

Sumário do Módulo 03 Parte 2 – Aprendizado Não-Supervisionado

- 1. Taxonomia dos Algoritmos de Aprendizado de Máquina
- 2. Aprendizado Não-Supervisionado
- 3. Agrupamento Particional K-Means
- 4. Agrupamento Hierárquico Agglomerative Clustering
- Considerações sobre os Agrupamentos







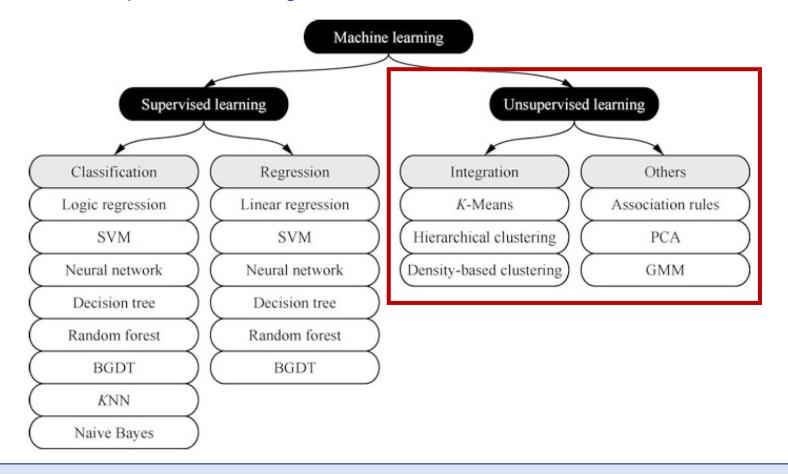


Objetivos do Módulo

- Este módulo tem por objetivo apresentar uma visão geral do aprendizado Não-Supervisionado para a tarefa de agrupamento de dados.
- Mostrar o funcionamento dos algoritmos de agrupamento Particionais, mais especificamente o método K-Means.
- Mostrar o funcionamento dos algoritmos de agrupamento Hierárquicos, mais especificamente o método Agglomerative Clustering.

1. Taxonomia dos Algoritmos de Aprendizado de Máquina

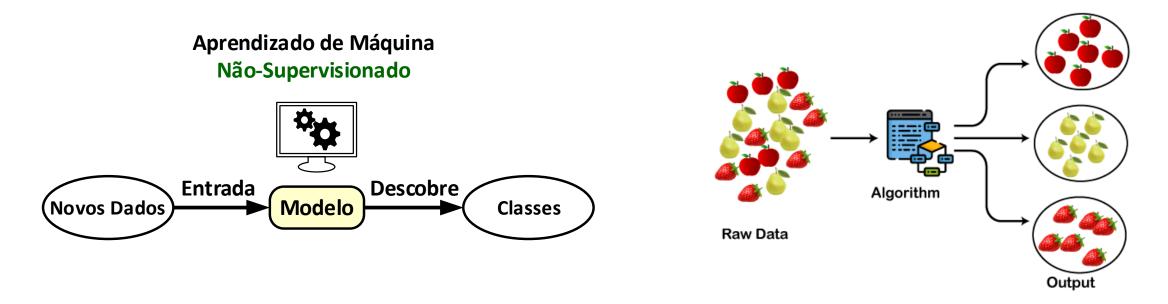
O aprendizado demanda uma interação entre um agente que receberá o conhecimento (aluno) e o ambiente (professor). Os paradigmas de aprendizagem variam conforme o papel desempenhado pelo agente, cujos métodos podem ser organizados conforme a natureza dessa interação.



Considera que não existem experiências prévias, sendo que a tarefa do modelo é descobrir as prováveis classes de objetos conforme um critério de similaridade.



Considera que não existem experiências prévias, sendo que a tarefa do modelo é descobrir as prováveis classes de objetos conforme um critério de similaridade.



Algoritmo de Agrupamento: Dado um conjunto de objetos de dados, no qual cada objeto é descrito por atributos numéricos, organize-os em um número de grupos (classes), de modo que aqueles que estejam no mesmo grupo sejam mais semelhantes, e diferente dos objetos de outros grupos.

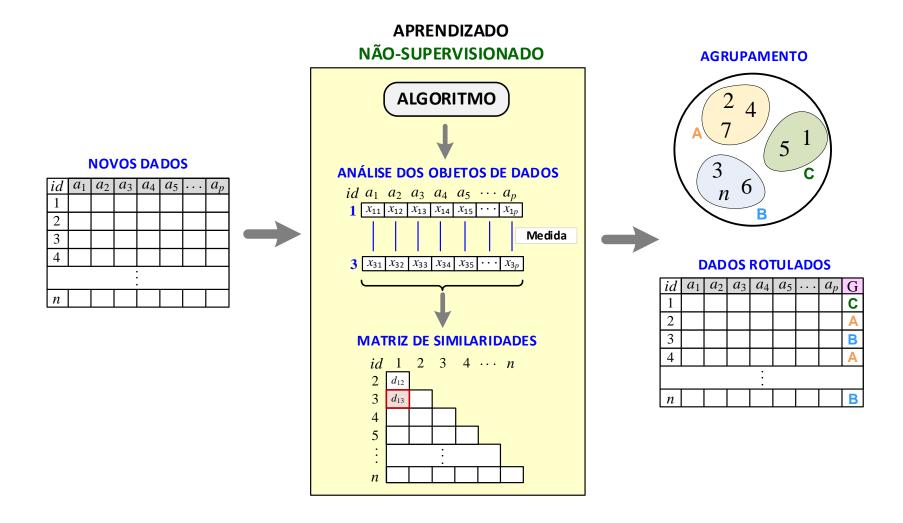
Características

- Abrange os métodos que realizam a tarefa de agrupamento (clustering) de dados.
- Útil quando não se conhece quais são as classes de objetos ou sua quantidade.
- O agrupamento tem propósito descritivo, pois auxilia o entendimento através de uma visão compacta dos dados.

Características

- Abrange os métodos que realizam a tarefa de agrupamento (clustering) de dados.
- Útil quando não se conhece quais são as classes de objetos ou sua quantidade.
- O agrupamento tem propósito descritivo, pois auxilia o entendimento através de uma visão compacta dos dados.
- Cada grupo corresponde a uma classe de objetos, a partir da qual são derivadas as regras que definem a similaridade entre eles.
- Os algoritmos de agrupamento examinam os dados para encontrar padrões em grupos de itens semelhantes com base em um critério.
- As classes descobertas podem ser usadas no treinamento de métodos supervisionados.

