

# Aprendizado de Máquina – IMD1101

## Aula 22 – Aprendizado Não Supervisionado 02

# Algoritmo Hierárquico Divisivo ou Aglomerativo

# Agrupamento Hierárquico

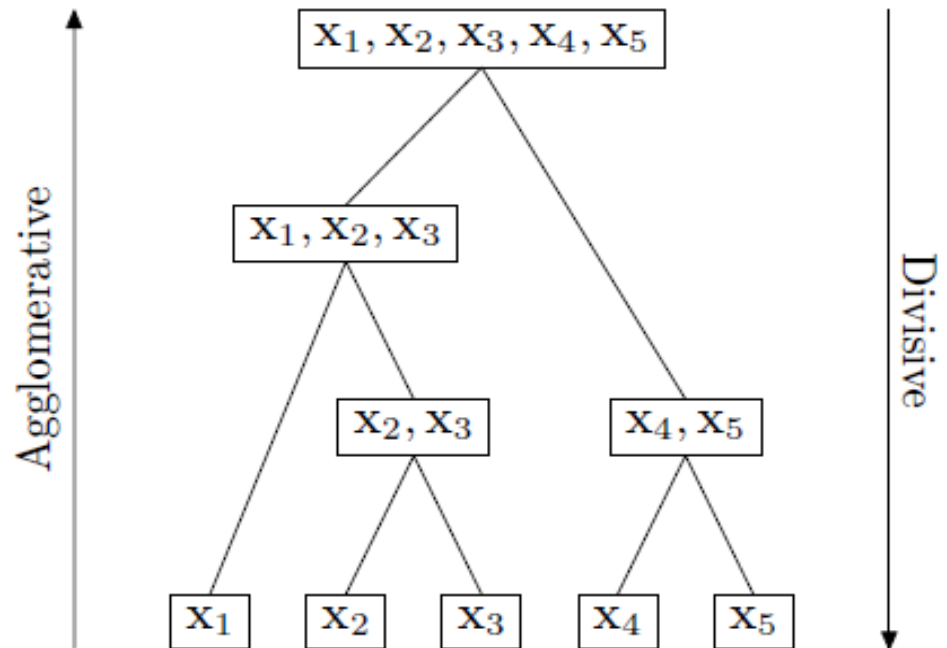
□ Envolvem a construção de uma hierarquia de uma estrutura do tipo árvore.

□ Tipos:

❖ Divisivo

❖ Aglomerativo

- Ligação Simples
- Ligação Completa
- Ligação Média.



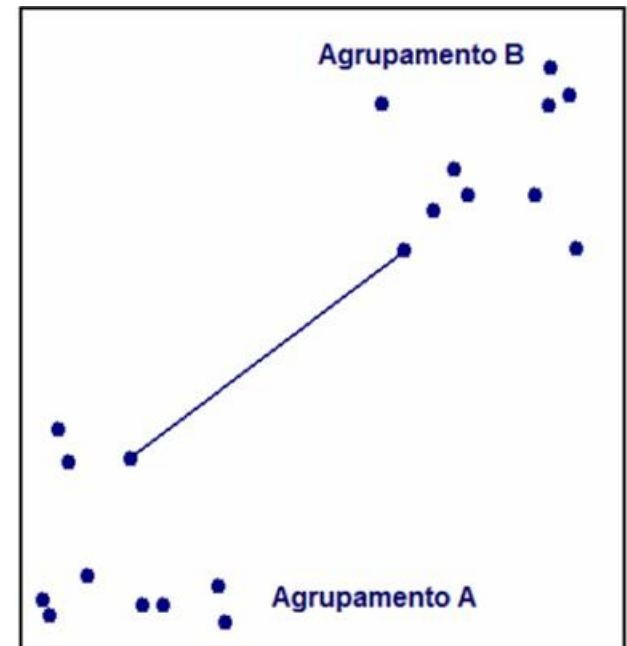
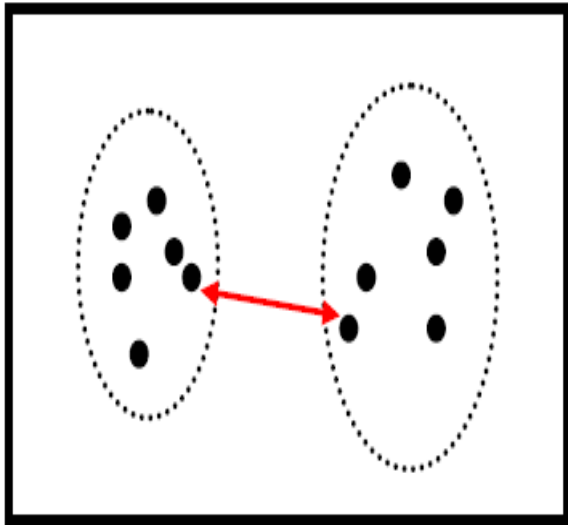
# Métodos Aglomerativos

- ❑ Cada **objeto** começa como seu **próprio grupo** (cluster).
- ❑ Em passos seguintes, os dois grupos (ou objetos) mais próximos (similares) **são combinados** em um novo agregado.
  - ❖ O número de grupos é reduzido em uma unidade em cada passo.
- ❑ Ao final, todos os elementos são reunidos em um grande agregado.

# Ligação Simples

## ❑ Definição:

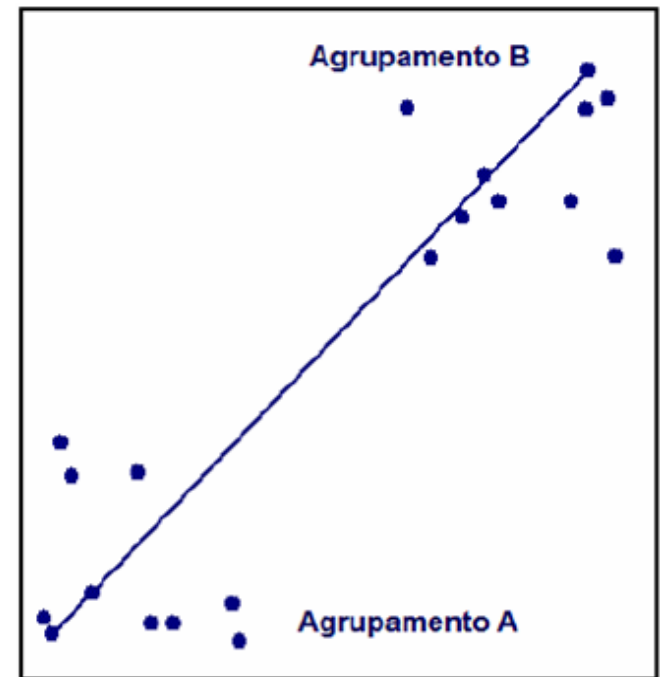
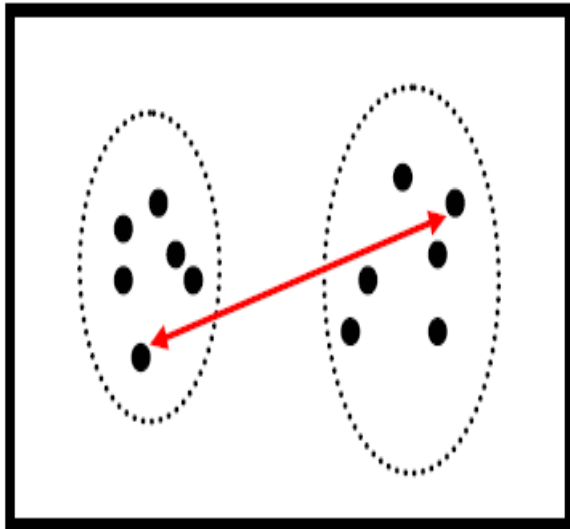
- ❖ Distância entre dois clusters é a distância entre os pontos mais próximos. Também chamado “agrupamento de vizinhos”.



