UNIFACS
Feira de Santana

Noberto Maciel
noberto.maciel@ulife.com.br

Sistemas de Controle e Inteligência Artificial

AULA 09
Prática

Aula de hoje:

1. K-Medoids
2. Algoritmo Hierárquico aglomerativo
3. DBScan (baseado em densidade)
4. Métodos de validação de agrupamentos

K-Medoids

O que é:

Algoritmo que usa uma técnica particional (partitional clustering) para encontrar um número K de clusters pré-determinado. Diferentemente do K-Means, o K-Medoids utiliza elementos reais (pertencentes ao conjunto de dados) como centros de cluster.

Objetivo:

Encontrar padrões ocultos nos dados, tentando gerar grupos com características similares.

Algoritmo 1: K-medoids Básico

```
1: Obter K = inteiro;
2: Escolher K pontos como os medoids iniciais, randomicamente
3: Enquanto (medoids mudarem de posição)
4:    Calcular a distância dos novos medoids para todos os pontos
5:    Formar K clusters atribuindo cada ponto ao medoide mais próximo
6:    Recalcular medoides de cluster (ponto central, menor dist. média)
7: Fim
```

Métricas de distância:

São formas de calcular a distância entre os medoids e os elementos do conjunto de dados.

- Distância Euclidiana
- Distância Geodésica
- Distância de Manhattan

Vantagens:

Fácil de implementar;

Rápido;

Trabalha bem com grandes conjuntos de dados;

Menos sensibilidade a outliers.

Desvantagens:

A qualidade dos agrupamentos gerados está diretamente ligada à inicialização randômica dos medoids;

Não trabalha bem com alta dimensionalidade;

Apresenta distorções em conjuntos de dados com agrupamentos desiguais.

Aplicações:

1) Segmentação de Clientes (Marketing e Vendas): agrupar clientes com base em comportamento de compra, permitindo campanhas mais personalizadas; identificar perfis de clientes semelhantes para estratégias de fidelização.

2) Análise de Imagens e Processamento de Vídeo: compressão de imagens, reduzindo cores ao agrupar pixels similares; segmentação de imagens, separando objetos de fundo.

3) Agrupamento de Dados em Redes e Telecomunicações: identificar padrões de uso da rede para otimizar tráfego e recursos; detectar anomalias no tráfego, ajudando na segurança cibernética.

4) Bioinformática e Medicina: agrupamento de genes com funções semelhantes; identificação de padrões em exames médicos, como na detecção de tumores em imagens.

5) Geolocalização e Planejamento Urbano: clustering de locais para planejamento de infraestrutura (ex.: estações de transporte); segmentação de áreas com base em dados socioeconômicos.

6) Análise Financeira e Detecção de Fraudes: agrupar transações para identificar padrões de comportamento financeiro; detectar atividades suspeitas em cartões de crédito e contas bancárias.

7) Recomendação de Conteúdo: identificar grupos de usuários com gostos similares para recomendações personalizadas (ex.: Netflix, Spotify);

8) Manutenção Preditiva na Indústria: agrupar sensores e máquinas com comportamentos semelhantes para prever falhas antes que ocorram.

Algoritmo Hierárquico

Algoritmo hierárquico

O que é:

Algoritmo que usa uma técnica hierárquica para encontrar padrões de agrupamento nos dados. Podem ser aglomerativos (bottom-up) ou divisivos (top-down).

Objetivo:

Encontrar padrões ocultos nos dados, tentando gerar grupos com características similares/dissimilares.

Algoritmo hierárquico

Algoritmo 1: Hierárquico Básico (aglomerativo)

01: **Entrada:** N elementos (dados) a serem agrupados;
02: **Inicializar** cada elemento como um cluster individual (N clusters);
03: **Construir** matriz de distâncias D com distâncias de todos clusters;
04: **Enquanto** o número de clusters for maior que 1, repetir **Faça:**
05: **Encontrar** dois clusters mais próximos (Ci, Cj) em D;
06: **Unir** os clusters Ci e Cj em um único cluster Cij;
07: **Atualizar** matriz D para refletir as novas distâncias

Dependendo do método de ligação:

i. *Ligação única*: distância mínima entre os elementos dos clusters
ii. *Ligação completa*: distância máxima entre os elementos dos clusters
iii. *Ligação média*: média das distâncias entre os elementos dos clusters

08: **Remover** entradas antigas dos clusters Ci e Cj de D;
09: **Adicionar** o novo cluster Cij à matriz de distâncias;
10: **Retornar** a estrutura hierárquica de agrupamento (dendrograma)

