

Inteligência Artificial ao alcance de todos!



Aula 08/2020: Clustering, K-Means e K-NN

Apoio:



Professor: Luiz Andrade

Short Bio



Engenheiro civil pela Escola Politécnica da USP
Msc em Engenharia de Sistemas Logísticos pela Escola Politécnica da USP
(in course) PhD em Engenharia de Transportes pela Escola Politécnica da USP



M.B.A. em Global Supply Chain e Logistics pelo Massachusetts Institute of Technology



13 anos de experiência com empreendedorismo e tecnologia. Fundação de duas start-ups.



Voluntário da Escola Livre de IA

Agenda

- O que cluster analysis
- Aplicações
- Representação e medidas de similaridade
- Algoritmo k-means
- Algoritmo k-means para classificação
- Algoritmo k-NN
- Relação entre k-means e k-NN (métodos de protótipo)
- Aplicação em caso prático

Cluster analysis

- Objetivo: “segmentar uma **coleção de objetos** em sub-conjuntos tal que os objetos de um conjunto são mais **semelhantes** entre si do que elementos de conjuntos diferentes”
- Conjuntos: o problema parte de um conjunto de observações e não é sabido a priori a qual cluster cada elemento pertence (aprendizado não supervisionado).
- Objetos: são observações descritas por diferentes variáveis ou medidas (ex.: usuários, filmes, restaurantes) ou mesmo pela distância relativa entre outros objetos da coleção.
- Semelhantes: a semelhança ou dessemelhança é definida como uma função de “distância” entre dois objetos. Específica de cada aplicação.

Aplicações

- Processamento de Textos
 - Identificação de tipos similares de arquivos/textos
 - Segmentação de conteúdo
 - Sumarização de textos (compressão)
- Processamento de imagens
 - Agrupamento de tipos de imagens
 - Classificação de imagens (hand written digits)
 - Compressão de imagens
- Sistemas de controle
 - Identificação de outliers
- E-commerce
 - Segmentação de clientes
 - Sistemas de recomendação (identificação de clusters de clientes e/ou produtos)
- Outros:
 - Segmentação e análise de sequenciamento genético
 - Agrupamento de perfis de redes sociais (marketing segmentation)
 - ...

Representação

Distância baseada em atributos

- Distância entre dois elementos é baseada em uma **função de similaridade**
- Depende da quantidade e tipo de dimensões das observações
 - Dimensões reais
 - Dimensões categóricas
 - ...
- Produz distâncias simétricas

obs	x	y
1	0.23	1.32
2	-0.87	2.44
3	-0.38	-3.96
...
n	0.74	2.07

Ex.: distância quadrática

$$d(o_i, o_j) = \sum_k (o_{ik} - o_{jk})^2$$

Matrizes de proximidade

- Distância entre dois elementos é baseada em uma **matriz quadrada semi-positiva definida**
- Se existem n elementos a serem considerados, é necessário uma matriz ($n \times n$)
- Em geral assumem-se matrizes simétricas com diagonal 0

obs	1	2	3	...	n
1	0	1.34	1.44	...	3.23
2	1.34	0	4.32	...	0.43
3	1.44	4.32	0	...	0.67
...
n	3.23	0.43	0.67	...	0

Representação

- Diferentes funções de similaridade para diferentes tipos de variáveis

Variáveis quantitativas

$$d(o_i, o_j) = \sum_k (o_{ik} - o_{jk})^2 \quad d(o_i, o_j) = \sum_k |(o_{ik} - o_{jk})|$$

Variáveis ordinais (Ex.: Alto/Medio/Baixo, muito ruim/ruim/neutro/bom/muito bom)

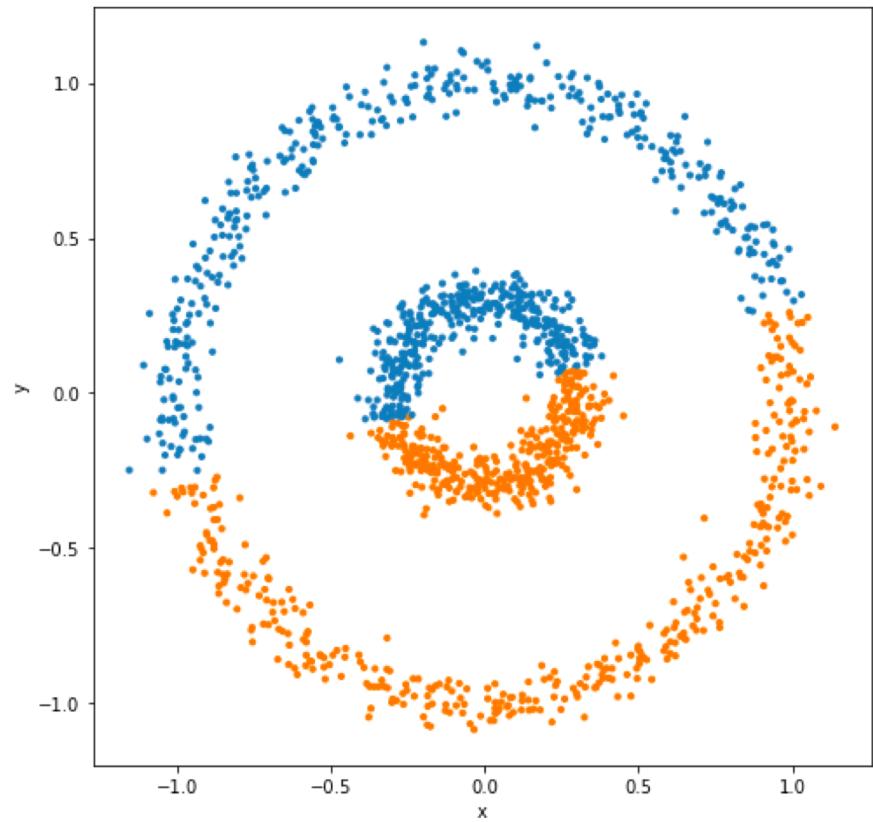
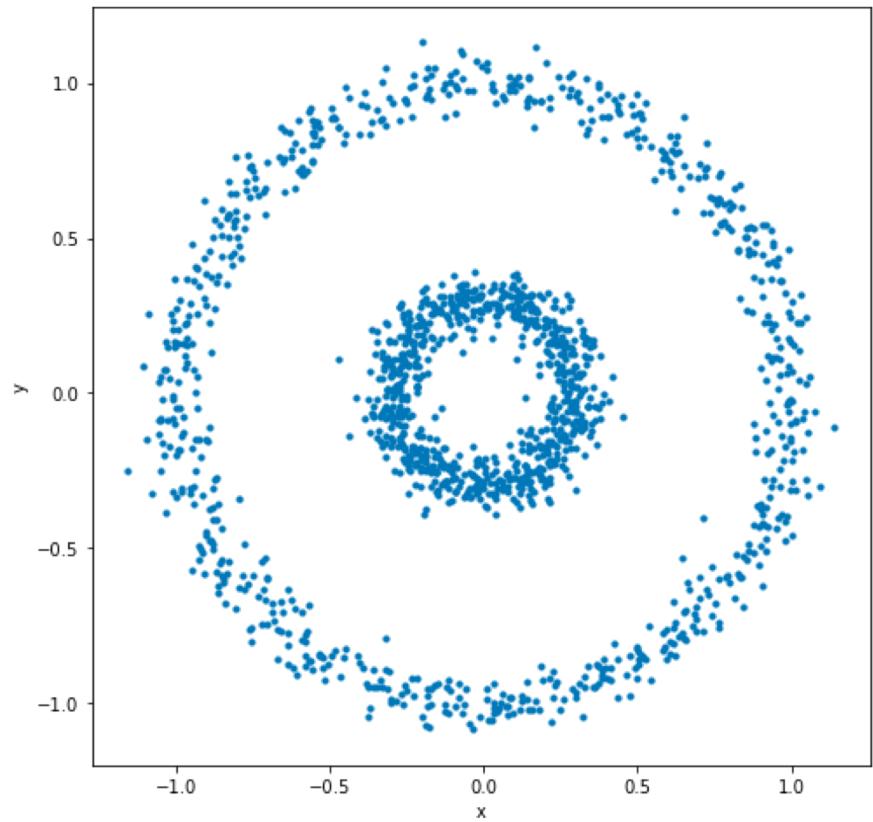
Transformação

$$o'_i = \frac{o_i - 1/2}{M} \quad \longrightarrow \quad \text{Trata-se } o'_i \text{ como uma variável quantitativa}$$