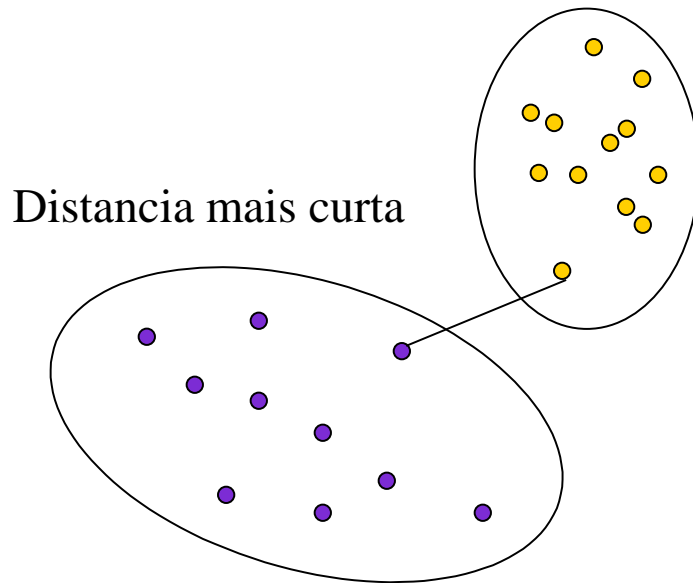
# Ligação Completa

## ❑ Definição:

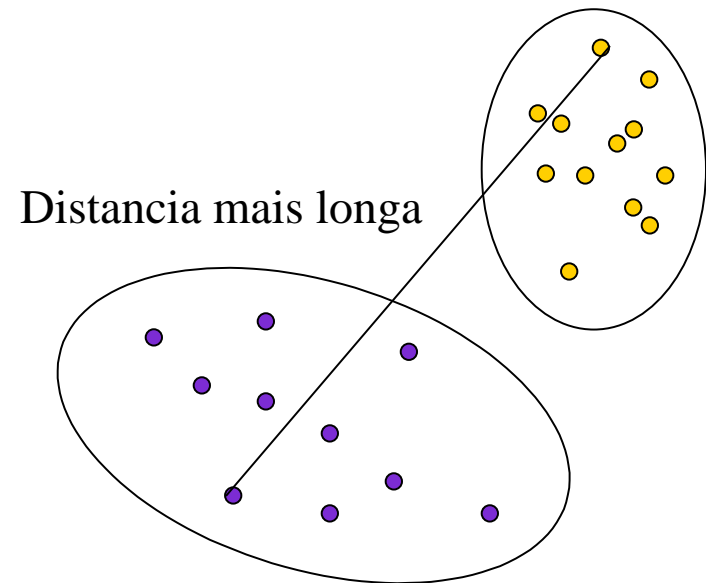
- ❖ Distância entre dois clusters é a distância entre os pontos mais distantes.



# Ligação Simples x Ligação Completa



Tendência a produzir  
grupos alongados



Tendência a produzir  
grupos compactos

# Ligação Média

## □ Ligação Média:

- ❖ Distância (similaridade) média de todos os objetos em um grupo para os demais em outro é usada como critério.
- ❖ São menos dependentes de valores extremos, como ocorre com a ligação simples ou completa.
- ❖ Tendem a combinar grupos com pequena variação interna.



# Método Hierárquico

## ❑ Ligação Simples: exemplo (1 / 6)

	X1	X2	X3	X4	X5
C1	7,000	10,000	9,000	7,000	10,000
C2	9,000	9,000	8,000	9,000	9,000
C3	5,000	5,000	6,000	7,000	7,000
C4	6,000	6,000	3,000	3,000	4,000
C5	1,000	2,000	2,000	1,000	2,000
C6	4,000	3,000	2,000	3,000	3,000
C7	2,000	4,000	5,000	2,000	5,000

Exemplo do uso do **Método Hierárquico Aglomerativo** (ligação simples), com Correlação de Pearson (-1 a 1 ) como medida de proximidade.

# Método Hierárquico

## ❑ Ligação Simples: exemplo (1/6) cont.

	C1	C2	C3	C4	C5	C6	C7
C1	1,000						
C2	-0,147	1,000					
C3	0,000	0,000	1,000				
C4	0,087	0,516	-0,824	1,000			
C5	0,963	-0,408	0,000	-0,060	1,000		
C6	-0,466	0,791	-0,354	0,699	-0,645	1,000	
C7	0,891	-0,516	0,165	-0,239	0,963	-0,699	1,000

Matriz de correlação entre os sete grupos.

# Método Hierárquico

## ❑ Ligação Simples: exemplo (2/6)

Formando o grupo (1,5):

	(C1,C5)	C2	C3	C4	C6	C7
(C1,C5)	1,000					
C2	-0,147	1,000				
C3	0,000	0,000	1,000			
C4	0,087	0,516	-0,824	1,000		
C6	-0,466	0,791	-0,354	0,699	1,000	
C7	0,963	-0,516	0,165	-0,239	-0,699	1,000

# Método Hierárquico

## ❑ Ligação Simples: exemplo (2/6) cont.

Verificando a maior correlação entre os grupos:

	(C1,C5)	C2	C3	C4	C6	C7
(C1,C5)	1,000					
C2	-0,147	1,000				
C3	0,000	0,000	1,000			
C4	0,087	0,516	-0,824	1,000		
C6	-0,466	0,791	-0,354	0,699	1,000	
C7	0,963	-0,516	0,165	-0,239	-0,699	1,000

# Método Hierárquico

## ❑ Ligação Simples: exemplo (3/6)


Formando o grupo (1,5,7):

	(C1,C5,C7)	C2	C3	C4	C6
(C1,C5,C7)	1,000				
C2	-0,147	1,000			
C3	0,165	0,000	1,000		
C4	0,087	0,516	-0,824	1,000	
C6	-0,466	0,791	-0,354	0,699	1,000

# Método Hierárquico

## ❑ Ligação Simples: exemplo (3/6) cont.

Verificando a maior correlação entre os grupos:

	(C1,C5,C7)	C2	C3	C4	C6
(C1,C5,C7)	1,000				
C2	-0,147	1,000			
C3	0,165	0,000	1,000		
C4	0,087	0,516	-0,824	1,000	
C6	 -0,466	0,791	-0,354	0,699	1,000

# Método Hierárquico

## ❑ Ligação Simples: exemplo (4/6)

Formando o grupo (2,6):

	(C1,C5,C7)	(C2,C6)	C3	C4
(C1,C5,C7)	1,000	-0,147	0,165	0,087
(C2,C6)	-0,147	1,000	-0,824	0,516
C3	0,165	0,000	1,000	-0,824
C4	0,087	0,699	-0,824	1,000

# Método Hierárquico

## ❑ Ligação Simples: exemplo (4/6) cont.

Verificando a maior correlação entre os grupos:

	(C1,C5,C7)	(C2,C6)	C3	C4
(C1,C5,C7)	1,000	-0,147	0,165	0,087
(C2,C6)	-0,147	1,000	-0,824	0,516
C3	0,165	0,000	1,000	-0,824
C4	→ 0,087 →	0,699	-0,824	1,000



# Método Hierárquico

## ❑ Ligação Simples: exemplo (5/6)

Formando o grupo (2,6,4):

	(C1,C5,C7)	(C2,C6,C4)	C3
(C1,C5,C7)	1,000	0,087	0,165
(C2,C6,C4)	0,087	1,000	-0,824
C3	0,165	0,000	1,000

# Método Hierárquico

## ❑ Ligação Simples: exemplo (5/6) cont.

Verificando a maior correlação entre os grupos:

	(C1,C5,C7)	(C2,C6,C4)	C3
(C1,C5,C7)	1,000	0,087	0,165
(C2,C6,C4)	0,087	1,000	-0,824
C3	0,165	0,000	1,000

# Método Hierárquico

## ❑ Ligação Simples: exemplo (6/6)

Formando o grupo (1,5,7,3):

	(C1,C5,C7,C3)	(C2,C6,C4)
(C1,C5,C7,C3)	1,000	0,087
(C2,C6,C4)	0,087	1,000

Formando o grupo (1,5,7,3, 2,6,4):

	(C1,C5,C7,C3) (C2,C6,C4)
(C1,C5,C7,C3) (C2,C6,C4)	1,000

# Métodos Hierárquicos

## ❑ Características:

- ❖ Abordagem aglomerativa.
- ❖ Determinístico.

## ❑ Dificuldades:

- ❖ A estrutura é sempre uma árvore.
- ❖ Os objetos só podem ser agrupados baseando-se em decisões locais, que, uma vez tomadas, não podem ser re-avaliadas.