Modelo genérico de um método de Agrupamento de Dados

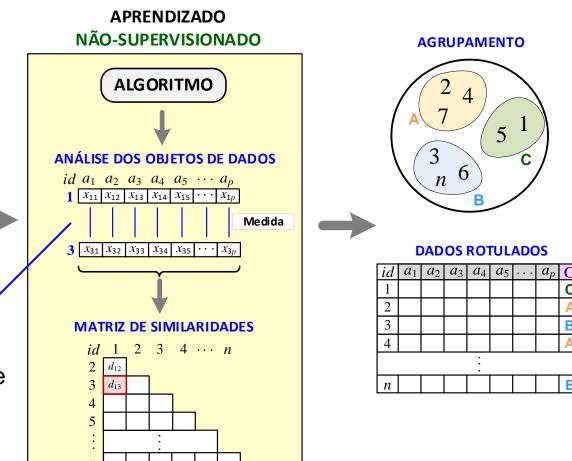


Modelo genérico de um método de Agrupamento de Dados

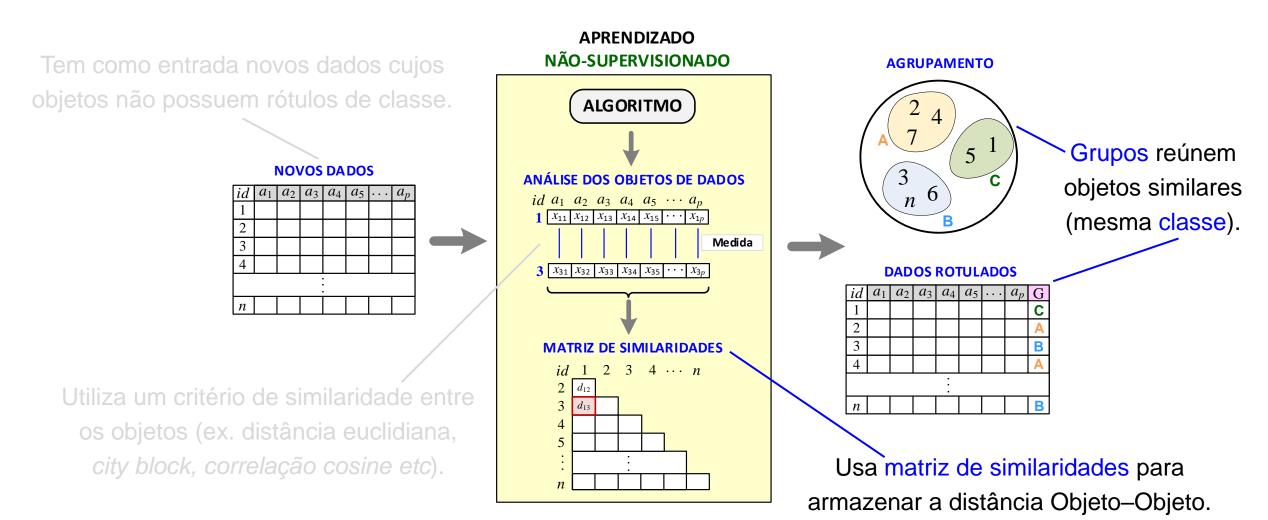
Tem como entrada novos dados cujos objetos não possuem rótulos de classe.



Utiliza um critério de similaridade entre os objetos (ex. distância euclidiana, city block, correlação cosine etc).



Modelo genérico de um método de Agrupamento de Dados

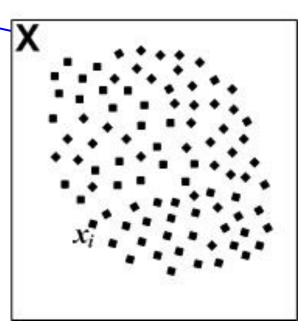


Análise de Agrupamentos

Seja $\mathbf{X} = \{x_1, x_2, ..., x_n\}$ um conjunto de n objetos x_i

Cada objeto x_i é um vetor com p dimensões de componentes numéricos (atributos).

$$x_1 = a_{11} a_{12} a_{13} \dots a_{1p}$$

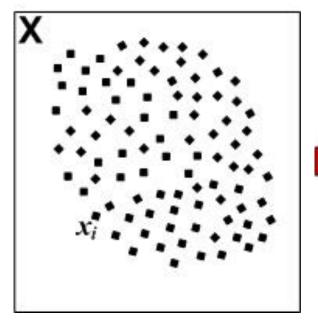


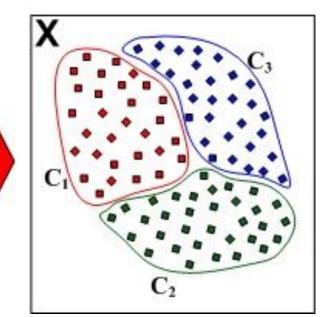
Análise de Agrupamentos

Seja $X = \{x_1, x_2, ..., x_n\}$ um conjunto de n objetos x_i

Cada objeto x_i é um vetor com p dimensões de componentes numéricos (atributos).

$$x_1 = \underbrace{\begin{bmatrix} a_{11} \mid a_{12} \mid a_{13} \mid \dots \mid a_{1p} \end{bmatrix}}_{p}$$





RESTRIÇÕES

- **1.** $C_1 \cup C_2 ... \cup C_K = X$
- **2.** $C_i \neq \emptyset$, $\forall i$
- **3.** $C_i \cap C_j = \emptyset$, $\forall i \neq j$

OBJETIVO: Encontrar agrupamentos $C = \{C_1, C_2, ..., C_K\}$, sendo K a quantidade de grupos.

Análise de Agrupamentos

Matriz de Dados (Objeto-Atributo)

SV	AADAC	AAMP	AANAT	AARS	AATF	ABAT	ABCA1	ABCA2	ABCA3
BR-01T	69,87	71,80	132,85	102,35	61,12	225,13	106,12	100,86	89,46
BR-02T	50,53	53,75	485,39	20,00	127,93	731,04	896,59	505,28	65,36
BR-03T	20,00	62,78	309,12	61,18	94,53	478,09	501,36	303,07	77,41
BR-04T	20,00	58,26	397,26	40,59	111,23	604,56	698,97	404,18	47,59
BR-05T	55,59	60,52	141,28	50,88	41,15	541,32	600,16	353,62	62,50
BR-06T	25,90	39,59	269,27	45,73	76,19	229,18	649,57	378,90	55,05
BR-07T	20,00	33,37	205,27	48,31	58,67	385,25	624,87	366,26	58,77
BR-08T	79,42	24,32	237,27	47,02	44,95	307,21	424,81	372,58	56,91
BR-09T	46,47	28,85	147,51	47,66	51,81	346,23	524,84	369,42	38,56
BR-10T	20,00	26,58	192,39	47,34	48,38	163,36	237,41	371,00	47,73
BR-11T	83,96	55,43	169,95	47,50	50,10	254,80	381,13	370,21	43,15
BR-12T	71,51	41,01	362,34	47,42	49,24	209,08	309,27	370,60	22,72
BR-13T	93,25	96,44	266,15	47,46	33,11	231,94	345,20	370,41	32,93