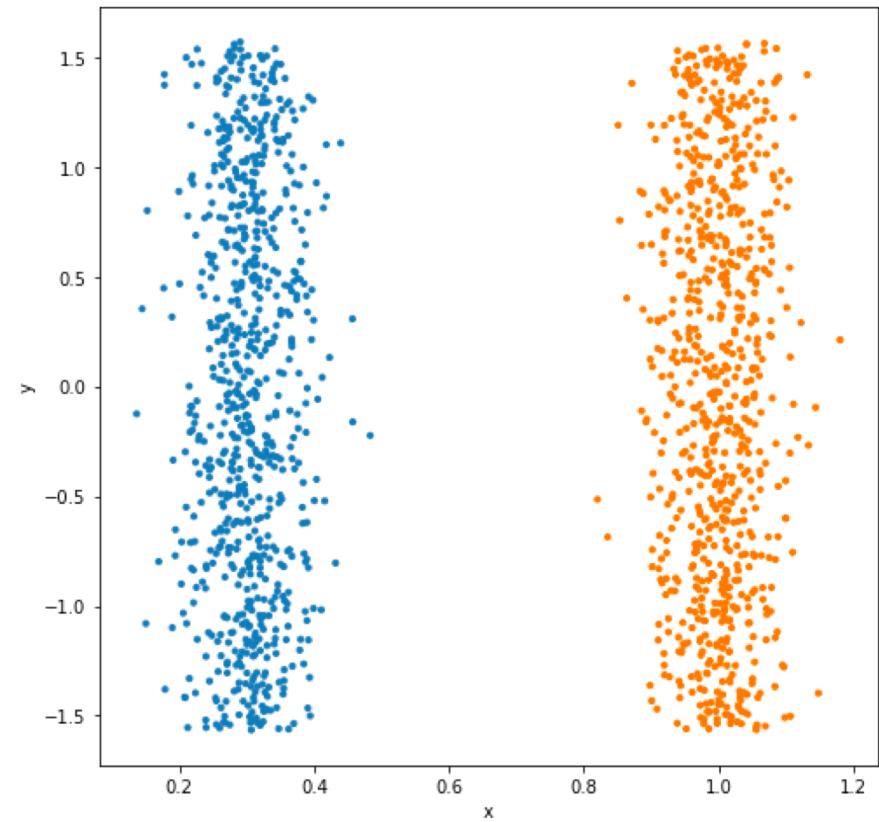
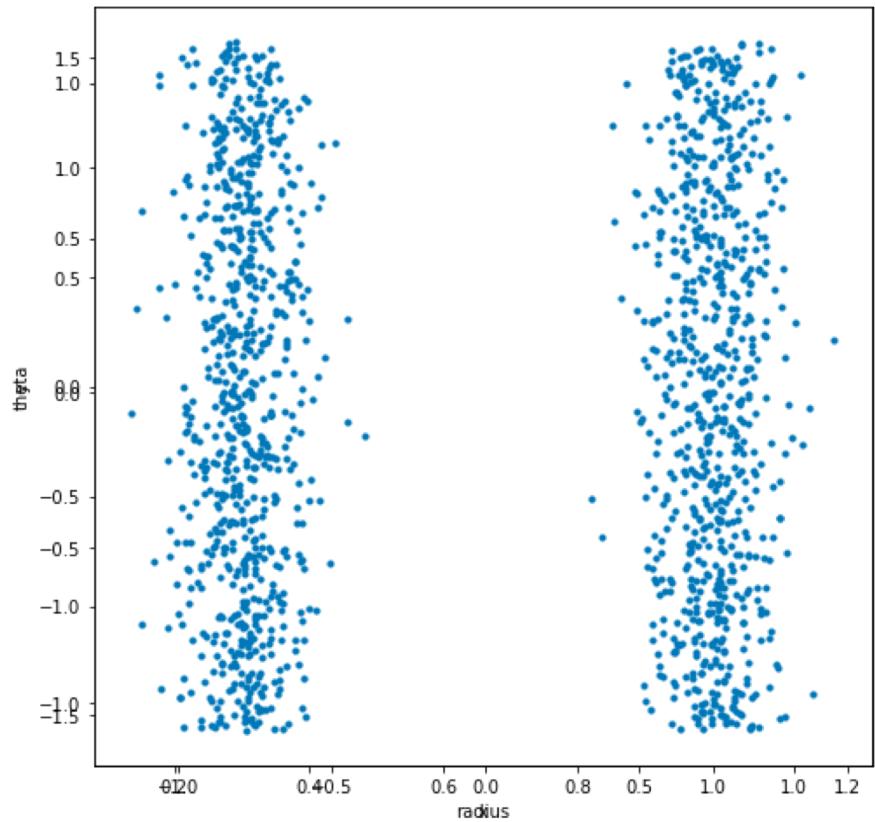
Variáveis categóricas

$$d(o_i, o_j) = \sum_k (0 \text{ se } o_{ik} = o_{jk}, 1 \text{ caso contrário})$$

Representação



Representação



K-Means

- Algoritmo simples e iterativo
- Funciona apenas quando todas as variáveis são quantitativas ou o problema é definido por uma matriz de distâncias
- Deseja-se minimizar a dissimilaridade entre observações do mesmo cluster:

$$W(C) = \frac{1}{2} \sum_{k=1}^K \sum_{C(x_i)=k} \sum_{C(x_j)=k} (x_i - x_j)^2$$

de clusters (parâmetro do algoritmo)

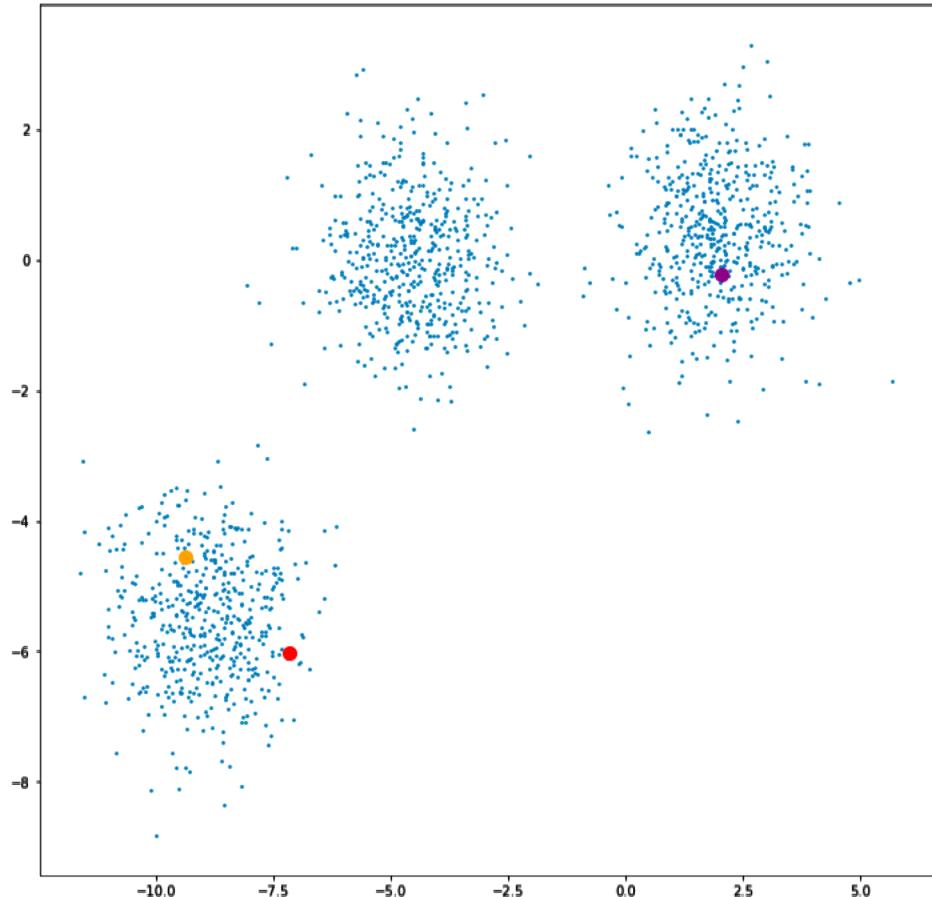
i e j pertencem ao mesmo cluster

$$= \sum_{k=1}^K \sum_{i=1}^N I(C(i) = k) \sum_{C(i)=k} (x_i - \bar{x}_k)^2$$

Distância de cada observação ao centro de cada cluster

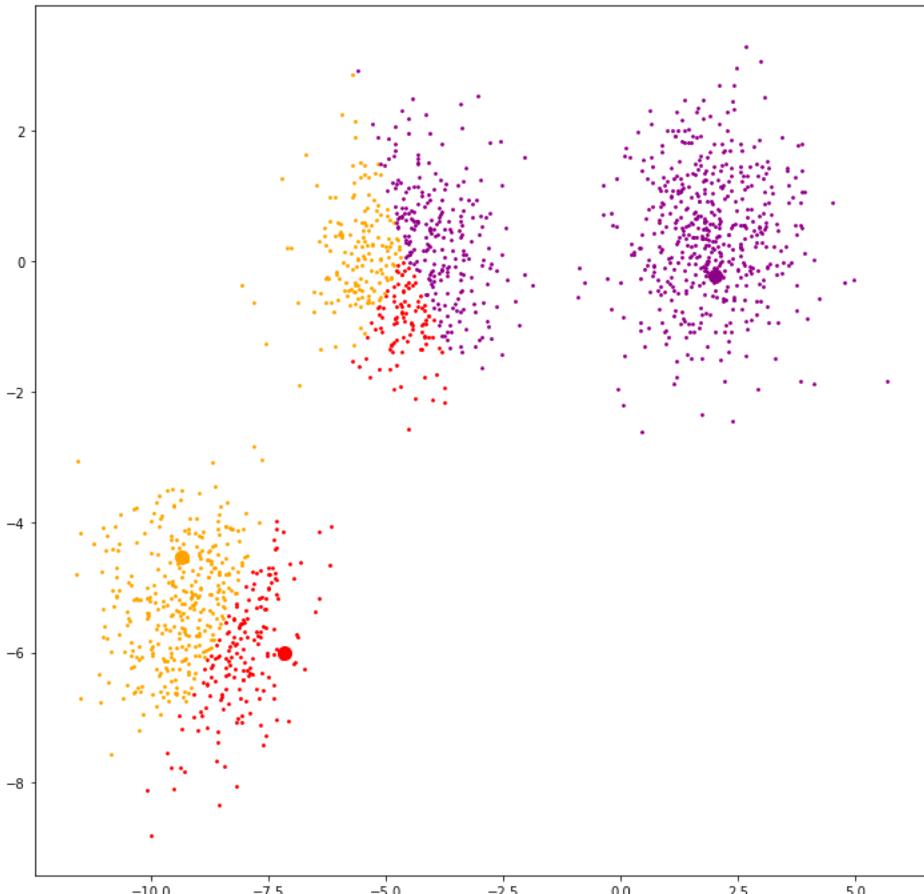
- O conjunto ótimo de clusters C e alocações ($C(i)$) é definido uma vez que $W(C)$ seja minimizado

