

# Aprendizado de Máquina

Aprendizado Não-supervisionado

Daniel Sabino A. de Araújo

# Aprendizado descritivo

# Aprendizado descritivo

- Aprendizado não supervisionado: identificação de informações relevantes nos dados sem elemento externo para guiar o aprendizado
- Identificação de propriedades intrínsecas aos dados
- Encontrar padrões ou tendências que auxiliem no entendimento dos dados

Não existem atributos meta

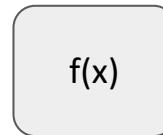
⇒ algoritmo aprende a representar entradas em conjunto de dados  $X$  segundo algum critério de qualidade

$x_{11}$	$x_{12}$	...	$x_{1m}$
$x_{21}$	$x_{22}$	...	$x_{2m}$
...	...	...	...
$x_{n1}$	$x_{n2}$		$x_{nm}$

Técnica de AM  
não  
supervisionado



Modelo  
Descritivo



# Tipos de tarefas descritivas

- Sumarização

- Encontrar descrição simples e compacta para dados
  - Ex. medidas estatísticas (média, desvio-padrão, mínimo), visualização, etc

- Associação

- Encontrar padrões frequentes de associações entre atributos
  - Ex. regras de associação

- Agrupamento

- Identificação de grupos nos dados de acordo com sua similaridade
  - Ex. k-médias, agrupamento hierárquico

# Sumarização

- Meta: encontrar descrição simples e compacta de um conjunto de dados
- Frequentemente usada para:
  - Exploração de dados
  - Geração automática de relatórios

# Sumarização

- Técnicas podem ser divididas em:
  - Simples:
    - Média
    - Mediana
    - Desvio padrão
  - Mais sofisticadas:
    - Regras de sumarização
    - Técnicas de visualização multivariadas

Nome	Idade	Sexo	Altura	Tem_filhos
João	32	M	180	S
Maria	30	F	160	N
Pedro	23	M	---	S
José	45	M	170	S
Sueli	18	F	175	N

Nome	Idade	Sexo	Altura	Tem_filhos
João	32	M	180	S
Maria	30	F	160	N
Pedro	23	M	---	S
José	45	M	170	S
Sueli	18	F	175	N

Menor altura: 160

Sexo mais presente: M

Mediana da idade: 30

Idade média: 29.6

# Associação

- Meta: encontrar um conjunto de associações entre os itens de uma base de dados
  - Também denotada por mineração de padrões frequentes
- Primeiros trabalhos: identificar grupos de produtos frequentemente comprados em conjunto
  - Para auxiliar campanhas de marketing

Transação	Itens comprados
1	Pão, queijo, manteiga, massa
2	Pão, geléia, suco
3	Queijo, arroz, massa
4	Massa, queijo
5	Massa, queijo, pão

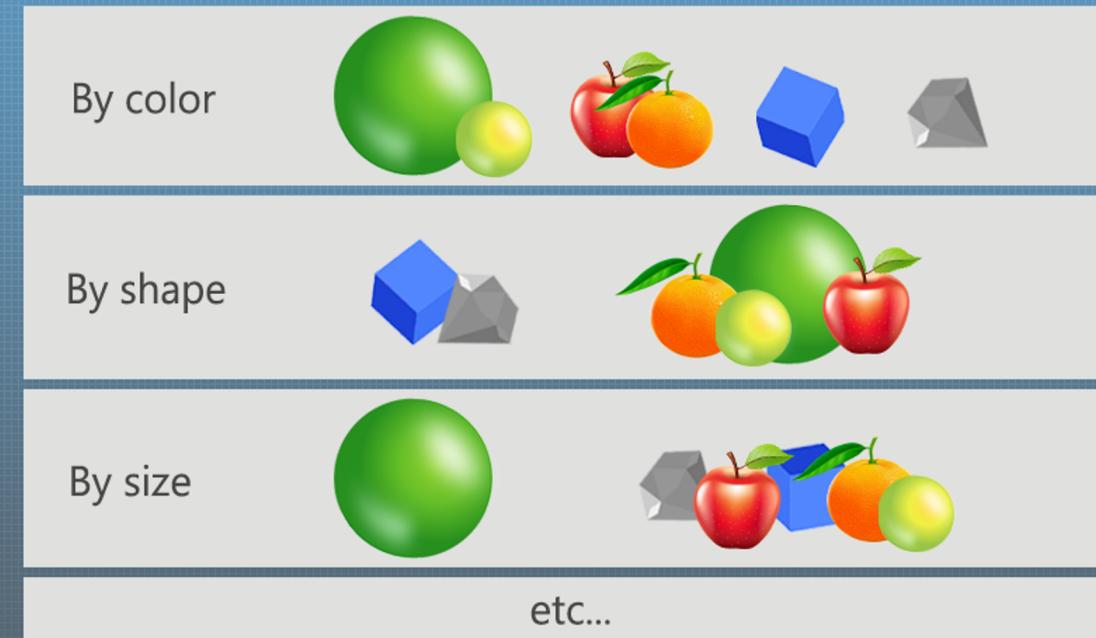
100% dos clientes que compram queijo também compram massa

