MÓDULO:
Clustering - Modelos Não Supervisionados

Todos os exercícios e colabs do módulo podem ser acessados

CLICANDO AQUI

Obs: os mesmos exercícios e colabs acima seguem anexados em cada aula ao longo do módulo.

data expert.®

Data Science & Machine Learning

CLS 22 - Retomando Conceitos

Consultor: Tulio Souza

Túlio Souza

Data Cientist
@Avenue Code

Consultor de Projetos de Machine Learning no Mercado Nacional e Internacional

Co-fundador da comunidade Machine Learning Experience

Machine Learn Experience: Milhares de pessoas impactadas com projetos em mais de 50 eventos.

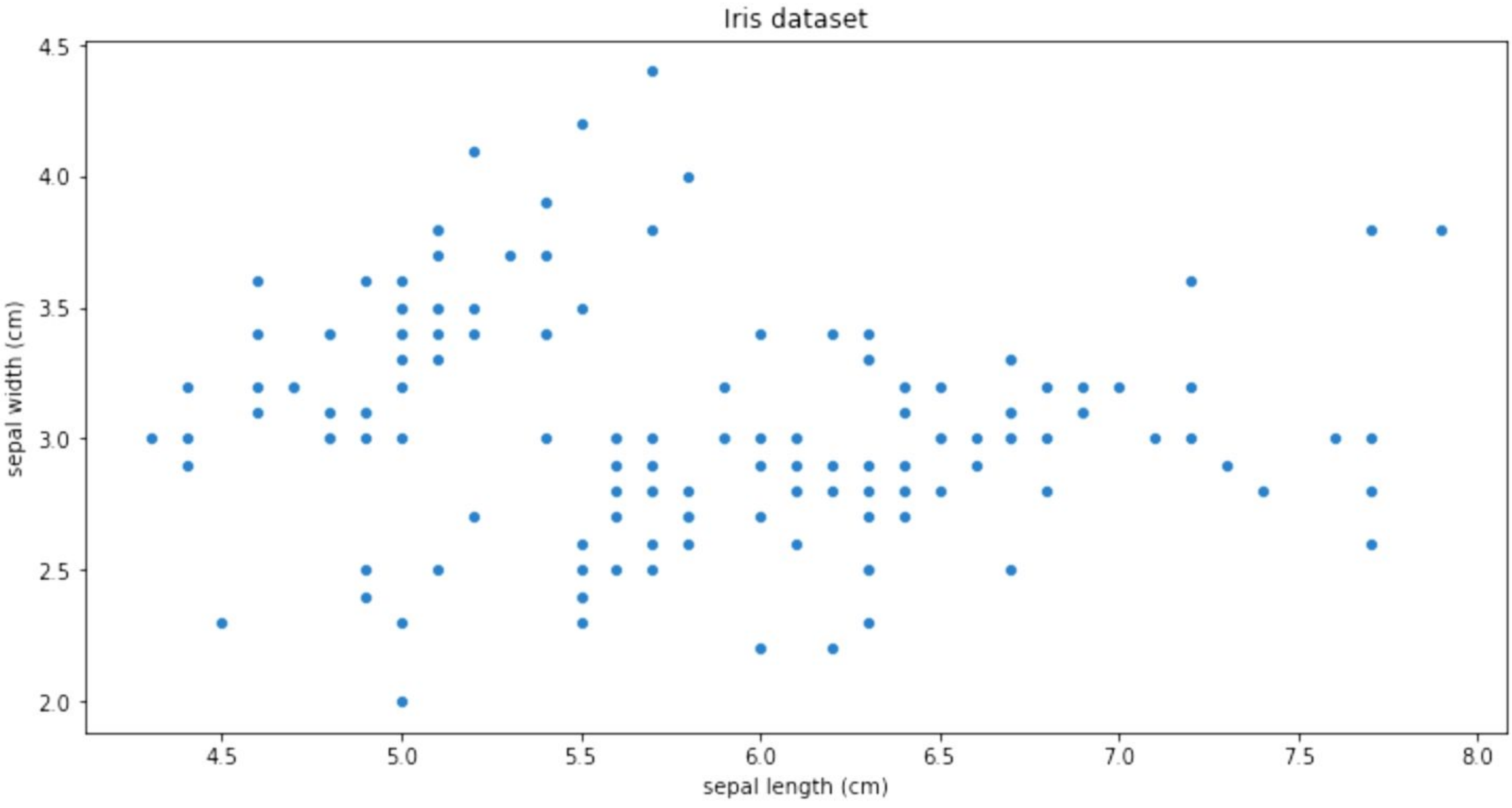
Análise de Cluster

Mineração de dados que visa fazer agrupamentos automáticos segundo o seu grau de semelhança.