# Dúvidas ...



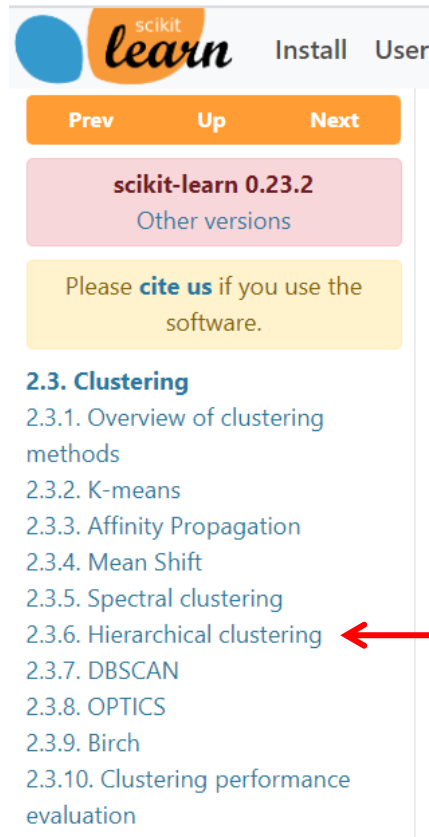
# Hierárquico

❑ Exemplo: dataset (pré-processado)

Nº	Idade	Genero_F	ECivil_D	ECivil_S	ECivil_V	ECivil_C	Filhos	Escola_S	Escola_M	Escola_F	Escola_P	CCorrente_N	Renda	CCredito_N	IProprio_N
1	0,43	0	1	0	0	0	0,50	1	0	0	0	0	0,46	0	0
2	0,30	1	0	1	0	0	0,00	0	1	0	0	1	0,29	0	1
3	1,00	0	0	0	1	0	1,00	0	0	1	0	0	1,00	0	1
4	0,03	1	0	0	0	1	0,50	1	0	0	0	1	0,08	1	0
5	0,77	1	0	0	0	1	0,25	1	0	0	0	0	0,23	0	0
6	0,57	0	0	0	0	1	0,75	0	1	0	0	1	0,25	0	1
7	0,13	1	0	1	0	0	0,25	1	0	0	0	0	0,37	0	1
8	0,23	1	0	0	0	1	0,75	0	0	0	1	1	0,73	0	0
9	0,37	0	1	0	0	0	0,00	1	0	0	0	0	0,52	1	1
10	0,00	0	0	1	0	0	0,00	0	1	0	0	1	0,00	1	1

# Hierárquico Aglomerativo

## ❑ Utilizando Hierarchical clustering (scikit-learn):



The screenshot shows the scikit-learn documentation sidebar. At the top is the scikit-learn logo and navigation links for 'Install' and 'User'. Below are buttons for 'Prev', 'Up', and 'Next'. A pink box highlights 'scikit-learn 0.23.2' with a link to 'Other versions'. A yellow box contains the text 'Please cite us if you use the software.' Below these are links to various clustering topics. A red arrow points to '2.3.6. Hierarchical clustering'.

scikit-learn Install User

Prev Up Next

scikit-learn 0.23.2  
Other versions

Please [cite us](#) if you use the software.

**2.3. Clustering**

- 2.3.1. Overview of clustering methods
- 2.3.2. K-means
- 2.3.3. Affinity Propagation
- 2.3.4. Mean Shift
- 2.3.5. Spectral clustering
- 2.3.6. Hierarchical clustering**
- 2.3.7. DBSCAN
- 2.3.8. OPTICS
- 2.3.9. Birch
- 2.3.10. Clustering performance evaluation

### 2.3.6. Hierarchical clustering

Hierarchical clustering is a general family of clustering algorithms that build nested clusters by merging or splitting them successively. This hierarchy of clusters is represented as a tree (or dendrogram). The root of the tree is the unique cluster that gathers all the samples, the leaves being the clusters with only one sample. See the [Wikipedia page](#) for more details.

The `AgglomerativeClustering` object performs a hierarchical clustering using a bottom up approach: each observation starts in its own cluster, and clusters are successively merged together. The linkage criteria determines the metric used for the merge strategy:

- **Ward** minimizes the sum of squared differences within all clusters. It is a variance-minimizing approach and in this sense is similar to the k-means objective function but tackled with an agglomerative hierarchical approach.
- **Maximum** or **complete linkage** minimizes the maximum distance between observations of pairs of clusters.
- **Average linkage** minimizes the average of the distances between all observations of pairs of clusters.
- **Single linkage** minimizes the distance between the closest observations of pairs of clusters.

`AgglomerativeClustering` can also scale to large number of samples when it is used jointly with a connectivity matrix, but is computationally expensive when no connectivity constraints are added between samples: it considers at each step all the possible merges.

# Hierárquico

- ❑ Abrir dataset **PessoaNormBinary.csv**;
- ❑ Utilizar o algoritmo Hierárquico Aglomerativo
  - ❖ numClusters = 3;
  - ❖ linkType = COMPLETE;
  - ❖ distanceFunction = EuclideanDistance.

<https://www.dropbox.com/sh/fhkqy2wybxjl0n5/AAABevgbnnM4HSdPgeUU6tgPa?dl=0>



# Hierárquico Aglomerativo

## ❑ Utilizando Hierarchical clustering (scikit-learn):

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from google.colab import files
import io

from sklearn.cluster import AgglomerativeClustering #Hierarchical
from sklearn import metrics

uploaded = files.upload()
dados = pd.read_csv(io.BytesIO(uploaded['PessoaNormBinary.csv']))
dados.info()
dados.head()

# ## Hierarchical - Linkage = complete
cluster = AgglomerativeClustering(n_clusters=3, affinity='euclidean', linkage='complete')
cluster.fit(dados)

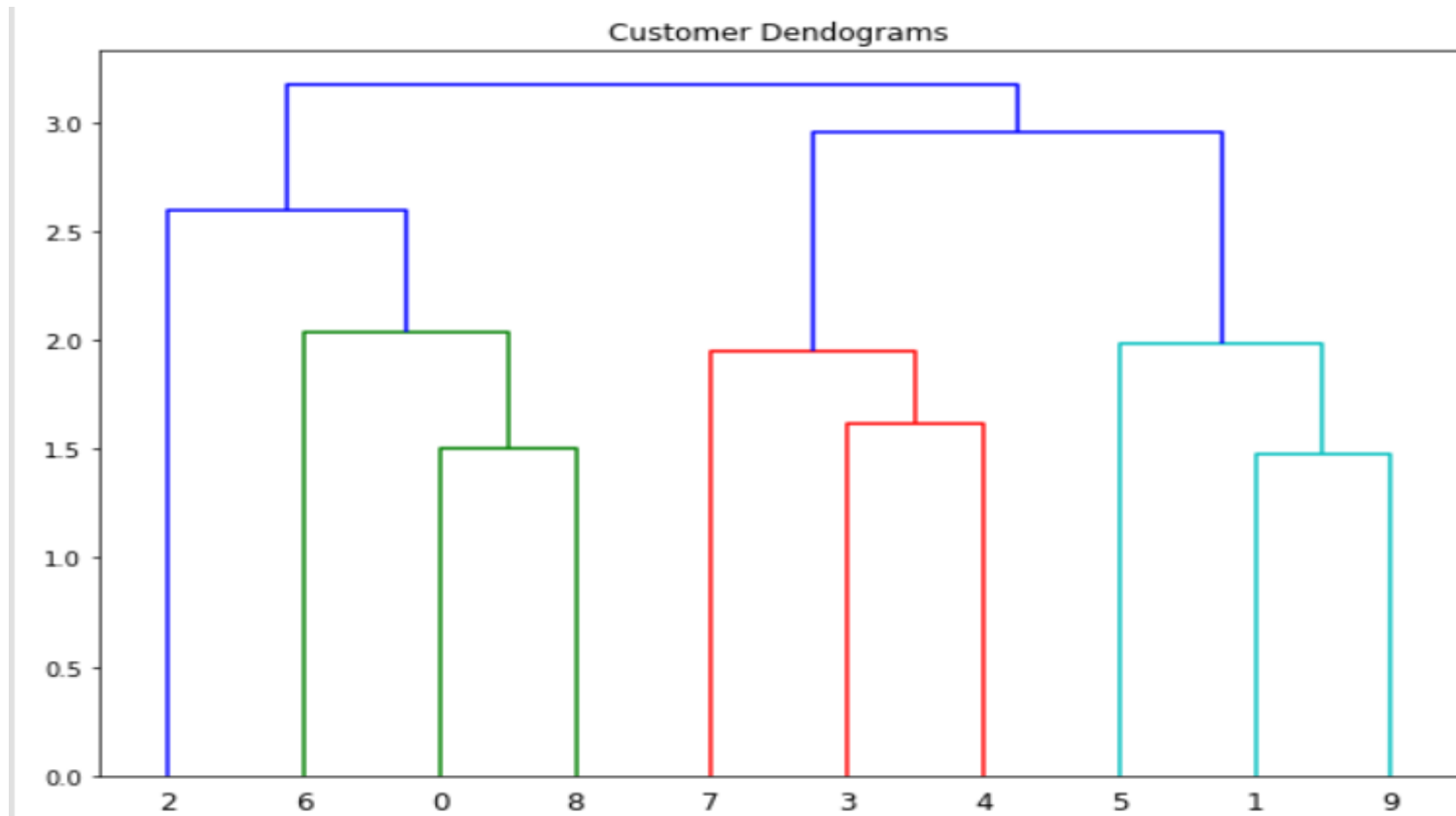
import scipy.cluster.hierarchy as shc

plt.figure(figsize=(10, 7))
plt.title("Customer Dendograms")
dend = shc.dendrogram(shc.linkage(dados, method='complete'))
```

Hierarquico.py

# Hierárquico Aglomerativo

❑ Utilizando Hierarchical clustering (scikit-learn):



# Hierárquico Aglomerativo

## ❑ Utilizando Hierarchical clustering (scikit-learn):

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from google.colab import files
import io

from sklearn.cluster import AgglomerativeClustering #Hierarchical
#from sklearn import metrics

uploaded = files.upload()
dados = pd.read_csv(io.BytesIO(uploaded['PessoaNormBinary.csv']))
dados.info()
#dados.head()

# ## Hierarchical - Linkage = complete
ahc = AgglomerativeClustering(n_clusters=3, affinity='euclidean', linkage='complete')
ahc.fit(dados)
ahc.fit_predict(dados)

# imprimir os labels
print(ahc.labels_)

# juntando os labels com o restante do dataset
dados["cluster"] = ahc.labels_

# Visualização dos novos atributos
dados.head()
```

