<https://www.youtube.com/c/PGCAPINPE>



Exemplos de SPAM

Algoritmo de classificação



Conjunto de regras generalizadas para
a identificação de SPAM

CLASSIFICAÇÃO

MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:

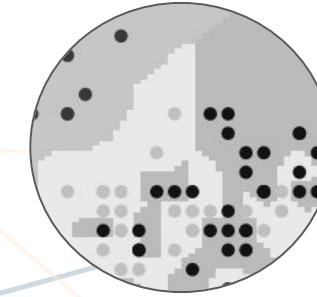
<https://www.youtube.com/c/PGCAPINPE>



Decision tree



Naive Bayes



k -Nearest Neighbors

k-Nearest Neighbors

MC2 - INTRODUÇÃO AO MACHINE LEARNING

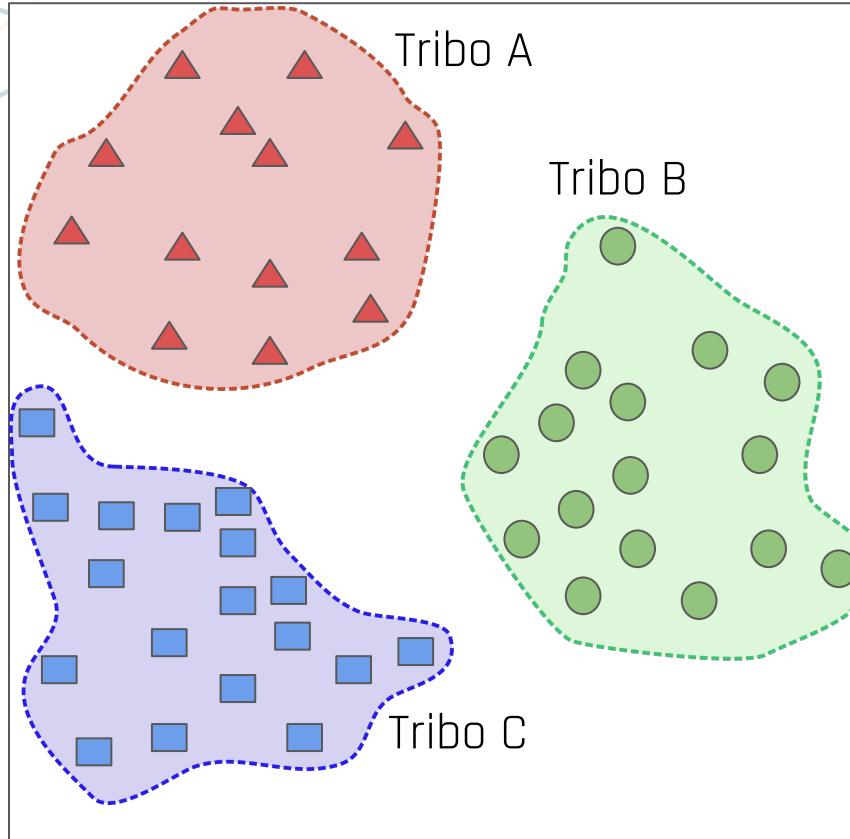
Transmissão pelo YouTube:
<https://www.youtube.com/c/PGCAPINPE>



O k-Nearest neighbors (kNN) é um algoritmo de aprendizado de máquina, supervisionado, utilizado para a **classificação e regressão**

- **Não-Paramétrico:** Não faz suposições sobre a distribuição dos dados
- **Baseado em instância:** O algoritmo não cria um modelo para mapear os dados. Os dados de treino só são utilizados quando uma amostra precisa ser classificada.
 - A etapa de treino é omitida (Aggarwal, 2015)

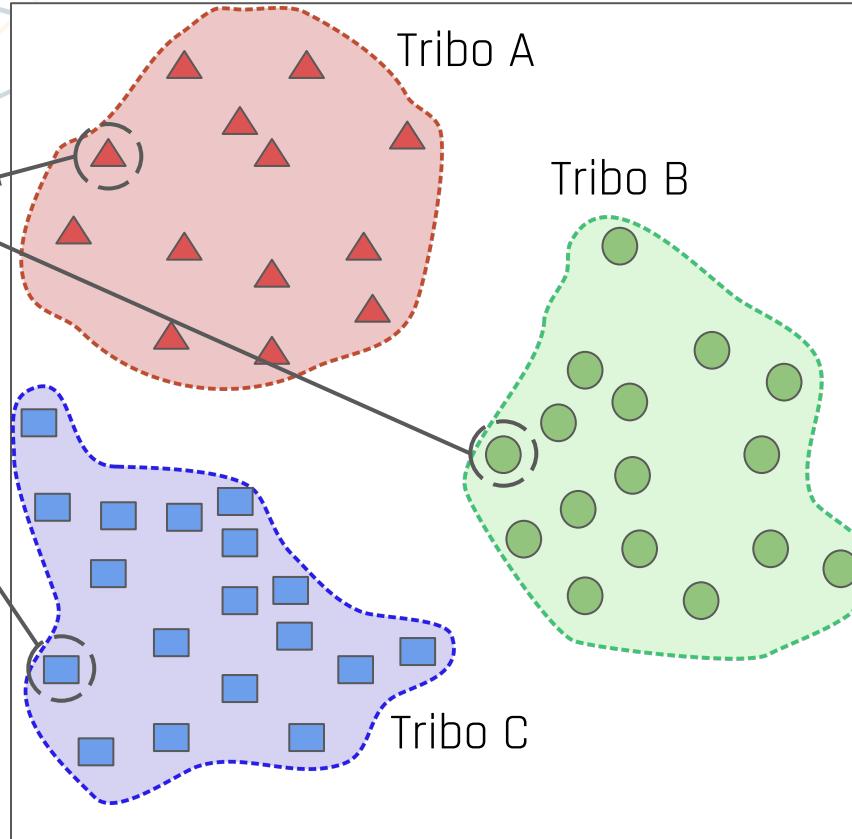
k-Nearest Neighbors



k-Nearest Neighbors



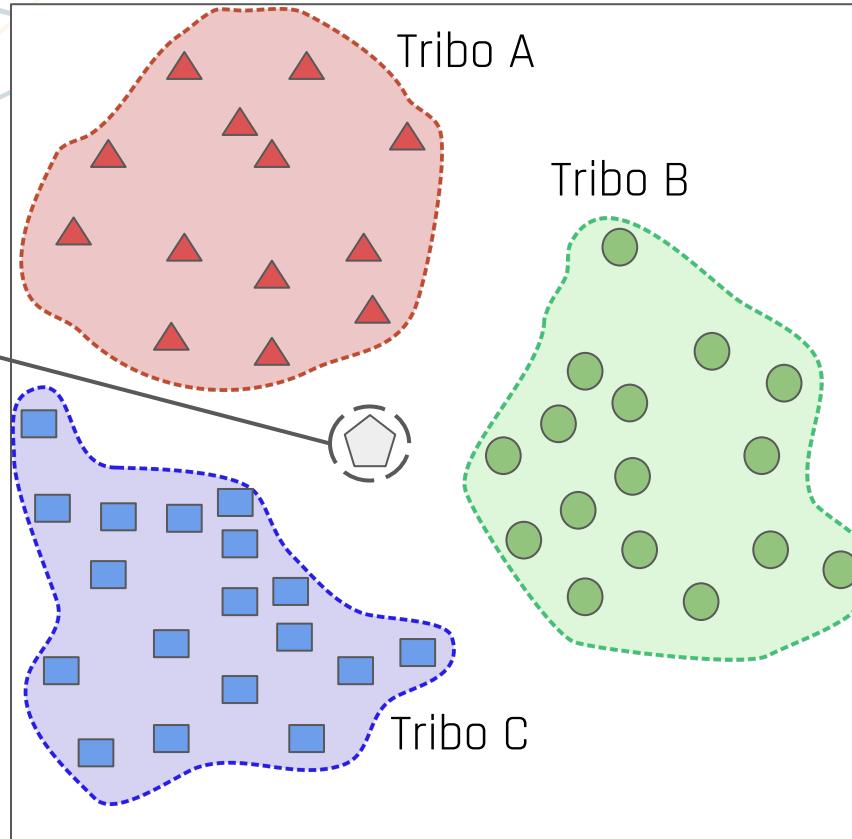
Pontos de plantio



k-Nearest Neighbors



Novo ponto de plantio

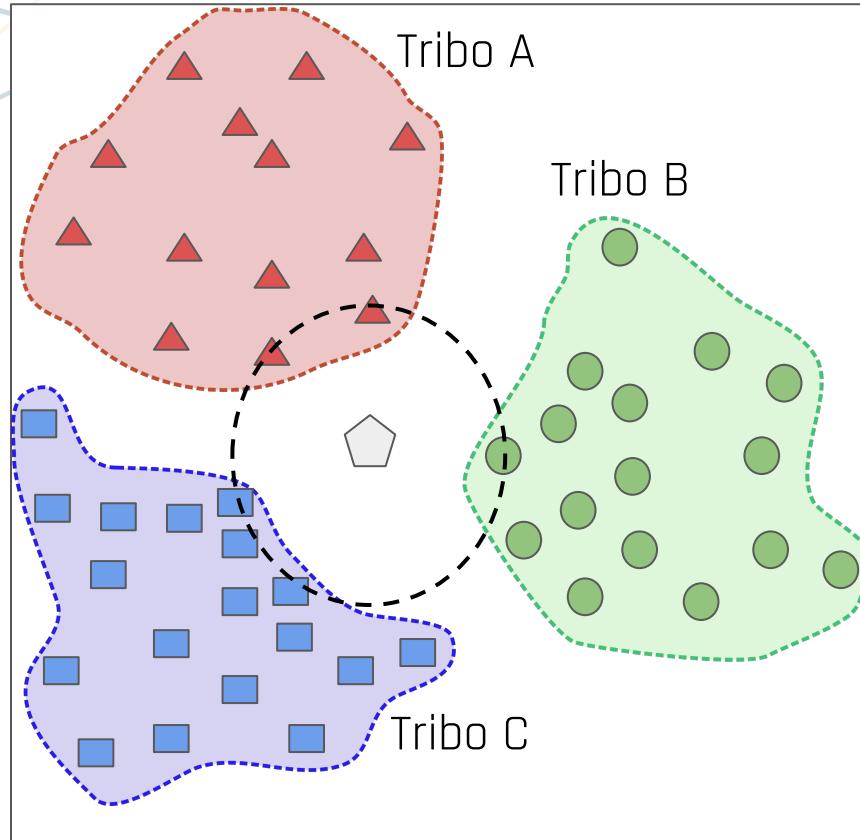




***k*-Nearest Neighbors**

Análise de vizinhança

Verifica o rótulo dos ***k*** vizinhos mais próximos para a determinação do rótulo do novo elemento

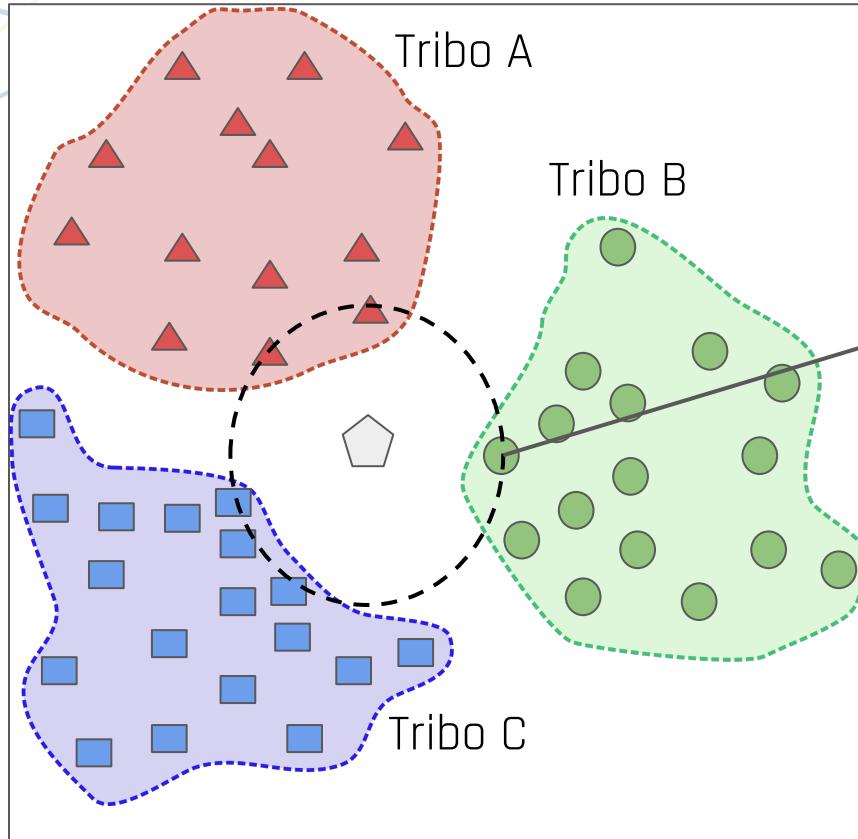




k-Nearest Neighbors

Análise de vizinhança

Verifica o rótulo dos ***k*** vizinhos mais próximos para a determinação do rótulo do novo elemento