K-Means



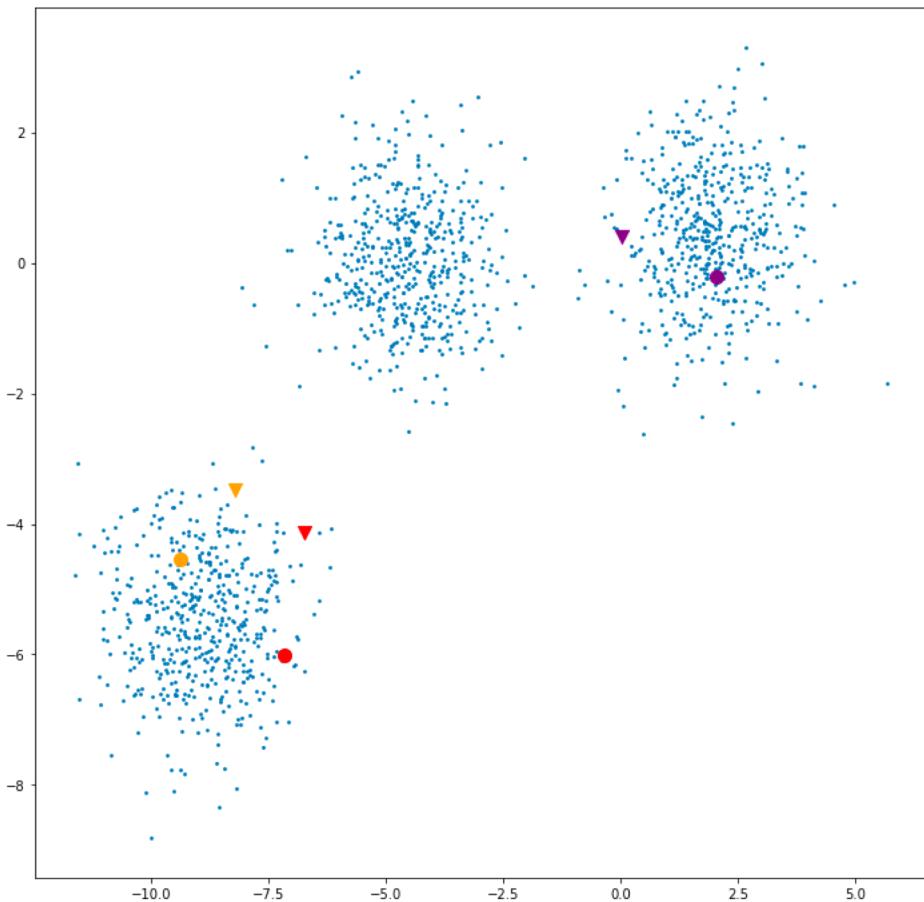
- Passo 1: seleciono K observações aleatórias*

K-Means



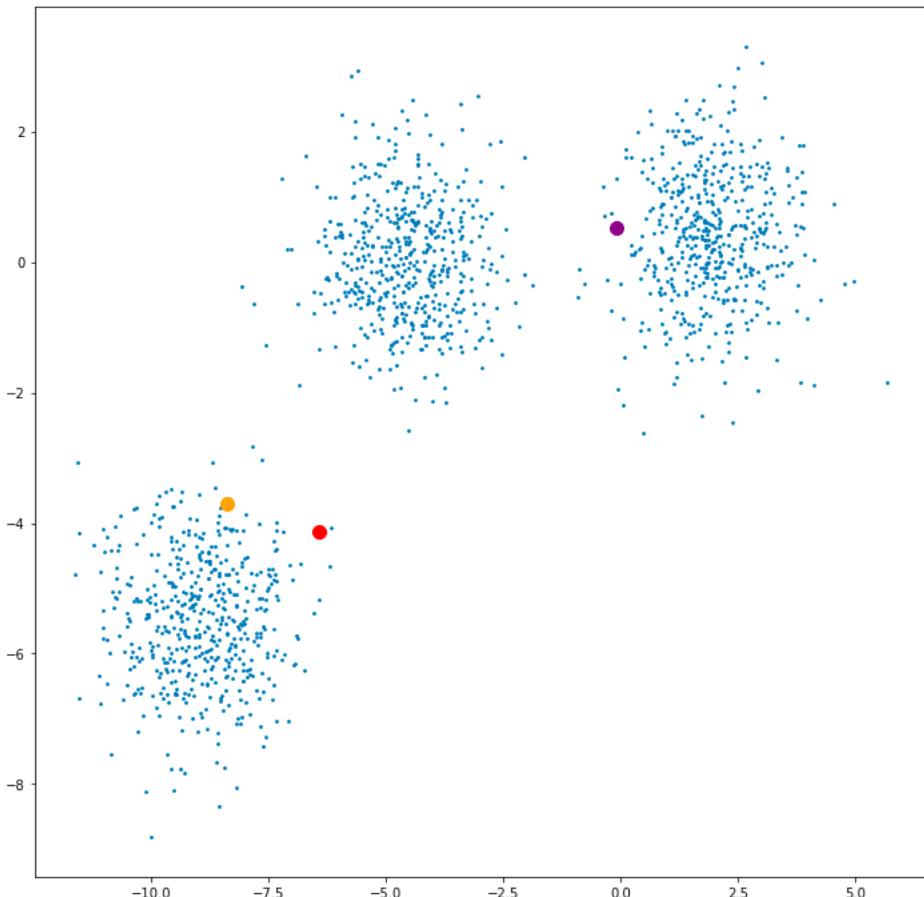
- Passo 1: seleciono K observações aleatórias* e defino como centros dos clusters
- Passo 2: aloco cada observação ao centro de cluster mais próximo

K-Means



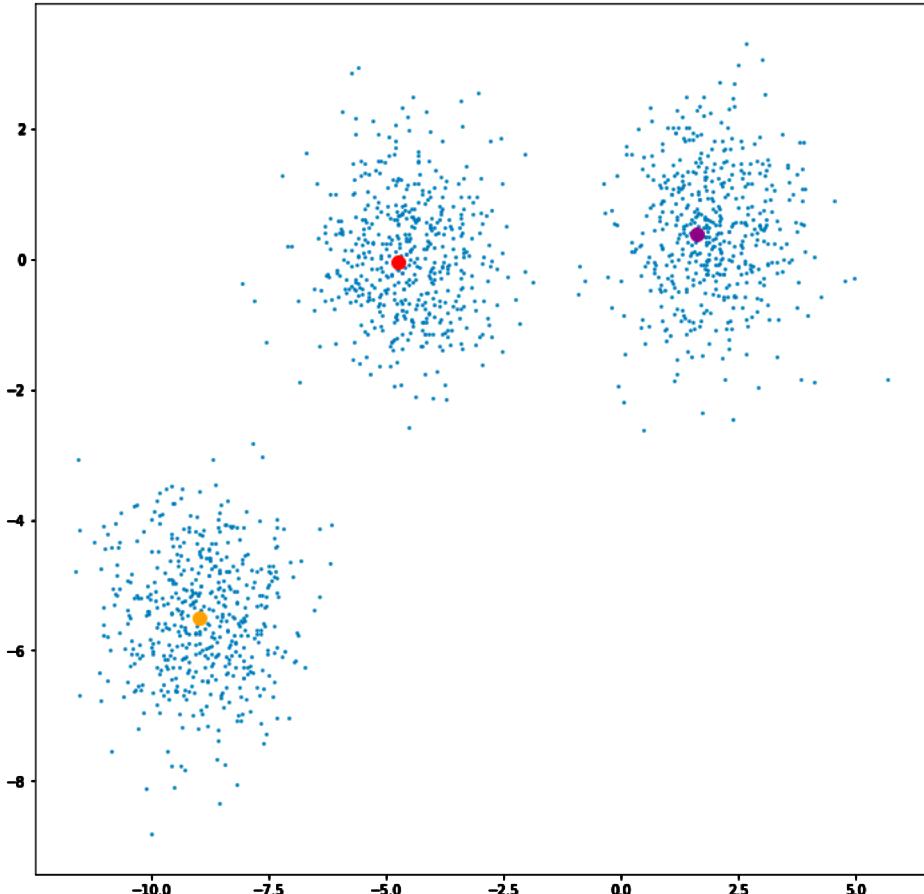
- Passo 1: seleciono K observações aleatórias* e defino como centros dos clusters
- Passo 2: aloco cada observação ao centro de cluster mais próximo
- Passo 3: calcula a mediana dos pontos de um cluster