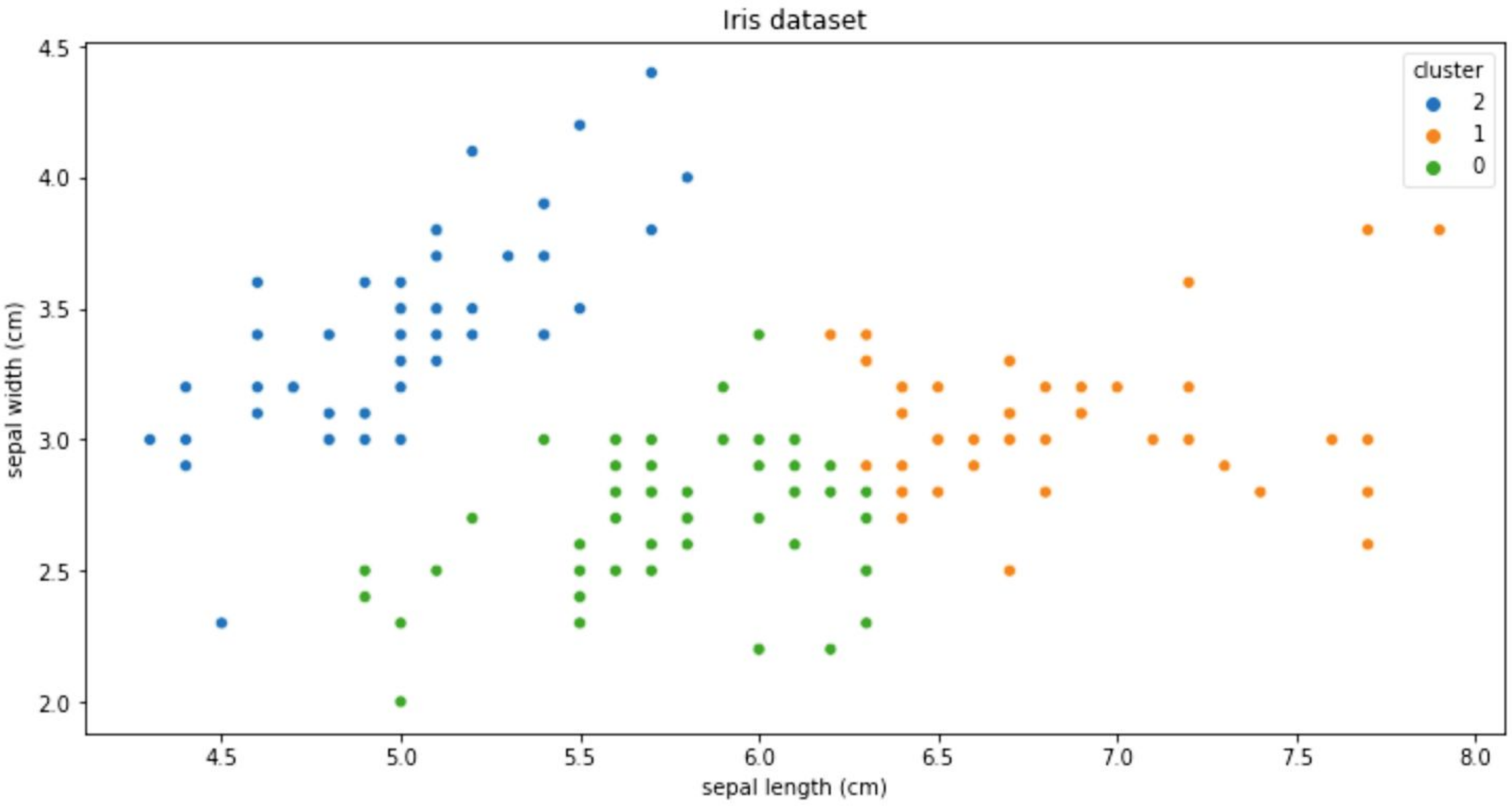
O critério de semelhança faz parte da definição do problema dependendo do algoritmo.

A cada conjunto dá-se o nome de grupo, aglomerado ou agrupamento (cluster).

Dados Brutos



Dados Clusterizados



Clustering Methods

01

Centroid Clustering

02

Distribution Clustering

03

Density Clustering

04

Hierarchical Clustering

data expert.®

Data Science & Machine Learning

CLS 23 - Mean Shift

Consultor: Tulio Souza

O que
é Mean Shift

Como funciona
o algoritmo?

Parâmetros

Escolhendo a
melhor
distância

Mean Shift

É um método de clusterização

Baseado em centróides.

Cada deslocamento é definido

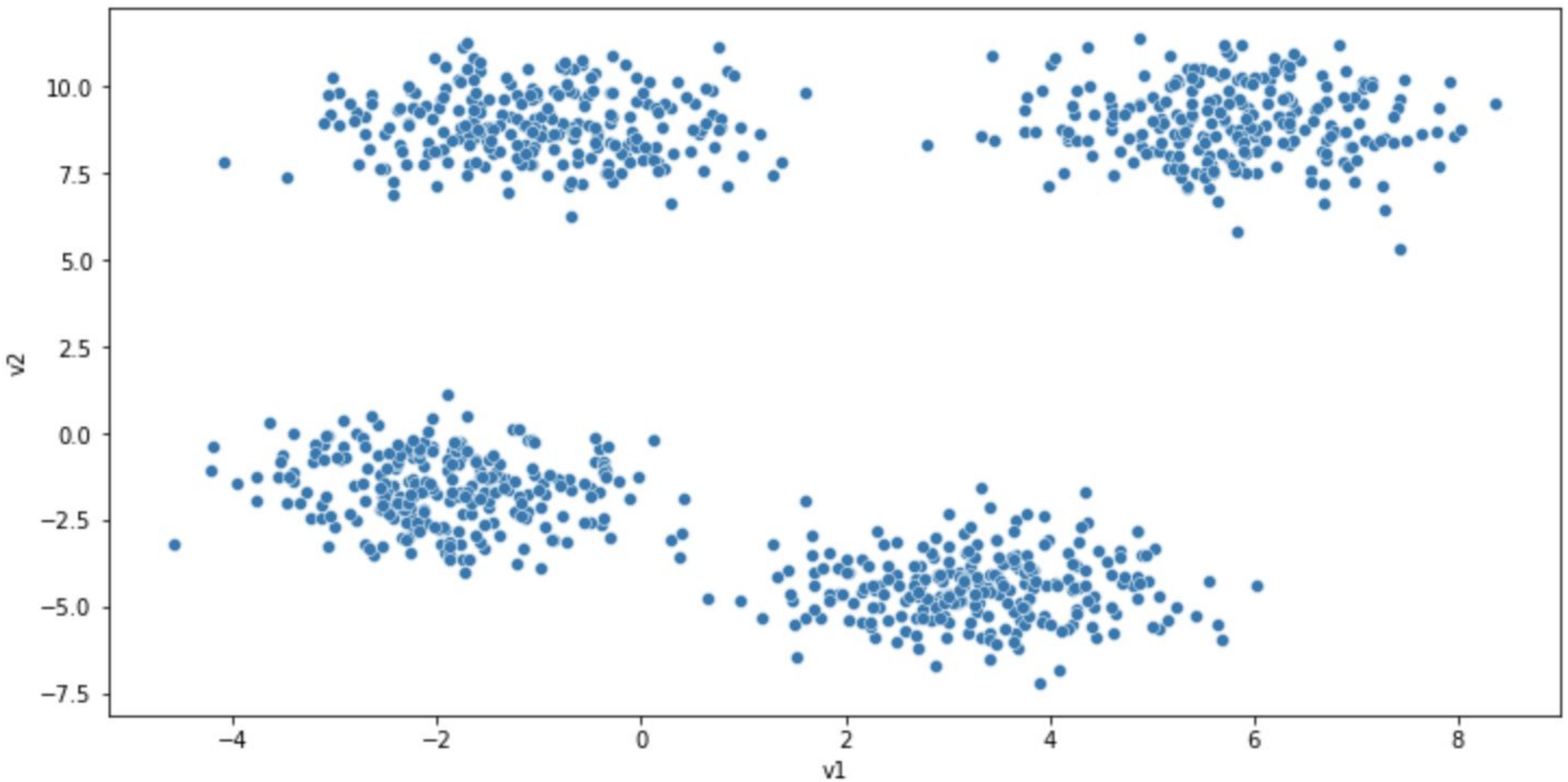
por um vetor de deslocamento médio.