Matriz de Similaridades (Objeto-Objeto)

	BR-02T	BR-03T	BR-04T	BR-05T	BR-06T	BR-07T	BR-08T	BR-09T	BR-10T	BR-11T	BR-12T
BR-13T	230,09	202,67	176,15	112,58	690,28	200,00	176,15	123,55	57,89	10,53	7,14
BR-12T	0,11	0,01	0,19	0,28	0,34	0,04	0,57	0,83	1,02	2,23	3,67
BR-11T	0,07	0,16	0,29	1,92	0,22	0,47	0,87	5,76	0,65	1,94	1,24
BR-10T	9,12	5,67	4,38	5,37	27,36	17,00	13,13	16,10	82,09	23,57	26,72
BR-09T	19,99	17,70	17,51	21,46	59,98	53,09	52,52	64,39	179,95	67,32	
BR-08T	12,56	0,16	0,12	1,56	37,68	0,48	0,37	4,69	113,04		
BR-07T	4,25	7,06	6,59	8,33	23,33	5,88	9,42	92,67			
BR-06T	0,10	0,14	0,19	1,16	11,55	9,56	0,76				
BR-05T	1,89	1,17	0,76	4,62	32,68	100,44					
BR-04T	22,62	26,13	26,36	66,29	10,67						
BR-03T	22,57	22,57	23,00	15,33							
BR-02T	105,75	99,57									
BR-01T	390,58										

Análise de Agrupamentos

Matriz de Dados (Objeto-Atributo)

535	AADAC	AAMP	AANAT	AARS	AATF	ABAT	ABCA1	ABCA2	ABCA3
BR-01T	69,87	71,80	132,85	102,35	61,12	225,13	106,12	100,86	89,46
BR-02T	50,53	53,75	485,39	20,00	127,93	731,04	896,59	505,28	65,36
BR-03T	20,00	62,78	309,12	61,18	94,53	478,09	501,36	303,07	77,41
BR-04T	20,00	58,26	397,26	40,59	111,23	604,56	698,97	404,18	47,59
BR-05T	55,59	60,52	141,28	50,88	41,15	541,32	600,16	353,62	62,50
BR-06T	25,90	39,59	269,27	45,73	76,19	229,18	649,57	378,90	55,05
BR-07T	20,00	33,37	205,27	48,31	58,67	385,25	624,87	366,26	58,77
BR-08T	79,42	24,32	237,27	47,02	44,95	307,21	424,81	372,58	56,91
BR-09T	46,47	28,85	147,51	47,66	51,81	346,23	524,84	369,42	38,56
BR-10T	20,00	26,58	192,39	47,34	48,38	163,36	237,41	371,00	47,73
BR-11T	83,96	55,43	169,95	47,50	50,10	254,80	381,13	370,21	43,15
BR-12T	71,51	41,01	362,34	47,42	49,24	209,08	309,27	370,60	22,72
BR-13T	93,25	96,44	266,15	47,46	33,11	231,94	345,20	370,41	32,93

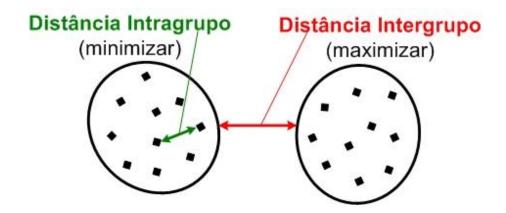
MÉTRICA

Matriz de Similaridades (Objeto-Objeto)

	BR-02T	BR-03T	BR-04T	BR-05T	BR-06T	BR-07T	BR-08T	BR-09T	BR-10T	BR-11T	BR-12T
BR-13T	230,09	202,67	176,15	112,58	690,28	200,00	176,15	123,55	57,89	10,53	7,14
BR-12T	0,11	0,01	0,19	0,28	0,34	0,04	0,57	0,83	1,02	2,23	3,67
BR-11T	0,07	0,16	0,29	1,92	0,22	0,47	0,87	5,76	0,65	1,94	1,24
BR-10T	9,12	5,67	4,38	5,37	27,36	17,00	13,13	16,10	82,09	23,57	26,72
BR-09T	19,99	17,70	17,51	21,46	59,98	53,09	52,52	64,39	179,95	67,32	
BR-08T	12,56	0,16	0,12	1,56	37,68	0,48	0,37	4,69	113,04		
BR-07T	4,25	7,06	6,59	8,33	23,33	5,88	9,42	92,67			
BR-06T	0,10	0,14	0,19	1,16	11,55	9,56	0,76				
BR-05T	1,89	1,17	0,76	4,62	32,68	100,44					
BR-04T	22,62	26,13	26,36	66,29	10,67						
BR-03T	22,57	22,57	23,00	15,33							
BR-02T	105,75	99,57									
BR-01T	390,58										

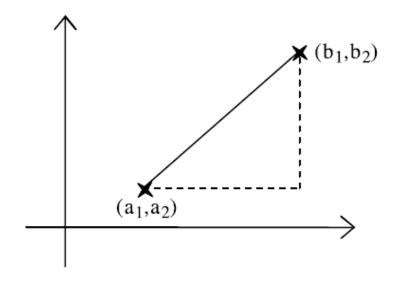
SITUAÇÃO IDEAL

Grupos homogêneos (coesão interna) e mais separados possíveis (isolamento externo).

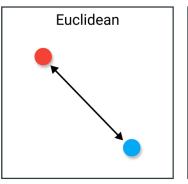


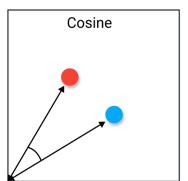
Medidas de Similaridade

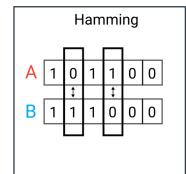
O agrupamento usa uma medida de similaridade para atribuir pontos no espaço (objetos) a um cluster.