K-Means



- Passo 1: seleciono K observações aleatórias* e defino como centros dos clusters
- Passo 2: aloco cada observação ao centro de cluster mais próximo
- Passo 3: calcula a mediana dos pontos de um cluster
- Passo 4: reposiciono os centros dos cluster com o ponto mais próximo de cada mediana calculada

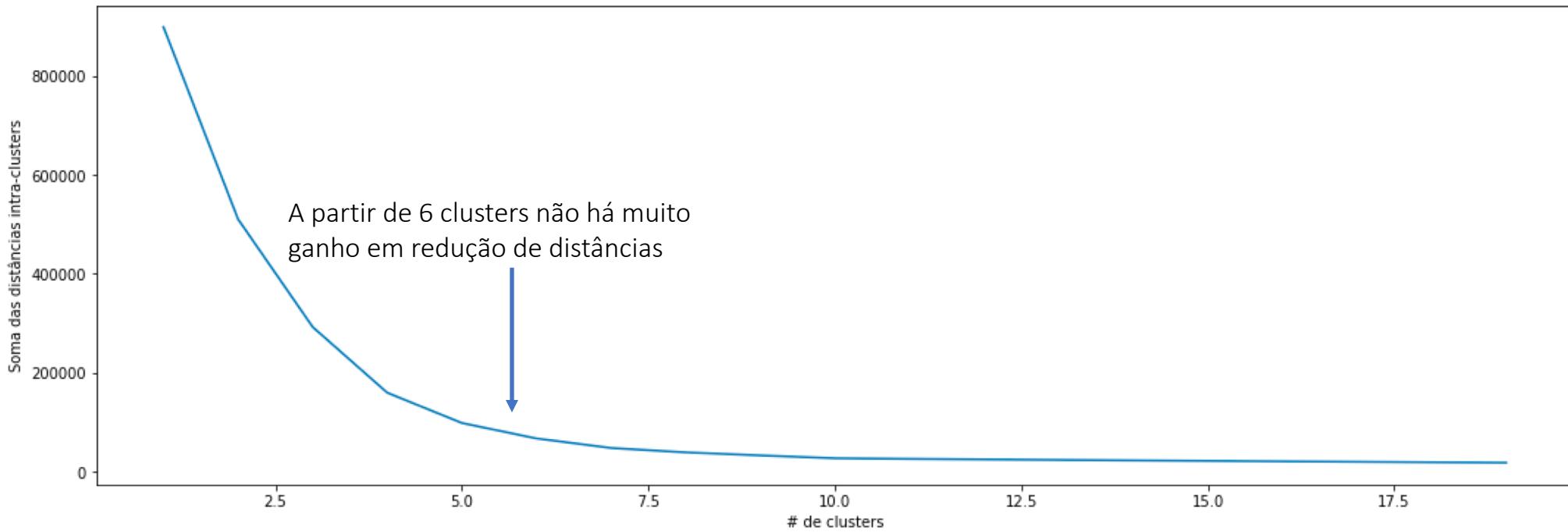
K-Means



- Passo 1: seleciono K observações aleatórias* e defino como centros dos clusters
- Passo 2: aloco cada observação ao centro de cluster mais próximo
- Passo 3: calcula a mediana dos pontos de um cluster
- Passo 4: reposiciono os centros dos cluster com o ponto mais próximo de cada mediana calculada
- Repete passos 2, 3 e 4 até que o centro dos clusters não mude mais

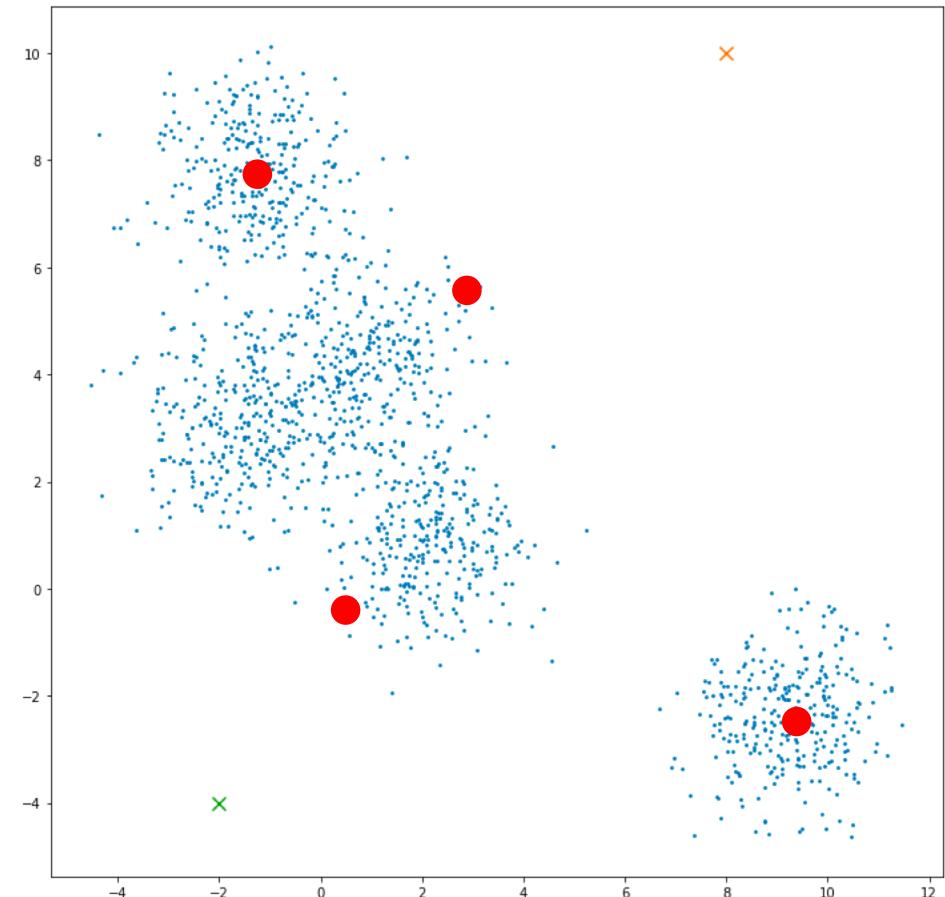
K-Means

- A quantidade de clusters é um parâmetro empírico e deve ser selecionado:
 - Ou com base em conhecimento de especialista no problema (Ex.: Número de classificações de usuários, categorias de produtos)
 - Ou com base em testes da **curva de distância total** (também chamado de inércia em alguns pacotes computacionais)



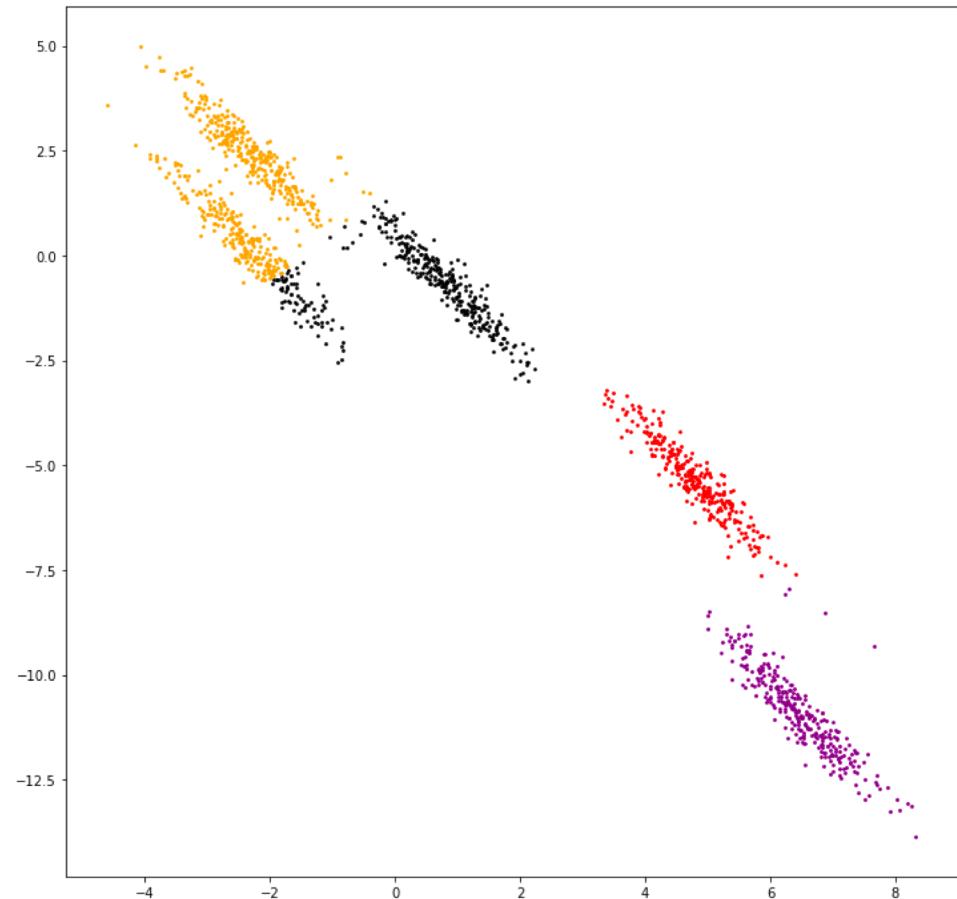
K-Means – Vantagens e Desvantagens

- Vantagens:
 - Simples e de fácil implementação
 - Eficiente $O(\text{iterações} \times n_{\text{observações}} \times n_{\text{clusters}})$
- Desvantagens:
 - Necessário especificar K
 - Sensível a *outliers*
 - Não adequado para encontrar clusters não convexos (distribuições anisotrópicas)