Mean Shift

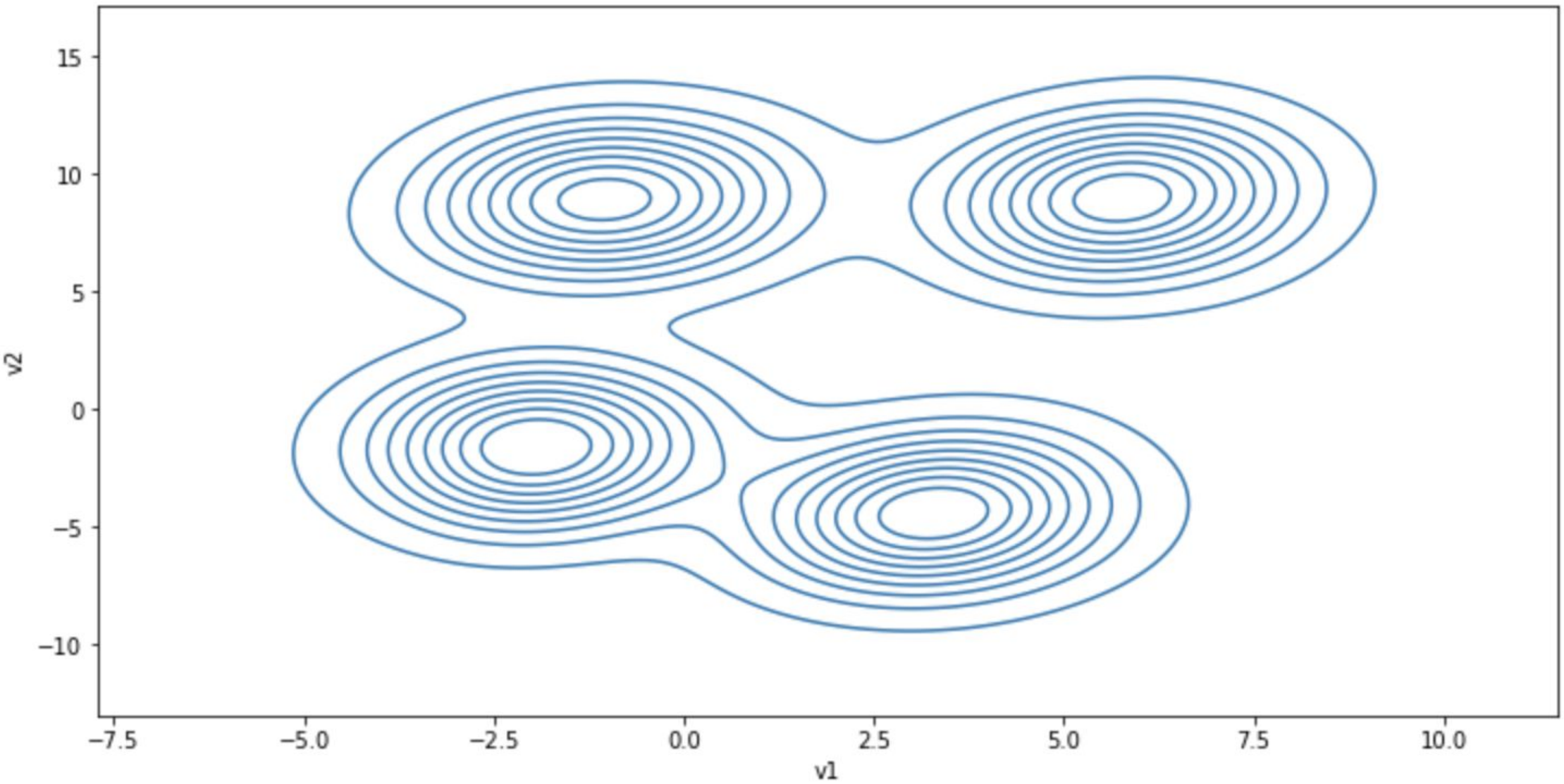
O vetor de deslocamento sempre aponta para o aumento máximo na densidade.