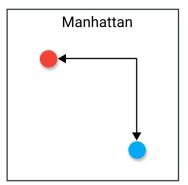


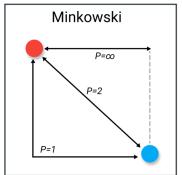
$$\sqrt{(a_1-b_1)^2+(a_2-b_2)^2}$$

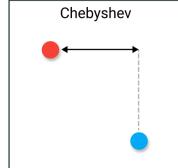


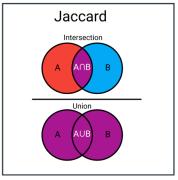


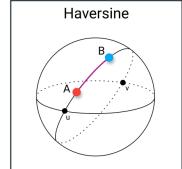


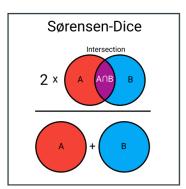












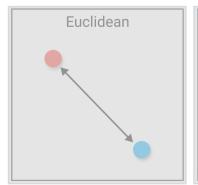
Medidas de Similaridade

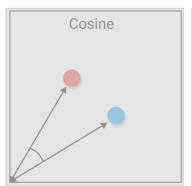
O agrupamento usa uma medida de similaridade para atribuir pontos no espaço (objetos) a um cluster.

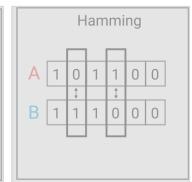
A similaridade é difícil de ser definida pois depende do atributo e do critério considerados.

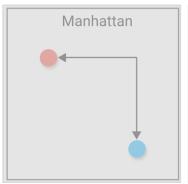


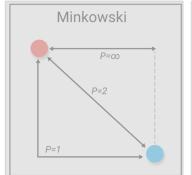


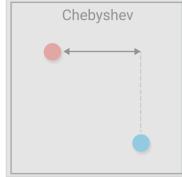


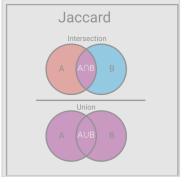




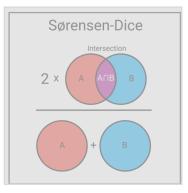




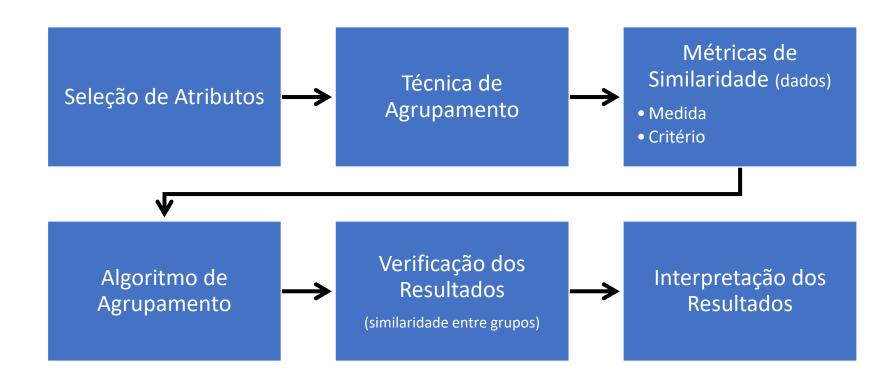




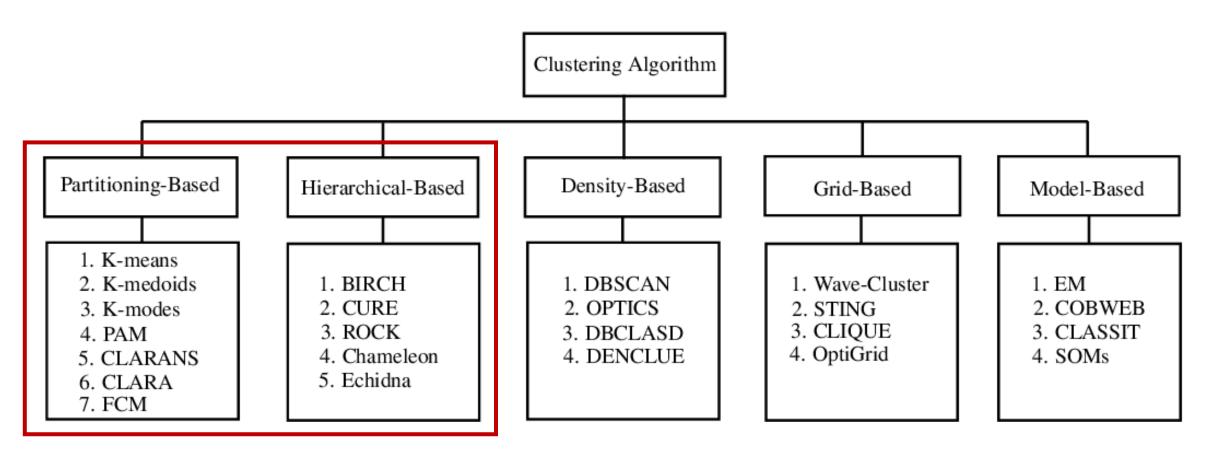




Workflow do Aprendizado Não-Supervisionado



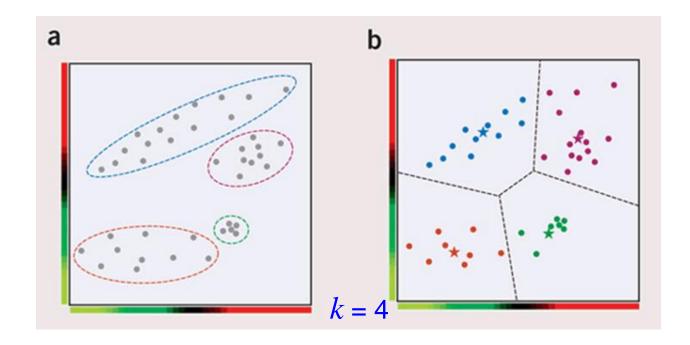
Taxonomia dos Algoritmos de Agrupamento



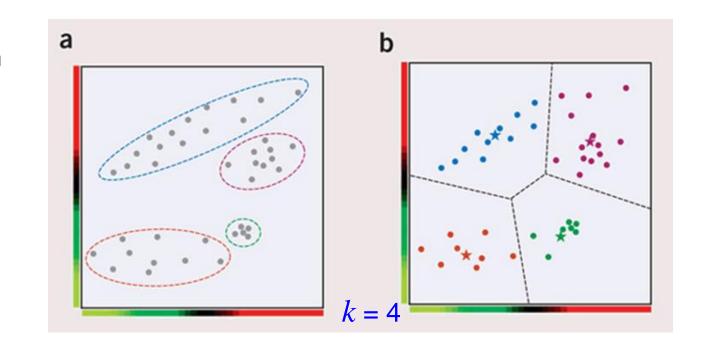
Fonte: A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis.

Disponível em: www.researchgate.net/figure/An-overview-of-clustering-taxonomy_fig1_320799077. Acessado em: 19/02/2023.