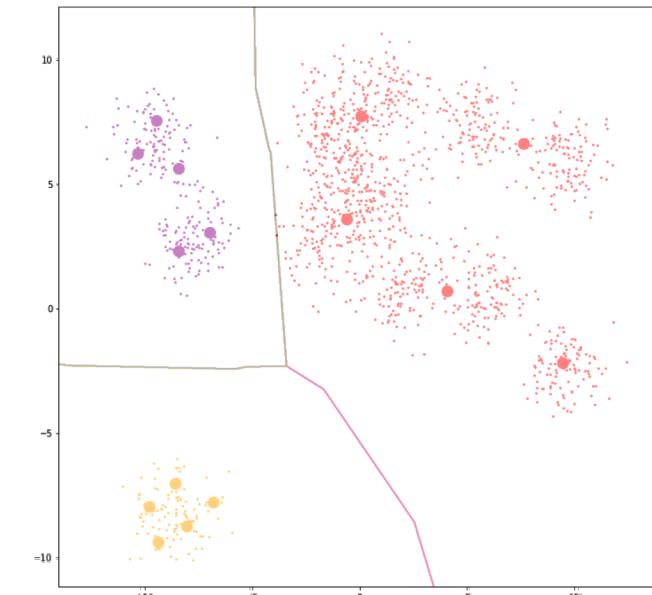
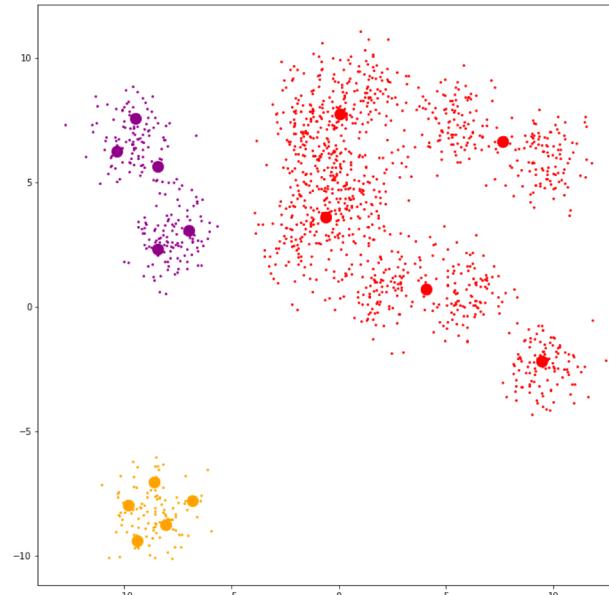
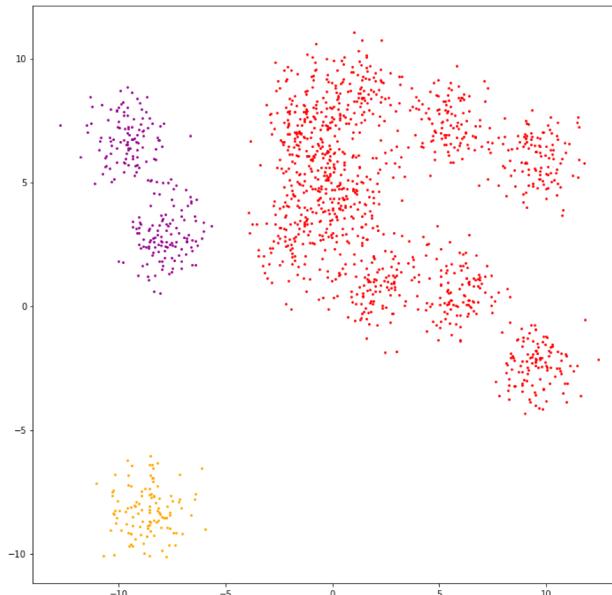
K-Means – Vantagens e Desvantagens

- Vantagens:
 - Simples e de fácil implementação
 - Eficiente $O(\text{iterações} \times n_{\text{observações}} \times n_{\text{clusters}})$
- Desvantagens:
 - Necessário especificar K
 - Sensível a *outliers*
 - Não adequado para encontrar clusters não convexos (distribuições anisotrópicas)



K-Means para classificação

- Podemos utilizar o algoritmo K-Means para problemas de classificação (aprendizado supervisionado):
 - Para cada classe do problema ajustamos R clusters (ou R-protótipos da classe)
 - Dada uma nova observação, sua classe é a mesma que o protótipo mais próximo em termos de uma função de similaridade
 - Com isso traçamos uma superfície de decisão.



K-Nearest Neighbours

- O algoritmo k-NN não é um algoritmo de clustering, mas usa ideias similares
- O algoritmo k-NN é um algoritmo de aprendizado **supervisionado**
- Dado um conjunto de treinamento com pares (x_i, y_i) , a classe ou valor previsto y' para uma amostra de teste x é dada por:

Regressão

$$y' = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

$N_k(x)$ São os k-vizinhos mais próximos de x

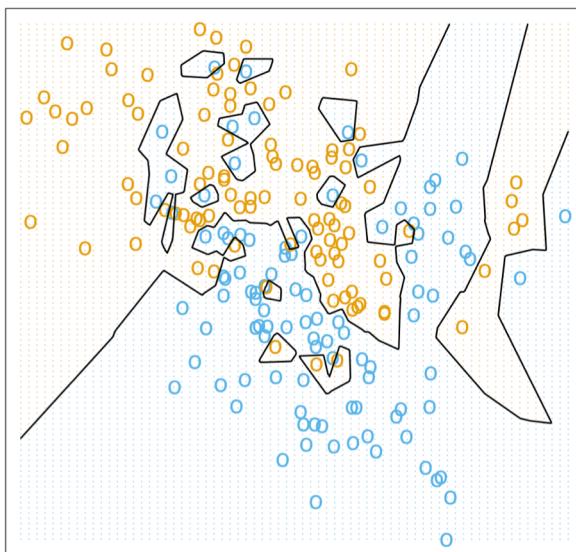
Classificação

$$y' = \text{moda } (y_i | x_i \in N_k(x))$$

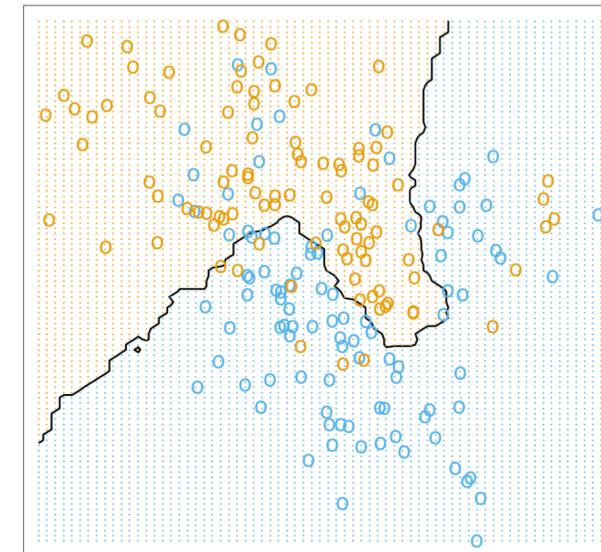
K-Nearest Neighbours

- O parâmetro K é definido para cada caso.
- Em geral:
 - se aumento K, aumento o bias e reduzo a variância
 - se reduzo K, reduzo o bias e aumento a variância