Envolve a mudança de um kernel iterativamente para uma região de densidade mais alta até a convergência.

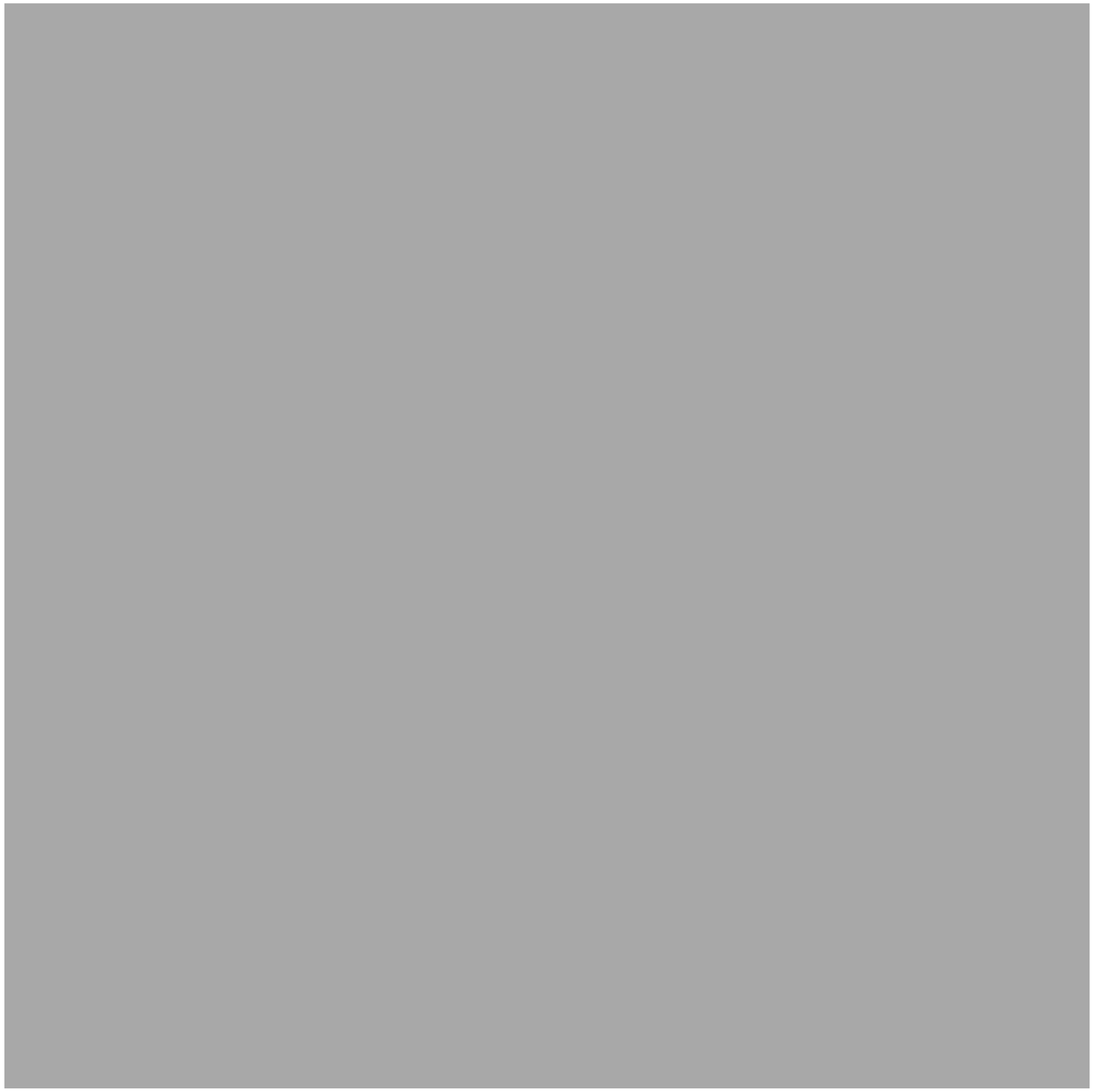
Mean Shift



Mean Shift



Mean Shift



Como Funciona?

Uma janela deslizante circular centrada em um ponto X (aleatório) e tendo o raio R como o núcleo.

A cada iteração, a janela deslizante é deslocada para regiões de maior densidade.

Deslocamos a janela deslizante de acordo com a média dos pontos .

Como Funciona?

Até que não haja uma direção na qual uma mudança possa acomodar mais pontos dentro do kernel.

As etapas 1 a 3 são repetidas, com muitas janelas deslizantes, até que todos os pontos fiquem dentro.

Quando as diferentes janelas deslizantes se sobrepõem, a que contém maior parte dos pontos é mantida.

Mean Shift

Bandwidth

Parâmetro que equivale
ao raio da janela deslizante
se escolhermos um valor grande.

Todos os pontos serão parte
da mesma janela,
formando um só cluster,
oposto também acontece.

Estimate_bandwidth

O sklearn já tem uma função

Implementada para estimativa.

O parâmetro quantile

deve ser ajustado com objetivo

de chegar na melhor

configuração de cluster.

Recapitulando

01

O que é Mean Shift?

02

Como funciona o algoritmo?

03

Parâmetros

04

Escolhendo a melhor distância

data expert.®

Data Science & Machine Learning

CLS 28 - Gaussian Mixture

Consultor: Tulio Souza

Agenda

01

Gaussian Mixture

02

Distribuição Normal

03

Como funciona?

04

Expectation-Maximization

05

Vantagens

06

Achando o número de clusters

Gaussian Mixture

É um modelo usado **universalmente**
para aprendizado ou agrupamento (clustering)
generativo **não supervisionado.**

São usados para representar
sub populações normalmente
distribuídas em uma população geral.