- Buscam o melhor agrupamento dos n objetos de entrada em k grupos de saída.
- Algoritmos tipicamente adotam estratégia divisiva, que inicia com um único grupo com todos os objetos, os quais são gradualmente particionados conforme os k grupos. No geral, buscam minimizar uma função associada ao agrupamento (ex. soma das distâncias intragrupo).



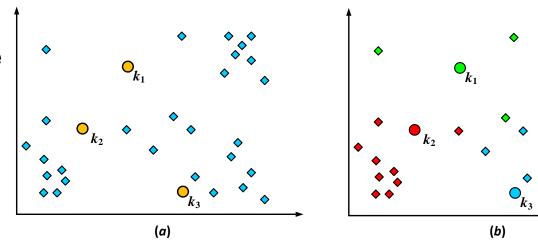
- Buscam o melhor agrupamento dos n objetos de entrada em k grupos de saída.
- Algoritmos tipicamente adotam estratégia divisiva, que inicia com um único grupo com todos os objetos, os quais são gradualmente particionados conforme os k grupos. No geral, buscam minimizar uma função associada ao agrupamento (ex. soma das distâncias intragrupo).
- O número grupos k é um hiperparâmetro, sendo que existem índices que auxiliam na sua determinação (ex. *Elbow Method*, índice Silhouette, Davies-Bouldin.
- Exemplos de algoritmos: K-Means, K-Prototypes, PAM (Partitioning Around Medoids), CLARA (Clustering Large Applications) entre outros.



K-Means

Usa k centroides para representar os grupos, com os objetos sendo associados ao centroide mais próximo.

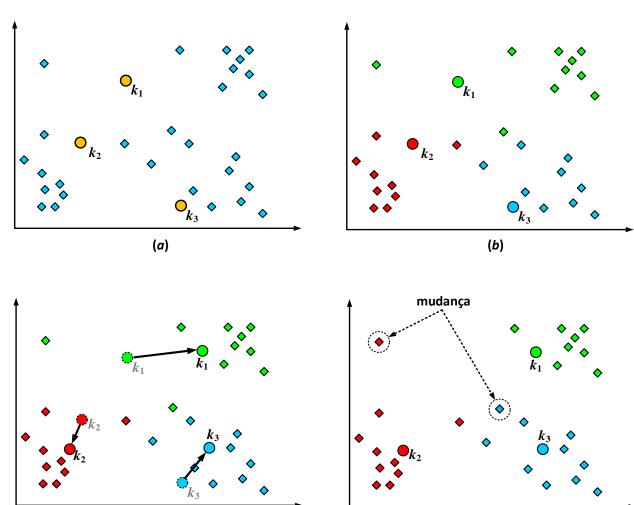
- (a) escolhe aleatoriamente um número k de centroides para iniciar os *clusters*;
- (b) calcula a distância entre cada objeto e o centroide mais próximo;



K-Means

Usa k centroides para representar os grupos, com os objetos sendo associados ao centroide mais próximo.

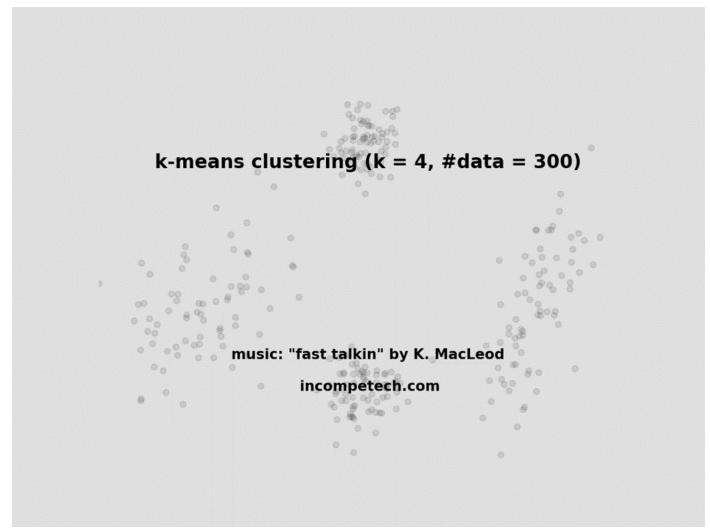
- (a) escolhe aleatoriamente um número k de centroides para iniciar os *clusters*;
- (b) calcula a distância entre cada objeto e o centroide mais próximo;
- (c) move cada centroide considerando a média dos objetos do *cluster* correspondente;
- (d) Repete os passos 2 e 3 alternadamente até a convergência do método.



(c)

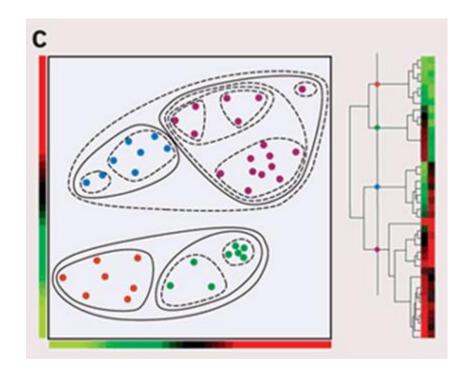
(d)

K-Means – Exemplo de Funcionamento



Disponível: https://www.youtube.com/watch?v=5I3Ei69I40s

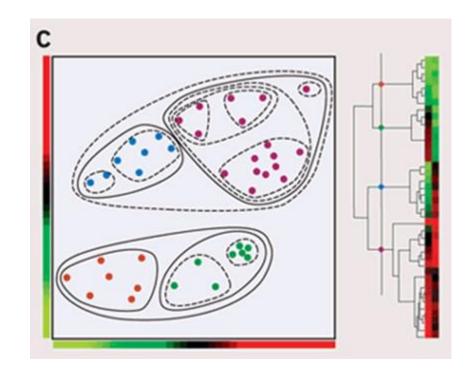
Os dados são organizados em uma representação hierárquica e sobreposta dos grupos, que facilita a visualização da formação do agrupamento, identificar os pontos de união dos objetos e seu grau de semelhança.



Os dados são organizados em uma representação hierárquica e sobreposta dos grupos, que facilita a visualização da formação do agrupamento, identificar os pontos de união dos objetos e seu grau de semelhança.

Algoritmos de agrupamento hierárquicos são organizados conforme duas perspectivas:

Aglomerativos (botton-up): realizam fusões sucessivas dos n objetos, que inicialmente constituem clusters isolados, até que se obtenha um único cluster com todos os objetos.