Algoritmo hierárquico

- **Single Linkage**

  $D(c_1,c_2) = \min D(x_1,x_2)$

  Minimum distance or distance between closest elements in clusters

- **Complete Linkage**

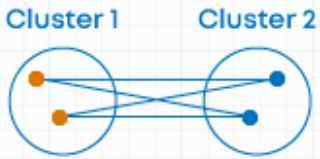
  $D(c_1,c_2) = \max D(x_1,x_2)$

  Maximum distance between elements in clusters

- **Average Linkage**

  $D(c_1,c_2) = \frac{1}{|c_1|}\frac{1}{|c_2|}\Sigma\Sigma D(x_1,x_2)$

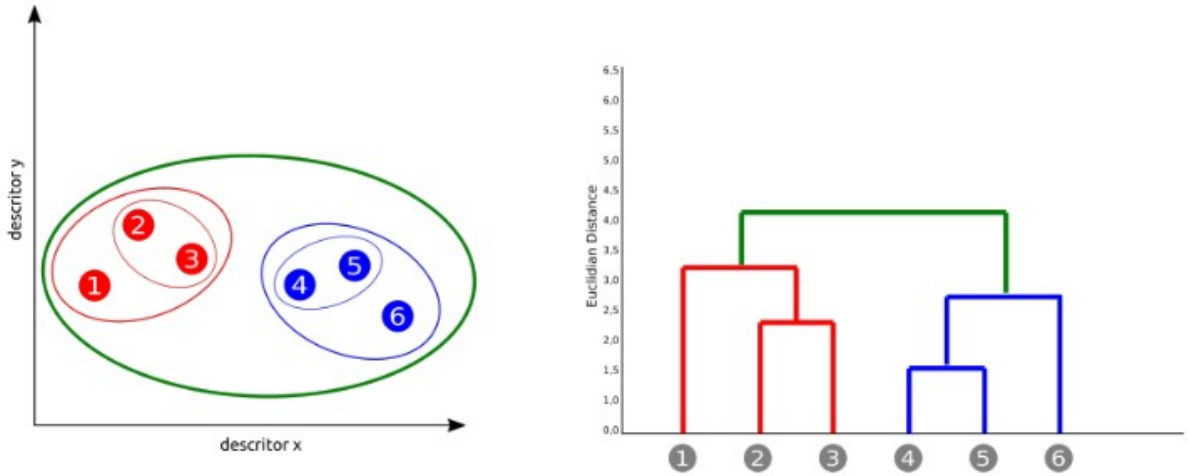
  Average of the distances of all pairs



- **Centroid Method**

  Combining clusters with minimum distance between the centroids of the two clusters

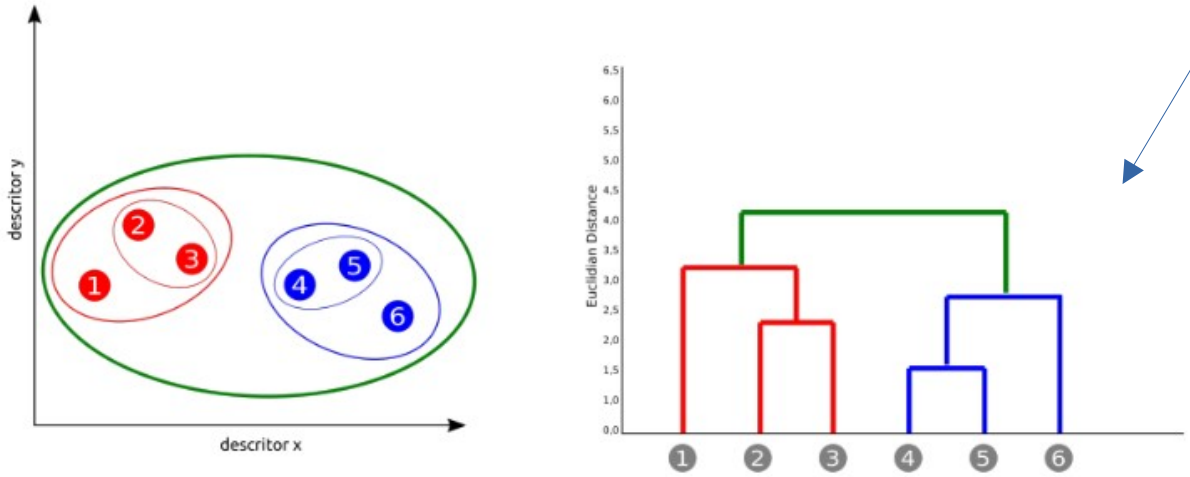
Algoritmo hierárquico



(a) (b)

Figura 2.1: Representação do resultado de um algoritmo hierárquico. a) representação de conjuntos. b) representação por dendrograma