Transação	Itens comprados
1	Pão, queijo, manteiga, massa
2	Pão, geléia, suco
3	Queijo, arroz, massa
4	Massa, queijo
5	Massa, queijo, pão

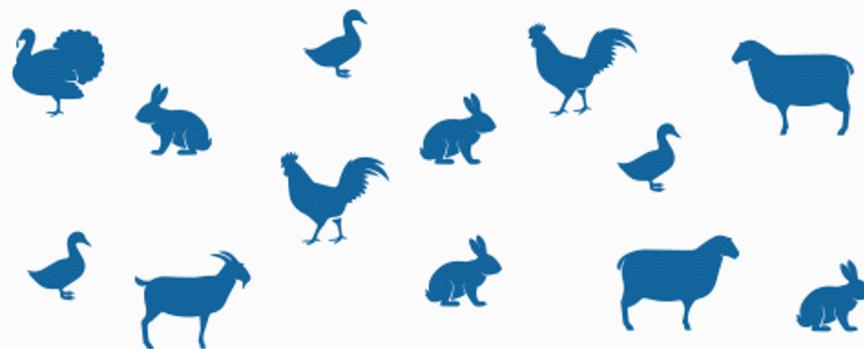
# Agrupamento

- Meta: identificar grupos nos dados de acordo com similaridade entre os objetos
  - Instrumentos valiosos na análise exploratória de dados
  - Apropriado para explorar e verificar estruturas presentes
  - Têm aplicações práticas em várias áreas
    - Bioinformática: identificar genes com padrão de expressão semelhante
    - Marketing: identificar clientes com perfil semelhante, segmentação de mercado
    - Visão computacional: segmentação de imagens