Não exigem a qual subpopulação
um ponto de dados pertence
e permite que o modelo **aprenda**
as subpopulações **automaticamente.**

É também chamado de
Expectation-Maximization Clustering
ou EM Clustering.

É baseado na **estratégia de otimização.**

Distribuição Normal

É uma distribuição de probabilidade **absolutamente contínua** parametrizada pela sua **média matemática e desvio padrão.**

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \ \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

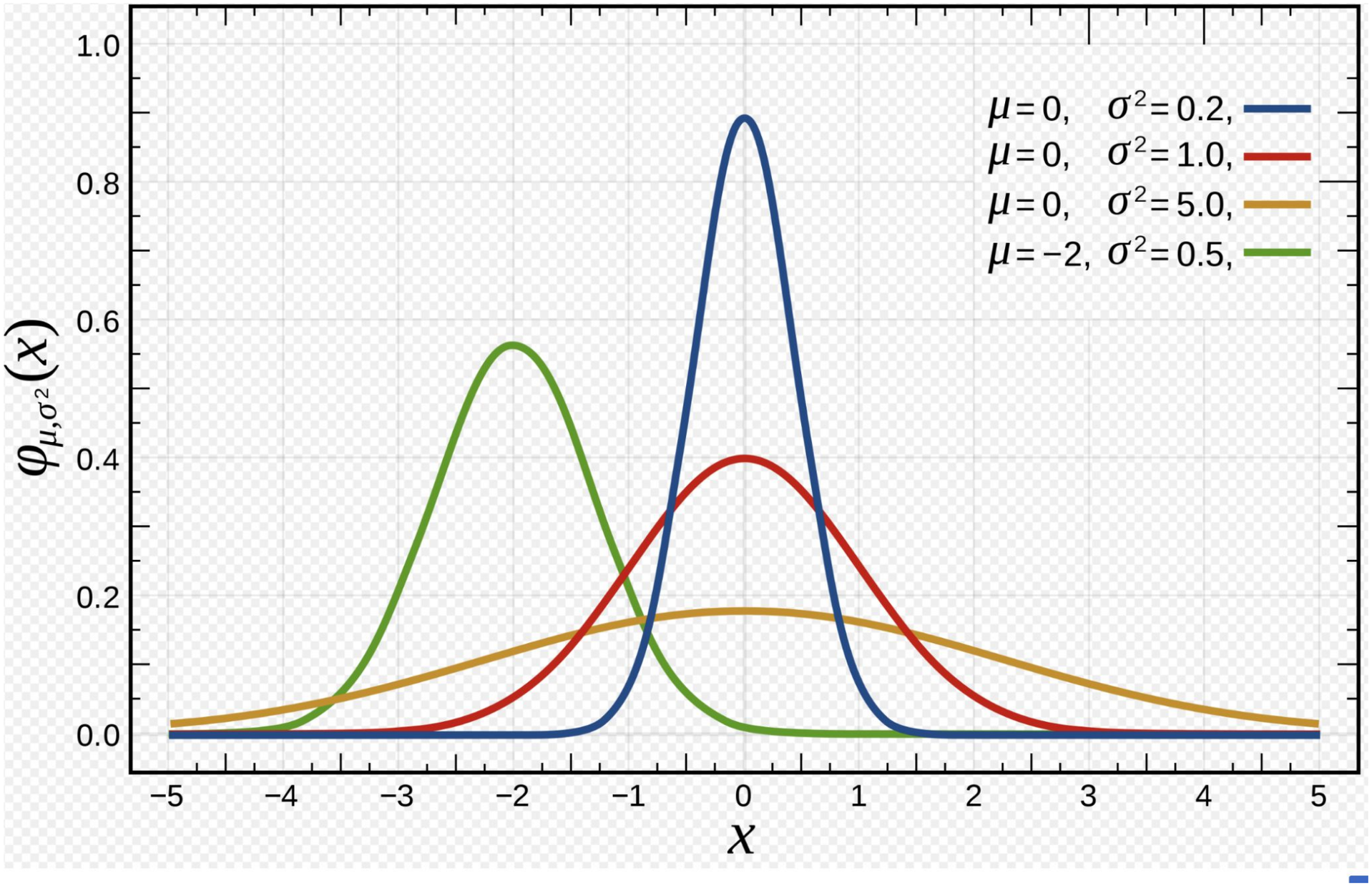
Distribuição Normal

Probabilisticamente a média das variáveis independentes de uma amostra aleatória

Corresponde ao comportamento do efeito agregado de experiências aleatórias independentes

Pode aproximar–se da distribuição de efeito agregado de outras distribuições

Distribuição Normal



Gaussian Mixture

Se houver Múltiplas distribuições, podemos construir
o que chamamos de Modelo de Mistura Gaussiana.

Três distribuições GD1, GD2, GD3 tendo média μ1, μ2, μ3
e variância 1,2,3 para um determinado conjunto de pontos.

O GMM identificará a probabilidade
de dados pertencentes
a cada uma dessas distribuições.

Como funciona?

- Seleciona-se o número de clusters (k).
- Calcula-se a probabilidade dos pontos a distribuição de cada um destes clusters.
- Recalcula-se os parâmetros das distribuições gaussianas (clusters).
- Itera-se até o algoritmo convergir.

Como funciona?