Pode ser calculado com a **distância euclidiana**, mas outras poderiam ser aplicadas

k-Nearest Neighbors

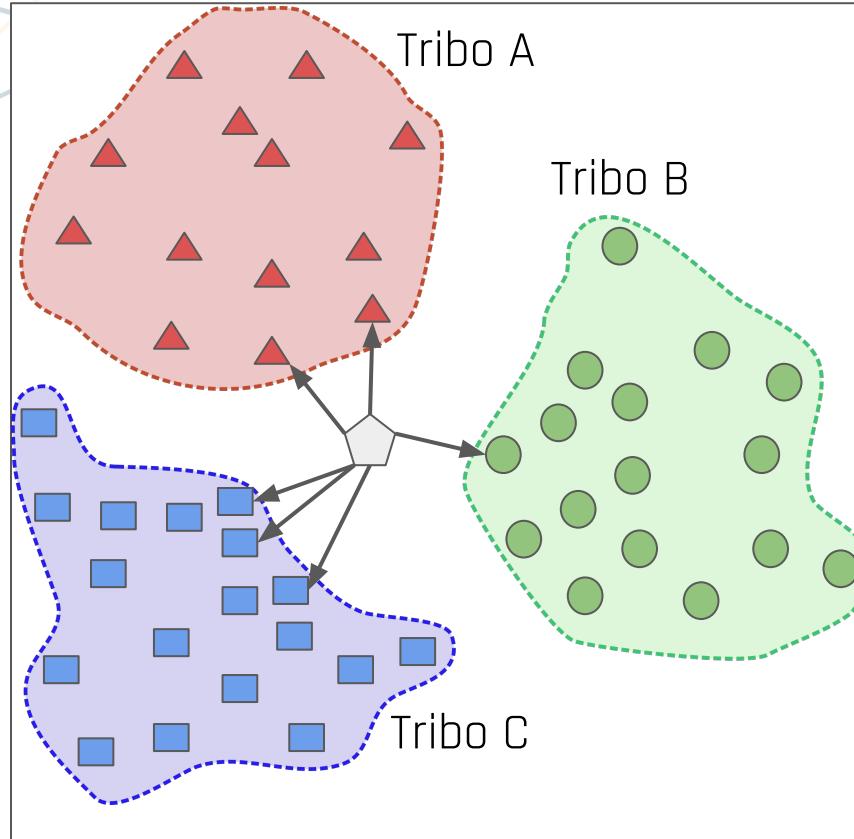


Análise de
vizinhança

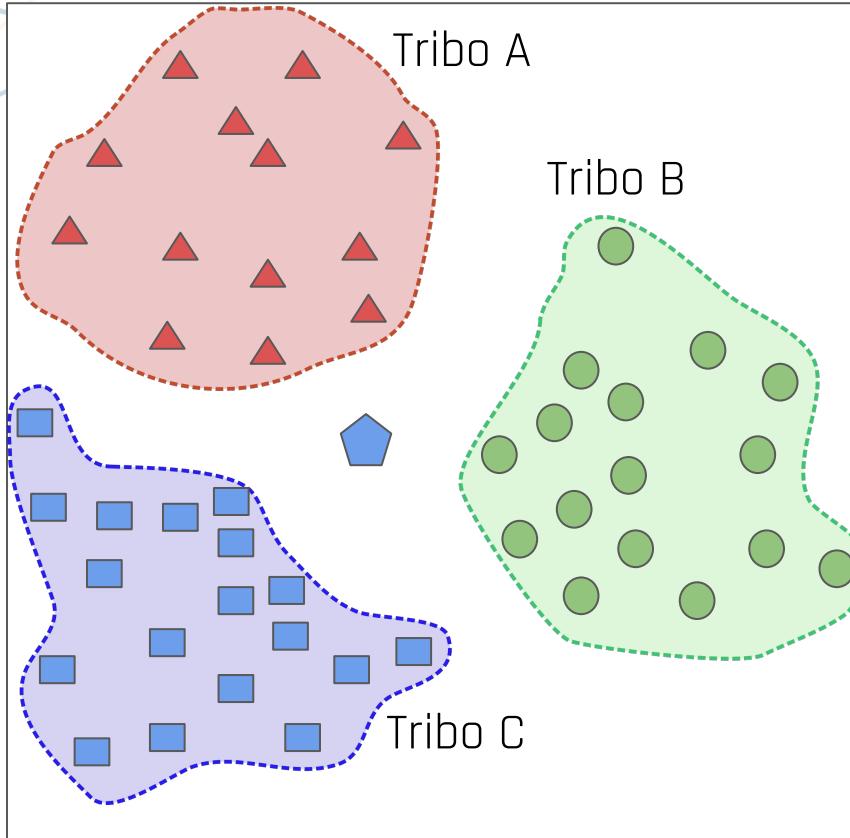
2

3

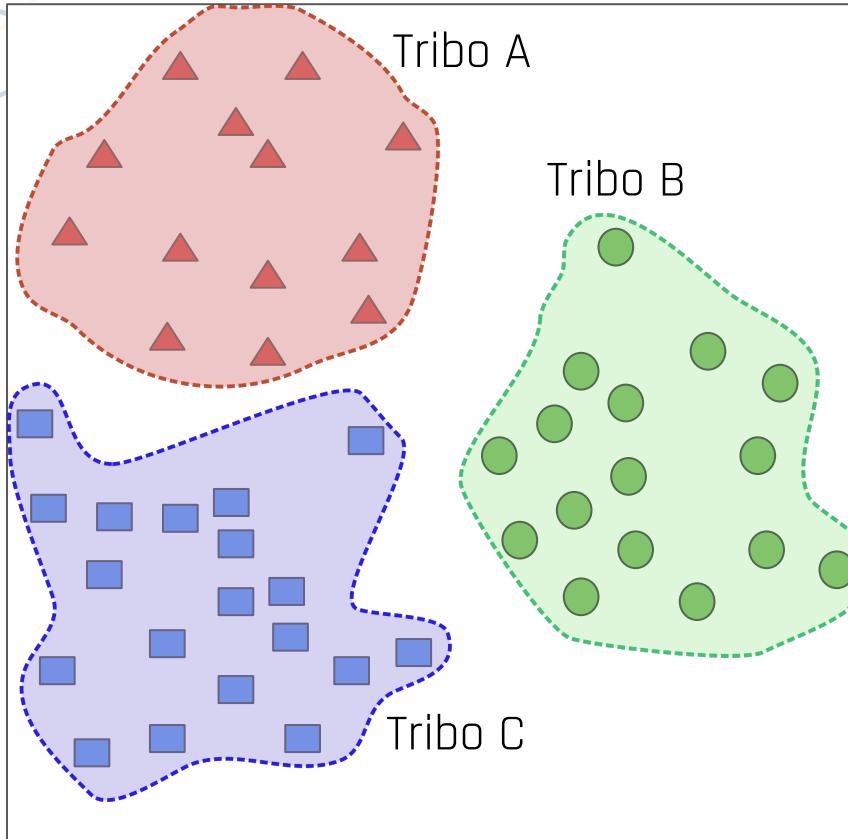
1



k-Nearest Neighbors



k-Nearest Neighbors



k-Nearest Neighbors



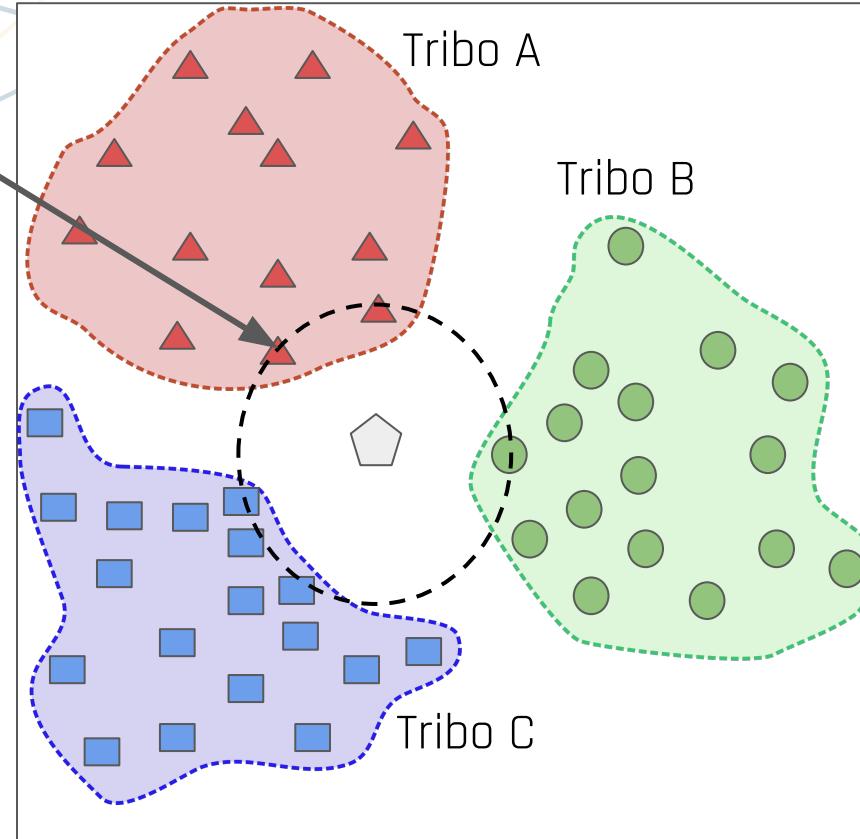
Resumindo os passos do kNN

- 1º Calcula a distância entre o ponto a ser classificado e **todos** demais pontos de dados
 - No exemplo: **Cálculo da distância do novo ponto de plantio para todos os demais**
- 2º Define os k pontos mais próximos
 - No exemplo: **Determina os pontos de plantio mais próximos do novo ponto**
- 3º Faz a análise de vizinhança
 - No exemplo: **Verifica o rótulo (Tribo) dos vizinhos**
- 4º Determina a rótulo do novo ponto