Hierarquico02.py

# Hierárquico Aglomerativo

❑ Utilizando Hierarchical clustering (scikit-learn):

Escolaridade_P	Correntista_N	RendaBruta	CartaoCredito_N	ImovelProprio_N	cluster
0	0	0.456522	0	0	0
0	1	0.293478	0	1	1
0	0	1.000000	0	1	0
0	1	0.076087	1	0	2
0	0	0.228261	0	0	2

# Hierárquico

- ❑ Abrir dataset **PessoaNormBinary.csv**;
- ❑ Utilizar o algoritmo Hierárquico Aglomerativo
  - ❖ numClusters = 3;
  - ❖ linkType = COMPLETE;
  - ❖ distanceFunction = EuclideanDistance.
- ❑ Salvar arquivo resultante:
  - ❖ PessoaNorm\_Hiera3k\_cLink.csv

# Hierárquico Aglomerativo

## ❑ Utilizando Hierarchical clustering (scikit-learn):

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from google.colab import files
import io
from sklearn.cluster import AgglomerativeClustering #Hierarchical

uploaded = files.upload()
dados = pd.read_csv(io.BytesIO(uploaded['PessoaNormBinary.csv']))
dados.info()

# ## Hierarchical - Linkage = complete
ahc = AgglomerativeClustering(n_clusters=3, affinity='euclidean', linkage='complete')
ahc.fit(dados)
ahc.fit_predict(dados)

# juntando os labels com o restante do dataset
dados["Cluster"] = ahc.labels_
dados["Cluster"] = 'cluster' + dados["Cluster"].astype(str)

