
INSIGHT
Data Science Laboratory
Federal University of Ceará

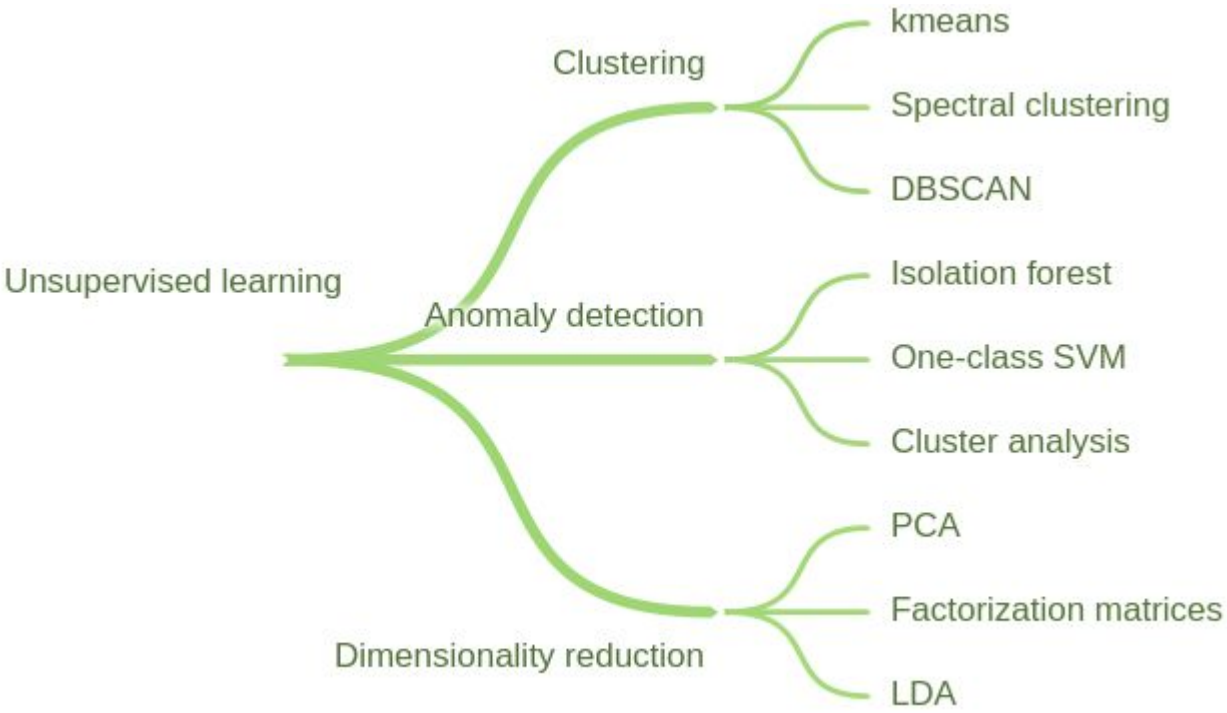
AGENDA

1. Aprendizado não supervisionado
2. Clusterização
3. K-means
4. Clusterização hierárquica
5. DBSCAN
6. Outros algoritmos
7. Referências
8. Agradecimentos

1. Aprendizado não supervisionado

Como aprender sobre dados sem rótulos?

Aprendizado não supervisionado



Mapa completo: https://storage.ning.com/topology/rest/1.0/file/get/135091853?profile=original

Aprendizado não supervisionado

No aprendizado **supervisionado**, os dados de treinamento **possuem rótulos**.

Exemplo:

- Classificação:
  [0.50, 0.78, 0.32, 0.89, 0.41] [“Bom”]

- Regressão:
  [0.34, 0.76, 0.48, 0.12, 0.43] [257]

Em muitas situações reais temos que lidar com dados **não supervisionados**, ou seja que **não possuem rótulos**

Aprendizado não supervisionado

Por que os dados não possuem rótulos?

- Rotular um grande conjunto de dados pode custar muito **tempo**, **esforço** e **dinheiro**
- Em muitas situações podemos querer descobrir as **similaridades** ou **diferenças** entre os padrões existentes nos dados.

Aprendizado não supervisionado

Exemplos:

- Seguro: identificar grupo de clientes que acionam sinistros com alta frequência;
- Classificação de documentos;
- Planejamento urbano: identificar grupos de casas conforme valor, tipo e localização;
- Organizar produtos em lojas;
- Detecção de fraudes.

2.
Clusterização

Criando grupos de dados

Clusterização

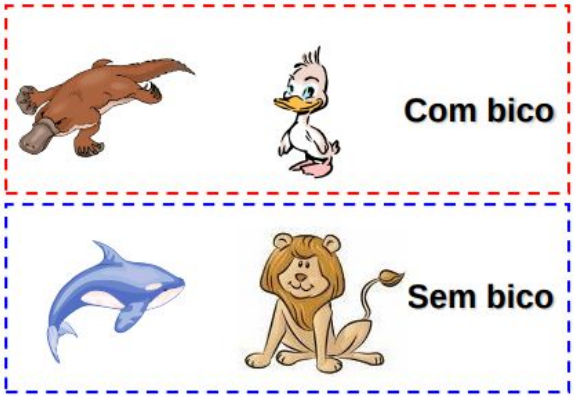
Clusterização é o **agrupamento** em conjuntos de dados, utilizando **similaridade** baseadas nas **características**.

Exemplo, como separar esse conjunto de animais?

Clusterização

Clusterização é o **agrupamento** em conjuntos de dados, utilizando **similaridade** baseadas nas **características**.