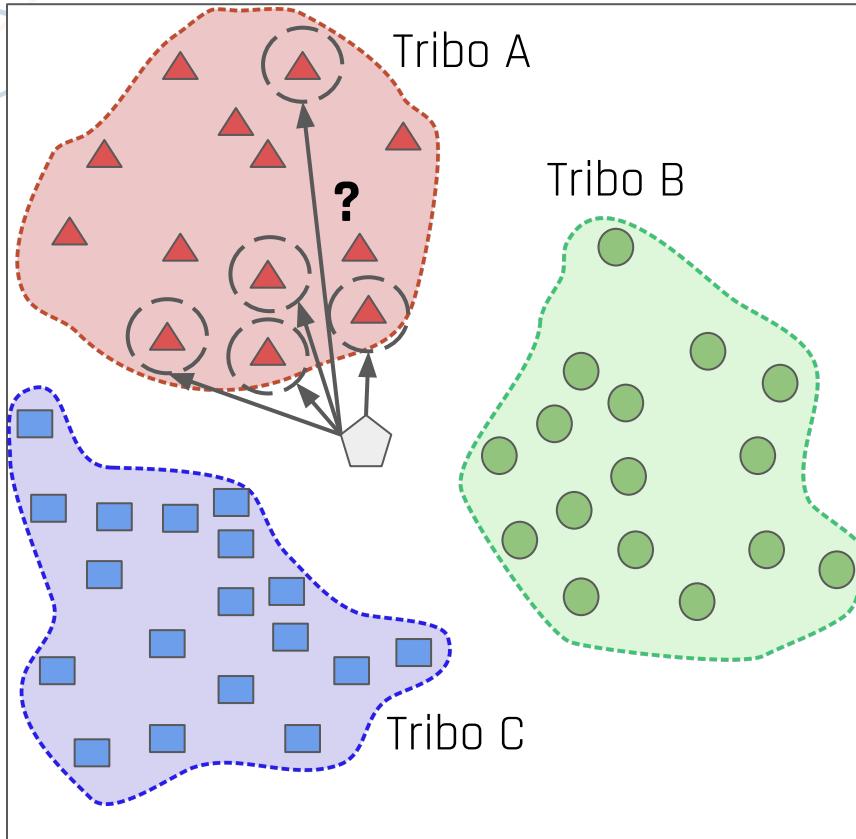
k-Nearest Neighbors - Complexidade



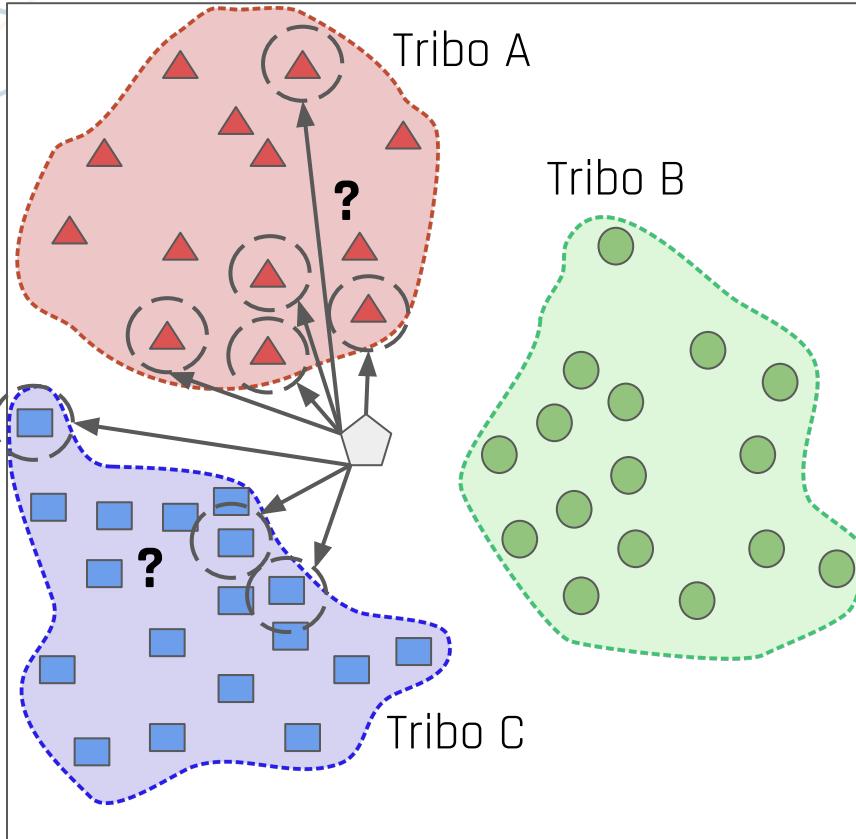
Como saber quem
são os elementos
mais próximos ?



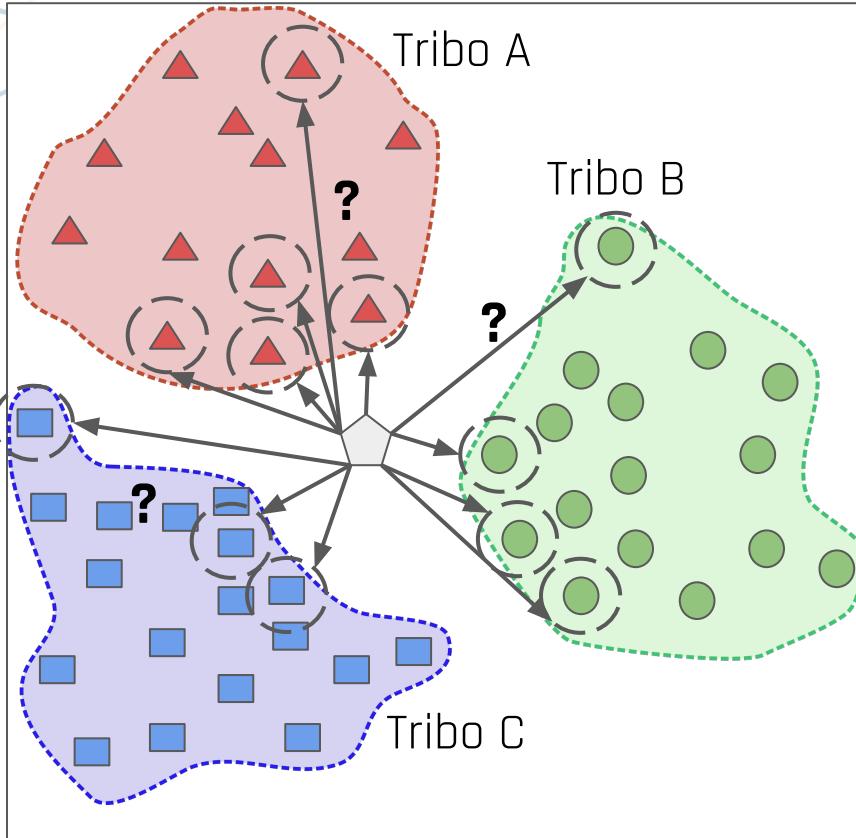
***k*-Nearest Neighbors - Complexidade**



***k*-Nearest Neighbors - Complexidade**



***k*-Nearest Neighbors - Complexidade**



***k*-Nearest Neighbors - Complexidade**

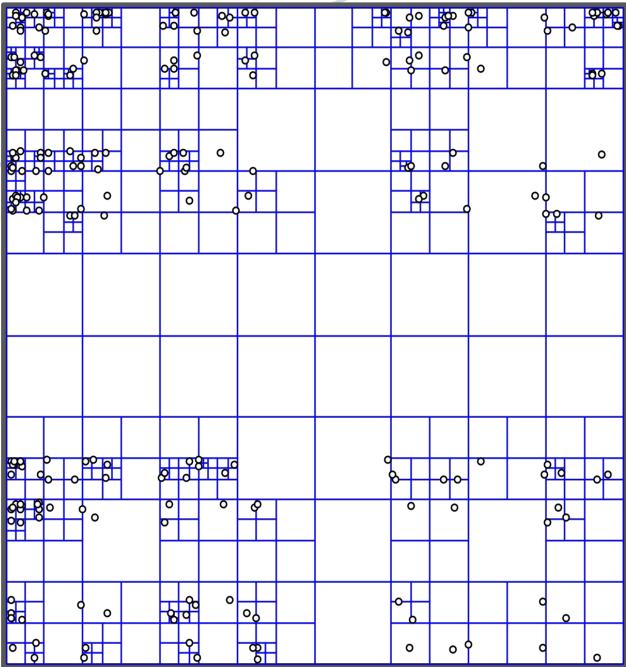
MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:

<https://www.youtube.com/c/PGCAPINPE>

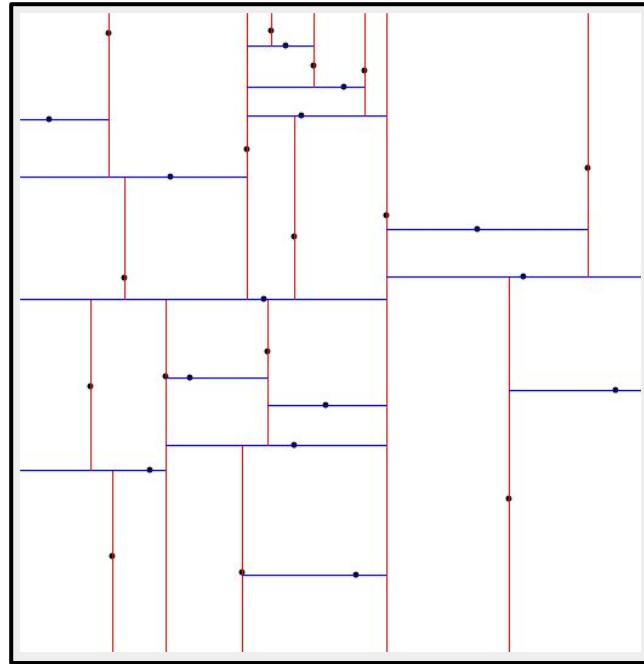


Quad-Tree



Fonte: <https://en.wikipedia.org/wiki/Quadtree>

KD-Tree

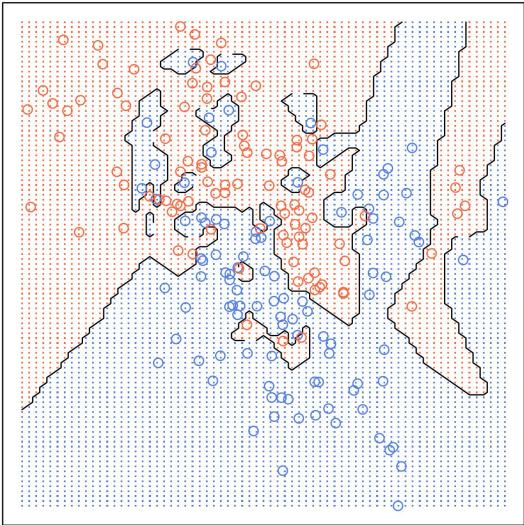


Fonte: <https://github.com/mgruben/Kd-Trees>

k*-Nearest Neighbors - Valor de *K



$K = 1$

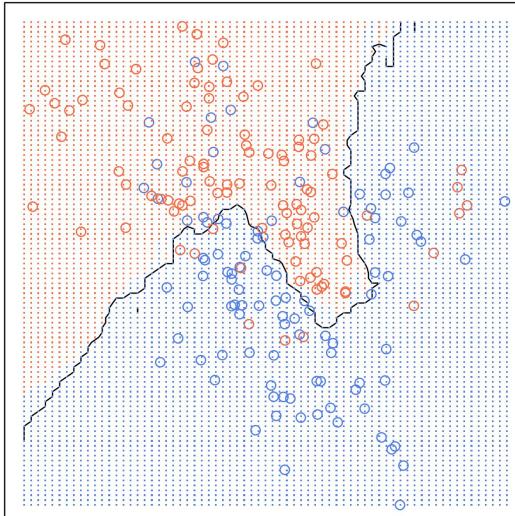


Fonte: <https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/>

Ajuste a características específicas dos dados de treinamento



$K = 20$



Fonte: <https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/>

Ajustes mais gerais, podendo não considerar características relevantes dos dados

k*-Nearest Neighbors - Valor de *K

MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:
<https://www.youtube.com/c/PGCAPINPE>



Lantz (2013) define que:

- A escolha do valor de k varia de acordo com a complexidade do problema e quantidade de amostras;
- Abordagens práticas para a escolha são:
 - Definir k como sendo a raiz quadrada da quantidade de amostras de treino
 - Testes com vários valores de K

Ma *et al* (2014) indicam a existência de subjetividade na definição do valor de k

k-Nearest Neighbors - Aplicações

MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:

<https://www.youtube.com/c/PGCAPINPE>



APLICAÇÃO DE K-NEAREST NEIGHBOR EM IMAGENS MULTISPECTRAIS PARA A ESTIMATIVA DE PARÂMETROS FLORESTAIS

Marcos Vinicius Giongo Alves¹, Ugo Chiavetta¹, Henrique Soares Koehler²,
Sebastião do Amaral Machado³, Flávio Felipe Kirchner³

¹Eng. Florestal, Dr., Università Del Molise, Pesche (IS), Itália - marcos@world-forestry.com; ugo.chiavetta@unimol.it

²Eng. Florestal, Dr., Depto. de Fitotecnia e Fitossanidade, UFPR, Curitiba, PR, Brasil - kochler@ufpr.br

³Eng. Florestal, Dr., Depto. de Ciências Florestais, UFPR, Curitiba, PR, Brasil - samachado@ufpr.br; kirchner@ufpr.br

Recebido para publicação: 23/07/2010 – Aceito para publicação: 03/06/2013

Resumo

A gestão dos recursos naturais requer a estimativa de uma série de parâmetros para o apoio da identificação de alternativas mais adequadas para a gestão e manejo das áreas florestais. Em particular, os ecossistemas florestais exigem um complexo e crescente conjunto de informações, e os inventários florestais fornecem informações preciosas, entretanto, espacialmente, de forma não contínua. Muitos trabalhos científicos vêm direcionando esforços para o desenvolvimento de metodologias que relacionam os dados da terra com informações de imagens multispectrais. A modelagem dessas relações pode estender as estimativas dos dados de inventário florestal em áreas não amostradas. Neste trabalho, foi avaliado o desempenho de uma análise não paramétrica, com a utilização do algoritmo K-Nearest Neighbor em imagens SPOT. Foram avaliados os resultados obtidos na espacialização de atributos florestais em uma área em Molise, na Itália. Entre as diversas metodologias para os cálculos das distâncias espaciais, o uso do método baseado nas distâncias multiregressivas não paramétricas apresentaram os melhores resultados. A densidade e número de espécies levantados em campo apresentaram um coeficiente de correlação de Pearson $\rho = 0.58$, comparativamente às informações obtidas nas imagens multispectrais, ligeiramente inferior aos obtidos para a área basal e volume, que obtiveram, respectivamente, $\rho = 0.62$ e 0.71 .

Palavras-chave: Inventário florestal; sensoriamento remoto; área basal; volume.