# Visualização dos novos atributos
dados.head()

# Salvando Pessoa.csv transformado
df = pd.DataFrame(dados)
df.to_csv('PessoaNorm_Hiera3k_cLink.csv')

# Importando arquivo transformado
from google.colab import files
files.download('PessoaNorm_Hiera3k_cLink.csv')
```

Hierarquico03.py

# Hierárquico Aglomerativo

## Utilizando Hierarchical clustering (scikit-learn):

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
	Idade	Genero_F	EstadoCivil_D	EstadoCivil_S	EstadoCivil_V	EstadoCivil_C	Filhos	Escolaridade_S	Escolaridade_M	Escolaridade_F	Escolaridade_P	Correntista_N	RendaBruta	CartaoCredito_N	ImovelProprio_N	Cluster
	00.433333	0	1	0	0	0	0.5	1	0	0	0	0	00.456522	0	0	0cluster0
	10.3	1	0	1	0	0	0.0	0	1	0	0	0	10.293478	0	0	1cluster1
	21.0	0	0	0	1	0	0.0	0	0	1	0	0	01.0	0	0	1cluster0
	30.033333	1	0	0	0	0	10.5	1	0	0	0	0	10.076087	1	0	0cluster2
	40.766667	1	0	0	0	0	10.25	1	0	0	0	0	00.228261	0	0	0cluster2
	50.566667	0	0	0	0	0	10.75	0	1	0	0	0	10.25	0	0	1cluster1
	60.133333	1	0	1	0	0	0.25	1	0	0	0	0	00.369565	0	0	1cluster0
	70.233333	1	0	0	0	0	10.75	0	0	0	1	0	10.728261	0	0	0cluster2
	80.366667	0	1	0	0	0	0.0	1	0	0	0	0	00.521739	1	0	1cluster0
	90.0	0	0	1	0	0	0.0	0	1	0	0	0	10.0	1	0	1cluster1

Este Computador > Downloads

Nome

Hoje

Fundos

PessoaNorm\_Hiera3k\_cLink

ProjetoFinal\_LP-II\_TipoA\_2022-2

ProjetoFinal\_LP-II\_TipoB\_2022-2

ProjetoFinal\_LP-II\_TipoC\_2022-2

# Hierárquico

- ❑ Abrir dataset **PessoaNormBinary.csv**
- ❑ Utilizar o algoritmo Hierárquico Aglomerativo
  - ❖ numClusters = 3;
  - ❖ linkType = SINGLE; e
  - ❖ linkType = AVERAGE;
  - ❖ distanceFunction = EuclideanDistance.
- ❑ Salvar os arquivos resultantes:
  - ❖ PessoaNormBinary\_Ha-sLink-3k.csv; e
  - ❖ PessoaNormBinary\_Ha-avLink-3k.csv



# Validação de Agrupamentos

# Validação de Agrupamentos

- ❑ Para a análise de agrupamentos, a questão é:
  - ❖ Como avaliar a “**qualidade**” dos grupos resultantes?
- ❑ Por que avaliá-los?
  - ❖ Comparar **diferentes algoritmos** de agrupamento.
  - ❖ Comparar **duas partições**.
  - ❖ Comparar **dois grupos** (clusters).

# Medidas para Validação

□ Medidas numéricas aplicadas para avaliar aspectos da validação de agrupamentos.

❖ Índices Externos:

- Avaliam o agrupamento gerado baseado em uma estrutura pré-especificada (conjunto de dados).
  - Índice Rand ajustado (adjusted Rand) e índice de Jaccard.

❖ Índices Internos:

- Medem a qualidade de um agrupamento usando apenas os dados originais (instâncias ou matriz de similaridade).
  - Índice Davies-Bouldin, Silhuetas, Índice Dunn, ...

# Índices Internos

- ❑ Não há um **rotulo** (classe) para os dados.
- ❑ Medem a qualidade de um agrupamento usando apenas os dados originais.
  - ❖ Utilizam alguma medida de similaridade (**compacticidade**).
- ❑ Principais medidas:
  - ❖ Índice Davies-Bouldin,
  - ❖ Silhuetas,
  - ❖ Índice Dunn, ...

---

# Davies-Bouldin (DB)

---

# Índice Davies-Bouldin (DB)

- Dada uma partição  $\{C_1, C_2, \dots, C_k\}$ , definimos a **similaridade relativa (RS)** entre dois grupos,  $C_i$  e  $C_j$ ,

como:

$$RS_{i,j} = \frac{E_i + E_j}{d(m_i, m_j)}$$

$$E_i = \frac{1}{C_i} \sum_{x \in C_i} (x - z_i)^2$$

Onde:

- $d(m_i, m_j)$  é a distância entre as médias do grupo  $i$  e grupo  $j$ ;
- $E_i$  é a distância quadrada média dos pontos no  $i$ -ésimo grupo para o centroide (média) desse grupo.

# Índice Davies-Bouldin (DB)

- Com  $RS_{i,j}$  podemos calcular a similaridade relativa máxima entre o grupo  $i$  e cada um dos outros ( $MRS_i$ ):

$$MRS_i = \max_{i,j} \{ RS_{i,j} \}$$

- O índice **Davies-Bouldin** (DB) para a partição  $\{C_1, C_2 \dots C_k\}$  é a média de  $MRS_i$  ( $i = 1, 2 \dots k$ ):

$$DB(k) = \frac{1}{k} \sum_{i=1}^k MRS_i$$

# Índice Davies-Bouldin (DB)

## ❑ Partição resultante:

#	Idade	Genero_F	ECivil_D	ECivil_S	ECivil_V	ECivil_C	Filhos	Escola_S	Escola_M	Escola_F	Escola_P	CCorrente_N	Renda	CCredito_N	IProprio_N	Grupo
1	0,43	0	1	0	0	0	0,50	1	0	0	0	0	0,46	0	0	cluster1
2	0,30	1	0	1	0	0	0,00	0	1	0	0	1	0,29	0	1	cluster2
3	1,00	0	0	0	1	0	1,00	0	0	1	0	0	1,00	0	1	cluster1
4	0,03	1	0	0	0	1	0,50	1	0	0	0	1	0,08	1	0	cluster3
5	0,77	1	0	0	0	1	0,25	1	0	0	0	0	0,23	0	0	cluster3
6	0,57	0	0	0	0	1	0,75	0	1	0	0	1	0,25	0	1	cluster2
7	0,13	1	0	1	0	0	0,25	1	0	0	0	0	0,37	0	1	cluster1
8	0,23	1	0	0	0	1	0,75	0	0	0	1	1	0,73	0	0	cluster3
9	0,37	0	1	0	0	0	0,00	1	0	0	0	0	0,52	1	1	cluster1
10	0,00	0	0	1	0	0	0,00	0	1	0	0	1	0,00	1	1	cluster2

## ❑ Partição resultante ordenada:

#	Idade	Genero_F	ECivil_D	ECivil_S	ECivil_V	ECivil_C	Filhos	Escola_S	Escola_M	Escola_F	Escola_P	CCorrente_N	Renda	CCredito_N	IProprio_N	Grupo
1	0,43	0	1	0	0	0	0,50	1	0	0	0	0	0,46	0	0	cluster1
3	1,00	0	0	0	1	0	1,00	0	0	1	0	0	1,00	0	1	cluster1
7	0,13	1	0	1	0	0	0,25	1	0	0	0	0	0,37	0	1	cluster1
9	0,37	0	1	0	0	0	0,00	1	0	0	0	0	0,52	1	1	cluster1
2	0,30	1	0	1	0	0	0,00	0	1	0	0	1	0,29	0	1	cluster2
6	0,57	0	0	0	0	1	0,75	0	1	0	0	1	0,25	0	1	cluster2
10	0,00	0	0	1	0	0	0,00	0	1	0	0	1	0,00	1	1	cluster2
4	0,03	1	0	0	0	1	0,50	1	0	0	0	1	0,08	1	0	cluster3
5	0,77	1	0	0	0	1	0,25	1	0	0	0	0	0,23	0	0	cluster3
8	0,23	1	0	0	0	1	0,75	0	0	0	1	1	0,73	0	0	cluster3



# Índice Davies-Bouldin (DB)

## Entendendo o índice:

#	Idade	Gen_F	ECivil_D	ECivil_S	ECivil_V	ECivil_C	Filhos	Escola_S	Escola_M	Escola_F	Escola_P	CC_N	Renda	CCredito_N	Imovel_N	Grupo
1	0,43	0	1	0	0	0	0,50	1	0	0	0	0	0,46	0	0	cluster1
3	1,00	0	0	0	1	0	1,00	0	0	1	0	0	1,00	0	1	cluster1
7	0,13	1	0	1	0	0	0,25	1	0	0	0	0	0,37	0	1	cluster1
9	0,37	0	1	0	0	0	0,00	1	0	0	0	0	0,52	1	1	cluster1
2	0,30	1	0	1	0	0	0,00	0	1	0	0	1	0,29	0	1	cluster2
6	0,57	0	0	0	0	1	0,75	0	1	0	0	1	0,25	0	1	cluster2
10	0,00	0	0	1	0	0	0,00	0	1	0	0	1	0,00	1	1	cluster2
4	0,03	1	0	0	0	1	0,50	1	0	0	0	1	0,08	1	0	cluster3
5	0,77	1	0	0	0	1	0,25	1	0	0	0	0	0,23	0	0	cluster3
8	0,23	1	0	0	0	1	0,75	0	0	0	1	1	0,73	0	0	cluster3

# Índice Davies-Bouldin (DB)

❑ Calculando o **centróide** do grupo1:

#	Idade	Gen_F	ECivil_D	ECivil_S	ECivil_V	ECivil_C	Filhos	Escola_S	Escola_M	Escola_F	Escola_P	CC_N	Renda	CCredito_N	Imovel_N	Grupo
1	0,43	0	1	0	0	0	0,50	1	0	0	0	0	0,46	0	0	cluster1
3	1,00	0	0	0	1	0	1,00	0	0	1	0	0	1,00	0	1	cluster1
7	0,13	1	0	1	0	0	0,25	1	0	0	0	0	0,37	0	1	cluster1
9	0,37	0	1	0	0	0	0,00	1	0	0	0	0	0,52	1	1	cluster1

centróide =>	0,48	0,25	0,50	0,25	0,25	0,00	0,44	0,75	0,00	0,25	0,00	0,00	0,59	0,25	0,75	:
--------------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	---

# Índice Davies-Bouldin (DB)

❑ Calculando o **centróide** do grupo2:

#	Idade	Gen_F	ECivil_D	ECivil_S	ECivil_V	ECivil_C	Filhos	Escola_S	Escola_M	Escola_F	Escola_P	CC_N	Renda	CCredito_N	Imovel_N	Grupo
2	0,30	1	0	1	0	0	0,00	0	1	0	0	1	0,29	0	1	cluster2
6	0,57	0	0	0	0	1	0,75	0	1	0	0	1	0,25	0	1	cluster2
10	0,00	0	0	1	0	0	0,00	0	1	0	0	1	0,00	1	1	cluster2

<b>centróide =&gt;</b>	0,29	0,33	0,00	0,67	0,00	0,33	0,25	0,00	1,00	0,00	0,00	1,00	0,18	0,33	1,00	
------------------------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	--

# Índice Davies-Bouldin (DB)

❑ Calculando o **centróide** do grupo3:

#	Idade	Gen_F	ECivil_D	ECivil_S	ECivil_V	ECivil_C	Filhos	Escola_S	Escola_M	Escola_F	Escola_P	CC_N	Renda	CCredito_N	Imovel_N	Grupo
4	0,03	1	0	0	0	1	0,50	1	0	0	0	1	0,08	1	0	cluster3
5	0,77	1	0	0	0	1	0,25	1	0	0	0	0	0,23	0	0	cluster3
8	0,23	1	0	0	0	1	0,75	0	0	0	1	1	0,73	0	0	cluster3

<b>centróide =&gt;</b>	0,34	1,00	0,00	0,00	0,00	0,00	1,00	0,50	0,67	0,00	0,00	0,33	0,67	0,34	0,33	0,00
------------------------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------

# Índice Davies-Bouldin (DB)

Calculando os  $E_{is}$ : 
$$E_i = \frac{1}{C_i} \sum_{x \in C_i} (x - z_i)^2$$

<b>centróide =&gt;</b>	0,48	0,25	0,50	0,25	0,25	0,00	0,44	0,75	0,00	0,25	0,00	0,00	0,59	0,25	0,75	Dist.QuaM	E1
	0,05	0,25	0,50	0,25	0,25	0,00	0,06	0,25	0,00	0,25	0,00	0,00	0,13	0,25	0,75	8,9577	
	0,52	0,25	0,50	0,25	0,75	0,00	0,56	0,75	0,00	0,75	0,00	0,00	0,41	0,25	0,25	27,4808	
	0,35	0,75	0,50	0,75	0,25	0,00	0,19	0,25	0,00	0,25	0,00	0,00	0,22	0,25	0,25	16,0392	
	0,12	0,25	0,50	0,25	0,25	0,00	0,44	0,25	0,00	0,25	0,00	0,00	0,07	0,75	0,25	11,3527	<b>15,9576</b>
<b>centróide =&gt;</b>	0,29	0,33	0,00	0,67	0,00	0,33	0,25	0,00	1,00	0,00	0,00	1,00	0,18	0,33	1,00	Dist.QuaM	E2
	0,01	0,67	0,00	0,33	0,00	0,33	0,25	0,00	0,00	0,00	0,00	0,00	0,11	0,33	0,00	4,1620	
	0,28	0,33	0,00	0,67	0,00	0,67	0,50	0,00	0,00	0,00	0,00	0,00	0,07	0,33	0,00	8,1032	
	0,29	0,33	0,00	0,33	0,00	0,33	0,25	0,00	0,00	0,00	0,00	0,00	0,18	0,67	0,00	5,6964	<b>5,9872</b>
<b>centróide =&gt;</b>	0,34	1,00	0,00	0,00	0,00	1,00	0,50	0,67	0,00	0,00	0,33	0,67	0,34	0,33	0,00	Dist.QuaM	E3
	0,31	0,00	0,00	0,00	0,00	0,00	0,00	0,33	0,00	0,00	0,33	0,33	0,27	0,67	0,00	5,0440	
	0,42	0,00	0,00	0,00	0,00	0,00	0,25	0,33	0,00	0,00	0,33	0,67	0,12	0,33	0,00	6,0262	
	0,11	0,00	0,00	0,00	0,00	0,00	0,25	0,67	0,00	0,00	0,67	0,33	0,38	0,33	0,00	7,5360	<b>6,2021</b>

# Índice Davies-Bouldin (DB)

Calculando os  $RS_s$ :  $RS_{i,j} = \frac{E_i + E_j}{d(m_i, m_j)}$

	Centróides calculos anteriormente														
1	0,48	0,25	0,50	0,25	0,25	0,00	0,44	0,75	0,00	0,25	0,00	0,00	0,59	0,25	0,75
2	0,29	0,33	0,00	0,67	0,00	0,33	0,25	0,00	1,00	0,00	0,00	1,00	0,18	0,33	1,00
3	0,34	1,00	0,00	0,00	0,00	1,00	0,50	0,67	0,00	0,00	0,33	0,67	0,34	0,33	0,00



																Soma	RS
[1,2]	0,19	0,08	0,50	0,42	0,25	0,33	0,19	0,75	1,00	0,25	0,00	1,00	0,41	0,08	0,25	5,7044	3,8470
[1,3]	0,14	0,75	0,50	0,25	0,25	1,00	0,06	0,08	0,00	0,25	0,33	0,67	0,24	0,08	0,75	5,3608	4,1336
[2-3]	0,06	0,67	0,00	0,67	0,00	0,67	0,25	0,67	1,00	0,00	0,33	0,33	0,16	0,00	1,00	5,8019	2,1009



$$\begin{aligned}
 [1-2] &= (15,9576 + 5,9872) / 5,7044 = 3,8470 \\
 [1-3] &= (15,9576 + 6,2021) / 5,3608 = 4,1336 \\
 [2-3] &= (5,9872 + 6,2021) / 5,8019 = 2,1009
 \end{aligned}$$

# Índice Davies-Bouldin (DB)

□ Calculando os  $MRS_s$ :  $MRS_i = \max_{i,j} \{RS_{i,j}\}$

$$\begin{aligned}[1-2] &= (15,9576 + 5,9872) / 5,7044 = 3,8470 \\ [1-3] &= (15,9576 + 6,2021) / 5,3608 = 4,1336 \\ [2-3] &= (5,9872 + 6,2021) / 5,8019 = 2,1009\end{aligned}$$



$$\begin{aligned}\text{Máximo entre } [1-2] \text{ e } [1-3] &= [3,8470 : \mathbf{4,1336}] \Rightarrow 4,1336 \\ \text{Máximo entre } [2-1] \text{ e } [2-3] &= [\mathbf{3,8470} : 2,1009] \Rightarrow 3,8470 \\ \text{Máximo entre } [3-1] \text{ e } [3-2] &= [\mathbf{4,1336} : 2,1009] \Rightarrow 4,1336\end{aligned}$$

# Índice Davies-Bouldin (DB)

❑ Calculando o **DB**:

$$DB(k) = \frac{1}{k} \sum_{i=1}^k MRS_i$$

$$MRS_1 = 4,1336$$

$$MRS_2 = 3,8470$$

$$MRS_3 = 4,1336$$

$$k = 3$$

$$\text{Média} = (4,1336 + 3,8470 + 4,1336) / 3$$

$$DB = 4,0381$$



# Dúvidas ...



# Índice Davies-Bouldin (DB)

## ❑ Calculando o **DB** (Scikit-Learn):

### 2.3.10.7. Davies-Bouldin Index

If the ground truth labels are not known, the Davies-Bouldin index (`sklearn.metrics.davies_bouldin_score`) can be used to evaluate the model, where a lower Davies-Bouldin index relates to a model with better separation between the clusters.

This index signifies the average 'similarity' between clusters, where the similarity is a measure that compares the distance between clusters with the size of the clusters themselves.

Zero is the lowest possible score. Values closer to zero indicate a better partition.

In normal usage, the Davies-Bouldin index is applied to the results of a cluster analysis as follows:

```
>>> from sklearn import datasets
>>> iris = datasets.load_iris()
>>> X = iris.data
>>> from sklearn.cluster import KMeans
>>> from sklearn.metrics import davies_bouldin_score
>>> kmeans = KMeans(n_clusters=3, random_state=1).fit(X)
>>> labels = kmeans.labels_
>>> davies_bouldin_score(X, labels)
0.6619...
```

<https://scikit-learn.org/stable/modules/clustering.html>

# Índice Davies-Bouldin (DB)

## ❑ Calculando o **DB** (Scikit-Learn):

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from google.colab import files
import io

from sklearn.cluster import AgglomerativeClustering #Hierarchical
from sklearn import metrics

uploaded = files.upload()
dados = pd.read_csv(io.BytesIO(uploaded['PessoaNormBinary.csv']))

# ## Hierarchical - Linkage = complete
cluster = AgglomerativeClustering(n_clusters=2, affinity='euclidean', linkage='complete')
cluster.fit(dados)

# DB
print("Davies Bouldin Coefficient (k = 2, seed = 10): %0.3f"
      % metrics.davies_bouldin_score(dados, cluster.labels_))

# ## Hierarchical - Linkage = complete
cluster = AgglomerativeClustering(n_clusters=3, affinity='euclidean', linkage='complete')
cluster.fit(dados)

# DB
print("Davies Bouldin Coefficient (k = 3, seed = 10): %0.3f"
      % metrics.davies_bouldin_score(dados, cluster.labels_))
```

```
• PessoaNormBinary.csv(text/csv) - 625 bytes, last modified: 09/03/2019 - 100% done
Saving PessoaNormBinary.csv to PessoaNormBinary (4).csv
Davies Bouldin Coefficient (k = 2, seed = 10): 1.821
Davies Bouldin Coefficient (k = 3, seed = 10): 1.311
```

DB\_Hierarquico03.py

# Índice Davies-Bouldin (DB)

## ❑ Calculando o **DB** (Scikit-Learn):

```
import pandas as pd
from google.colab import files
import io

from sklearn.cluster import AgglomerativeClustering #Hierarchical
from sklearn import metrics

uploaded = files.upload()
dados = pd.read_csv(io.BytesIO(uploaded['NurseryNominalToBinary_NoClass.csv']))

for i in [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]:
    cluster = AgglomerativeClustering(n_clusters=i, affinity='euclidean', linkage='complete')
    cluster.fit(dados)

    print("Davies Bouldin Coefficient (k = %d): %0.3f" % (i, metrics.davies_bouldin_score(dados, cluster.labels_)))
```

DB\_Hierarquico02.py

# Índice Davies-Bouldin (DB)

## ❑ Calculando o **DB** (Scikit-Learn):

Escolher arquivos NurseryNo...NoClass.csv

- **NurseryNominalToBinary\_NoClass.csv**(text/csv) - 687306 bytes, last modified: 20/10/2022 - 100% done

Saving NurseryNominalToBinary\_NoClass.csv to NurseryNominalToBinary\_NoClass.csv

```
Davies Bouldin Coefficient (k = 2): 3.576
Davies Bouldin Coefficient (k = 3): 3.017
Davies Bouldin Coefficient (k = 4): 3.323
Davies Bouldin Coefficient (k = 5): 3.305
Davies Bouldin Coefficient (k = 6): 3.182
Davies Bouldin Coefficient (k = 7): 2.818
Davies Bouldin Coefficient (k = 8): 2.990
Davies Bouldin Coefficient (k = 9): 3.009
Davies Bouldin Coefficient (k = 10): 2.952
Davies Bouldin Coefficient (k = 11): 2.711
Davies Bouldin Coefficient (k = 12): 2.887
Davies Bouldin Coefficient (k = 13): 2.914
Davies Bouldin Coefficient (k = 14): 2.889
Davies Bouldin Coefficient (k = 15): 2.739
Davies Bouldin Coefficient (k = 16): 2.791
Davies Bouldin Coefficient (k = 17): 2.774
Davies Bouldin Coefficient (k = 18): 2.674
Davies Bouldin Coefficient (k = 19): 2.722
Davies Bouldin Coefficient (k = 20): 2.711
```



# Silhouette



# Silhouette

- ❑ O índice Silhouette combina a ideia de coesão e separação. Para cada instância  $i$ .
- ❑ Seja  $\text{diss}(i, \mathbf{C})$  a dissimilaridade média entre  $i$  e cada elemento no grupo  $\mathbf{C}$ .
- ❑ Seja  $\mathbf{A}$  o grupo ao qual a instância  $i$  pertence:
  - ❖  $a(i)$  distância média de  $i$  para as outras instâncias do seu grupo.
- ❑ Seja  $\mathbf{B}$  um outro grupo tal que  $\text{diss}(i, \mathbf{B})$  é menor de todas:
  - ❖  $b(i)$  distância mínima  $i$  para as instâncias dos outros grupos.