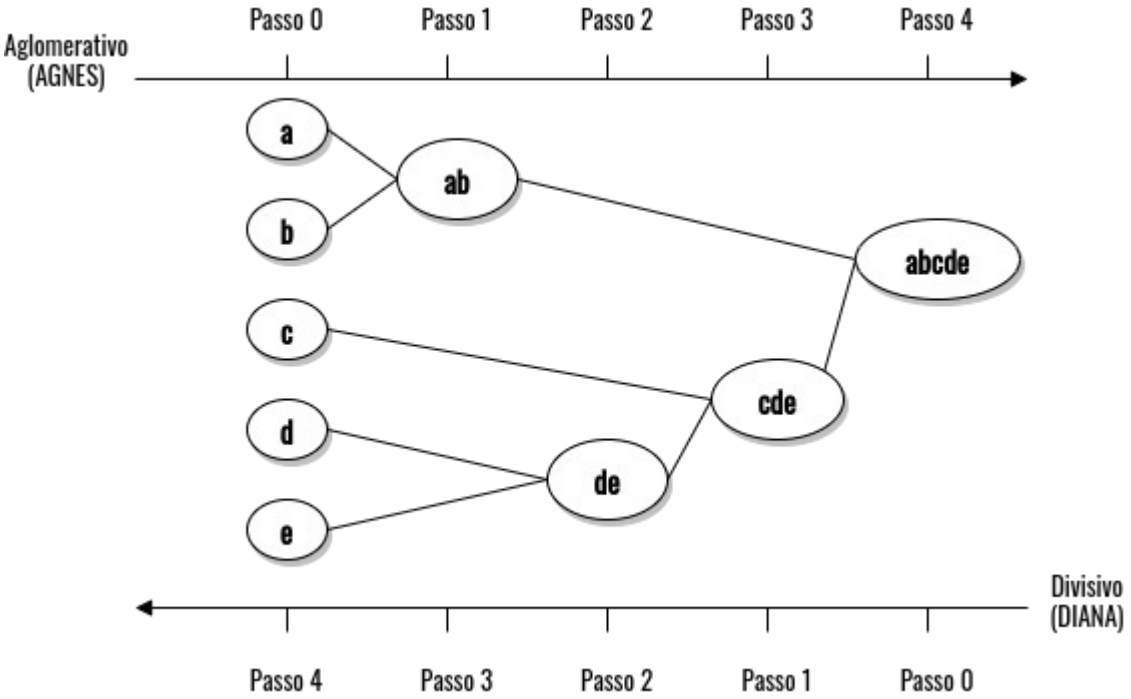
Dendrograma



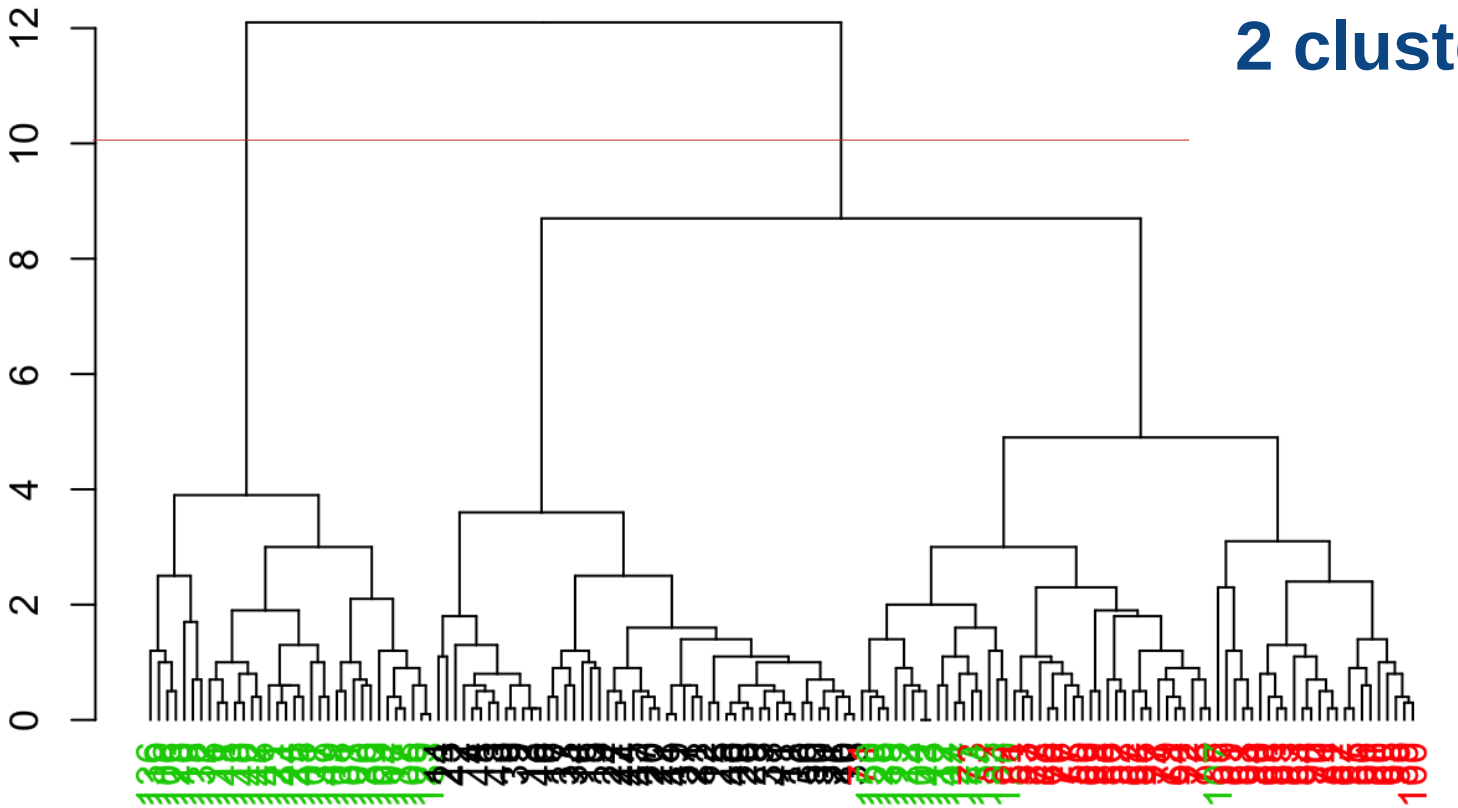
(a) (b)

Figura 2.1: Representação do resultado de um algoritmo hierárquico. a) representação de conjuntos. b) representação por dendrograma