# Machine Learning: Clustering



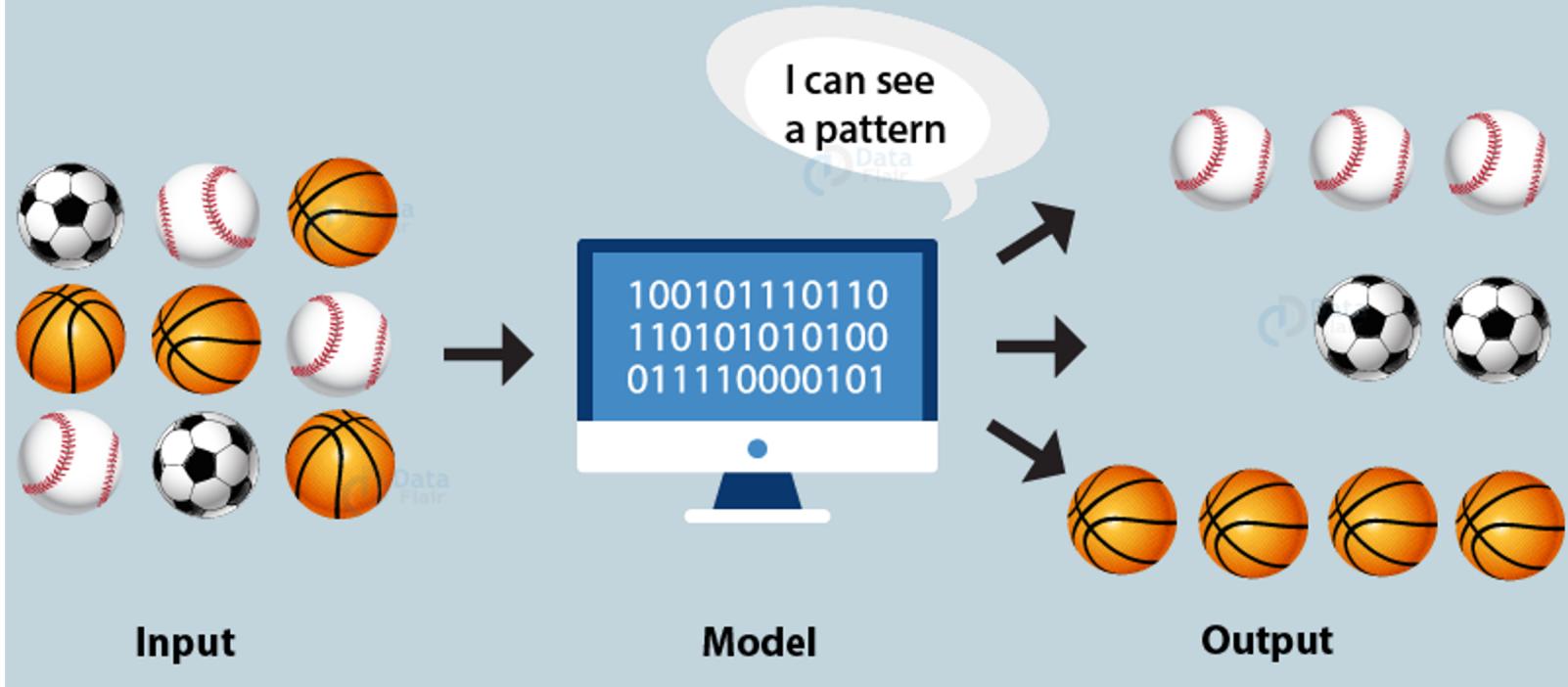
**ENSTOA**



Classification

Clustering

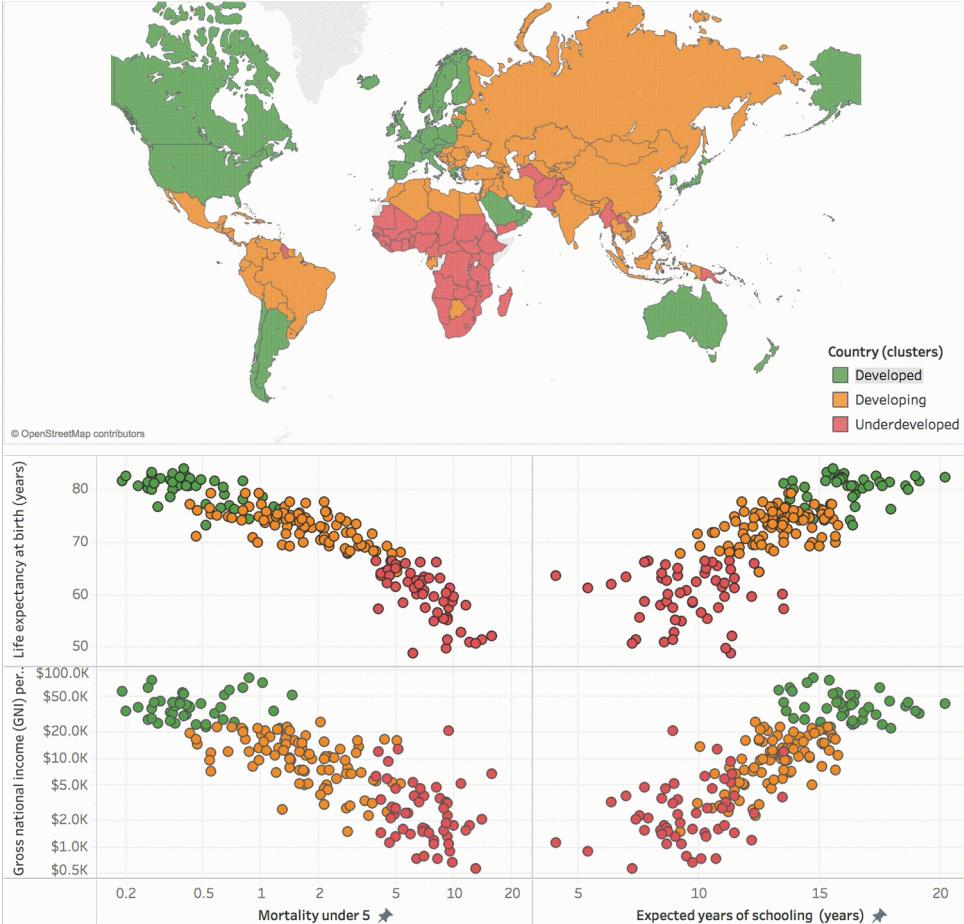
# Introduction to Clustering



# Análise de agrupamento

# Análise de Agrupamentos

- Objetivo de técnica de agrupamento: encontrar uma estrutura de clusters (grupos) nos dados
  - Objetos em cada cluster compartilham alguma característica ou propriedade
  - São de alguma maneira similares