Abstract

Application of k-nearest neighbor on multispectral images to estimate forest parameters. Natural resources management requires several parameters estimates in order to support the identification of the best alternatives to forest area management. In particular, forest ecosystems require a complex and increasing set of descriptive information, where forest inventories put up important information, however not in a continuous spatial way. Lately, several scientific researches have been focusing on establishing methodologies to relate data from field to those obtained from multispectral images. Modeling these relations can extend the estimates of forest inventory data to not sampled areas. This research evaluated performance of non-parametric analysis using the K-Nearest Neighbor (k-NN) on SPOT 5 images. It evaluated the results obtained from the spatialization of some forest attributes in a forest area located at Molise, Italy. Among several methodologies for spatial distance calculations, the use of multiregressive non-parametric distances revealed the best results. Density and number of species on the ground revealed a Pearson correlation coefficient of $\rho = 0.58$ as compared to data obtained from multispectral images, lightly lower than the obtained for basal area and volume, which were $\rho = 0.62$ and 0.71 , respectively.

Keywords: Forest inventory; remote sensing; basal area; volume.

Anais do XIX Simpósio Brasileiro de Sensoriamento Remoto

ISBN: 978-85-17-00097-3

14 a 17 de Abril de 2019

INPE - Santos-SP, Brasil

CLASSIFICAÇÃO DE ÁREAS QUEIMADAS POR MACHINE LEARNING USANDO DADOS DE SENSORIAMENTO REMOTO

Cícero Alves dos Santos Júnior¹, Olga Oliveira Bittencourt¹, Fabiano Morelli¹ e Rafael Santos¹

¹ INPE – National Institute for Space Research

Av. dos Astronautas, 1758 - 12227-010 - São José dos Campos - SP, Brazil
{cicero.alves; olga.bittencourt; fabiano.morelli; rafael.santos}@inpe.br

RESUMO

Apresentamos um estudo para melhorar a automação do processo de classificação de áreas queimadas usando dados de sensoriamento remoto. Mostramos os atributos mais relevantes para enriquecer a base de conhecimento e o resultado da aplicação deles em uma comparação de modelos de classificação de *machine learning*. Validamos nosso estudo com dados de queimadas do Cerrado feitos por especialistas. Os melhores resultados foram obtidos com os modelos *Random Forest* e *Neural Networks* e indicam a viabilidade de utilização da abordagem no processo de classificação de áreas queimadas.

Palavras-chave – áreas queimadas, classificação, machine learning, dados de sensoriamento remoto.

vegetação tanto no bioma Cerrado quanto no restante do território brasileiro e parte da América Latina. O monitoramento é realizado por sensoriamento remoto de duas formas independentes, dependendo da resolução das imagens do satélite. Imagens de baixa resolução espacial (pixels maiores que 300m) são usadas para gerar produtos de dados diariamente, foco de incêndio e previsão de risco de fogo. Imagens de média resolução espacial (pixels em torno de 30m) são usados para análises mais precisas e menos frequentes como as estimativas periódicas de emissão de poluentes e, mais recentemente, estimativas de superfícies queimadas. Seus resultados são utilizados, por exemplo, como subsídios de políticas públicas como o Código Florestal Brasileiro e para contribuir para que as metas de redução das emissões de gases assumidas pelo Governo brasileiro na Convenção do Clima [5] possam ser atingidas.

Um desafio nesse monitoramento é combinar eficiência e

HORA DE PRATICAR!



Workshop em
Computação
Aplicada

8-11 e 14-17 de setembro
Evento online

MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:

<https://www.youtube.com/c/PGCAPINPE>



Exemplo de aplicação do K-Nearest Neighbors (kNN)

- **Instrutores:** Adriano, Felipe Carvalho e Felipe Menino
- **Realização:** Dia 15/09
- **Descrição:** Objetiva-se apresentar aos alunos exemplos de aplicação de algoritmos de classificação utilizando K-Nearest Neighbors (kNN).
- **Sumário:**
 - Livro [Introdução ao Machine Learning](#)
 - Exemplo de [Classificação](#) em Python
 - Exemplo de [Regressão](#) em Python
 - Exemplo de [Agrupamento](#) em R

<https://www.kaggle.com/phelpsmemo/intro-ml-python-knn-worcap2020>

MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:

<https://www.youtube.com/c/PGCAPINPE>



Workshop em
Computação
Aplicada

8-11 e 14-17 de setembro
Evento online

Exemplo de aplicação da Árvore de decisão

- **Instrutores:** Adriano, Felipe Carvalho e Felipe Menino
- **Realização:** Dia 15/09
- **Descrição:** Objetiva-se apresentar aos alunos exemplos de aplicação de algoritmos de classificação utilizando Árvores de decisão.
- **Sumário:**
 - Livro [Introdução ao Machine Learning](#)
 - Exemplo de [Classificação](#) em Python
 - Exemplo de [Regressão](#) em Python
 - Exemplo de [Agrupamento](#) em R

<https://www.kaggle.com/phelpsmemo/intro-ml-python-decisiontree-worcap2020>



Workshop em
Computação
Aplicada

8-11 e 14-17 de setembro

Evento online

AGRUPAMENTO

<https://dataat.github.io/introducao-ao-machine-learning/agrupamento.html>

04



Aprendizado de Máquina



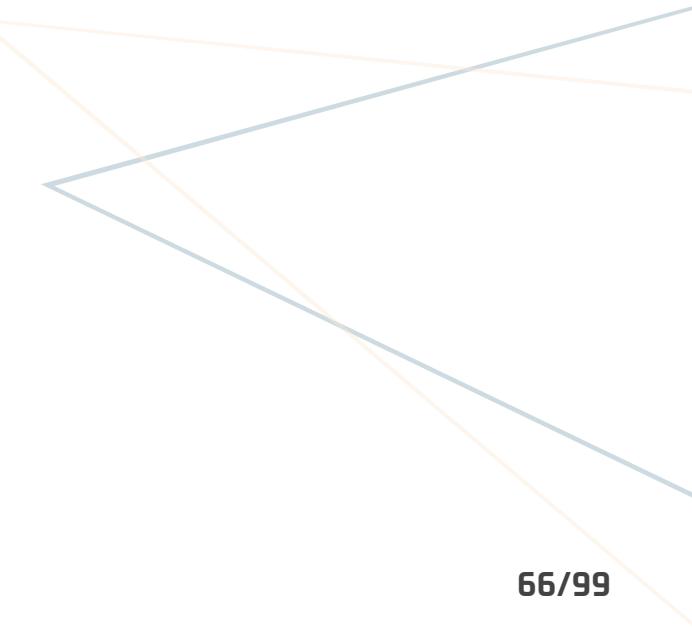


O que é um agrupamento?

Vamos aprender com ... cocos



Fonte: Stanford cs221 - by Chris Piech



O que é um agrupamento?

MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:

<https://www.youtube.com/c/PGCAPINPE>



- Cada coco é um observação;
- Cada observação tem os atributos:
 - Tamanho
 - Composição de fibras
 - Corte em cima (sem cabeça)



Fonte: Stanford cs221 - by Chris Piech

O que é um agrupamento?



id	tamanho	fibras	cortado
1	20	False	True
2	13	True	False
3	25	False	False
4	15	True	False
⋮	⋮	⋮	⋮



Fonte: Stanford cs221 - by Chris Piech



O que é um agrupamento?

id	tamanho	fibras	cortado
1	30	False	True
2	23	True	False
3	45	False	False
4	34	True	False
⋮	⋮	⋮	⋮



Fonte: Stanford cs221 - by Chris Piech

O que é um agrupamento?

id	tamanho	fibras	cortado
1	30	False	True
2	23	True	False
3	45	False	False
4	34	True	False
⋮	⋮	⋮	⋮



Fonte: Stanford cs221 - by Chris Piech

O que é um agrupamento?

MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:
<https://www.youtube.com/c/PGCAPINPE>



id	tamanho	fibras	cortado
1	30	False	True
2	23	True	False
3	45	False	False
4	34	True	False
⋮	⋮	⋮	⋮



Fonte: Stanford cs221 - by Chris Piech

Como podemos
separar os
cocos?

Como podemos separar os cocos?

MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:

<https://www.youtube.com/c/PGCAPINPE>



- Criando grupos;



Fonte: Stanford cs221 - by Chris Piech

Como podemos separar os cocos?

MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:
<https://www.youtube.com/c/PGCAPINPE>



- Criando grupos;



Semelhantes
entre si

Fonte: Stanford cs221 - by Chris Piech

Como podemos separar os cocos?

MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:
<https://www.youtube.com/c/PGCAPINPE>

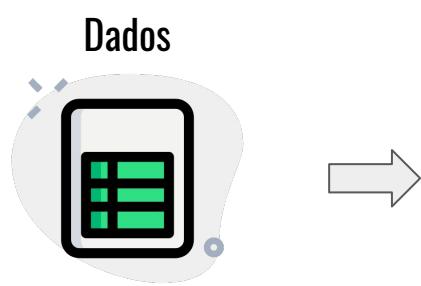
- Criando grupos;



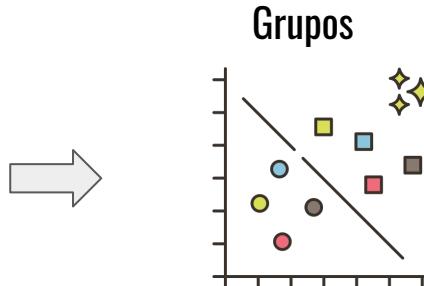
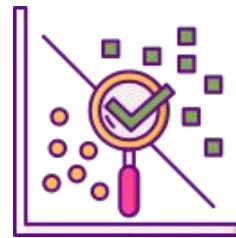
Diferentes entre
outros grupos

Fonte: Stanford cs221 - by Chris Piech

"Tarefa de criar grupos em conjuntos de dados, em que objetos do mesmo grupo sejam semelhantes entre si e diferentes dos objetos de outros grupos."



Técnica de agrupamento



Por que criar grupos em dados?

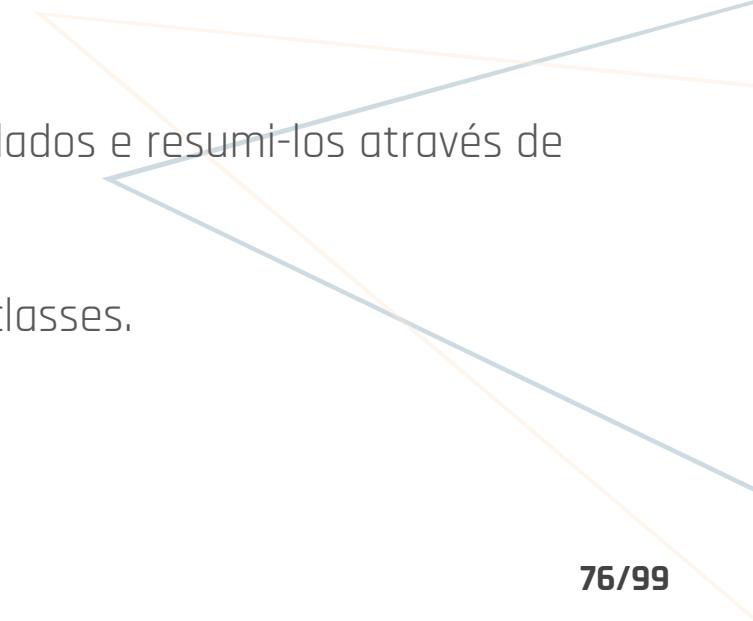
MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:

<https://www.youtube.com/c/PGCAPINPE>



- **Descoberta de conhecimento** em estruturas intrínsecas:
 - informações sobre dados, gerar hipóteses, detectar anomalias.
- **Classificação natural**: por exemplo, na Biologia, para identificar o grau de semelhança entre formas ou organismos (relação filogenética).
- **Compressão**: como um método para organizar os dados e resumi-los através de protótipos de agregados.
- **Pré-processamento**: Usar os grupos criados como classes.