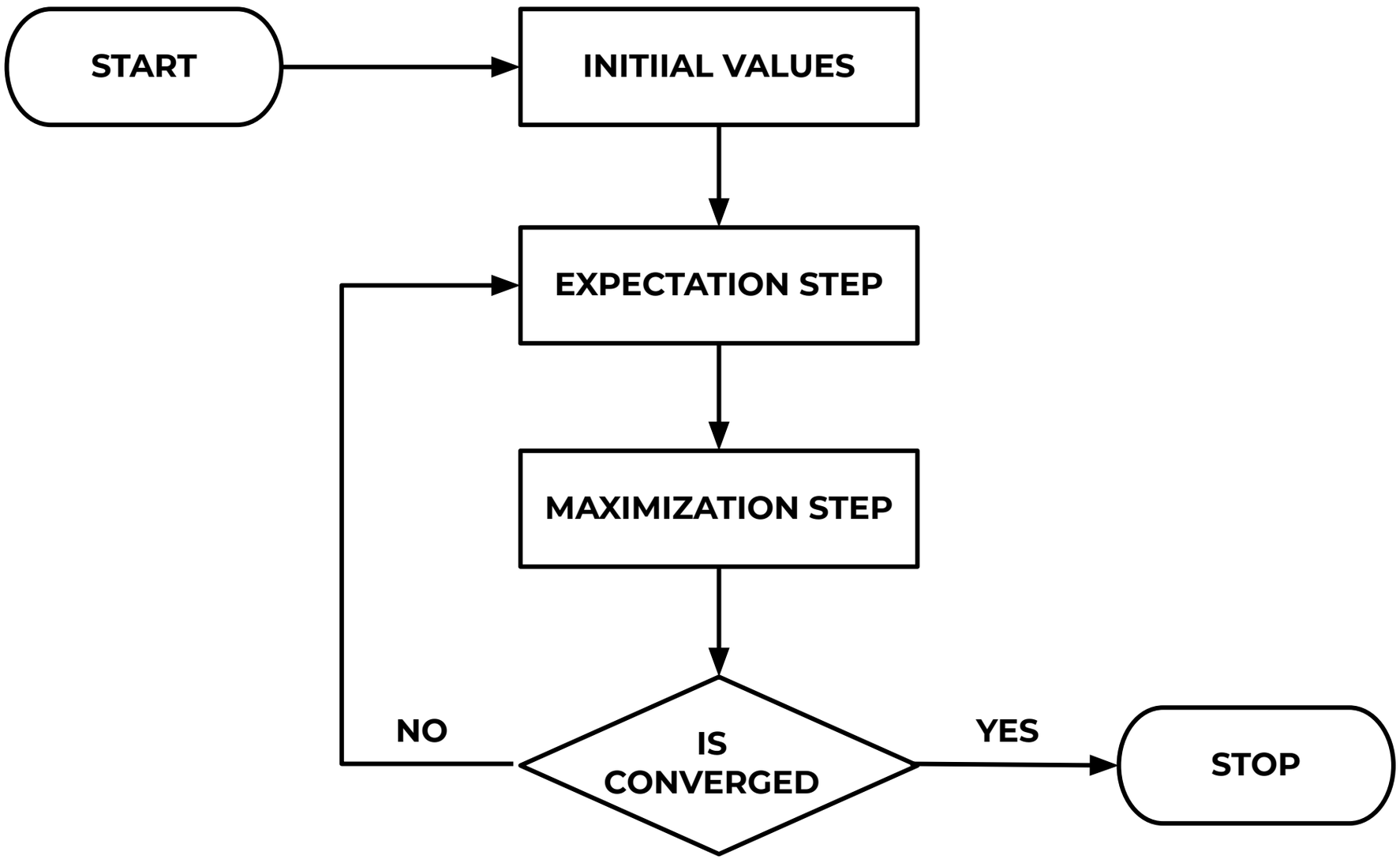
Expectation Maximization

Idéias básicas: dado um conjunto de dados incompletos e um conjunto de parâmetros iniciais.

E-Step: Usando os dados fornecidos e o valor atual dos parâmetros, estime o valor dos dados ocultos.

Passo-M: após o passo-E, é usado para maximizar a variável oculta e a distribuição conjunta dos dados.

Expectation Maximization



Vantagens

01

Os clusters podem ter diferentes formatos em diferentes direções.

02

Os clusters podem se sobrepor!

03

É possível gerar dados utilizando a técnica.

04

Cada observação pode ser descrita com a sua probabilidade.

Número de Clusters

Critério de informação Bayesiano (BIC) ou

Critério de informação de Schwarz (também SIC, SBC, SBIC)