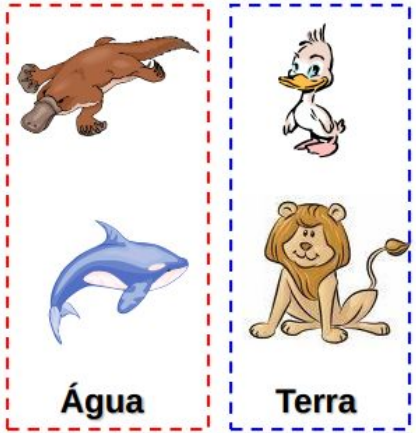
Exemplo, como separar esse conjunto de animais?



Clusterização

Clusterização é o **agrupamento** em conjuntos de dados, utilizando **similaridade** baseadas nas **características**.

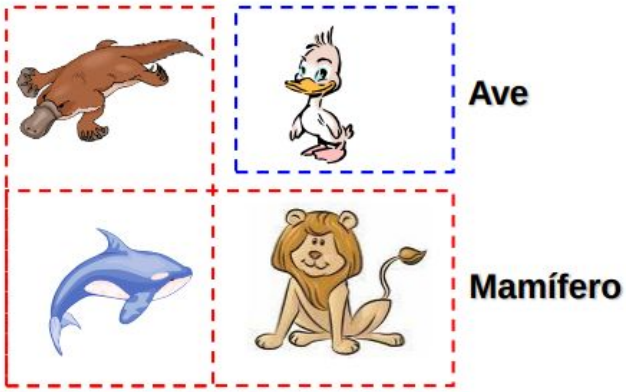
Exemplo, como separar esse conjunto de animais?



Clusterização

Clusterização é o **agrupamento** em conjuntos de dados, utilizando **similaridade** baseadas nas **características**.

Exemplo, como separar esse conjunto de animais?



3.
K-means

clusterização em k partições

ETAPAS PRINCIPAIS

1

2

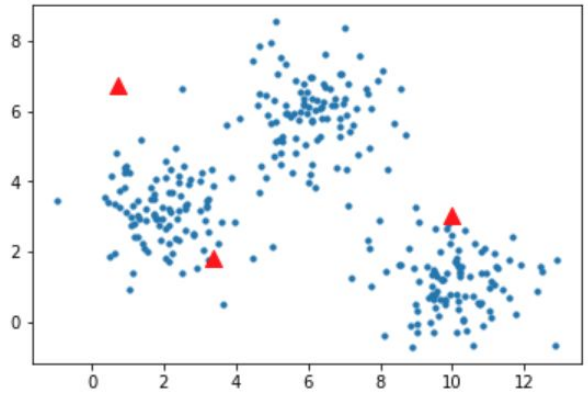
3

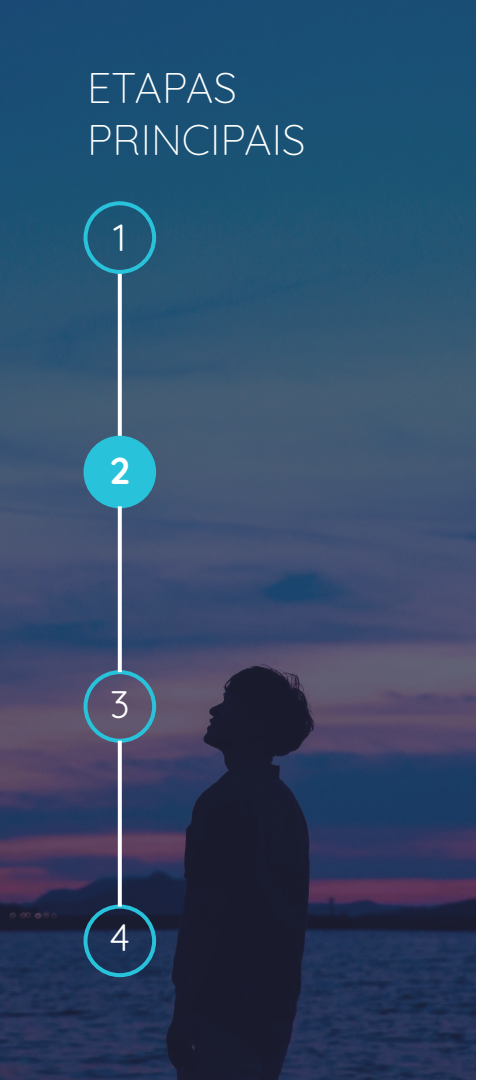
4

1. INICIALIZAÇÃO

A primeira etapa consiste em escolher randomicamente K pontos para representar os centróides iniciais.

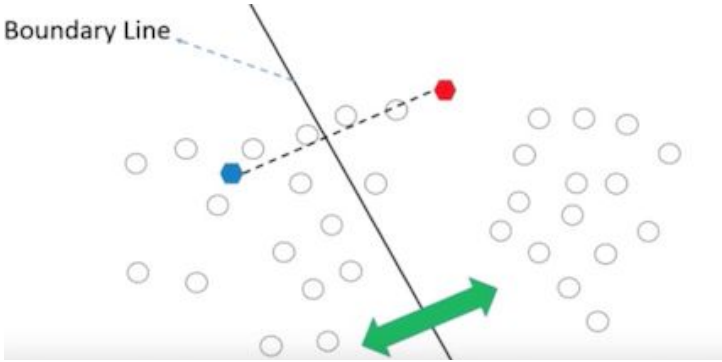
Uma boa maneira para inicializar os centróides, é utilizar as próprias amostras para criar pontos próximos ao conjunto de dados e esparsos entre si.