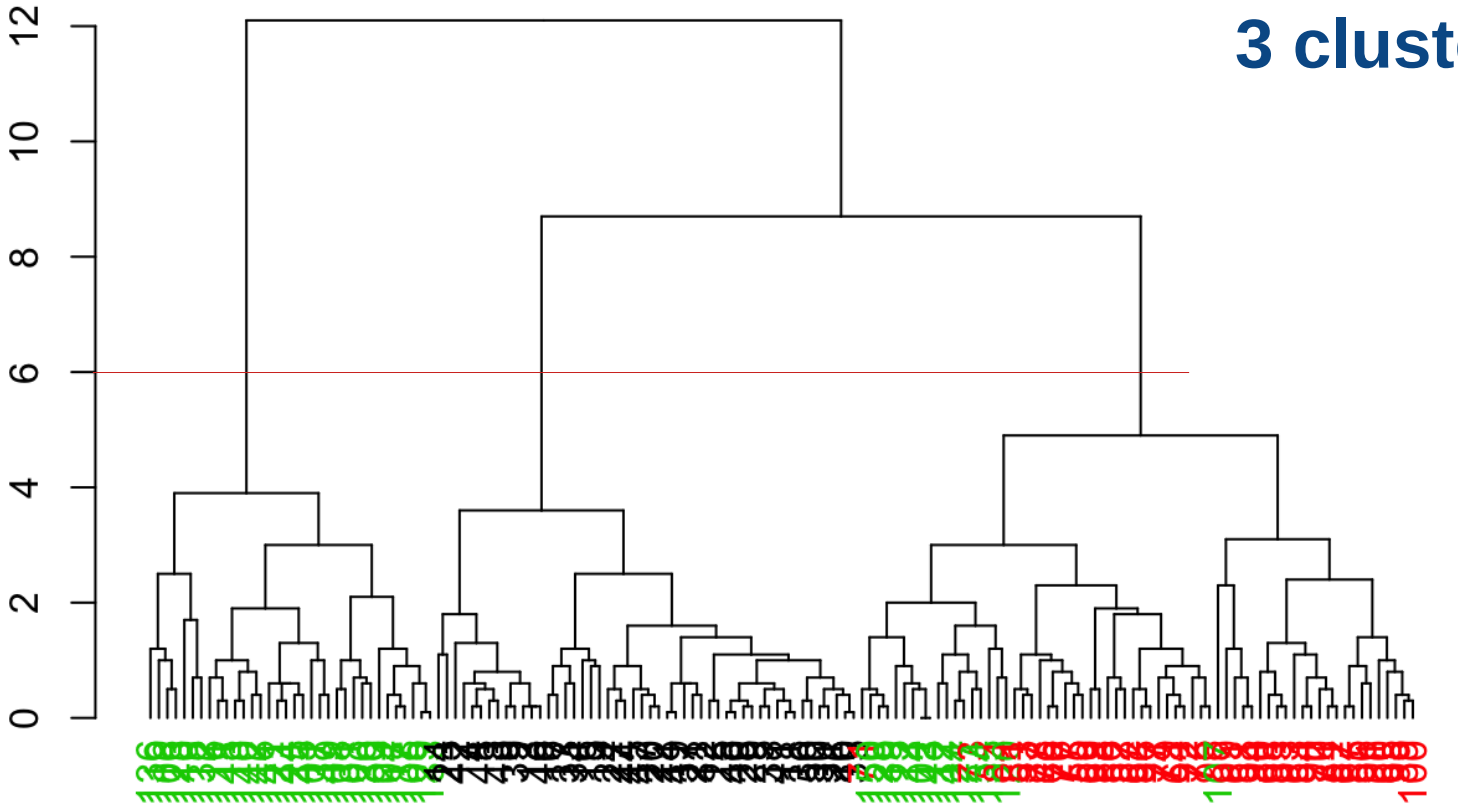
Algoritmo hierárquico



Algoritmo hierárquico



Algoritmo hierárquico



Algoritmo hierárquico

Vantagens:

Fácil de implementar;

Gera dendogramas que permitem excelente visualização e interpretação dos clusters formados;

Não necessita definir o número de clusters (K) na inicialização;

Não depende da inicialização randômica;

Menos sensibilidade a outliers;

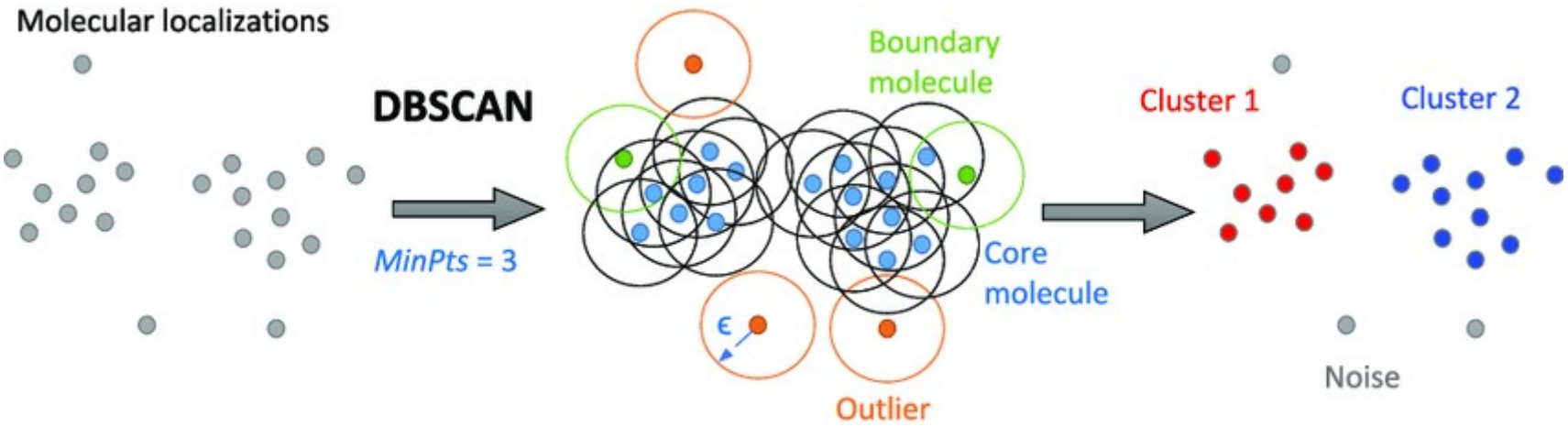
Desvantagens:

Não escala bem para grandes conjuntos de dados, prejudicando a visualização;

Resultados são sensíveis à escolha da métrica de distância e do método de linkage utilizados;

É computacionalmente mais intensivo do que métodos como K-means para grandes bancos de dados.