1-NN

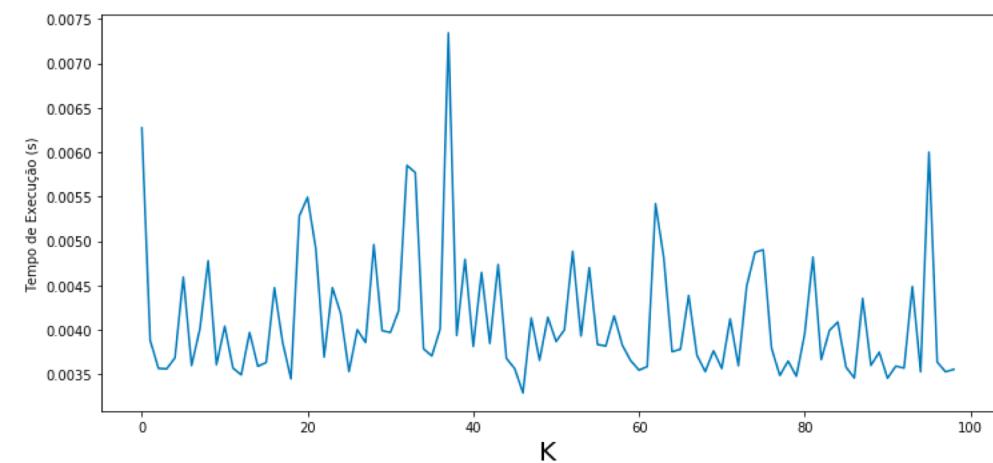
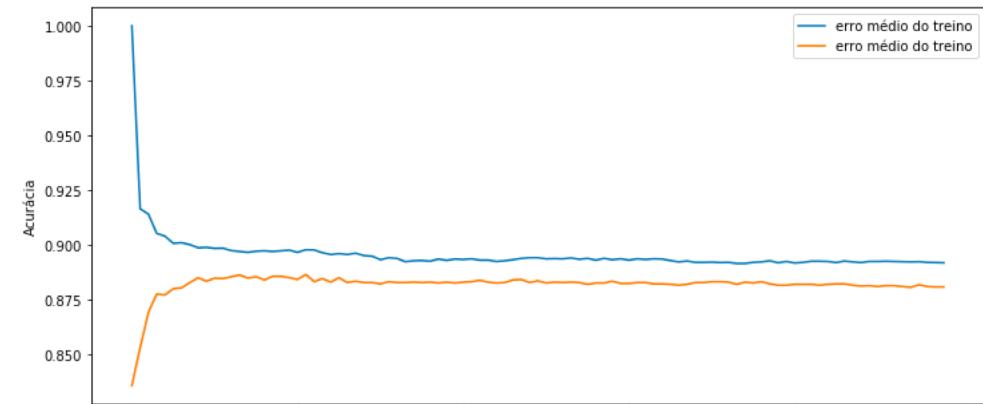
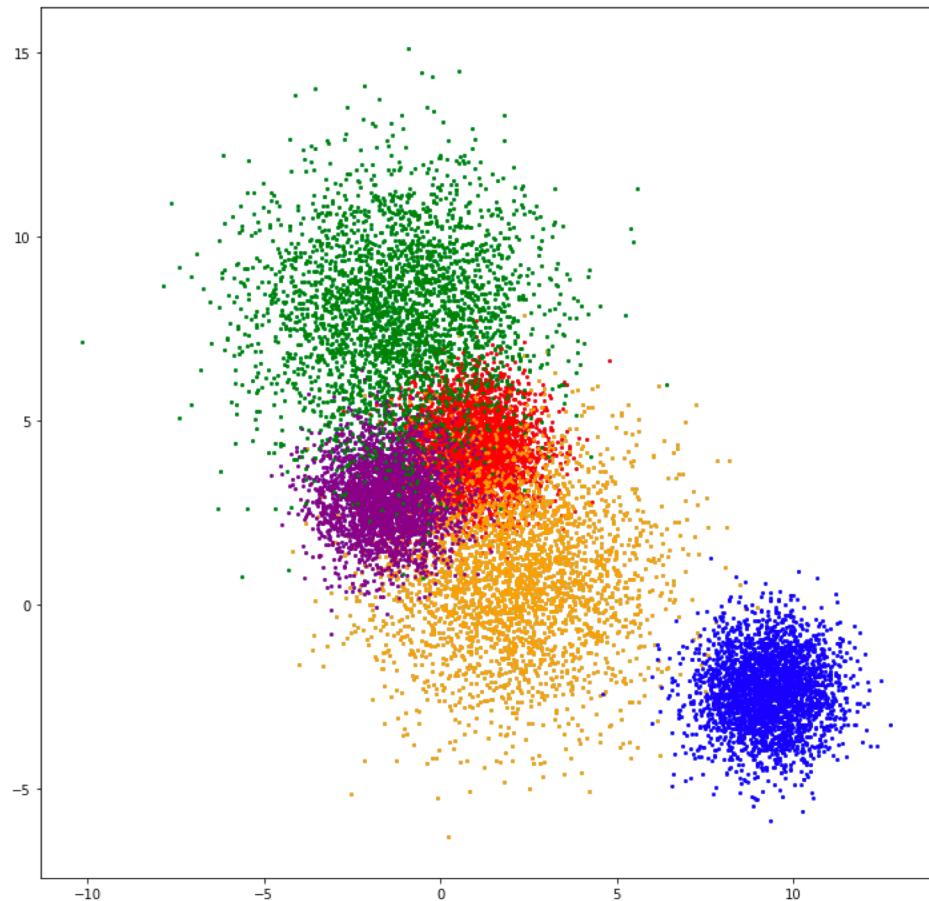


15-NN



K-Nearest Neighbours

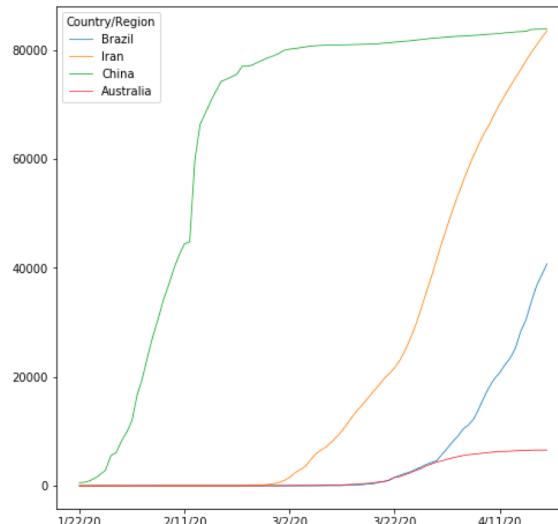
N=15.000 Classes=5



Aplicação prática – Países similares COVID-19

- Vamos analisar as curvas de contágio e óbito de diferentes países e tentar formar clusters;
- Objetivo: se encontrarmos grupos bem definidos, podemos ajudar as organizações de saúde a estabelecer políticas públicas similares para países similares;
- Fonte de dados: <https://github.com/CSSEGISandData/COVID-19>
- A partir das curvas de contágio e óbitos as features que foram construídas são:

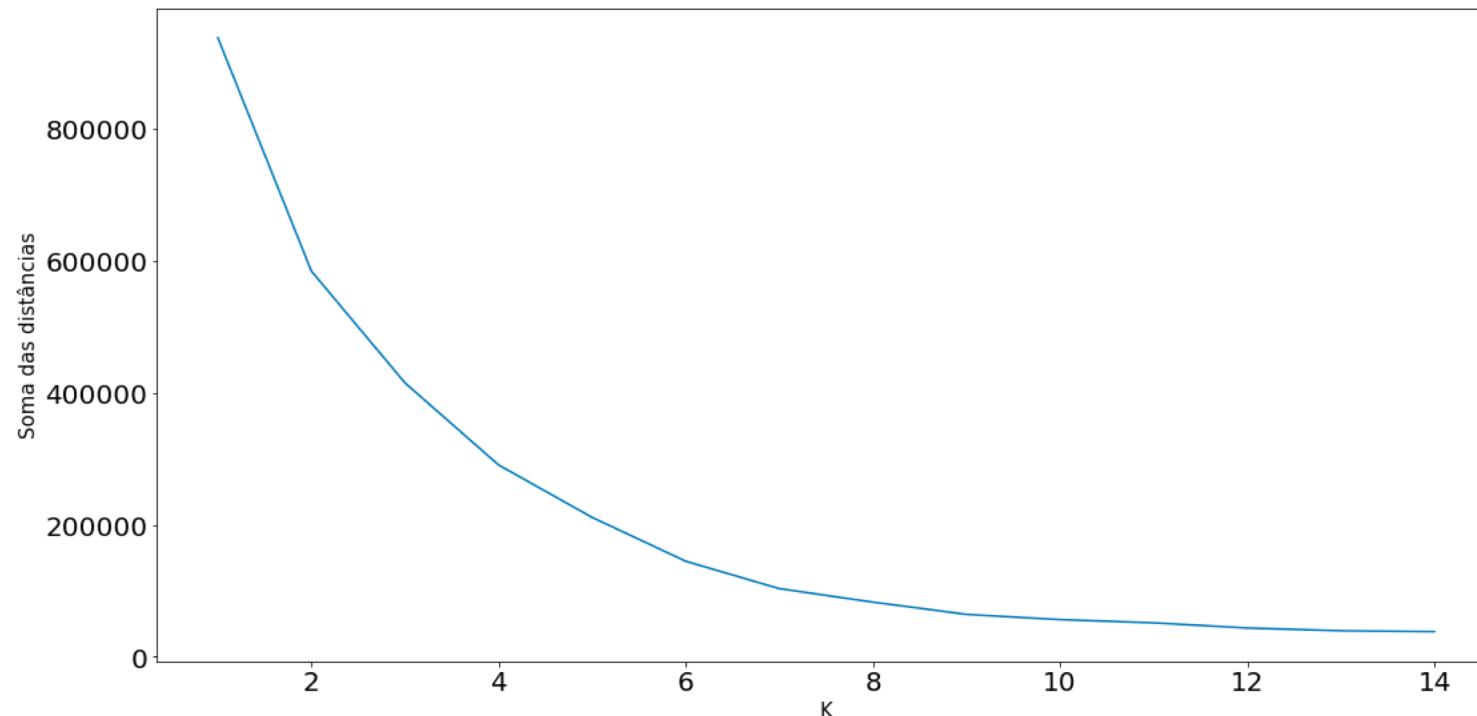
- Dias até primeiro caso
- Dias até 100 casos
- Dias até 1000 casos
- Dias até 10000 casos
- Dias até primeiro óbito
- Dias até 100 óbitos
- Dias até 1000 óbitos
- Dias até 10000 óbitos