2. ATRIBUIÇÃO AOS CLUSTERS

Na segunda etapa, cada dado será atribuído a um cluster, que será o centróide mais próximo de acordo com uma função de distância.



ETAPAS PRINCIPAIS

1

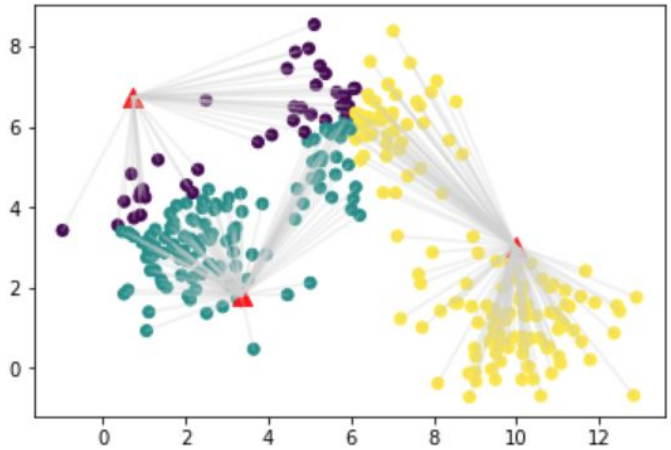
2

3

4

3. ATUALIZAR OS CENTRÓIDES

Após a atribuição dos dados aos clusters, a etapa de atualização consiste em calcular novos centróides. O novo valor de cada centróide será a média de todos os dados pertencentes ao cluster.



ETAPAS PRINCIPAIS

1

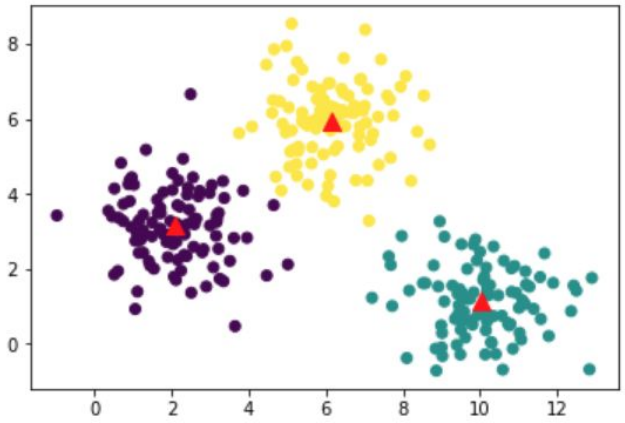
2

3

4

4. FINALIZAÇÃO

O algoritmo repete os passos 2 e 3 até não haver mais mudança na atualização dos centróides.



Exemplo

Fonte: http://shabal.in/visuals/kmeans/6.html

Número de clusters

Como escolher o valor de K?

A princípio o algoritmo do K-means parece ser um pouco ingênuo, pois ele divide os dados em K clusters, mesmo que não existam K clusters. Alguns métodos podem ajudar na escolha do valor de K.

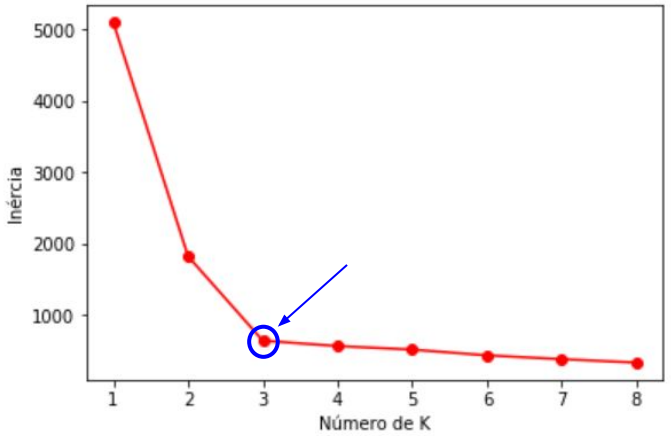
Exemplo:

- Método do cotovelo
- Dendrograma

Método do cotovelo

Executar o algoritmo K-means para um intervalo de valores de K ($1 \leq K \leq 20$, por exemplo), para cada valor de K é calculado a soma dos quadrados das distâncias dos dados para o centróide do cluster.

A ideia é analisar a variação intra-cluster para diferentes valores de K, buscando o número ideal da quantidade de clusters.



Colocar a mão na massa!

Regras:

- Codificação Individual
- Pode pesquisar na internet a vontade

Pontuação:

- Inicializar os centróides ---------- (1 ponto)
- Função de distância ---------- (1 ponto)
- Calcular o centróide mais próximo ---------- (1 ponto)
- Centróide mais próximo para todos os dados -- (1 ponto)
- Métrica de avaliação ---------- (1 ponto)
- Atualizar os clusters ---------- (1 ponto)
- Algoritmo completo ---------- (2 pontos)
- Método do cotovelo ---------- (2 pontos)

https://github.com/InsightLab/data-science-cookbook/blob/master/2019/09-clustering/Notebook_KMeans_Assignment.ipynb

K-means
Complexidade

Complexidade de espaço: o espaço necessário para armazenar os dados e os centróides.

Complexidade de espaço = O((m+k)*n), no qual **m** é a quantidade de dados, **k** é o número de centróides e **n** é o número de atributos.

Complexidade de tempo: é um problema NP-difícil, porém executando um número fixo de iterações, o algoritmo padrão apenas faz uma aproximação do ótimo local.

Complexidade de tempo = para um número fixo de **t** iterações, O(t*k*m*n), no qual **m** é a quantidade de dados, **k** é o número de centróides e **n** é o número de atributos.