DBScan


Molecular localizations
DBSCAN
MinPts = 3
Boundary molecule
Core molecule
ϵ
Outlier
Cluster 1
Cluster 2
Noise

Validação dos agrupamentos

Qual o número ideal de clusters?

Métodos validação de esquemas de agrupamentos:

- Dunn Index;
- Davies Bouldin (DB Index);
- Soma dos Erros Quadráticos (SSE);
- Silhouette
- DTRS (Decision Theoretic Rough Set)

...

Validação dos agrupamentos

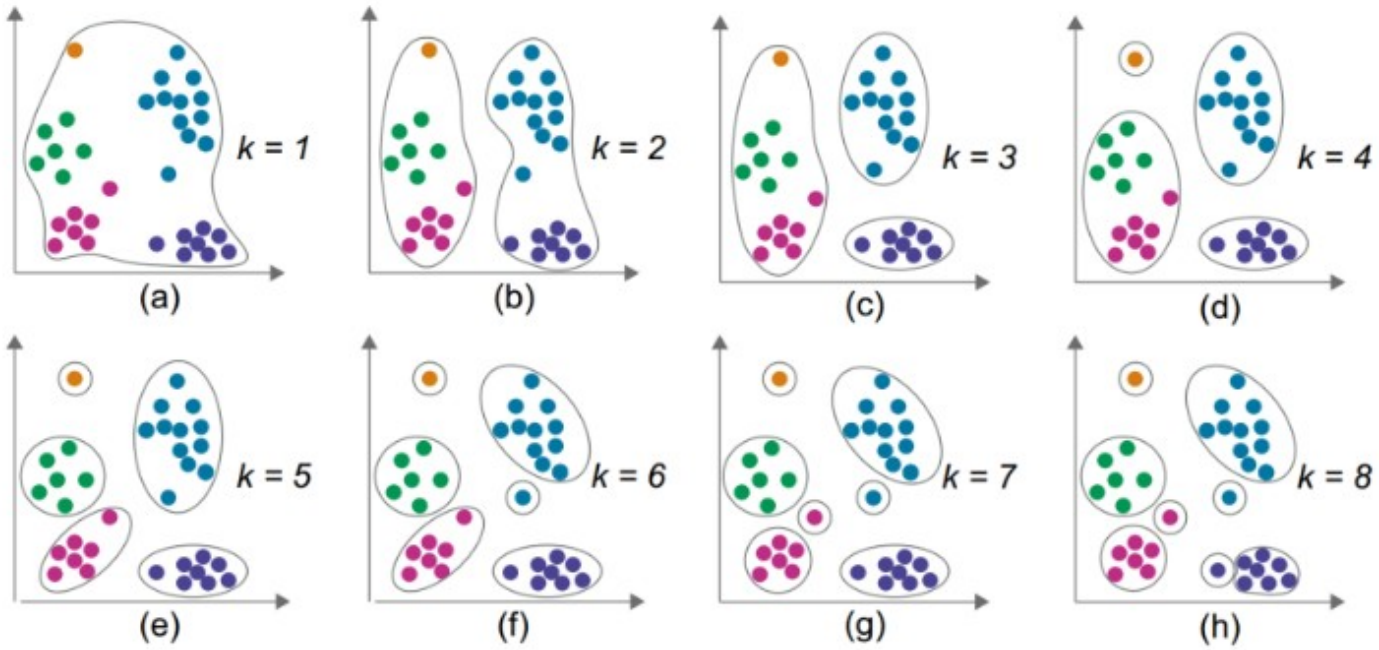
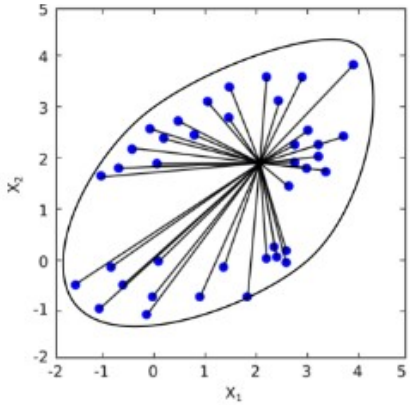
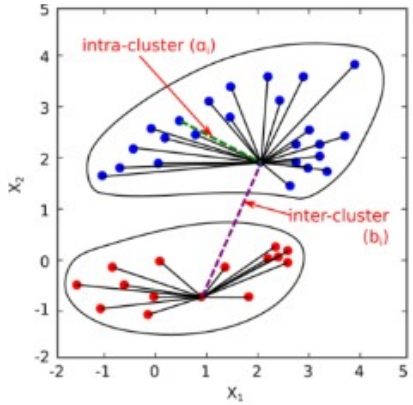


Figura 1.6: Problema da definição do número de *clusters* (Figuerêdo e Calumby, 2022)