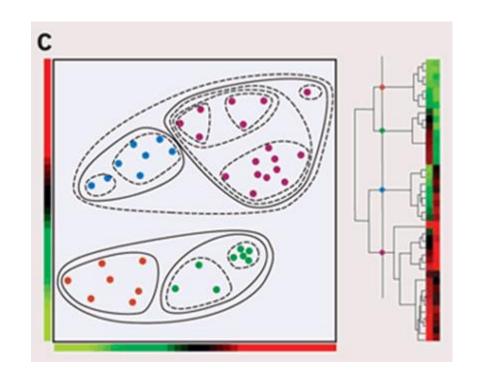


Os dados são organizados em uma representação hierárquica e sobreposta dos grupos, que facilita a visualização da formação do agrupamento, identificar os pontos de união dos objetos e seu grau de semelhança.

Algoritmos de agrupamento hierárquicos são organizados conforme duas perspectivas:

Aglomerativos (botton-up): realizam fusões sucessivas dos n objetos, que inicialmente constituem clusters isolados, até que se obtenha um único cluster com todos os objetos.

Divisivos (top-down): iniciam por um único *cluster* que contém os n objetos, os quais são sucessivamente divididos em *clusters* menores até o nível de clusters isolados.



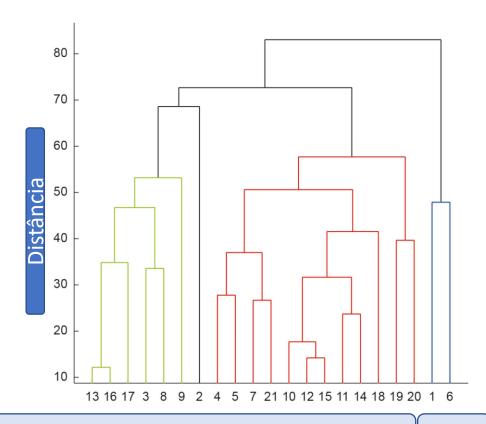
Os dados são organizados em uma representação hierárquica e sobreposta dos grupos, que facilita a visualização da formação do agrupamento, identificar os pontos de união dos objetos e seu grau de semelhança.

Algoritmos de agrupamento hierárquicos são organizados conforme duas perspectivas:

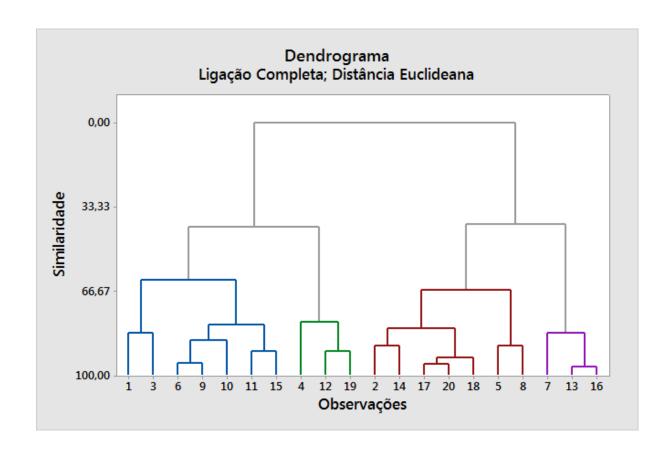
Aglomerativos (botton-up): realizam fusões sucessivas dos n objetos, que inicialmente constituem clusters isolados, até que se obtenha um único cluster com todos os objetos.

Divisivos (*top-down*): iniciam por um único *cluster* que contém os *n* objetos, os quais são sucessivamente divididos em *clusters* menores até o nível de clusters isolados.

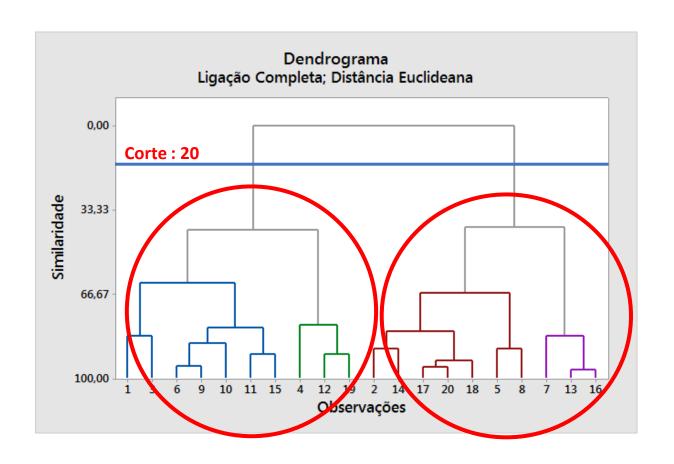
O dendrograma é uma representação visual em forma de árvore, que mostra a decomposição do conjunto de dados.



Algoritmos de agrupamento hierárquicos geralmente não requerem a definição prévia do número de *clusters*. O dendrograma pode ser usado para inferir o número adequado de agrupamentos.



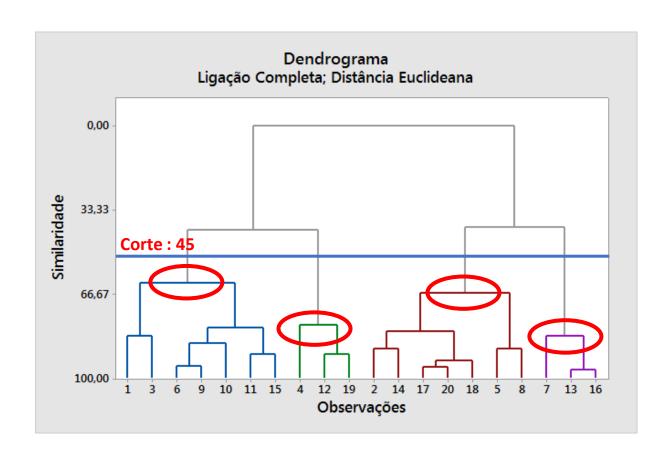
Algoritmos de agrupamento hierárquicos geralmente não requerem a definição prévia do número de *clusters*. O dendrograma pode ser usado para inferir o número adequado de agrupamentos.



O agrupamento mais adequado e dado pelo corte do dendrograma, que corresponde a uma linha horizontal no dendrograma para especificar a quantidade de grupos.