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

K-means
Vantagens e Desvantagens

Vantagens

1. Fácil de implementar;
2. Com grande número de atributos, o K-means é computacionalmente mais rápido que a clusterização hierárquica;
3. K-means pode produzir clusters mais concêntricos;
4. Uma amostra pode mudar de cluster, quando os centróides são recalculados.

Desvantagens

1. Inicialização dos centróides tem um grande impacto no resultado final;
2. Sensível a escala dos dados;
3. Todos os dados pertencem a um grupo;
4. É necessário definir o número de **k**.

3. Clusterização Hierárquica

Clusterização baseada em similaridade

Clusterização hierárquica

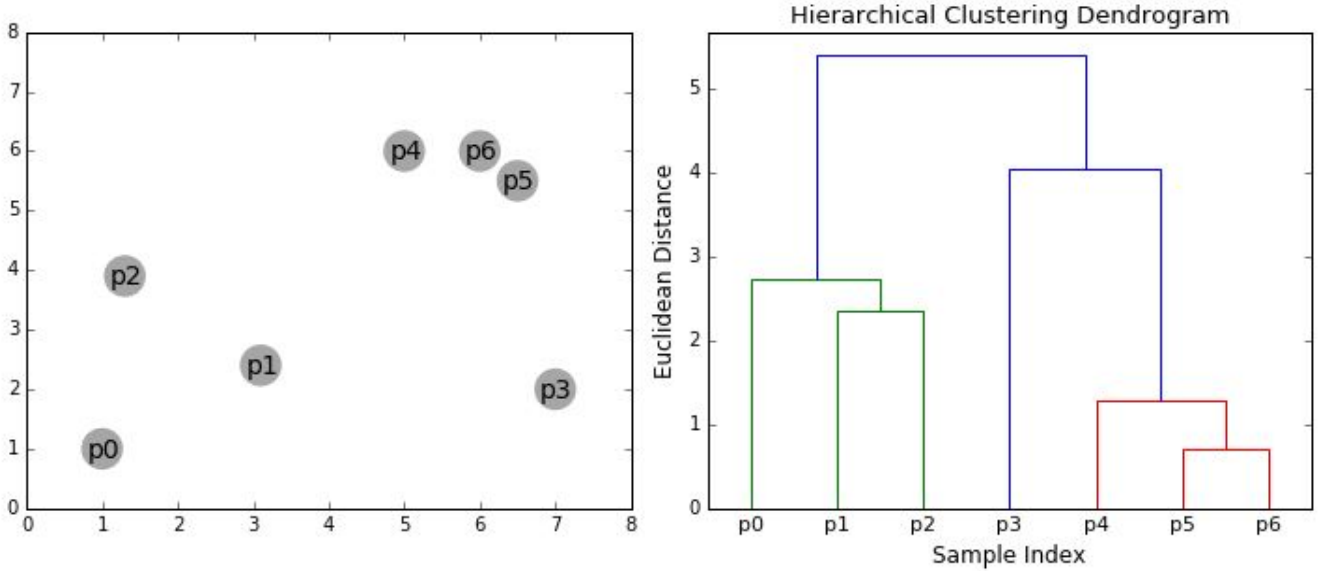
Aglomerativa (bottom-up):

Nesta técnica, inicialmente **cada dado** é considerado **um cluster** individual. Em cada iteração, os clusters **se juntam** de acordo com uma **métrica de similaridade** com outros clusters até que apenas **um ou K clusters** sejam formados.

Divisiva (up-bottom):

Nesta técnica, inicialmente **todos os dados** são considerados apenas **um cluster**. Em cada iteração, os dados diferentes **se separam** de acordo uma **métrica de similaridade** formando outros clusters até que cada dado permaneça em **um ou K clusters** sejam formados.

Clusterização hierárquica aglomerativa - Exemplo



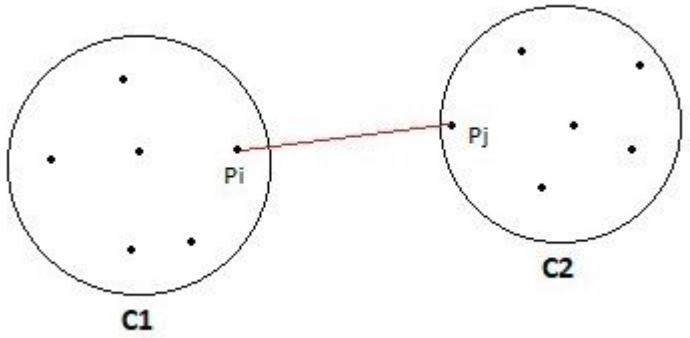
Fonte: https://dashee87.github.io/images/hierarch.gif

Clusterização hierárquica aglomerativa - Similaridade

- **MIN ou Single**: mínimo das distâncias entre todas as observações dos dois conjuntos.
- **MAX ou Complete**: máximo das distâncias entre todas as observações dos dois conjuntos
- **Average**: calcula a média das distâncias para a combinação em par de todos os dados
- **Ward**: similar ao average, mas utilizando a soma do quadrado das distâncias