Country/Region	first_case	100_cases	1000_cases	10000_cases	first_death	100_deaths	1000_deaths	10000_deaths
Afghanistan	33	65	89	0	60	0	0	0
Albania	47	61	0	0	49	0	0	0
Algeria	34	59	72	0	50	72	0	0
Andorra	40	60	0	0	60	0	0	0
Angola	58	0	0	0	67	0	0	0
Antigua and Barbuda	51	0	0	0	76	0	0	0
Argentina	41	58	69	0	46	83	0	0
Armenia	39	57	81	0	64	0	0	0
Australia	4	48	59	0	39	0	0	0
Austria	34	46	54	69	50	68	0	0
Azerbaijan	39	64	80	0	51	0	0	0
Bahamas	54	0	0	0	70	0	0	0
Bahrain	33	48	80	0	54	0	0	0
Bangladesh	46	75	83	0	56	89	0	0
Barbados	55	0	0	0	74	0	0	0

Aplicação prática – Países similares COVID-19

- Passo 1: vamos escolher a quantidade de clusters



Aplicação prática – Países similares COVID-19

- Passo 1: vamos escolher a quantidade de clusters
- Passo 2: vamos ajustar o modelo K-Means ($K=9$) para os dados dos países

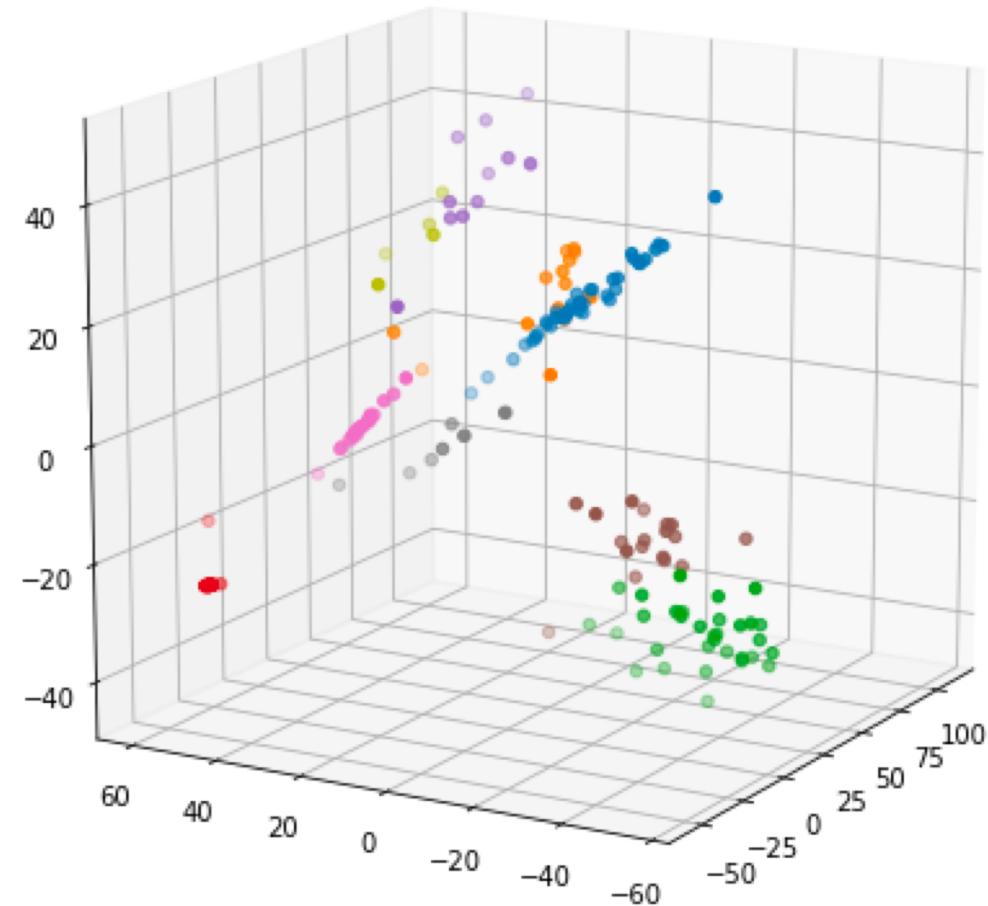
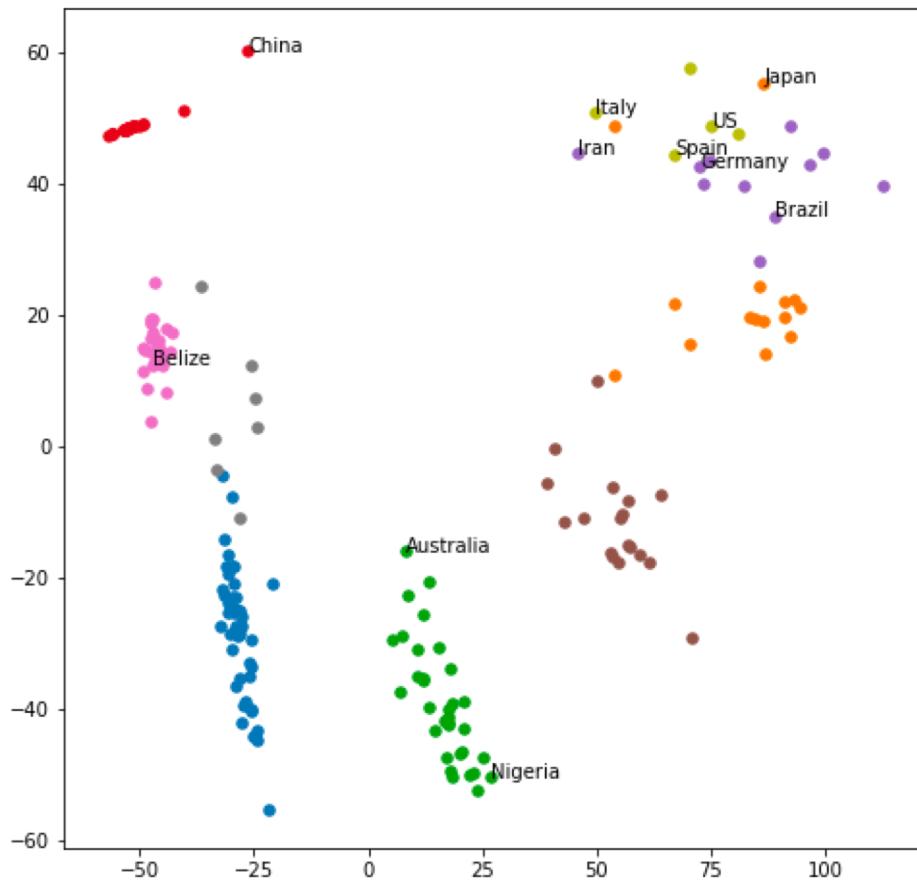
Aplicação prática – Países similares COVID-19

- Passo 1: vamos escolher a quantidade de clusters
- Passo 2: vamos ajustar o modelo K-Means ($K=9$) para os dados dos países
- Passo 3: como o espaço de variáveis possui 8 dimensões ($n_features=8$) precisamos de reduzir o espaço do problema. Utilizaremos a Análise de Componentes Principais (PCA)