O corte do dendrograma com similaridade igual a 20 geraria um agrupamento de **2 grupos**

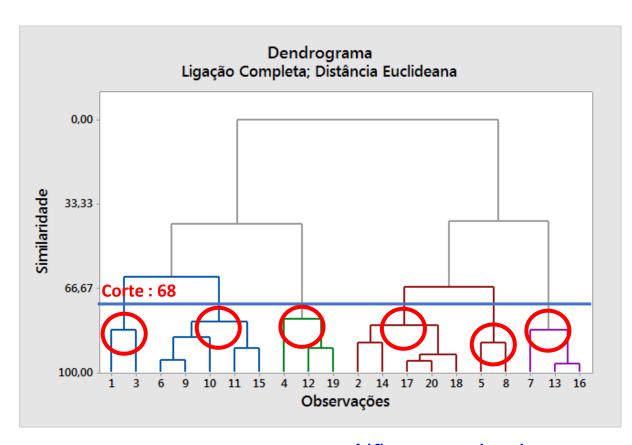
Algoritmos de agrupamento hierárquicos geralmente não requerem a definição prévia do número de *clusters*. O dendrograma pode ser usado para inferir o número adequado de agrupamentos.



O agrupamento mais adequado e dado pelo corte do dendrograma, que corresponde a uma linha horizontal no dendrograma para especificar a quantidade de grupos.

O corte do dendrograma com similaridade igual a 45 geraria um agrupamento de **4 grupos**

Algoritmos de agrupamento hierárquicos geralmente não requerem a definição prévia do número de *clusters*. O dendrograma pode ser usado para inferir o número adequado de agrupamentos.



O agrupamento mais adequado e dado pelo corte do dendrograma, que corresponde a uma linha horizontal no dendrograma para especificar a quantidade de grupos.

O corte do dendrograma com similaridade igual a 68 geraria um agrupamento de **6 grupos**

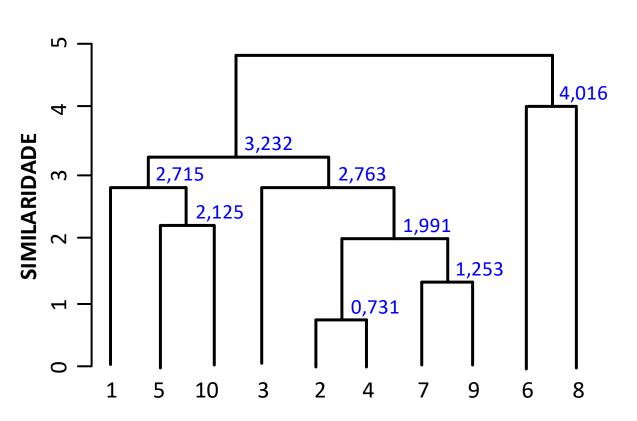
Técnicas para definição do corte:

- Critério de Mojena
- Método de Tocher
 - RMSSTD (Root Mean Square Standard Deviation)

Não usam dendrograma

Determinação do Corte do Dendrograma – Critério de Mojena

$$P_{corte} = M\acute{e}dia(PF) + DesvioPadrao(PF) * w$$



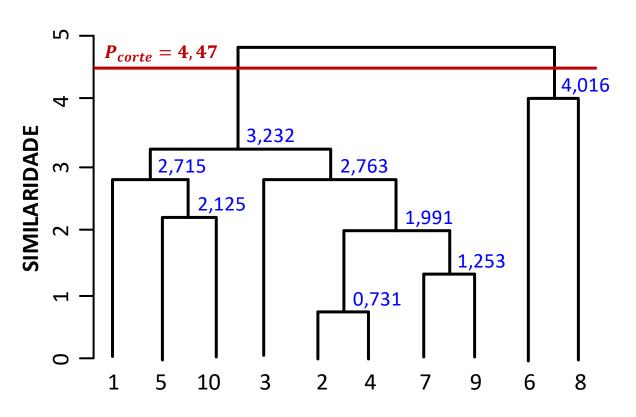
onde:

PF : Pontos de Fusão. Valor de similaridade onde os *clusters* se unem (linhas horizontais)

w: constante definida na literatura como 1,25 ou 2.

Determinação do Corte do Dendrograma – Critério de Mojena

$$P_{corte} = M\acute{e}dia(PF) + DesvioPadrao(PF) * w$$



onde:

PF : Pontos de Fusão. Valor de similaridade onde os *clusters* se unem (linhas horizontais)

 \boldsymbol{w} : constante definida na literatura como 1,25 ou 2.

Exemplo:

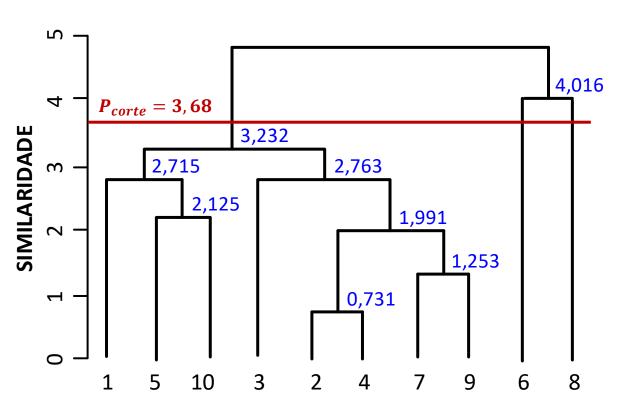
$$P_{corte}$$
 = 2,35 + 1,06 x 2

$$P_{corte} = 4,47$$

2 grupos

Determinação do Corte do Dendrograma - Critério de Mojena

$$P_{corte} = M\acute{e}dia(PF) + DesvioPadrao(PF) * w$$



onde:

PF : Pontos de Fusão. Valor de similaridade onde os *clusters* se unem (linhas horizontais)

w: constante definida na literatura como 1,25 ou 2.

Exemplo:

$$P_{corte}$$
 = 2,35 + 1,06 x 2

$$P_{corte} = 4,47$$

2 grupos

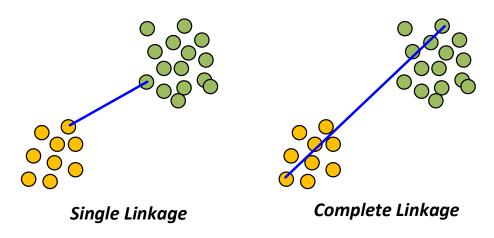
$$P_{corte}$$
 = 2,35 + 1,06 x 1,25

$$P_{corte} = 3,68$$

3 grupos

Critérios de Ligação de Grupos

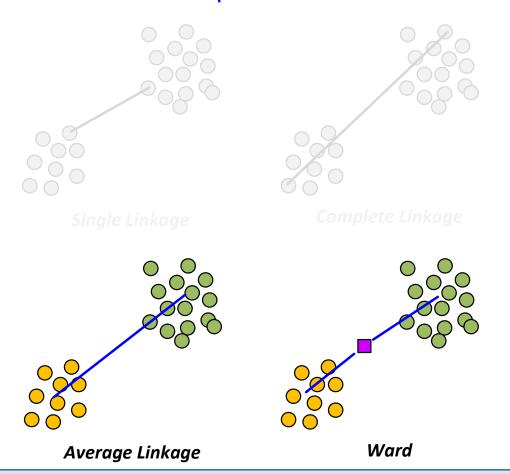
Os métodos aglomerativos são os mais adotados, se diferenciam no modo como definem a proximidade entre pares clusters:



- Single Linkage: vizinho mais próximo. A proximidade corresponde a distância entre dois objetos mais próximos.
- Complete Linkage: a proximidade corresponde a de dois objetos mais distantes.