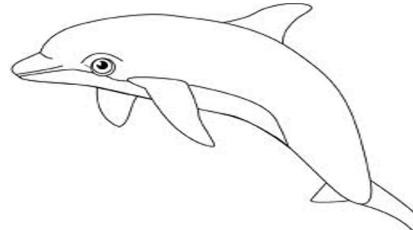
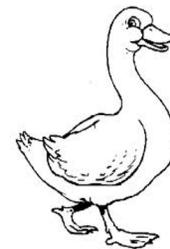
# Agrupamento

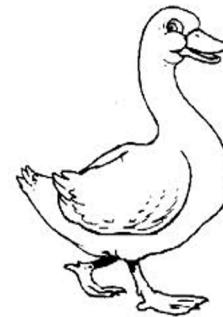
- Dificuldades:
  - Falta de alvo explícito para guiar a busca
    - De conhecimento das estruturas verdadeiras dos dados
  - Grande diversidade de critérios de agrupamento
  - Possibilidade de haver várias estruturas que descrevem um mesmo conjunto de dados
  - Escolha de algoritmo de agrupamento apropriado
  - Avaliação dos resultados obtidos

Análise de agrupamento tem caráter subjetivo

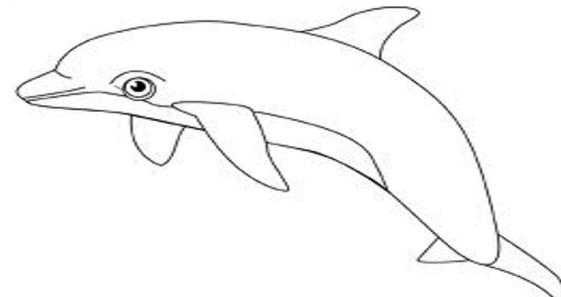
# Agrupamento

- Exemplo: como agrupar os seguintes animais?