# Silhouette

- A silhouette para a instância  $i$  é dada por:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$



# Silhouette

- ❑ O valor da silhouette de uma instância está no intervalo  $[-1, 1]$ :
  - ❖ Se uma instância está bem **situada** dentro de seu grupo, sua silhouette apresentará um **valor próximo de 1**.
  - ❖ Por outro lado, um **valor próximo de -1**, indica que a instância deveria ser associado a **outro grupo**.

# Silhouette

❑ Calculando as Distâncias entre os grupos (C1-C2):

#	Idade	Gen_F	ECivil_D	ECivil_S	ECivil_V	ECivil_C	Filhos	Escola_S	Escola_M	Escola_F	Escola_P	CC_N	Renda	CCredito_N	Imovel_N	Grupo
1	0,43	0	1	0	0	0	0,50	1	0	0	0	0	0,46	0	0	cluster1
3	1,00	0	0	0	1	0	1,00	0	0	1	0	0	1,00	0	1	cluster1
7	0,13	1	0	1	0	0	0,25	1	0	0	0	0	0,37	0	1	cluster1
9	0,37	0	1	0	0	0	0,00	1	0	0	0	0	0,52	1	1	cluster1
#	Idade	Gen_F	ECivil_D	ECivil_S	ECivil_V	ECivil_C	Filhos	Escola_S	Escola_M	Escola_F	Escola_P	CC_N	Renda	CCredito_N	Imovel_N	Grupo
2	0,30	1	0	1	0	0	0,00	0	1	0	0	1	0,29	0	1	cluster2
6	0,57	0	0	0	0	1	0,75	0	1	0	0	1	0,25	0	1	cluster2
10	0,00	0	0	1	0	0	0,00	0	1	0	0	1	0,00	1	1	cluster2

d1(1)	d1(2)	d1(3)	d1(4)	Média 1
2,701	2,474	2,765		2,647
2,827	2,411	3,000		2,746
1,760	2,540	2,284		2,195
2,656	2,5839	2,325		2,522

# Silhouette

❑ Calculando as Distâncias entre os grupos (C1-C3):

#	Idade	Gen_F	ECivil_D	ECivil_S	ECivil_V	ECivil_C	Filhos	Escola_S	Escola_M	Escola_F	Escola_P	CC_N	Renda	CCredito_N	Imovel_N	Grupo
1	0,43	0	1	0	0	0	0,50	1	0	0	0	0	0,46	0	0	cluster1
3	1,00	0	0	0	1	0	1,00	0	0	1	0	0	1,00	0	1	cluster1
7	0,13	1	0	1	0	0	0,25	1	0	0	0	0	0,37	0	1	cluster1
9	0,37	0	1	0	0	0	0,00	1	0	0	0	0	0,52	1	1	cluster1

#	Idade	Gen_F	ECivil_D	ECivil_S	ECivil_V	ECivil_C	Filhos	Escola_S	Escola_M	Escola_F	Escola_P	CC_N	Renda	CCredito_N	Imovel_N	Grupo
4	0,03	1	0	0	0	1	0,50	1	0	0	0	1	0,08	1	0	cluster3
5	0,77	1	0	0	0	1	0,25	1	0	0	0	0	0,23	0	0	cluster3
8	0,23	1	0	0	0	1	0,75	0	0	0	1	1	0,73	0	0	cluster3

d2(1)	d2(2)	d2(3)	d2(4)	Média 2
2,303	1,796	2,485		2,195
3,168	2,686	2,779		2,878
2,271	1,850	2,528		2,216
2,358	2,304	2,936		2,533

# Silhouette

❑ Calculando as Distâncias entre os grupos (C2-C1):

#	Idade	Gen_F	ECivil_D	ECivil_S	ECivil_V	ECivil_C	Filhos	Escola_S	Escola_M	Escola_F	Escola_P	CC_N	Renda	CCredito_N	Imovel_N	Grupo
1	0,43	0	1	0	0	0	0,50	1	0	0	0	0	0,46	0	0	cluster1
3	1,00	0	0	0	1	0	1,00	0	0	1	0	0	1,00	0	1	cluster1
7	0,13	1	0	1	0	0	0,25	1	0	0	0	0	0,37	0	1	cluster1
9	0,37	0	1	0	0	0	0,00	1	0	0	0	0	0,52	1	1	cluster1
#	Idade	Gen_F	ECivil_D	ECivil_S	ECivil_V	ECivil_C	Filhos	Escola_S	Escola_M	Escola_F	Escola_P	CC_N	Renda	CCredito_N	Imovel_N	Grupo
2	0,30	1	0	1	0	0	0,00	0	1	0	0	1	0,29	0	1	cluster2
6	0,57	0	0	0	0	1	0,75	0	1	0	0	1	0,25	0	1	cluster2
10	0,00	0	0	1	0	0	0,00	0	1	0	0	1	0,00	1	1	cluster2

2,701	2,827	1,760	2,656	2,486
2,474	2,411	2,540	2,584	2,502
2,765	3,000	2,284	2,325	2,594

# Silhouette

❑ Calculando as Distâncias entre os grupos (C2-C3):

#	Idade	Gen_F	ECivil_D	ECivil_S	ECivil_V	ECivil_C	Filhos	Escola_S	Escola_M	Escola_F	Escola_P	CC_N	Renda	CCredito_N	Imovel_N	Grupo
2	0,30	1	0	1	0	0	0,00	0	1	0	0	1	0,29	0	1	cluster2
6	0,57	0	0	0	0	1	0,75	0	1	0	0	1	0,25	0	1	cluster2
10	0,00	0	0	1	0	0	0,00	0	1	0	0	1	0,00	1	1	cluster2
#	Idade	Gen_F	ECivil_D	ECivil_S	ECivil_V	ECivil_C	Filhos	Escola_S	Escola_M	Escola_F	Escola_P	CC_N	Renda	CCredito_N	Imovel_N	Grupo
4	0,03	1	0	0	0	1	0,50	1	0	0	0	1	0,08	1	0	cluster3
5	0,77	1	0	0	0	1	0,25	1	0	0	0	0	0,23	0	0	cluster3
8	0,23	1	0	0	0	1	0,75	0	0	0	1	1	0,73	0	0	cluster3

2,524	2,507	2,399	2,477
2,319	2,300	2,083	2,234
2,501	2,950	2,854	2,769

# Silhouette

❑ Calculando as Distâncias entre os grupos (C3-C1):

#	Idade	Gen_F	ECivil_D	ECivil_S	ECivil_V	ECivil_C	Filhos	Escola_S	Escola_M	Escola_F	Escola_P	CC_N	Renda	CCredito_N	Imovel_N	Grupo
1	0,43	0	1	0	0	0	0,50	1	0	0	0	0	0,46	0	0	cluster1
3	1,00	0	0	0	1	0	1,00	0	0	1	0	0	1,00	0	1	cluster1
7	0,13	1	0	1	0	0	0,25	1	0	0	0	0	0,37	0	1	cluster1
9	0,37	0	1	0	0	0	0,00	1	0	0	0	0	0,52	1	1	cluster1

#	Idade	Gen_F	ECivil_D	ECivil_S	ECivil_V	ECivil_C	Filhos	Escola_S	Escola_M	Escola_F	Escola_P	CC_N	Renda	CCredito_N	Imovel_N	Grupo
4	0,03	1	0	0	0	1	0,50	1	0	0	0	1	0,08	1	0	cluster3
5	0,77	1	0	0	0	1	0,25	1	0	0	0	0	0,23	0	0	cluster3
8	0,23	1	0	0	0	1	0,75	0	0	0	1	1	0,73	0	0	cluster3

2,303	3,1683	2,271	2,358	2,525
1,796	2,6856	1,850	2,304	2,159
2,485	2,7792	2,528	2,936	2,682

# Silhouette


❑ Calculando as Distâncias entre os grupos (C3-C2):

#	Idade	Gen_F	ECivil_D	ECivil_S	ECivil_V	ECivil_C	Filhos	Escola_S	Escola_M	Escola_F	Escola_P	CC_N	Renda	CCredito_N	Imovel_N	Grupo
2	0,30	1	0	1	0	0	0,00	0	1	0	0	1	0,29	0	1	cluster2
6	0,57	0	0	0	0	1	0,75	0	1	0	0	1	0,25	0	1	cluster2
10	0,00	0	0	1	0	0	0,00	0	1	0	0	1	0,00	1	1	cluster2
#	Idade	Gen_F	ECivil_D	ECivil_S	ECivil_V	ECivil_C	Filhos	Escola_S	Escola_M	Escola_F	Escola_P	CC_N	Renda	CCredito_N	Imovel_N	Grupo
4	0,03	1	0	0	0	1	0,50	1	0	0	0	1	0,08	1	0	cluster3
5	0,77	1	0	0	0	1	0,25	1	0	0	0	0	0,23	0	0	cluster3
8	0,23	1	0	0	0	1	0,75	0	0	0	1	1	0,73	0	0	cluster3