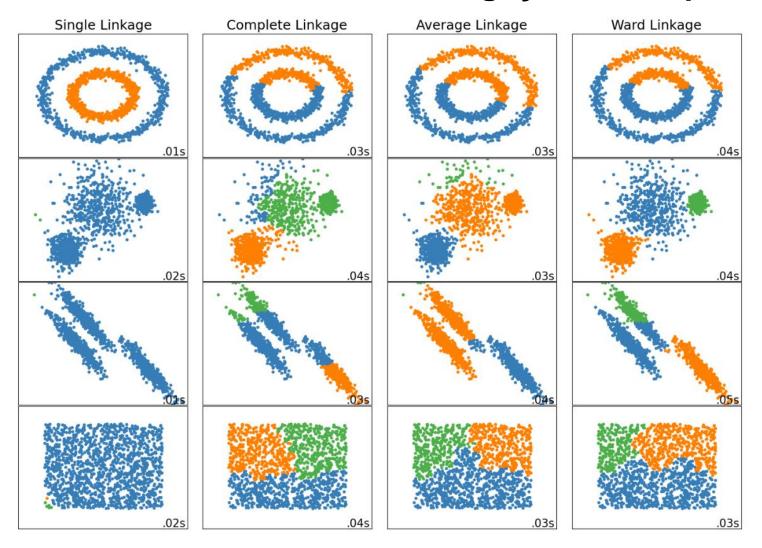
Critérios de Ligação de Grupos

Os métodos aglomerativos são os mais adotados, se diferenciam no modo como definem a proximidade entre pares clusters:



- Single Linkage: vizinho mais próximo. A proximidade corresponde a distância entre dois objetos mais próximos.
- Complete Linkage: a proximidade corresponde a de dois objetos mais distantes.
- Average Linkage: ligação média entre grupos (Unweighted Pair Group Method using Arithmetic averages, UPGMA). A proximidade é a média de todas as proximidades entre os objetos dos clusters.
- Ward: ou aumento mínimo da soma de quadrados (Minimal Increase of Sum-of-Squares, MISSQ). Agrupa as distâncias menores e recalcula novas distâncias através do erro quadrático médio das distâncias entre os objetos.

Resultados dos Critérios de Ligação de Grupos



5. Considerações sobre os Agrupamentos

Questões e demais aspectos relevantes na análise dos resultados do agrupamento:

- Os dados são de fato separáveis? Existem estruturas não-aleatórias nos dados?
- Qual o número ideal de clusters?
- Os clusters encontrados realmente correspondem a classes reais?
- Dados diversos agrupamentos para o mesmo conjunto de dados como definir o melhor?

5. Considerações sobre os Agrupamentos

Questões e demais aspectos relevantes na análise dos resultados do agrupamento:

- Os dados são de fato separáveis? Existem estruturas não-aleatórias nos dados?
- Qual o número ideal de *clusters*?
- Os clusters encontrados realmente correspondem a classes reais?
- Dados diversos agrupamentos para o mesmo conjunto de dados como definir o melhor?
- Quando se conhece os rótulos das classes, as medidas externas (ex. ARI, CRand etc) podem ser usadas para avaliar a qualidade dos agrupamentos, ou comparação entre métodos.
- A análise humana pode ser necessária quando não se conhece os rótulos de classe.
- Algoritmos de aprendizado supervisionado tipo caixa branca (ex. árvore de decisão) são úteis para o entendimento das regras consideradas no agrupamento.

Resumo

- Foram revistos os conceitos e as características do aprendizado de máquina Não-Supervisionado para a tarefa de agrupamento de dados.
- Foi descrito o funcionamento de um modelo genérico Não-Supervisionado, sendo detalhado o modo como é determinada a similaridade de novos objetos de dados até obtenção dos grupos.
- Descrita a taxonomia dos algoritmos de agrupamento.

Resumo

- Foram revistos os conceitos e as características do aprendizado de máquina Não-Supervisionado para a tarefa de agrupamento de dados.
- Foi descrito o funcionamento de um modelo genérico Não-Supervisionado, sendo detalhado o modo como é determinada a similaridade de novos objetos de dados até obtenção dos grupos.
- Descrita a taxonomia dos algoritmos de agrupamento.
- Foi explicado o funcionamento dos algoritmos Particionais, sendo dada ênfase para o método K-Means e suas etapas realizadas na sua execução.
- Foi explicado o funcionamento dos algoritmos Hierárquicos, seus tipos (Aglomerativo e Divisivo), os critérios de ligação de grupos (Single, Complete, Average, Ward), interpretação do dendrograma e métodos para definição da quantidade de grupos.
- Foram levantadas questões relevantes para o sucesso da tarefa de agrupamento.

Referências Bibliográficas

- [1] Huawei Technologies Co. Artificial Intelligence Technology Official Textbooks for Huawei ICT Academy. Morgan Kaufmann, 3rd edition, 2023.
- [2] Marc Peter Deisenroth, A Aldo Faisal, and Cheng Soon Ong. Mathematics for Machine Learning. Cambridge University Press, 2020.
- [3] Aurélien Géron. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media, Inc., 3rd edition, 2022.
- [4] Trevor Hastie, Robert Tibshirani, JeromeHFriedman, and JeromeHFriedman. The elements of statistical learning: data mining, inference, and prediction, volume 2. Springer, 2009.
- [5] Han Jiawei, Kamber Micheline, and Pei Jian. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd edition, 2012.
- [6] Andreas C Müller and Sarah Guido. Introduction to Machine Learning with Python: a guide for data scientists. O'Reilly Media, Inc., 2016.
- [7] Shai Shalev-Shwartz and Shai Ben-David. Understanding Machine Learning: From theory to algorithms. Cambridge university press, 3rd edition, 2014.

Apoio

Projeto Residência 8 (TPA N° 068/SOFTEX/UNIFEI)

Este projeto é apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei nº 8.248, de 23 de outubro de 1991, no âmbito do [PPI-Softex| PNM-Design], coordenado pela Softex.