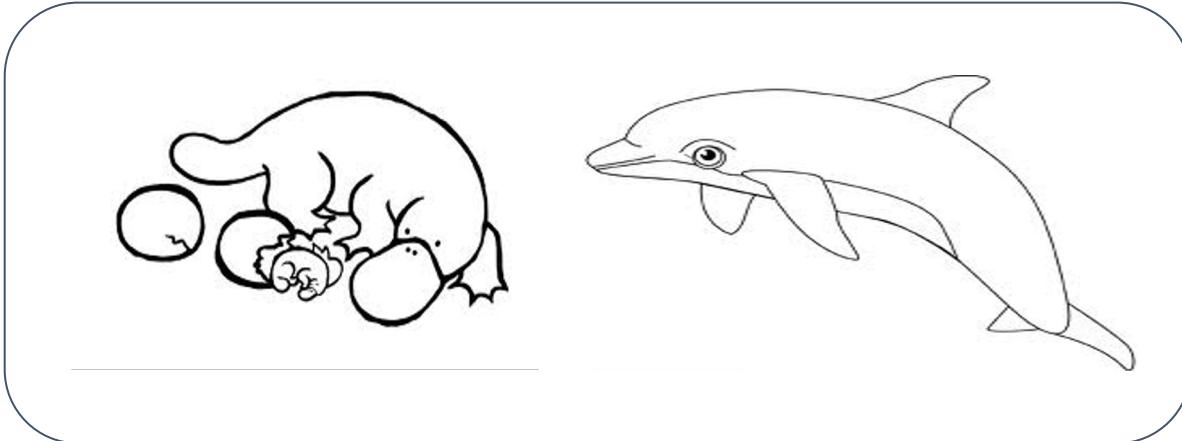




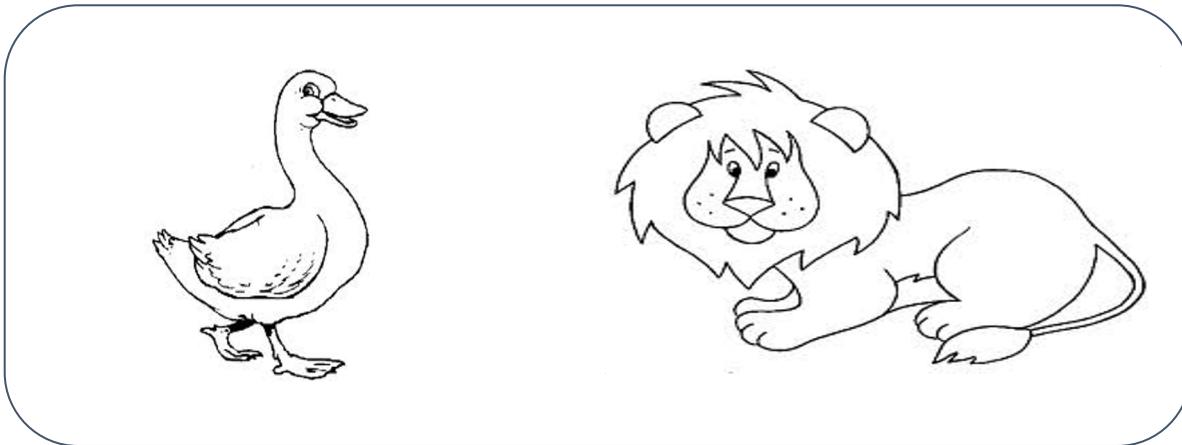
Com bico



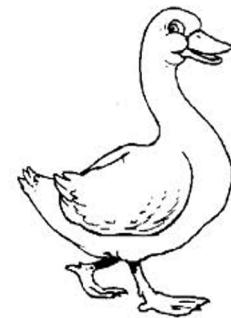
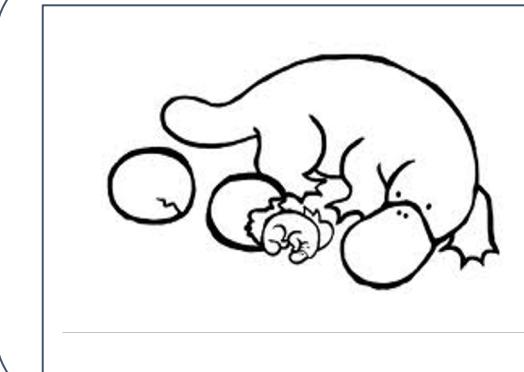
Sem bico



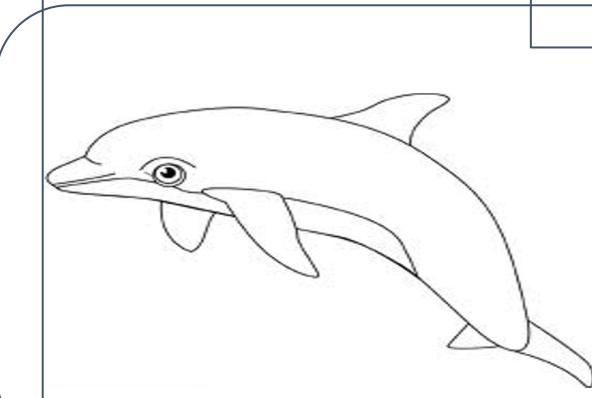
Aquático



Terrestre



Ovíparo



Mamífero

# Análise de Agrupamentos

- Considere os objetos pontos em um espaço de dimensão  $d$ 
  - $d$  = número de atributos
- Cluster: coleção de objetos próximos ou que satisfazem alguma relação espacial
  - Definição intuitiva, não existe definição formal única e precisa

# Clusters

- Algumas definições comuns:
- Cluster bem separado
  - Conjunto de pontos tal que qualquer ponto está mais próximo (é mais similar) a cada outro nesse cluster do que a qualquer outro ponto não pertencente a ele
- Cluster baseado em centro
  - Conjunto de pontos tal que qualquer ponto está mais próximo (é mais similar) ao centro desse cluster do que ao centro de qualquer outro cluster
  - Centroide: média aritmética dos pontos do cluster
  - Medoide: ponto mais representativo do cluster