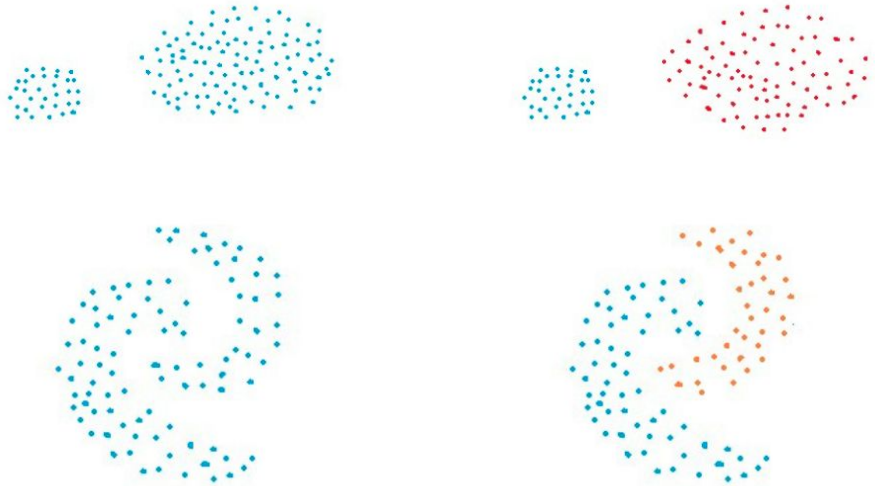
Similaridade
MIN ou Single

A semelhança entre dois clusters, vai ser a distância entre os dados mais próximos entre um cluster e outro.



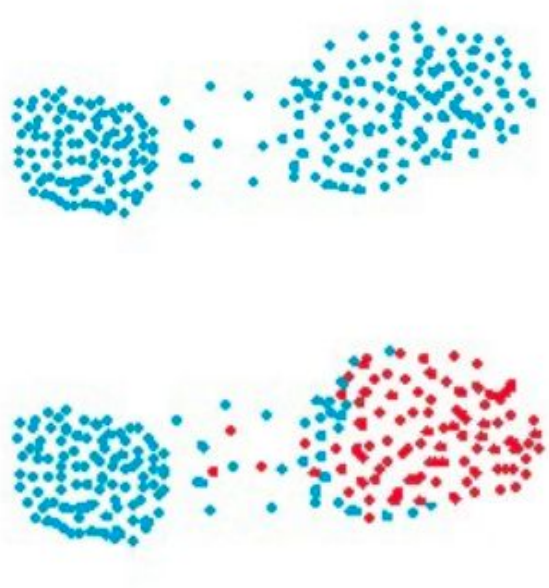
Similaridade
MIN ou Single

Vantagens: pode separar formas não elípticas, quando separadas por uma certa distância



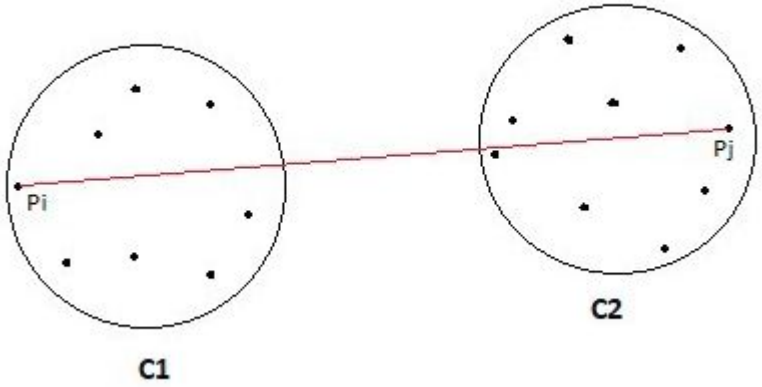
Similaridade
MIN ou Single

Contrapartida: pode não separar os clusters adequadamente se houver ruído entre os clusters



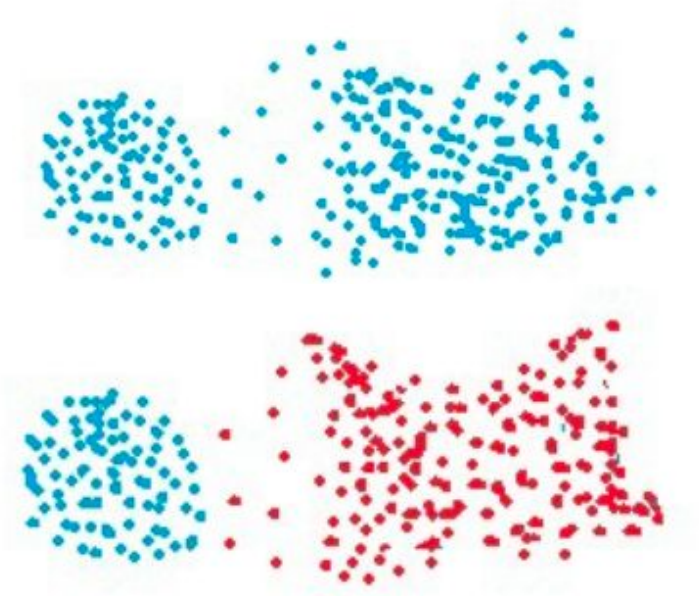
Similaridade
MAX ou Complete

A semelhança entre dois clusters, vai ser a distância entre os dados mais afastados entre um cluster e outro.