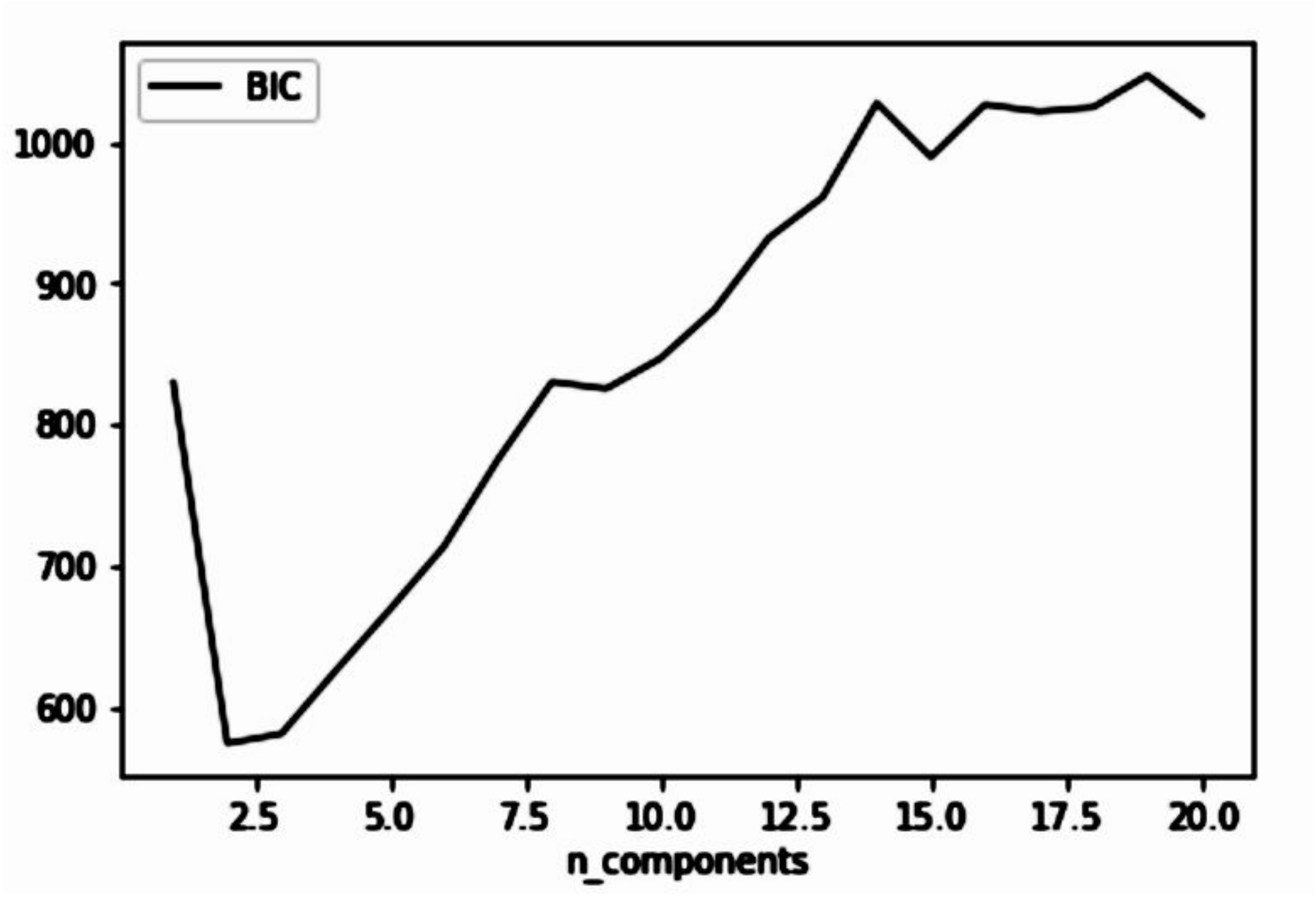
Seleção de modelo entre um conjunto finito de modelos.

Modelos com BIC mais baixo são geralmente preferidos.

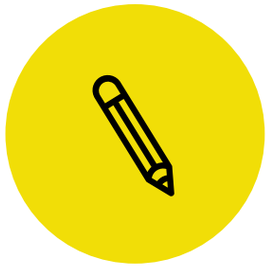
Baseia-se na função de verossimilhança.

Relacionado ao critério de informação de Akaike (AIC).

Número de clusters



Recapitulando



Como funciona?