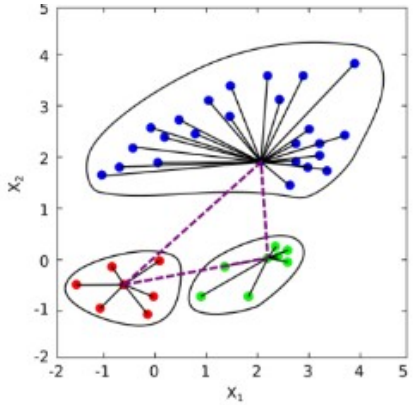
Validação dos agrupamentos



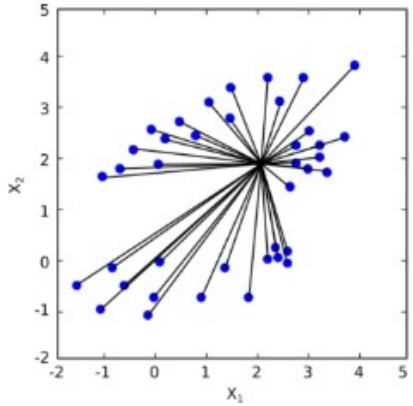
(a) $k = 1$



(b) $k = 2$



(c) $k = 3$



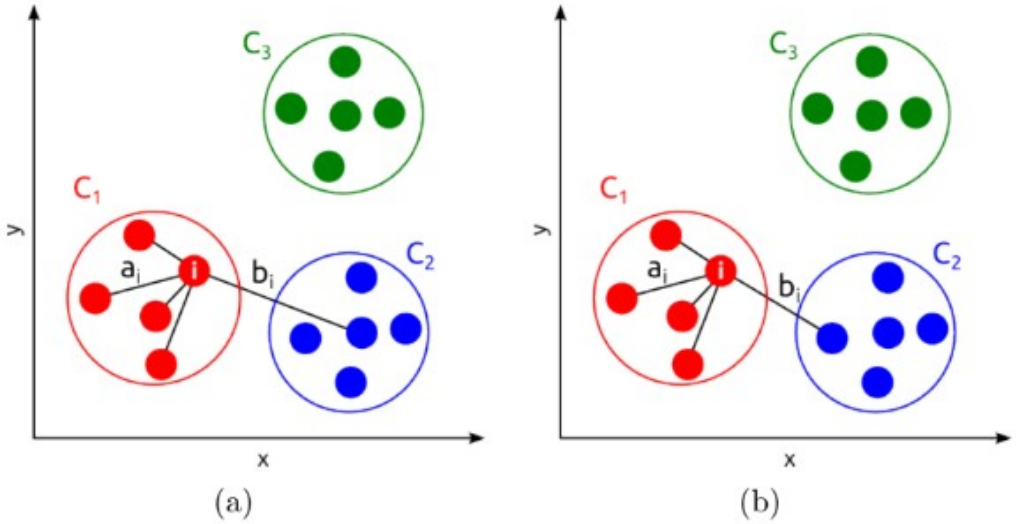
(d) $k = U$

Silhouette

Método matemático para encontrar o melhor esquema de clusters (melhor quantidade e qualidade de agrupamentos)

$$S_i = \begin{cases} 1 - a_i/b_i, & \text{se } a_i < b_i \\ 0, & \text{se } a_i = b_i \\ b_i/a_i - 1, & \text{se } a_i > b_i \end{cases}$$