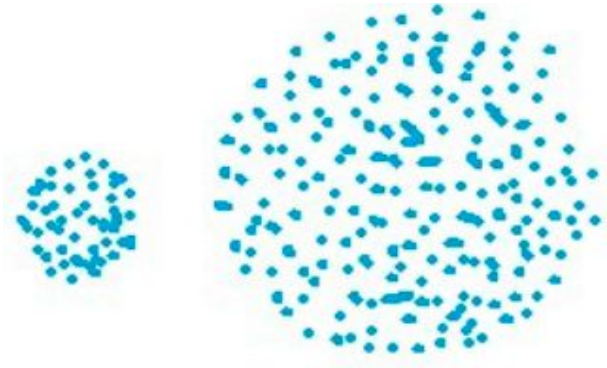
Similaridade
MAX ou Complete

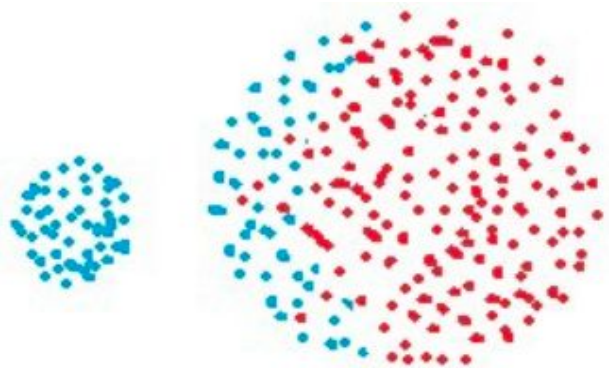
Vantagens: tem boa performance separando clusters mesmo com ruído entre os dados



Similaridade
MAX ou Complete

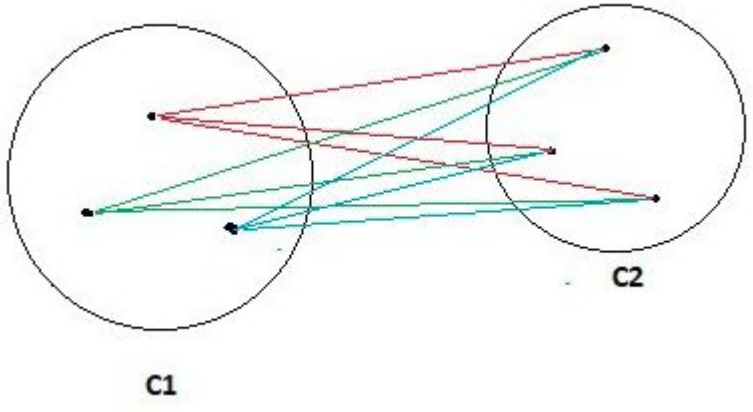
Contrapartida: tem tendência em quebrar grandes clusters; é enviesada para clusters globulares.





Similaridade Average

A semelhança entre dois clusters é a média das distâncias para a combinação em par de todos os dados



Similaridade Average

Vantagens: tem boa performance separando clusters mesmo com ruído entre os dados

Contrapartida: é enviesada para clusters globulares.

Similaridade Ward

A semelhança entre dois clusters é a média da soma das distâncias quadradas para a combinação em par de todos os dados

Vantagens: tem boa performance separando clusters mesmo com ruído entre os dados

Contrapartida: é enviesada para clusters globulares.

Clusterização Hierárquica
Complexidade

Complexidade de espaço: o espaço necessário para utilização do algoritmo é muito alto para grande grande quantidade de dados, pois é necessário armazenar a matriz de similaridade na memória RAM.

Complexidade de espaço = $O(m^2)$, no qual **m** é a quantidade de dados

Complexidade de tempo: são executadas **n** iterações e em cada iteração é necessário atualizar e restaurar a matriz de similaridade, logo a complexidade de tempo também é muito alta.

Complexidade de tempo = $O(n^3)$, no qual **n** é a quantidade de dados

Clusterização Hierárquica
Vantagens e desvantagens

Vantagens

1. Fácil de implementar;
2. Gera uma árvore de hierarquia, que é uma estrutura mais informativa.
3. Pode utilizar o dendrograma para decidir o número de clusters e tomar uma decisão de parada.

Desvantagens

1. Não existe um objetivo matemático para a clusterização;
2. Todos os métodos de ligação para calcular a similaridade têm suas próprias desvantagens;
3. Alta complexidade de espaço e tempo;
4. O algoritmo não pode ser utilizado para uma enorme quantidade de dados;
5. Sensível a outliers.

4.
DBSCAN

Clusterização baseada em densidade

DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) é um algoritmo de clusterização comumente utilizado em mineração de dados e aprendizado de máquina que utiliza uma abordagem baseada em densidade.

O DBSCAN agrupa dados que possuem outros dados próximos, baseado em uma **função de distância** e um número **mínimo de dados**. Em regiões com baixa densidade de dados, estes são considerado outliers.

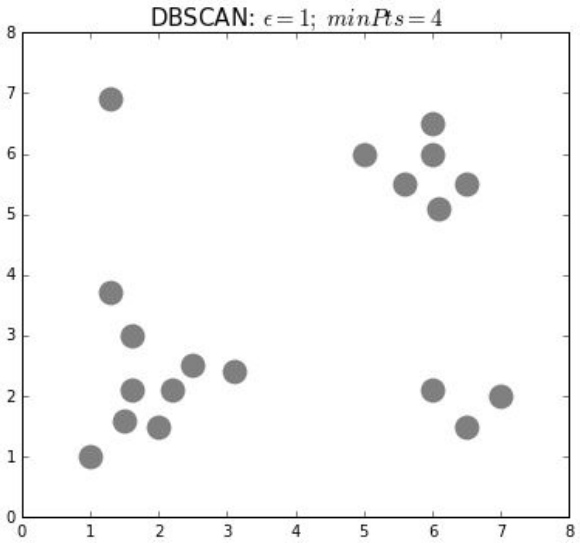
DBSCAN

Parâmetros:

eps: a distância mínima entre dois dados. Isso significa que, se a distância entre dois dados for menor ou igual ao valor de eps, os dados são considerados vizinhos.