Algoritmos de agrupamento

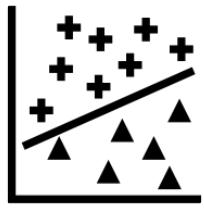
MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:

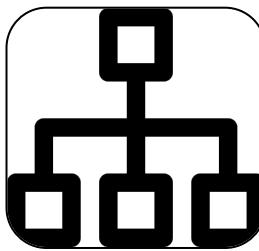
<https://www.youtube.com/c/PGCAPINPE>



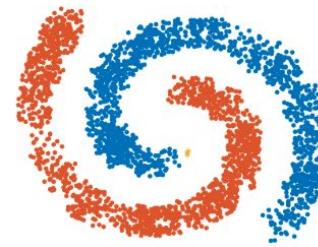
Kmeans



Hierarquico



DBSCAN



Algoritmos de agrupamento

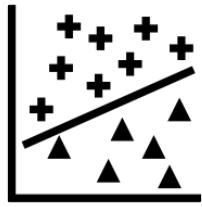
MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:

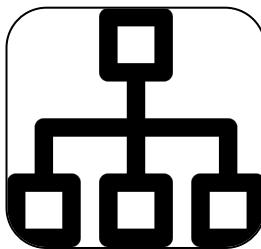
<https://www.youtube.com/c/PGCAPINPE>



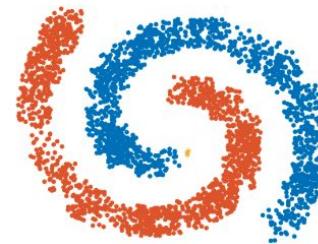
Kmeans



Hierarquico



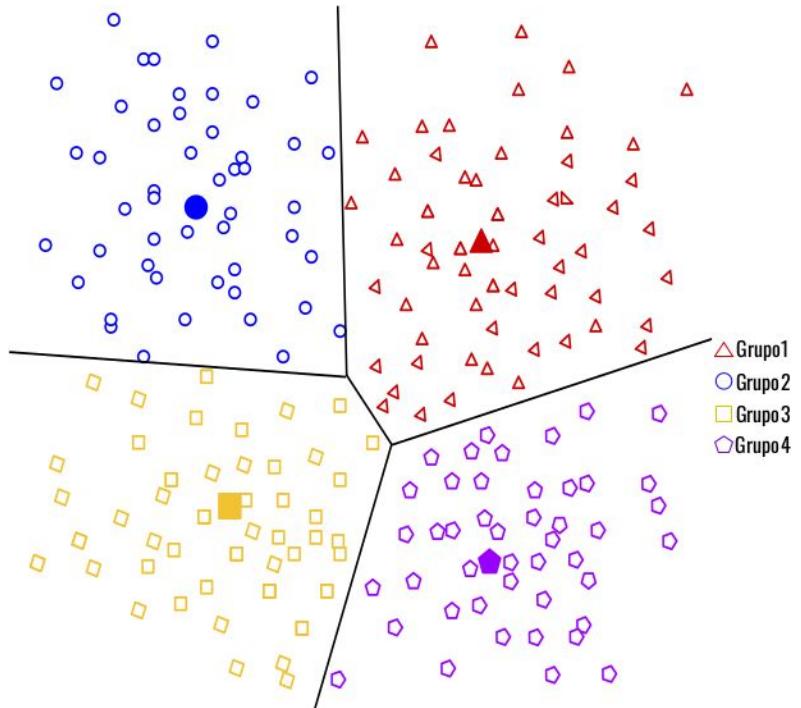
DBSCAN



Introdução Kmeans

- Baseado em Partição
- Usa centróide para representar grupos;
- Medidas de distâncias para determinar o grau de similaridade;
- Número de **K** à priori;

MC2 - INTRODUÇÃO AO MACHINE LEARNING
Transmissão pelo YouTube:
<https://www.youtube.com/c/PGCAPINPE>



Fonte: Adaptado de Developers (2020)

Funcionamento do Kmeans

MC2 - INTRODUÇÃO AO MACHINE LEARNING

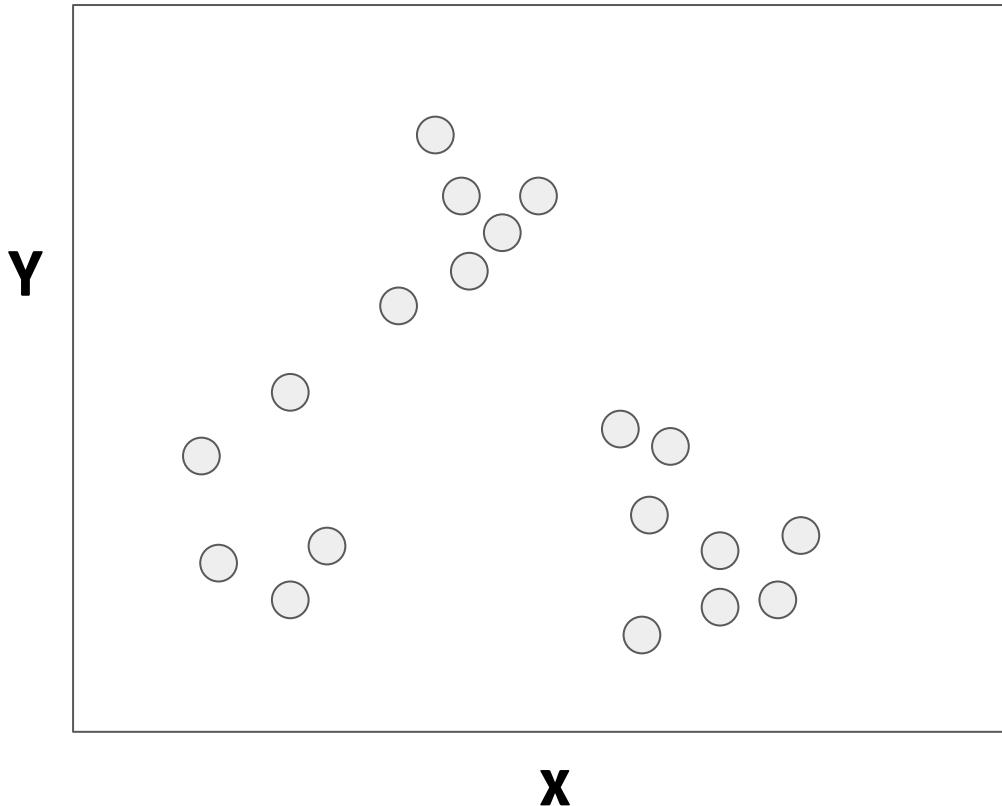
Transmissão pelo YouTube:

<https://www.youtube.com/c/PGCAPINPE>



- 1 etapa - Definição da quantidade de grupos;
- 2 etapa - Sorteio dos centróides;
- 3 etapa - Atribuição dos objetos a cada grupo;
- 4 etapa - Atualização dos centróides de cada grupo;
- 5 etapa - Caso os centróides sejam atualizados, volte ao passo (3), caso não, o algoritmo pára.

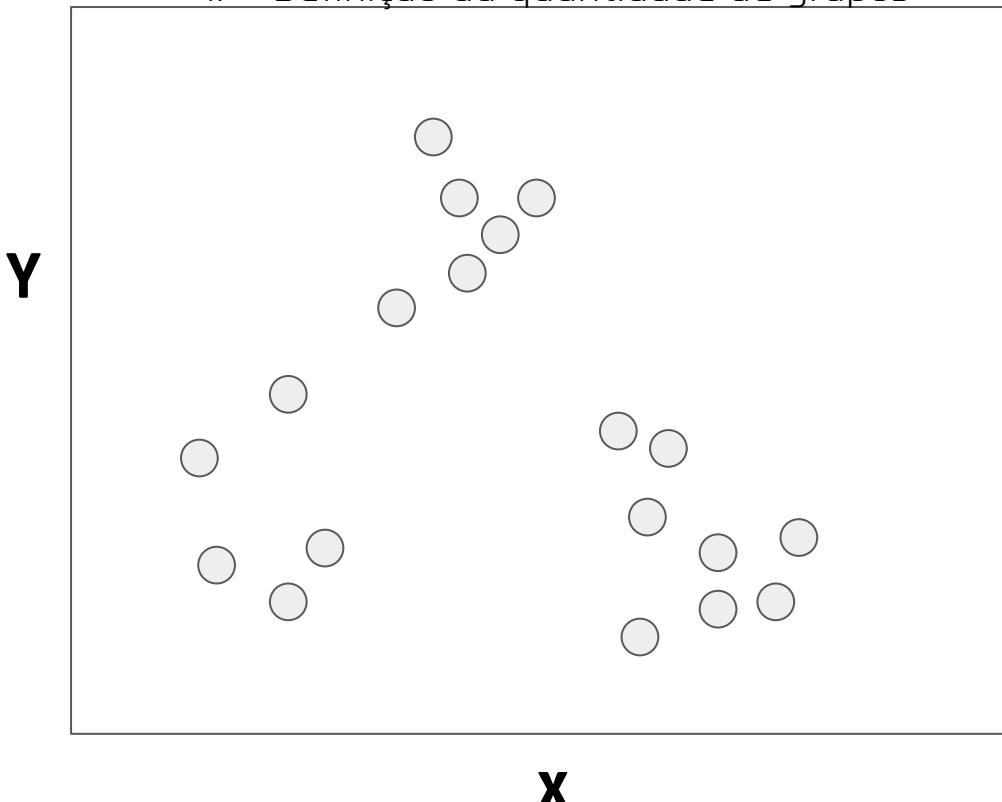
Funcionamento do Kmeans





Funcionamento do Kmeans

1. Definição da quantidade de grupos



Funcionamento do Kmeans

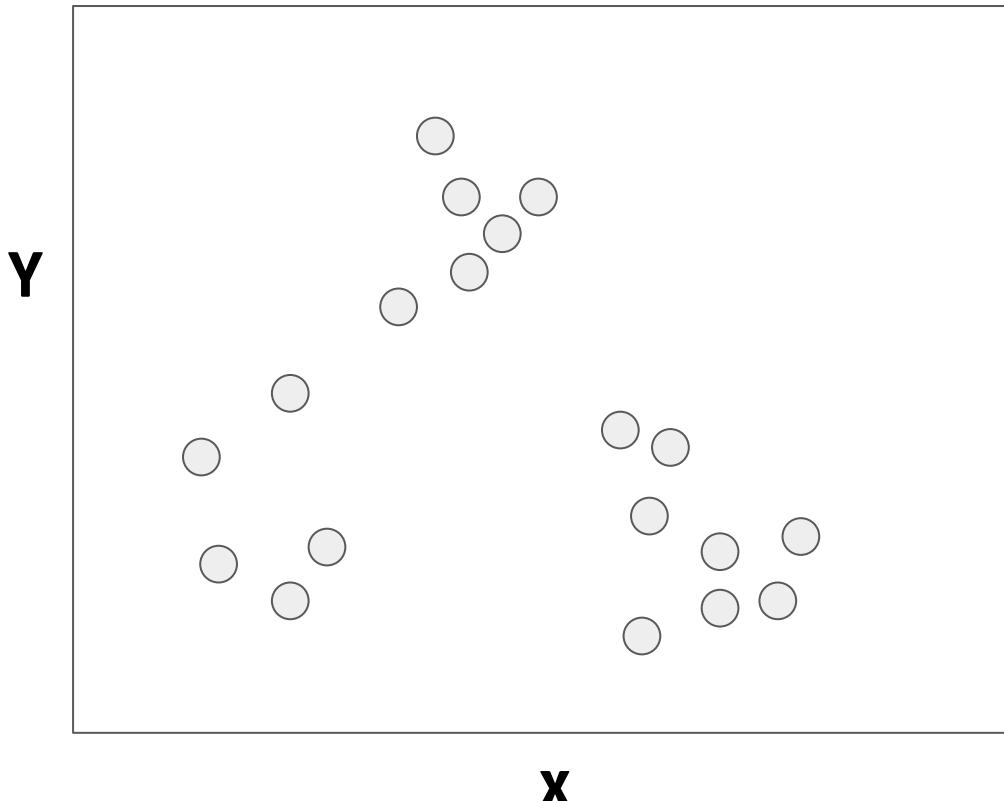
MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:

<https://www.youtube.com/c/PGCAPINPE>



K = 3



Funcionamento do Kmeans

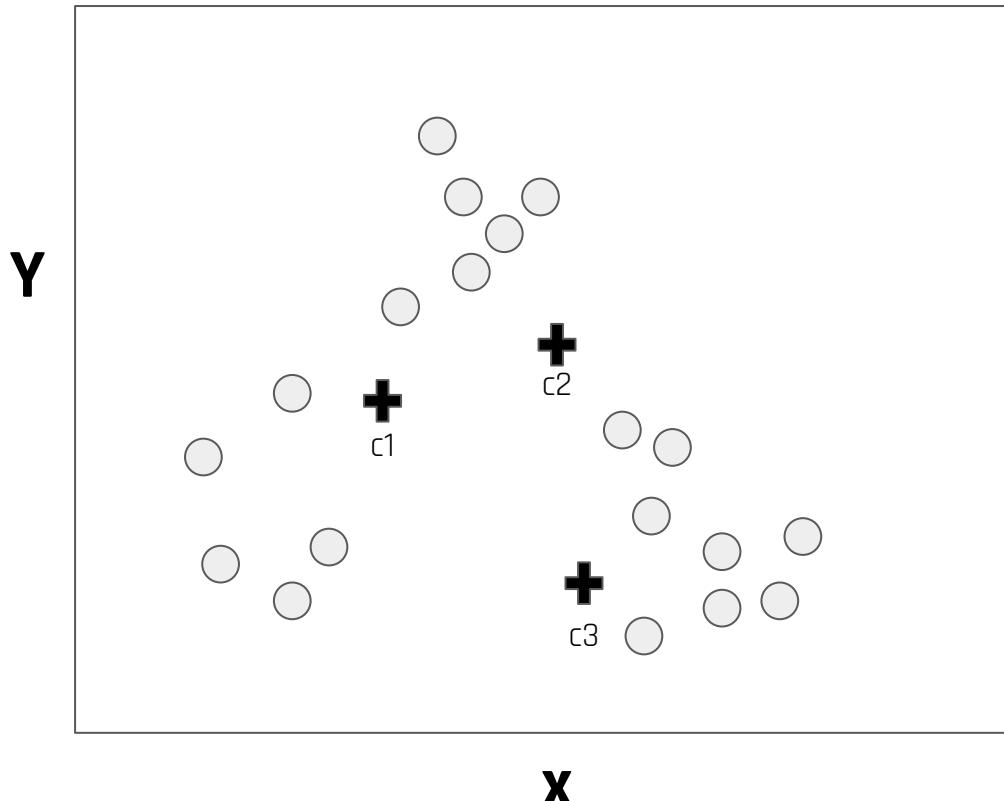
MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:

<https://www.youtube.com/c/PGCAPINPE>



2. Sorteio dos centróides



Funcionamento do Kmeans

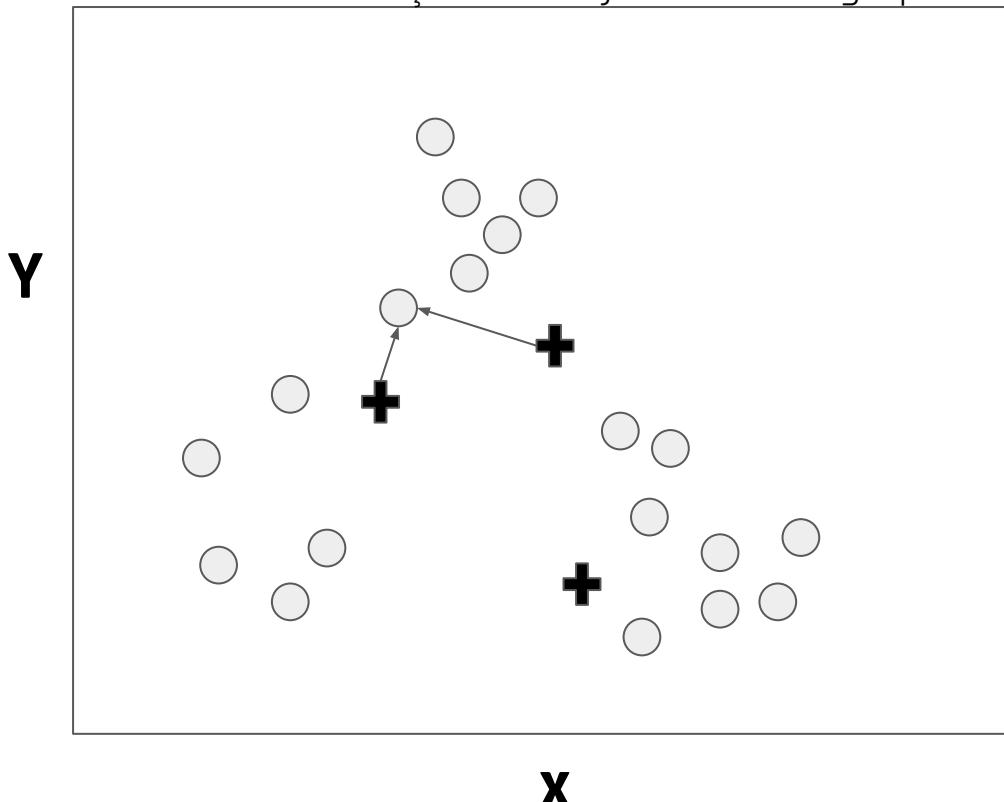
MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:

<https://www.youtube.com/c/PGCAPINPE>



3. Atribuição dos objetos a cada grupo



Funcionamento do Kmeans

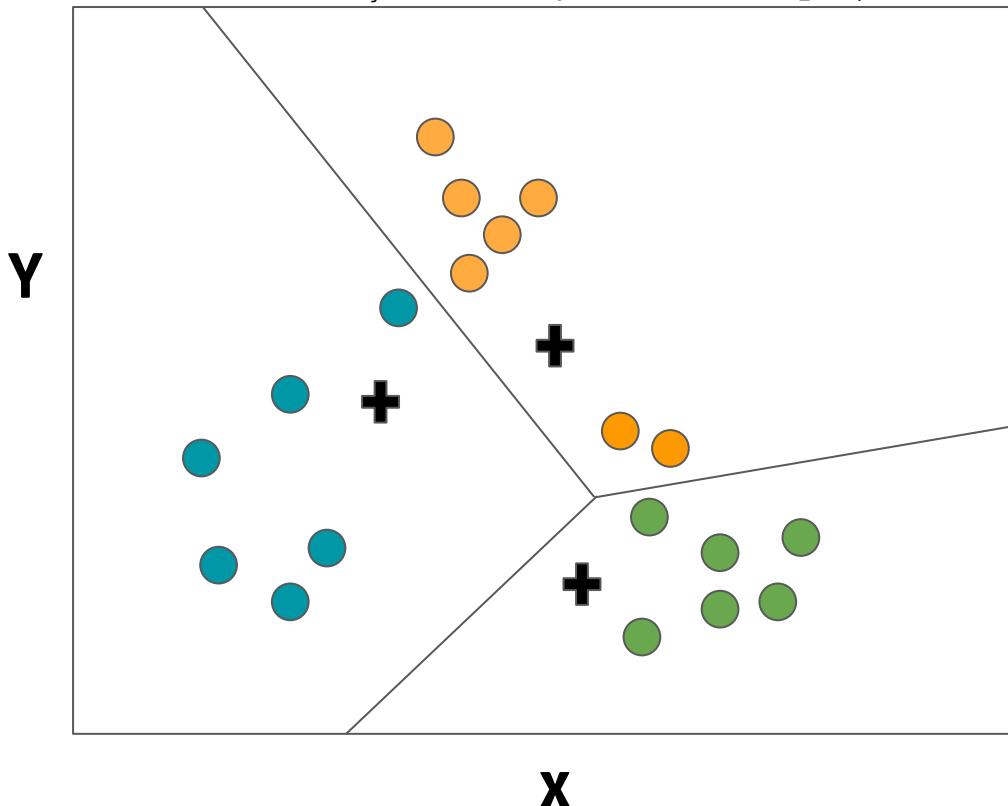
MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:

<https://www.youtube.com/c/PGCAPINPE>



3. Atribuição dos objetos a cada grupo



Funcionamento do Kmeans

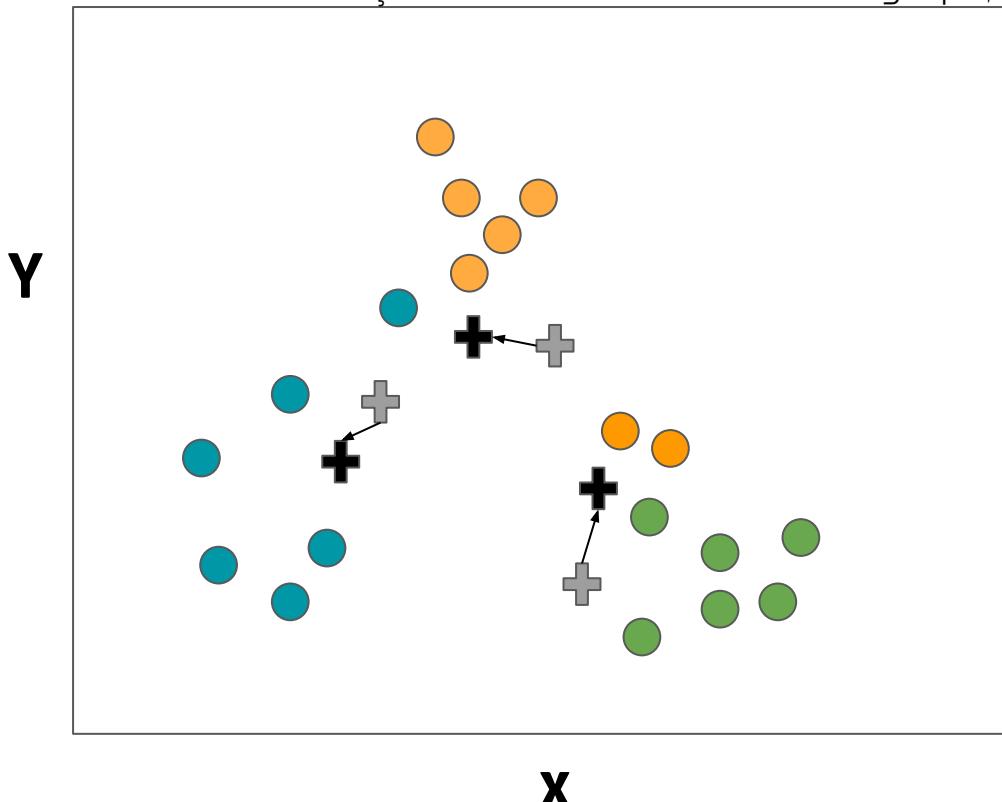
MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:

<https://www.youtube.com/c/PGCAPINPE>



4. Atualização dos centróides de cada grupo;



Funcionamento do Kmeans

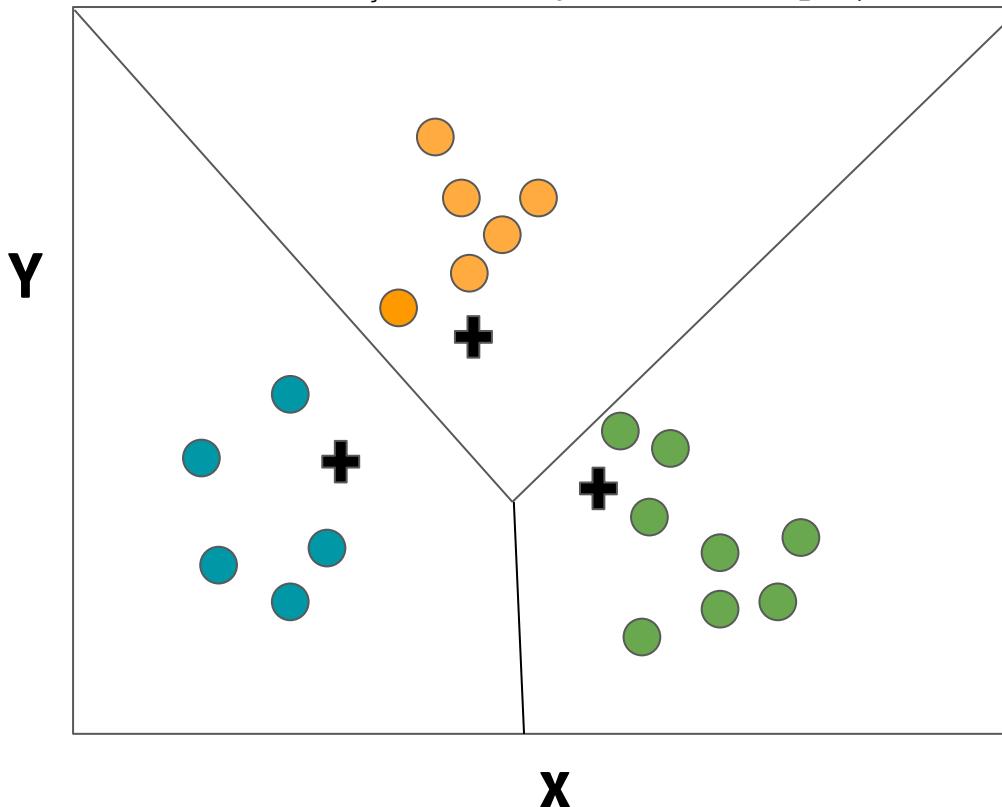
MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:

<https://www.youtube.com/c/PGCAPINPE>



3. Atribuição dos objetos a cada grupo



15 de setembro de 2020

88/99

Funcionamento do Kmeans

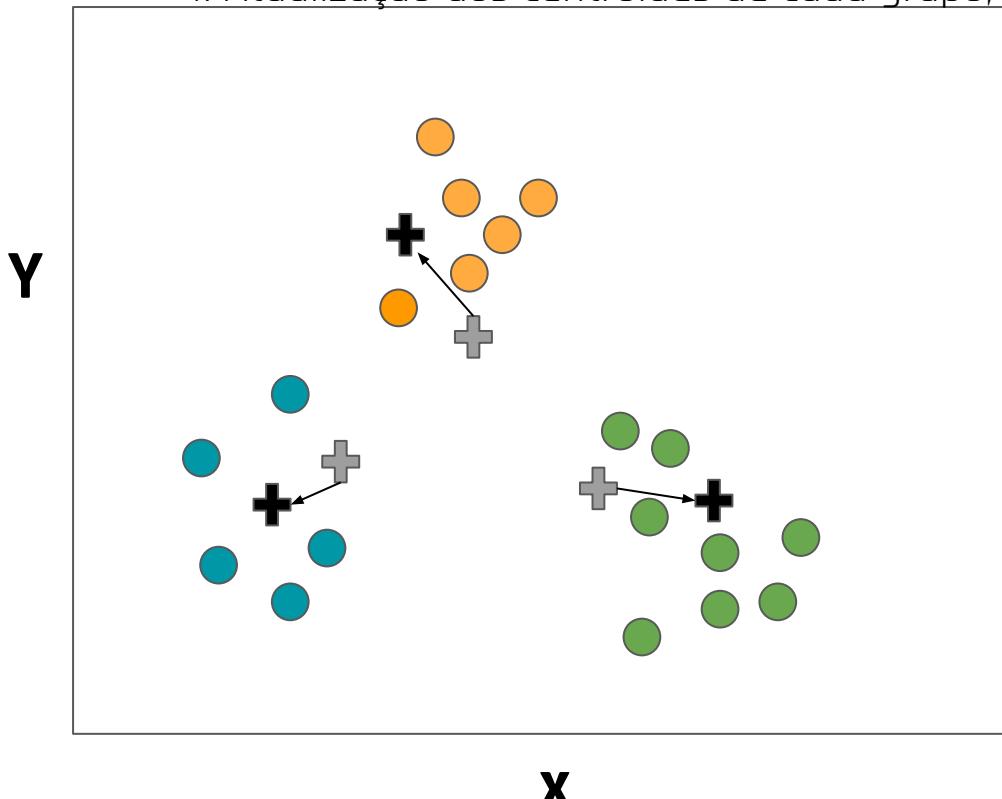
MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:

<https://www.youtube.com/c/PGCAPINPE>



4. Atualização dos centróides de cada grupo;



Funcionamento do Kmeans

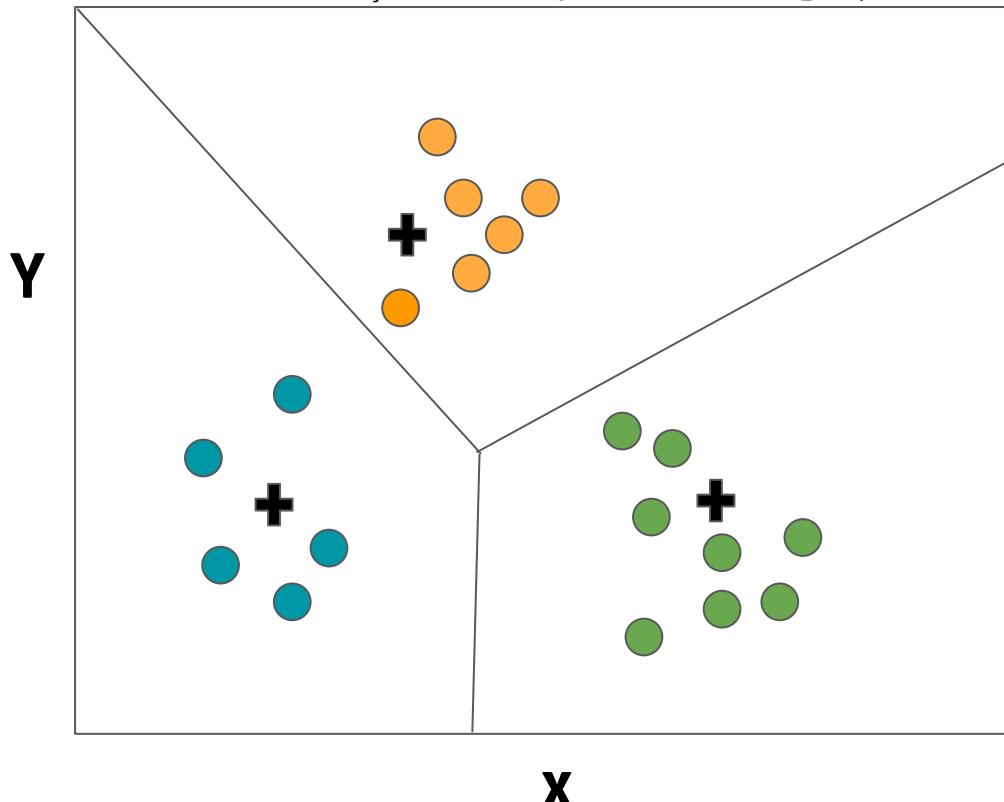
MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:

<https://www.youtube.com/c/PGCAPINPE>



3. Atribuição dos objetos a cada grupo



Funcionamento do Kmeans

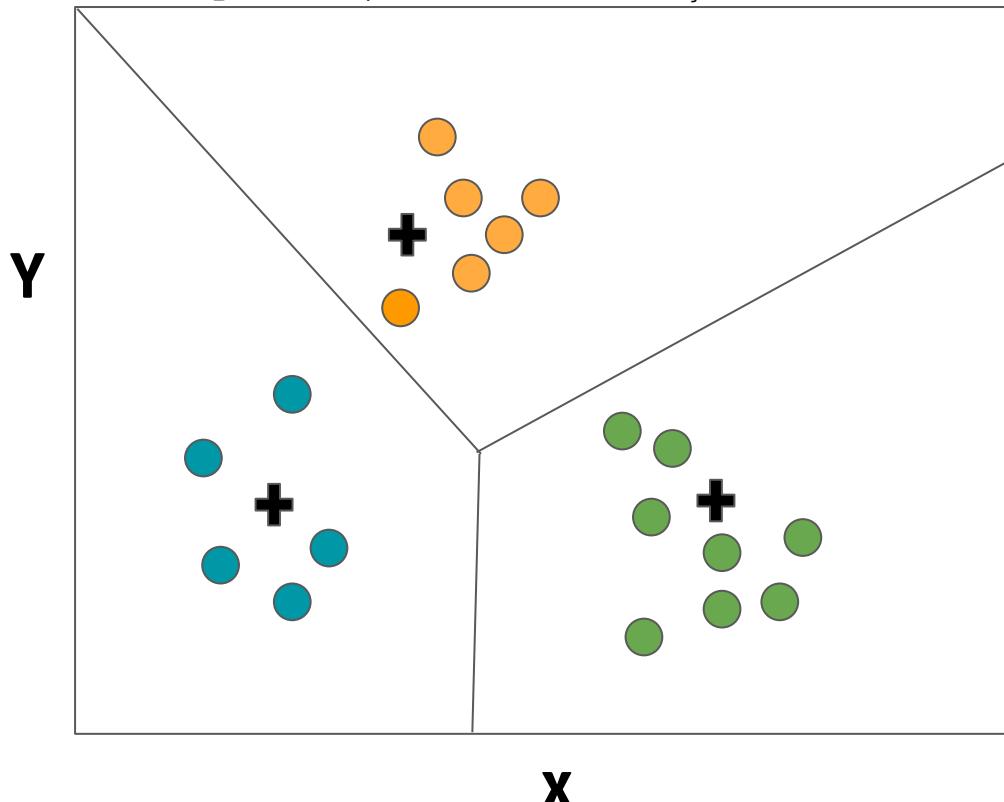
MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:

<https://www.youtube.com/c/PGCAPINPE>



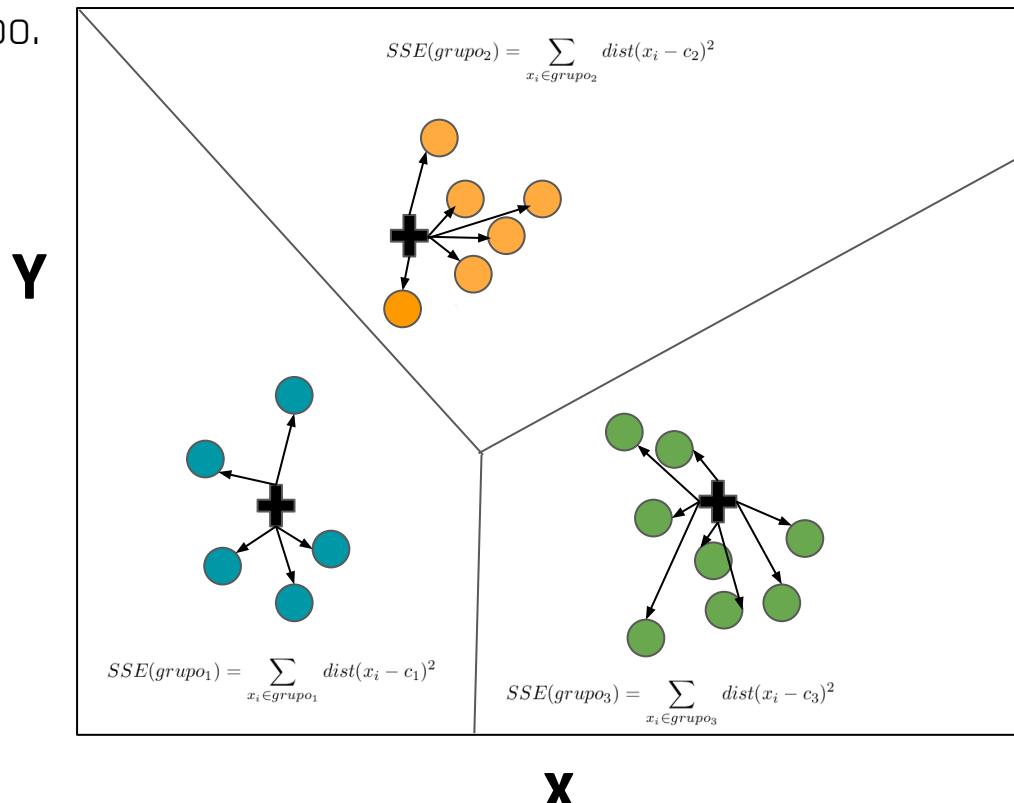
5. Algoritmo para, sem mudança no centróide.



Como avaliar o Kmeans?

- Somatório do erro de cada grupo.

$$SSE_{total} = SSE(grupo_1) + SSE(grupo_2) + SSE(grupo_3)$$

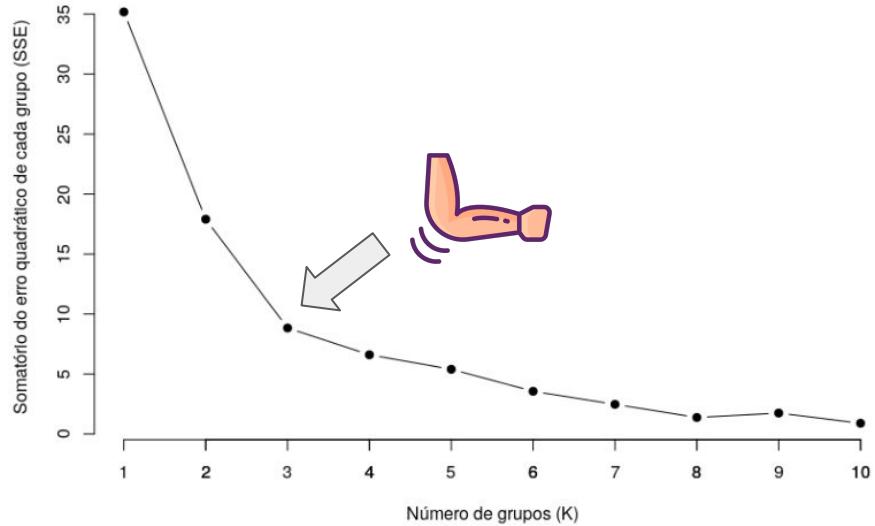


X

Como escolher o K?