2,524	2,319	2,501		2,448
2,507	2,300	2,950		2,586
2,399	2,083	2,854		2,446

# Silhouette

❑ Calculando  $b(i) = \min(m_1, m_2)$ :

Média 1	d2(1)	d2(2)	d2(3)	d2(4)	Média 2				b(i)
2,647	2,303	1,796	2,485		2,195				2,195
2,746	3,168	2,686	2,779		2,878				2,746
2,195	2,271	1,850	2,528		2,216				2,195
2,522	2,358	2,304	2,936		2,533				2,522
2,486	2,524	2,507	2,399		2,477				2,477
2,502	2,319	2,300	2,083		2,234				2,234
2,594	2,501	2,950	2,854		2,769				2,594
2,525	2,524	2,319	2,501		2,448				2,448
2,159	2,507	2,300	2,950		2,586				2,159
2,682	2,399	2,083	2,854		2,446				2,446



# Silhouette

❑ Calculando as Distâncias internas(C1):

#	Idade	Gen_F	ECivil_D	ECivil_S	ECivil_V	ECivil_C	Filhos	Escola_S	Escola_M	Escola_F	Escola_P	CC_N	Renda	CCredito_N	Imovel_N	Grupo
1	0,43	0	1	0	0	0	0,50	1	0	0	0	0	0,46	0	0	cluster1
3	1,00	0	0	0	1	0	1,00	0	0	1	0	0	1,00	0	1	cluster1
7	0,13	1	0	1	0	0	0,25	1	0	0	0	0	0,37	0	1	cluster1
9	0,37	0	1	0	0	0	0,00	1	0	0	0	0	0,52	1	1	cluster1

d1(1)	d1(2)	d1(3)
2,422	2,040	1,503
2,422	2,591	2,575
2,040	2,591	2,035
1,503	2,575	2,035

# Silhouette

❑ Calculando as Distâncias internas(C2):

#	Idade	Gen_F	ECivil_D	ECivil_S	ECivil_V	ECivil_C	Filhos	Escola_S	Escola_M	Escola_F	Escola_P	CC_N	Renda	CCredito_N	Imovel_N	Grupo
2	0,30	1	0	1	0	0	0,00	0	1	0	0	1	0,29	0	1	cluster2
6	0,57	0	0	0	0	1	0,75	0	1	0	0	1	0,25	0	1	cluster2
10	0,00	0	0	1	0	0	0,00	0	1	0	0	1	0,00	1	1	cluster2

1,620	1,878
1,620	1,945
1,878	1,945

# Silhouette

❑ Calculando as Distâncias internas(C3):

#	Idade	Gen_F	ECivil_D	ECivil_S	ECivil_V	ECivil_C	Filhos	Escola_S	Escola_M	Escola_F	Escola_P	CC_N	Renda	CCredito_N	Imovel_N	Grupo
4	0,03	1	0	0	0	1	0,50	1	0	0	0	1	0,08	1	0	cluster3
5	0,77	1	0	0	0	1	0,25	1	0	0	0	0	0,23	0	0	cluster3
8	0,23	1	0	0	0	1	0,75	0	0	0	1	1	0,73	0	0	cluster3

1,620	1,878
1,620	1,945
1,878	1,945

# Silhouette

❑ Calculando  $a(i) = \text{média}(d_1, d_2, d_n)$ :

d1(1)	d1(2)	d1(3)		a(i)
2,422	2,040	1,503	→	1,988
2,422	2,591	2,575		2,529
2,040	2,591	2,035		2,222
1,503	2,575	2,035		2,037
1,907	1,475			1,691
1,907	1,986			1,947
1,475	1,986			1,731
1,620	1,878			1,749
1,620	1,945			1,783
1,878	1,945			1,912

# Silhouette

❑ Calculando  $s(i)$  :

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

a(i)	b(i)	s(i)
1,988	2,195	0,094
2,529	2,746	0,079
2,222	2,195	-0,012
2,037	2,522	0,192
1,691	2,477	0,317
1,947	2,234	0,129
1,731	2,594	0,333
1,749	2,448	0,286
1,783	2,159	0,174
1,912	2,446	0,218

# Silhouette

□ A silhouette média por grupo  $s(C1)$ ,  $s(C2)$ ,  $s(Cn)$ :

s(i)	sC(1)
0,094	0,0882
0,079	
-0,012	
0,192	
0,317	sC(2)
0,129	0,2595
0,333	
0,286	sC(3)
0,174	0,2260
0,218	

# Silhouette

❑ Calculando a Silhouette Global (***G<sub>s</sub>***):

s(i)	sC(1)		
0,094	0,0882	←	
0,079			
-0,012			
0,192			
0,317	sC(2)		
0,129	0,2595	←	
0,333			
0,286	sC(3)		GS(k=3)
0,174	0,2260	←	0,1913
0,218			

# Dúvidas ...





# Silhouette

## ❑ Calculando o Silhouette (Scikit-Learn):

### `sklearn.metrics.silhouette_score`

```
sklearn.metrics.silhouette_score(X, labels, *, metric='euclidean', sample_size=None, random_state=None, **kwargs)
```

[\[source\]](#)

Compute the mean Silhouette Coefficient of all samples.

The Silhouette Coefficient is calculated using the mean intra-cluster distance ( $a$ ) and the mean nearest-cluster distance ( $b$ ) for each sample. The Silhouette Coefficient for a sample is  $(b - a) / \max(a, b)$ . To clarify,  $b$  is the distance between a sample and the nearest cluster that the sample is not a part of. Note that Silhouette Coefficient is only defined if number of labels is  $2 \leq n_{\text{labels}} \leq n_{\text{samples}}$ .

This function returns the mean Silhouette Coefficient over all samples. To obtain the values for each sample, use `silhouette_samples`.

The best value is 1 and the worst value is -1. Values near 0 indicate overlapping clusters. Negative values generally indicate that a sample has been assigned to the wrong cluster, as a different cluster is more similar.

Read more in the [User Guide](#).

<b>Parameters:</b>	<p><b>X</b> : <i>array</i> [<i>n_samples</i>, <i>a</i>, <i>n_samples</i>, <i>a</i>] if <i>metric</i> == "precomputed", or, [<i>n_samples</i>, <i>a</i>, <i>n_features</i>] otherwise Array of pairwise distances between samples, or a feature array.</p> <p><b>labels</b> : <i>array</i>, <i>shape</i> = [<i>n_samples</i>] Predicted labels for each sample.</p> <p><b>metric</b> : <i>string</i>, or <i>callable</i> The metric to use when calculating distance between instances in a feature array. If <i>metric</i> is a string, it must be one of the options allowed by <code>metrics.pairwise.pairwise_distances</code>. If <i>X</i> is the distance array itself, use <code>metric="precomputed"</code>.</p>
--------------------	---

# Silhouette

## ❑ Calculando o **Silhouette** (Scikit-Learn):

```
from sklearn.metrics import silhouette_samples, silhouette_score
from sklearn.cluster import AgglomerativeClustering #Hierarchical
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from google.colab import files
import io

uploaded = files.upload()
dados = pd.read_csv(io.BytesIO(uploaded['NurseryNominalToBinary_NoClass.csv']))

#Hierarquico
fig, ax = plt.subplots(2, 2, figsize=(15,8))
for i in [2, 3, 4, 5]:
    ha = AgglomerativeClustering(n_clusters=i, affinity='euclidean', linkage='complete')
    ha.fit_predict(dados)
    ha.fit(dados)
    indice = silhouette_score(dados, ha.labels_, metric='euclidean')
    print('Hierarquico Complete %dk - Silhouette=> %.3f' % (i, indice))
```

Silhouette\_Hierarquico02.py

# Silhouette

## ❑ Calculando o **Silhouette** (Scikit-Learn):

Escolher arquivos NurseryNo...NoClass.csv

- **NurseryNominalToBinary\_NoClass.csv**(text/csv) - 687306 bytes, last modified: 20/10/2022 - 100% done  
Saving NurseryNominalToBinary\_NoClass.csv to NurseryNominalToBinary\_NoClass (2).csv  
Hierarquico Complete 2k - Silhouette=> 0.065  
Hierarquico Complete 3k - Silhouette=> 0.097  
Hierarquico Complete 4k - Silhouette=> 0.068  
Hierarquico Complete 5k - Silhouette=> 0.063

# Obrigado!!!