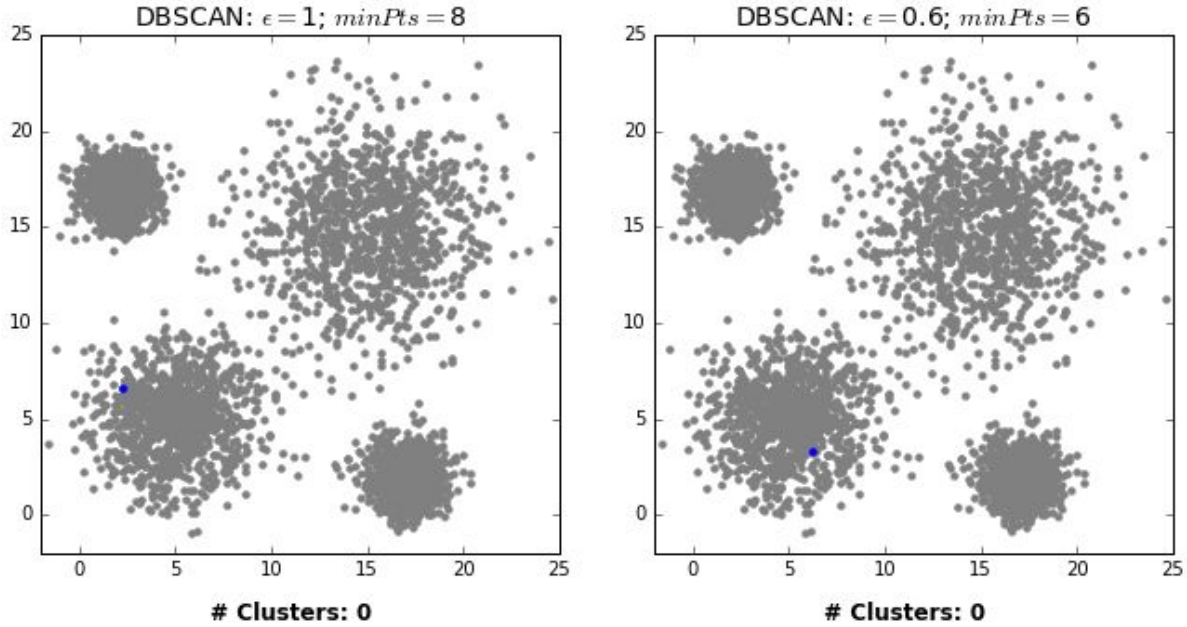
minPoints: o número mínimo de dados para formar uma região densa. Por exemplo, se for definida o número mínimo de dados como 3, será necessário pelo menos 3 dados vizinhos para formar uma região densa.

DBSCAN - Algoritmo e exemplo



Fonte: https://dashee87.github.io/images/DBSCAN_tutorial.gif

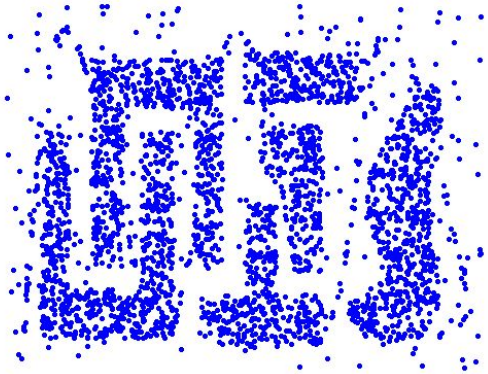
DBSCAN - Exemplo



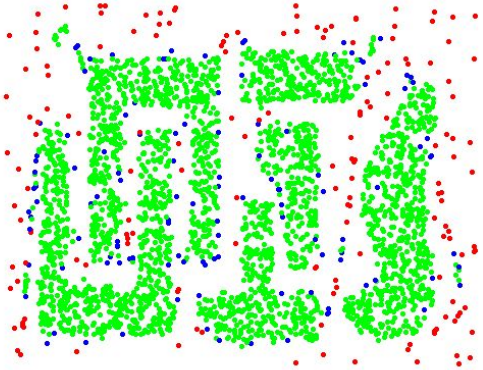
Fonte: https://dashee87.github.io/images/DBSCAN_search.gif