Aplicação prática – Países similares COVID-19

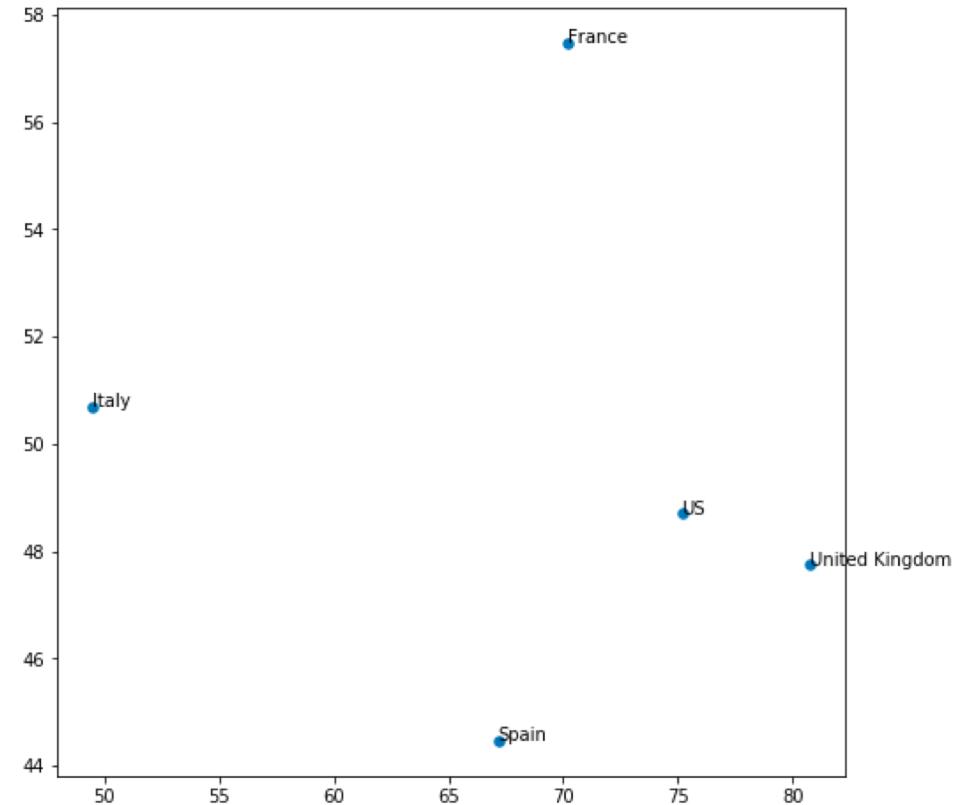
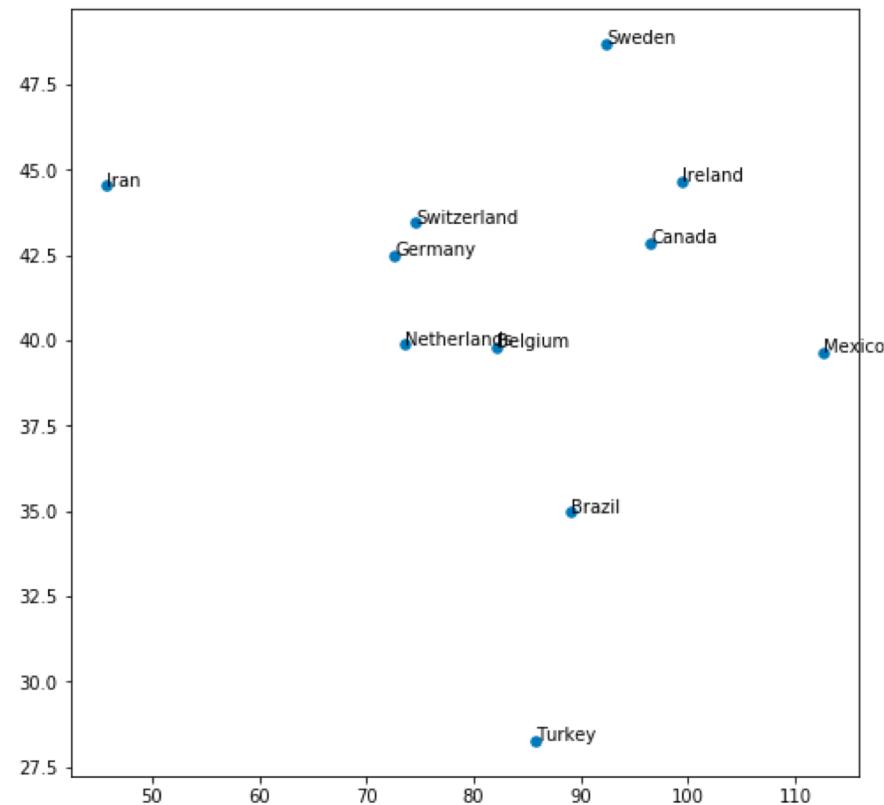
- Passo 1: vamos escolher a quantidade de clusters
- Passo 2: vamos ajustar o modelo K-Means ($K=9$) para os dados dos países
- Passo 3: como o espaço de variáveis possui 8 dimensões ($n_features=8$) precisamos de reduzir o espaço do problema. Utilizaremos a Análise de Componentes Principais (PCA)
- Passo 4: Visualizando resultados

Aplicação prática – Países similares COVID-19

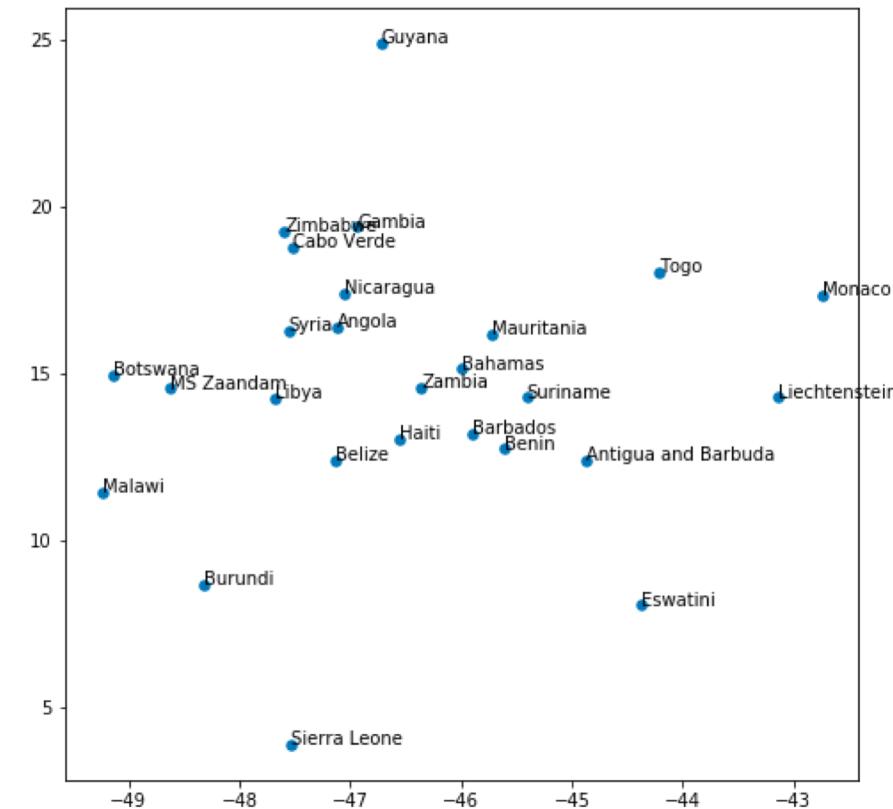
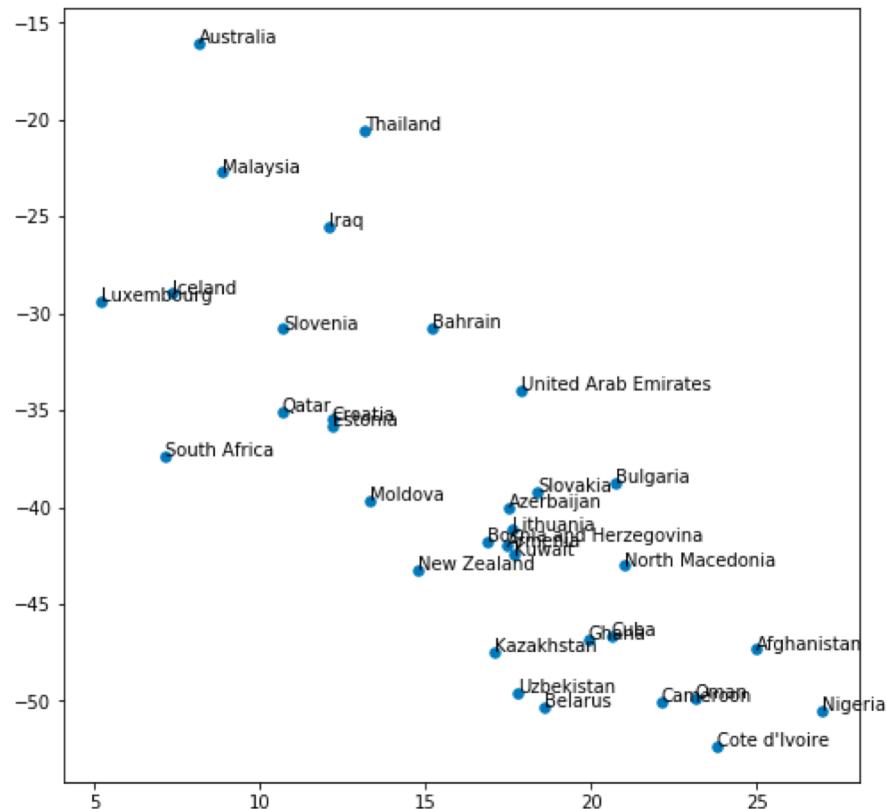


Os resultados apresentados, apesar de baseados em dados reais, tem finalidade educacional.

Aplicação prática – Países similares COVID-19



Aplicação prática – Países similares COVID-19



Referências

T. Hastie, R. Tibshirani, and J. Friedman. , The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, 2009

Tou J. T. , Advances in Information Systems Science VIII. Plenum Press New York Inc., New York, NY, USA, (1981)

<https://github.com/CSSEGISandData/COVID-19>

https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_assumptions.html

Contatos



luiz.andrade@tevec.com.br



+55 11 97163 2619



<https://www.linkedin.com/in/luiz-augusto-andrade/>