Vantagens

K-Ótimo

data expert.®

Data Science & Machine Learning

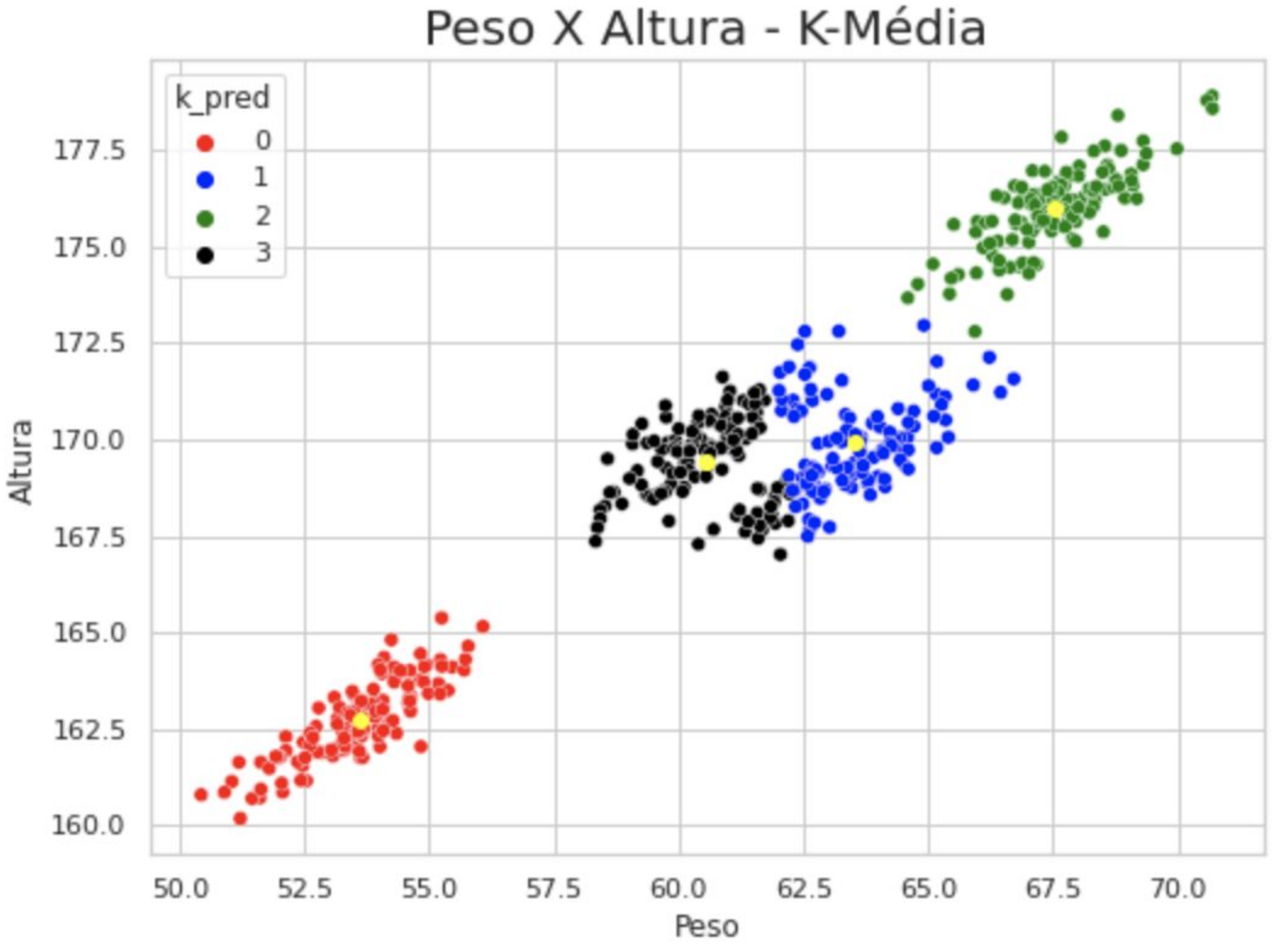
CLS 41 - Fechamento do modulo

Consultor: Tulio Souza

Agenda

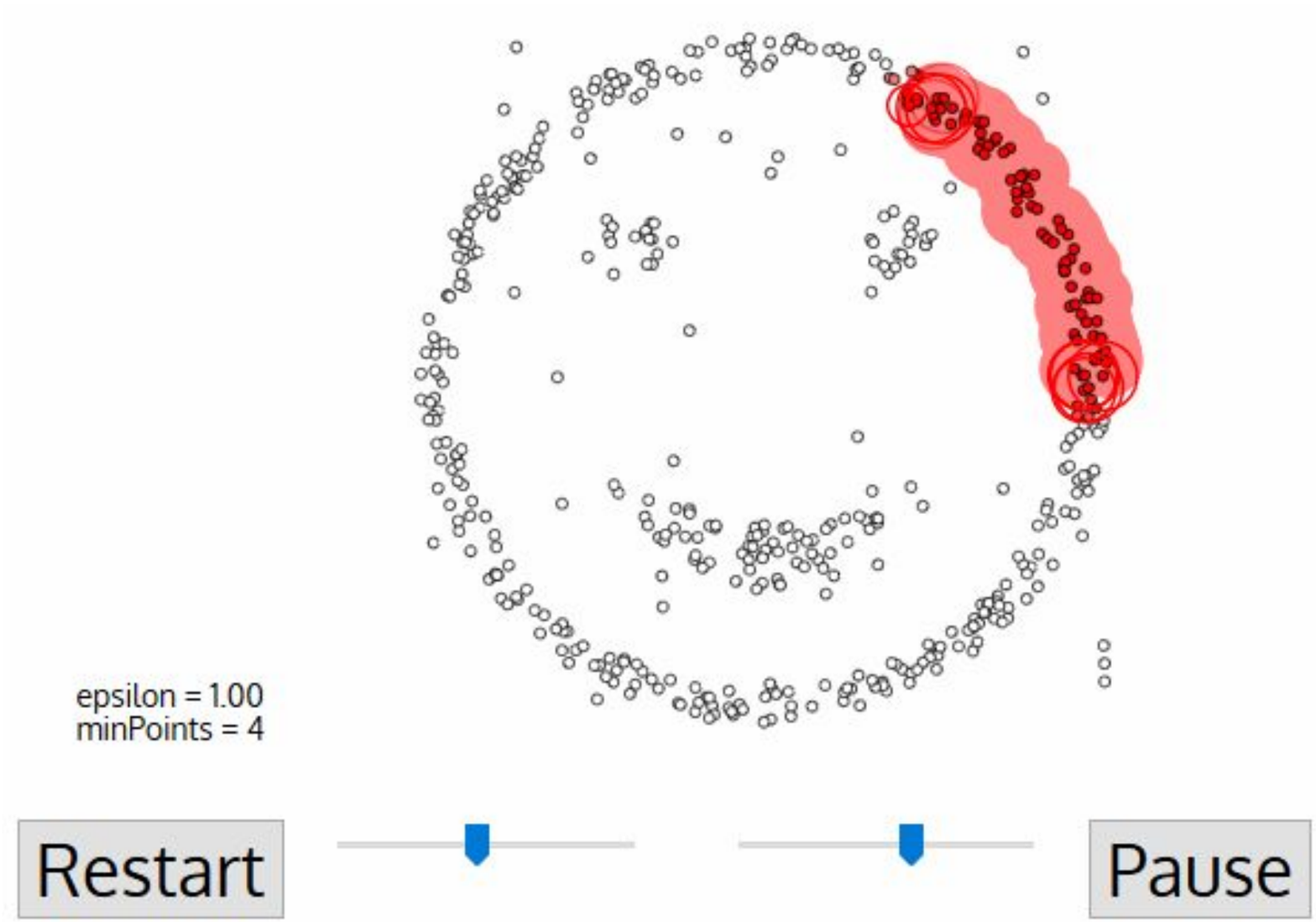
- KMeans
- MeanShift
- Gaussian Mixture
- DBScan
- Hierarchical Clustering

K-Means


Peso X Altura - K-Média
k_pred
0
1
2
3
177.5
175.0
172.5
170.0
167.5
165.0
162.5
160.0
Altura
50.0
52.5
55.0
57.5
60.0
62.5
65.0
67.5
70.0
Peso

MeanShift

DBSCAN



Gaussian Mixture

Hierarchical Clustering