DBSCAN - Exemplo

Quando o algoritmo não funciona bem



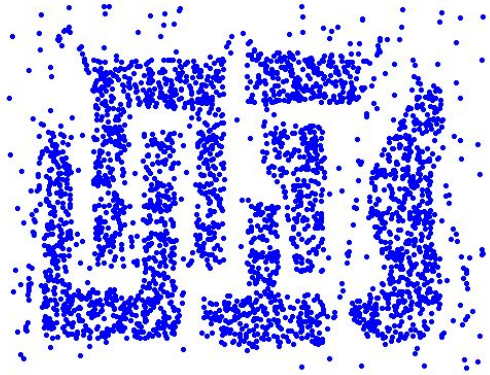
Dados originais



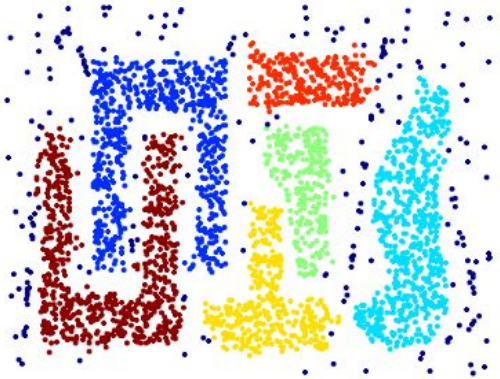
Dados clusterizados

DBSCAN - Exemplo

Quando o algoritmo funciona bem



Dados originais



Dados clusterizados

DBSCAN
Vantagens e desvantagens

Vantagens

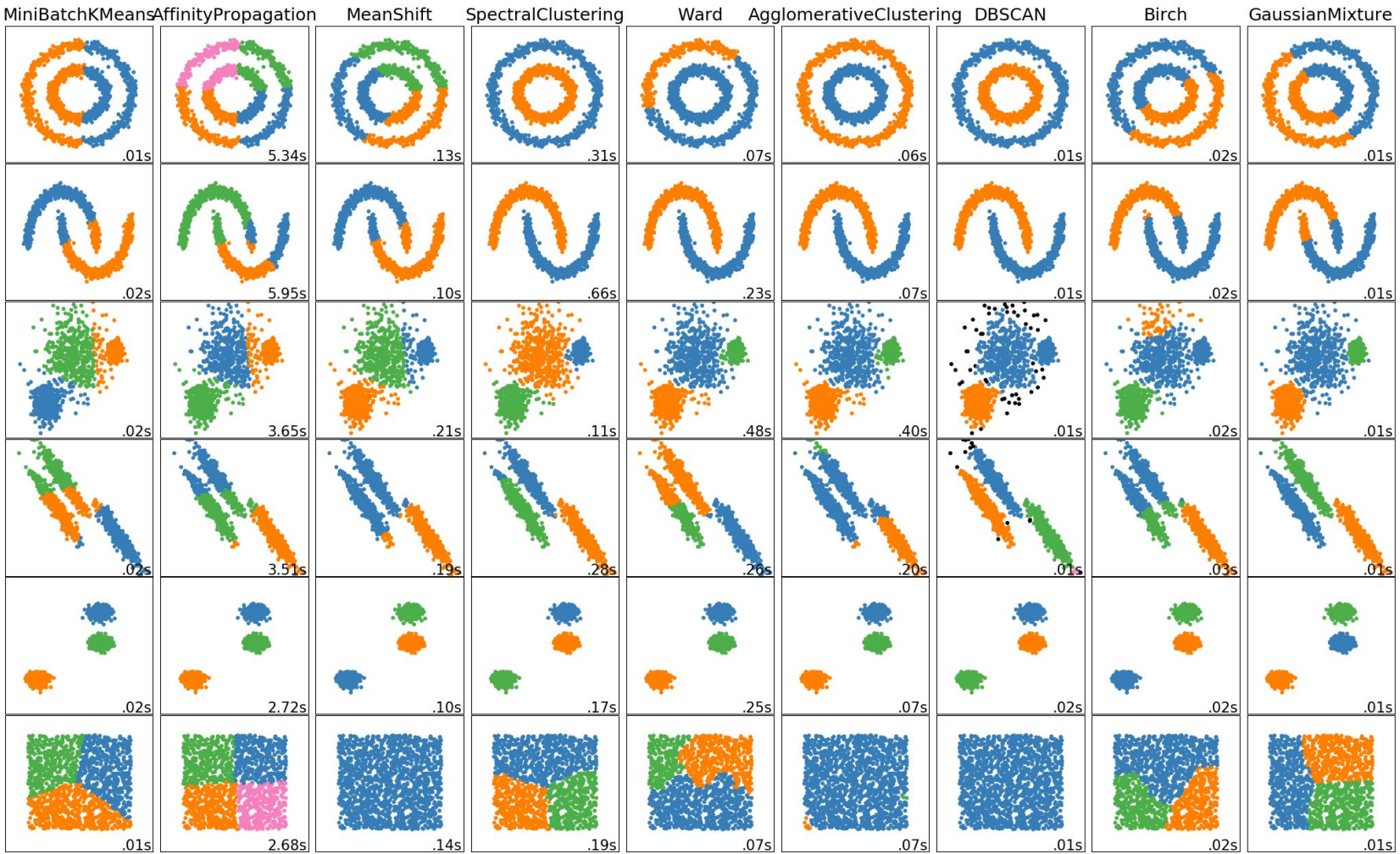
1. Cria grupos com formatos arbitrários;
2. Identificar outliers;
3. Uso de qualquer medida de similaridade.
4. Não é necessário especificar a quantidade de clusters;

Desvantagens

1. Alta complexidade computacional;
2. Sensível aos parâmetros de entrada (eps e minPoints);
3. Não funciona bem se os grupos têm densidades muito diferentes;
4. Não gera bons resultados para conjuntos multidimensionais.

4.
Outros algoritmos

Visão geral dos algoritmos de clusterização



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

REFERÊNCIAS

- Doutorando Lucas Cambuim - UFPE
  http://www.cin.ufpe.br/~lfsc/cursos/introducaoainteligenciaartificial/IA-Aula12-Clusterizacao.pdf
- Mestrando Felipe Zschornack R. Saraiva - UFC
  https://docs.google.com/presentation/d/10SnrYrevdnGF2JoYkIes2oBFa-Ttz7cZkgq_czH3AzI/edit#slide=id.g1726f05f0e_0_66
- Professor Edirlei Soares de Lima - UERJ
  http://edirlei.3dgb.com.br/aulas/ia_2011_2/IA_Aula_18_Aprendizado_Nao_Supervisionado.pdf
- Scikit-Learn - Machine Learning in Python
  https://scikit-learn.org/stable/modules/clustering.html
- GitHub David Sheehan - Cientista de dados
  https://dashee87.github.io/data%20science/general/Clustering-with-Scikit-with-GIFs/
- Medium - Towards Data Science
  https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec

OBRIGADO!

Dúvidas?

Você pode me encontrar em

- carlos@insightlab.ufc.br
- Telegram: @CarlosJun