$$S_i = \frac{(b_i - a_i)}{max(a_i, b_i)}$$



Validação dos agrupamentos

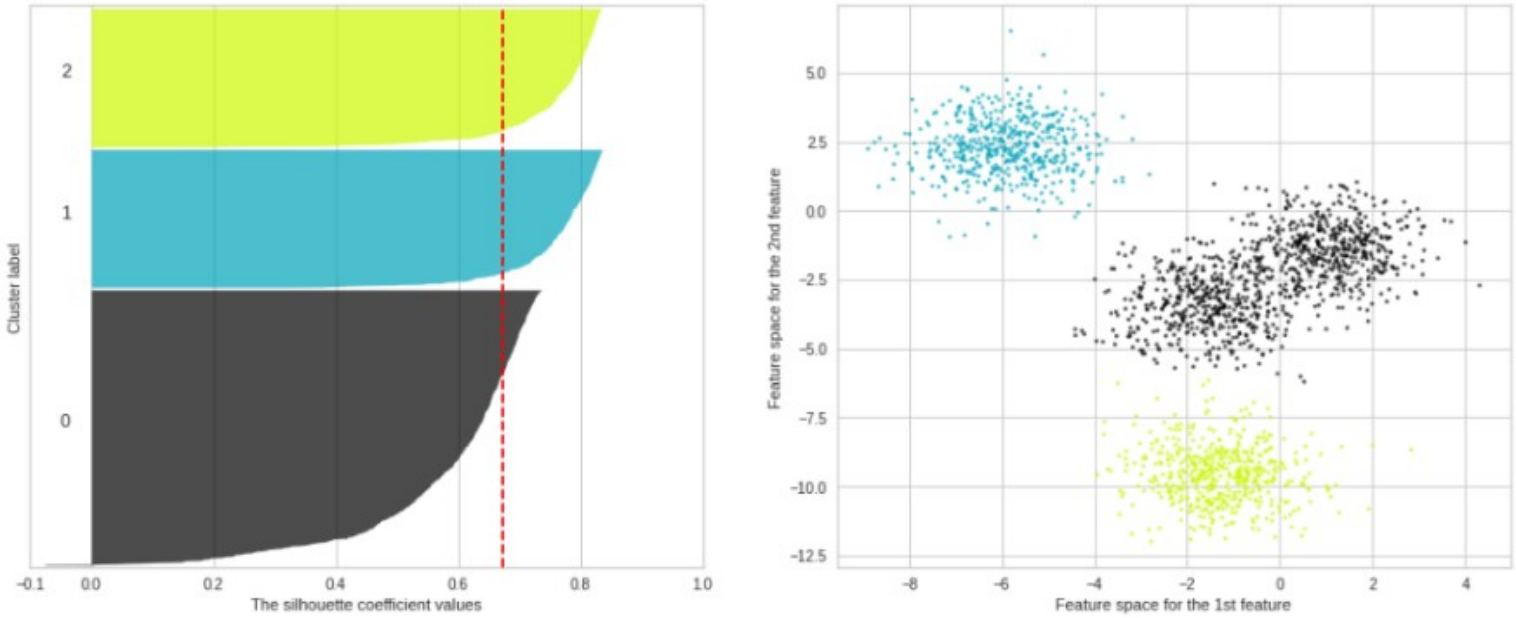
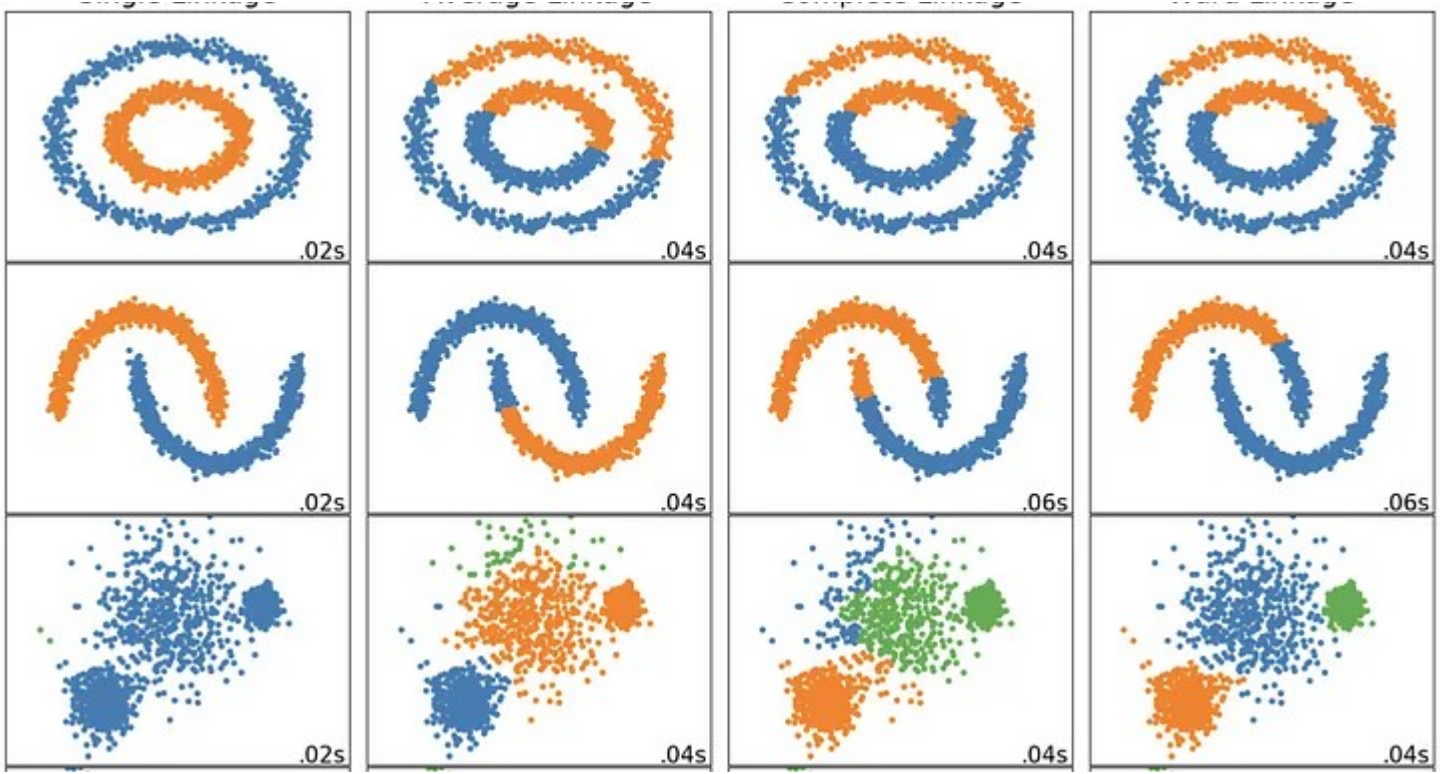
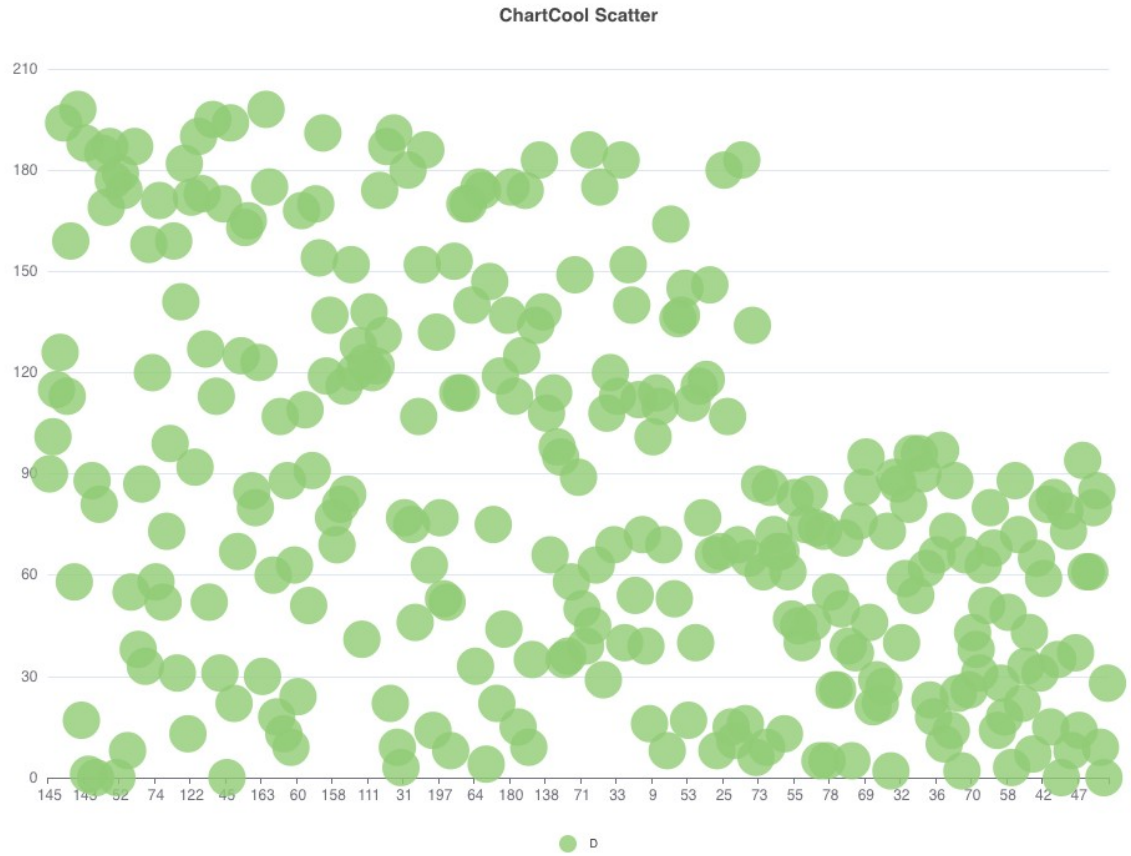