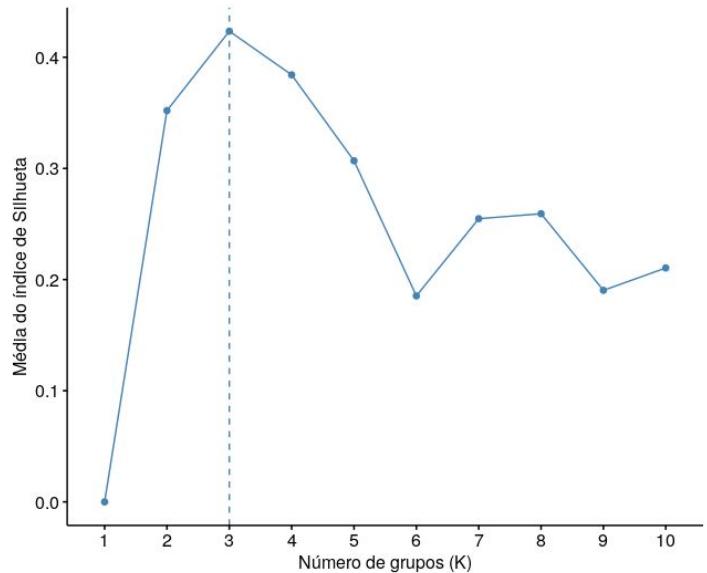
- Método do Cotovelo
- Método da Silhueta



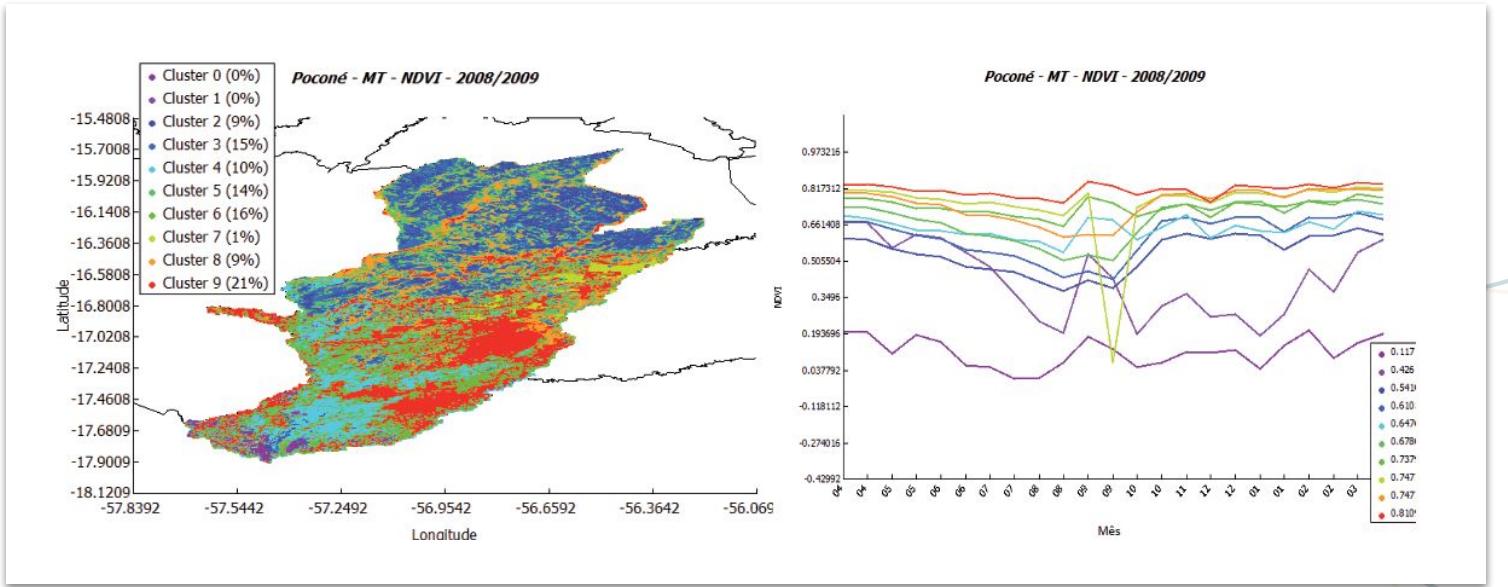
Como escolher o K?



- Método do Cotovelo
- Método da Silhueta
 - O índice de Silhueta varia de -1 a 1



Aplicações do Kmeans



Fonte: (SCRIVANI et al., 2014)

HORA DE PRATICAR!

MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:

<https://www.youtube.com/c/PGCAPINPE>



WORCAP
2020
Workshop em
Computação
APLICADA
8-11 e 14-17 de setembro
Evento online

MC2 - INTRODUÇÃO AO MACHINE LEARNING
Transmissão pelo YouTube
<https://www.youtube.com/c/PGCAPINPE>

COMPUTAÇÃO
APLICADA

Exemplo de aplicação do Método Kmeans

- Instrutores: Adriano, Felipe Carvalho e Felipe Menino
- Realização: Dia 15/09
- Descrição: Objetiva-se apresentar aos alunos exemplos de aplicação de algoritmos de agrupamento.
- Sumário:
 - Descrição do conjunto de dados
 - Agrupamento dos dados genéticos
 - Agrupamentos dos dados reais
 - Referências
- Links Úteis:
 - Livro [Introdução ao Machine Learning](#)
 - Exemplo de [Classificação](#) em Python
 - Exemplo de [Regressão com Máquina de Vetores de Suporte](#) em Python
 - Exemplo de [Agrupamento Hierárquico](#) em R

Descrição do Notebook

Neste notebook vamos apresentar um exemplo de aplicação do **Kmeans** usando a linguagem R. Para isso, vamos usar dois conjuntos de dados: dados genéticos com base em uma distribuição normal e dados sobre os **Países do mundo**.



O conjunto de dados de **Países do mundo** contém informações sobre o número de habitantes, a área do país, renda per capita, entre outros. No entanto, para fins didáticos, estamos usando o dado tratado e limpo, deste repositório: [Machine Learning aplicado a dados espaciais](#). A pergunta que queremos responder neste agrupamento é: Será que conseguimos obter grupos de países que possuem características semelhantes? Tudo indica que sim, né? Mas, vamos ver que essa tarefa não é nada trivial!

<https://www.kaggle.com/oldlipe/intro-ml-r-kmeans-worcap2020>

WORCAP
2020
Workshop em
Computação
APLICADA
8-11 e 14-17 de setembro
Evento online

MC2 - INTRODUÇÃO AO MACHINE LEARNING
Transmissão pelo YouTube
<https://www.youtube.com/c/PGCAPINPE>

COMPUTAÇÃO
APLICADA

Exemplo de aplicação do Método Hierarquico

- Instrutores: Adriano, Felipe Carvalho e Felipe Menino
- Realização: Dia 15/09
- Descrição: Objetiva-se apresentar aos alunos exemplos de aplicação de algoritmos de agrupamento.
- Sumário:
 - Descrição do conjunto de dados
 - Leitura dos dados
 - Aplicação do **Método Aglomerativo**
 - Visualização dos Dendrograma
 - Referências
- Links Úteis:
 - Livro [Introdução ao Machine Learning](#)
 - Exemplo de [Classificação](#) em Python
 - Exemplo de [Regressão com Máquina de Vetores de Suporte](#) em Python
 - Exemplo de [Agrupamento Kmeans](#) em R

Agrupando dados de países

Neste notebook vamos apresentar um exemplo de aplicação do Kmeans usando a linguagem R. Para isso, vamos usar os dados sobre os **Países do mundo**.



O conjunto de dados de **Países do mundo** contém informações sobre o número de habitantes, a área do país, renda per capita, entre outros. No entanto, para fins didáticos, estamos usando o dado tratado e limpo, deste repositório: [Machine Learning aplicado a dados espaciais](#). A pergunta que queremos responder neste agrupamento é: Será que conseguimos obter grupos de países que possuem características semelhantes? Tudo indica que sim, né? Mas, vamos ver que essa tarefa não é nada trivial!

<https://www.kaggle.com/oldlipe/intro-ml-r-hiererquico-worcap2020>



COMPUTAÇÃO
APLICADA

WORCAP
2020
Workshop em
Computação
APLICADA
8-11 e 14-17 de setembro
Evento online

REFERÊNCIAS BIBLIOGRÁFICAS

MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:
<https://www.youtube.com/c/PGCAPINPE>



- Aggarwal, Charu C. n.d. Data Classification: Algorithms and Applications. CRC Press.
- Aghabozorgi, Saeed, Ali Seyed Shirkhorshidi, and Teh Ying Wah. 2015. "Time-Series Clustering-a Decade Review." *Information Systems* 53. Elsevier: 16-38.
- Alpaydin, Ethem. 2020. Introduction to Machine Learning. MIT press.
- Bhavsar, Parth, Ilya Safro, Nidhal Bouaynaya, Robi Polikar, and Dimah Dera. 2017. "Machine Learning in Transportation Data Analytics." In *Data Analytics for Intelligent Transportation Systems*, 283-307. Elsevier.
- Blanzieri, Enrico, and Anton Bryl. 2008. "A survey of learning-based techniques of email spam filtering." *Artificial Intelligence Review* 29 (1): 63-92. <https://doi.org/10.1007/s10462-009-9109-6>.
- Boser, Bernhard E, Isabelle M Guyon, and Vladimir N Vapnik. 1992. "A Training Algorithm for Optimal Margin Classifiers." In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 144-52.
- Developers, Google. 2020. "Clustering Algorithms." Google inc.
- Drucker, Harris, Christopher JC Burges, Linda Kaufman, Alex J Smola, and Vladimir Vapnik. 1997. "Support Vector Regression Machines." In *Advances in Neural Information Processing Systems*, 155-61.

REFERÊNCIAS BIBLIOGRÁFICAS

MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:
<https://www.youtube.com/c/PGCAPINPE>

- Esling, Philippe, and Carlos Agon. 2012. "Time-Series Data Mining." ACM Computing Surveys (CSUR) 45 (1). ACM New York, NY, USA: 1–34.
- Gao, Jing, and Aidong Zhang. 2012. "CSE 601: Data Mining and Bioinformatics." University at Buffalo.
- Gasch, Audrey P, and Michael B Eisen. 2002. "Exploring the Conditional Coregulation of Yeast Gene Expression Through Fuzzy K-Means Clustering." Genome Biology 3 (11). Springer: research0059-1.
- Goodfellow, Ian, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. Deep Learning. Vol. 1. MIT press Cambridge.
- Han, Jiawei, Jian Pei, and Micheline Kamber. 2011. Data Mining: Concepts and Techniques. Elsevier.
- He, Tao, Yu-Jun Sun, Ji-De Xu, Xue-Jun Wang, and Chang-Ru Hu. 2014. "Enhanced Land Use/Cover Classification Using Support Vector Machines and Fuzzy K-Means Clustering Algorithms." Journal of Applied Remote Sensing 8 (1). International Society for Optics; Photonics: 083636.
- Helliwell, John F, Haifang Huang, Shun Wang, and Max Norton. 2020. "Social Environments for World Happiness." World Happiness Report 2020.

REFERÊNCIAS BIBLIOGRÁFICAS

MC2 - INTRODUÇÃO AO MACHINE LEARNING

Transmissão pelo YouTube:

<https://www.youtube.com/c/PGCAPINPE>



- Jain, Anil K. 2010. "Data Clustering: 50 Years Beyond K-Means." *Pattern Recognition Letters* 31 (8). Elsevier: 651-66.
- Kirch, Wilhelm, ed. 2008. "Pearson's Correlation Coefficient." In *Encyclopedia of Public Health*, 1090-1. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-1-4020-5614-7_2569.
- Lantz, Brett. 2013. *Machine Learning with R : Learn How to Use R to Apply Powerful Machine Learning Methods and Gain an Insight into Real-World Applications*. Birmingham, UK: Packt Publishing.
- Mitchell, Tom M. 1997. *Machine Learning*. First. McGraw-Hill Science/Engineering/Math.
- Russell, Stuart, and Peter Norvig. 2002. *Artificial Intelligence: A Modern Approach*. Second. Prentice Hall.
- Samuel, Arthur L. 1959. "Some Studies in Machine Learning Using the Game of Checkers." *IBM Journal of Research and Development* 3 (3). IBM: 210-29.
- Soto, Timothy. 2013. "Regression Analysis." In *Encyclopedia of Autism Spectrum Disorders*, edited by Fred R. Volkmar, 2538-8. New York, NY: Springer New York.
- TURING, A. M. 1950. "COMPUTING MACHINERY AND INTELLIGENCE." *Mind* LIX (236): 433-60. <https://doi.org/10.1093/mind/LIX.236.433>.
- Vapnik, Vladimir N, and Alexey Y Chervonenkis. 1963. "On a Class of Pattern-Recognition Learning Algorithms." *Automation and Remote Control* 25 (6). PLENUM PUBL CORP CONSULTANTS BUREAU, 233 SPRING ST, NEW YORK, NY 10013: 838.
- Von Ahn, Luis, Manuel Blum, Nicholas J Hopper, and John Langford. 2003. "CAPTCHA: Using Hard Ai Problems for Security." In *International Conference on the Theory and Applications of Cryptographic Techniques*, 294-311. Springer.
- Wang, Jiaqi, Xindong Wu, and Chengqi Zhang. 2005. "Support Vector Machines Based on K-Means Clustering for Real-Time Business Intelligence Systems." *International Journal of Business Intelligence and Data Mining* 1 (1). Inderscience Publishers: 54-64.

MINICURSO

INTRODUÇÃO AO MACHINE LEARNING

15 de setembro de 2020

OBRIGADO PELA ATENÇÃO!

Adriano Almeida
Felipe Carvalho
Felipe Menino

Transmissão pelo YouTube:
<https://www.youtube.com/c/PGCAPINPE>



Palestras

Minicursos

Hackathon



COMPUTAÇÃO
APLICADA