Figura 2.5: *Silhouette*: um valor próximo de 1 indica um bom esquema de agrupamento, ao passo que um valor -1 indica um esquema ruim. Gráficos obtidos através da biblioteca *SciKit Learn* (Buitinck et al., 2013) com dados fictícios. A linha pontilhada em vermelho é o *score* médio global.

Validação dos agrupamentos



Validação dos agrupamentos

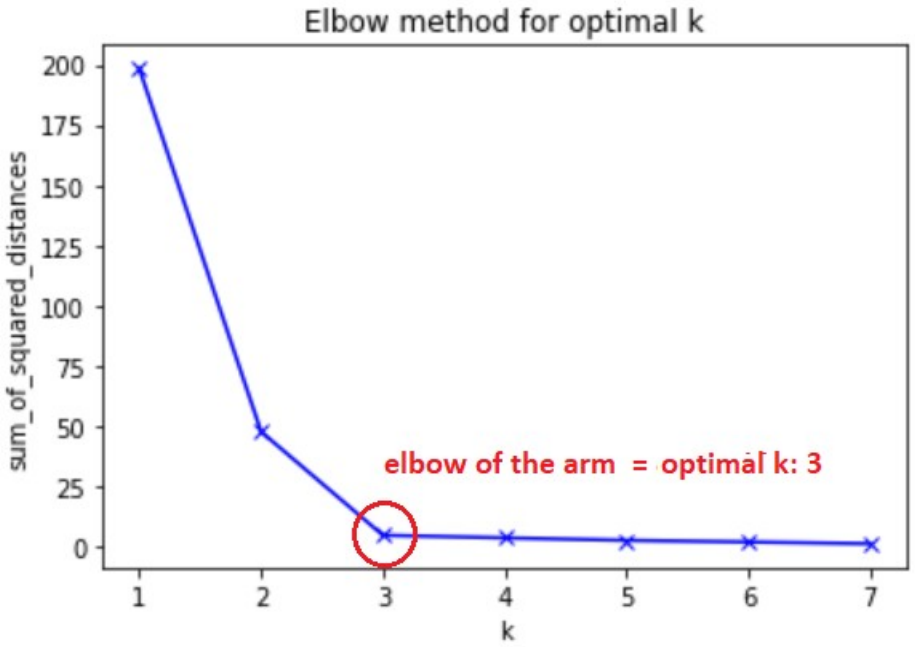


Método do Cotovelo (Elbow's Curve)

Método auxiliar e visual para encontrar o melhor esquema de clusters (melhor quantidade e qualidade de agrupamentos).

Normalmente, utilizamos algum outro método de validação para obter os índices de qualidade dos agrupamentos.

Devemos plotar uma curva usando o valor do índice obtido contra o número de grupos testado.



K-Medoids + Silhouette + Elbow’s Curve

Implemente o algoritmo k-medoids para encontrar os padrões no dataset iris. Faça 10 execuções para valores de k distintos. Para cada execução, calcule o score Silhouette. Use os valores gerados pelo silhouette (score médio) para plotar um gráfico do valor (y) pelo número de agrupamentos testado (k). Observe e escolha o melhor esquema de agrupamentos segundo o método do cotovelo visual.

Utilize a biblioteca Sklearn (pesquise e verifique a documentação)

Referências

[1] CHOWDHARY, K. R. Fundamentals of artificial intelligence. New Delhi: Springer India, 2020.

[2] NORVIG, Peter; RUSSELL, Stuart. Inteligência Artificial: uma abordagem moderna. Tradução da 4ª Ed. Rio de Janeiro: Elsevier, 2022.

[3] DOS DA SILVA, Fabrício M; LENZ, Maikon L.; FREITAS, Pedro H C.; SANTOS, Sidney C. Bispo. Inteligência artificial. Porto Alegre: SAGAH, 2019. E-book.

[4] MACIEL, Noberto Pires. Métodos de descoberta adaptativa de subconsultas para busca diversificada de imagens, 2024, 160 f., Dissertação (mestrado) - Programa de Pós-Graduação em Ciência da Computação, Universidade Estadual de Feira de Santana, Feira de Santana, 2024. Disponível em: http://tede2.uefs.br:8080/handle/tede/1847

[5] https://www.datacamp.com/pt/tutorial/dbscan-clustering-algorithm

[6] https://www.datacamp.com/pt/tutorial/hierarchical-clustering

[7] https://medium.com/@prasanth32888/k-medoids-clustering-cee6042155c6

[8] https://medium.com/cwi-software/entendendo-clusters-e-k-means-56b79352b452