# Clusters

- Cluster contínuo ou encadeado
  - Conjunto de pontos tal que qualquer ponto está mais próximo (é mais similar) a um ou mais pontos nesse cluster do que a qualquer outro ponto não pertencente a ele
- Cluster baseado em densidade
  - Região densa de pontos, separada por outras regiões de alta densidade por regiões de baixa densidade
- Cluster baseado em similaridade
  - Conjunto de pontos que são similares, enquanto pontos em clusters diferentes não são similares

# Critérios de agrupamento

- Cada definição resulta em um critério de agrupamento
  - Forma de selecionar uma estrutura de clusters (modelo) que melhor se ajuste aos dados
- Cada algoritmo de agrupamento:
  - Baseado em um critério de agrupamento
  - Usa uma medida de proximidade
  - Usa um método de busca para encontrar uma estrutura
    - De acordo com o critério de agrupamento adotado

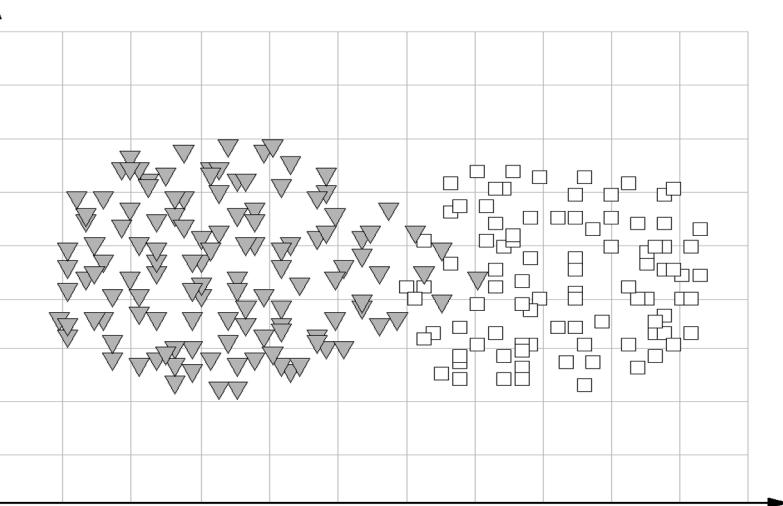
# Critérios de agrupamento

- Categorias:
- Compactação ou homogeneidade
  - Associada a variação intra-cluster pequena
  - Efetivos na descoberta de clusters esféricos e/ou bem separados
  - Podem falhar para estruturas mais complexas
- Encadeamento ou ligação
  - Conceito mais local (objetos vizinhos devem compartilhar o mesmo cluster)
  - Apropriado para detectar clusters de formas arbitrárias
  - Não robusto para quando há pouca separação espacial entre os clusters

# Critérios de agrupamento

- Categorias:
- Separação espacial
  - Considera distâncias entre os clusters
  - Fornece pouca orientação durante o agrupamento, podendo levar a soluções triviais
  - É comumente empregado em conjunto com outros

# Critérios de agrupamento

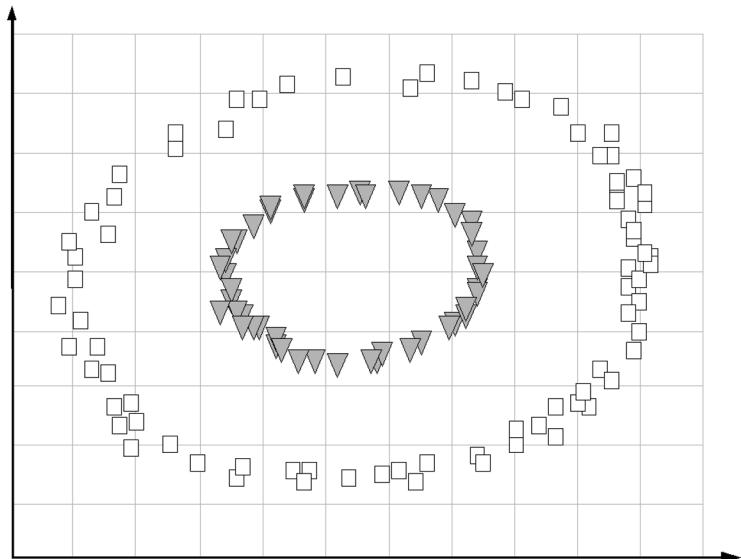


Conjunto de dados  
globular:

Dois clusters esféricos  
**bem separados**

Algoritmos baseados  
em **compactação**  
conseguem captar essa  
estrutura

# Critérios de agrupamento



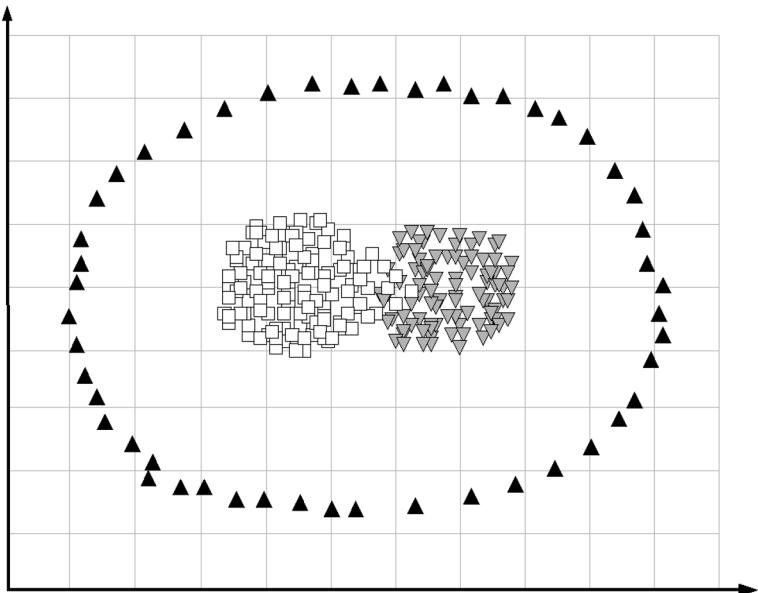
Conjunto de dados  
anel:

Dois clusters **bem distintos** na  
forma de anel

Algoritmos baseados  
em **encadeamento**  
conseguem captar essa  
estrutura

# Critérios de agrupamento

E quando conjunto é heterogêneo?



Conjunto de dados com clusters em conformidade com critérios de agrupamento diferentes.

- Um cluster em anel
- Dois clusters globulares.

*Não existe um único algoritmo de agrupamento capaz de encontrar todos os tipos de agrupamentos*

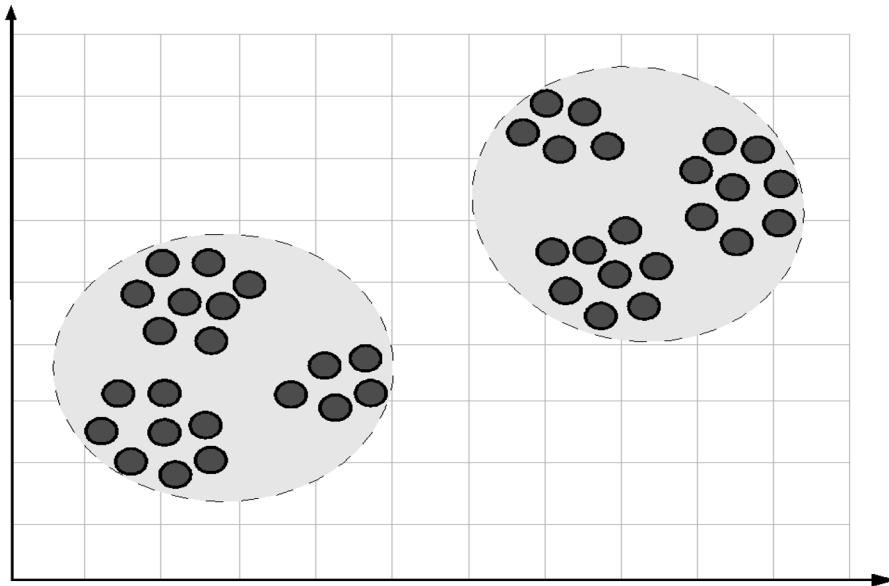
# Algoritmos de Agrupamento

- Existe um grande número
  - Cada um buscando clusters de acordo com um critério diferente
    - critério de agrupamento representa principal aspecto de um algoritmo de agrupamento
    - Técnicas de avaliação também seguem algum critério
  - Exemplos:
    - Algoritmo k-médias: procura clusters compactos
    - Algoritmo hierárquico ligação simples: otimizam critério baseado em encadeamento

# Nível de Refinamento

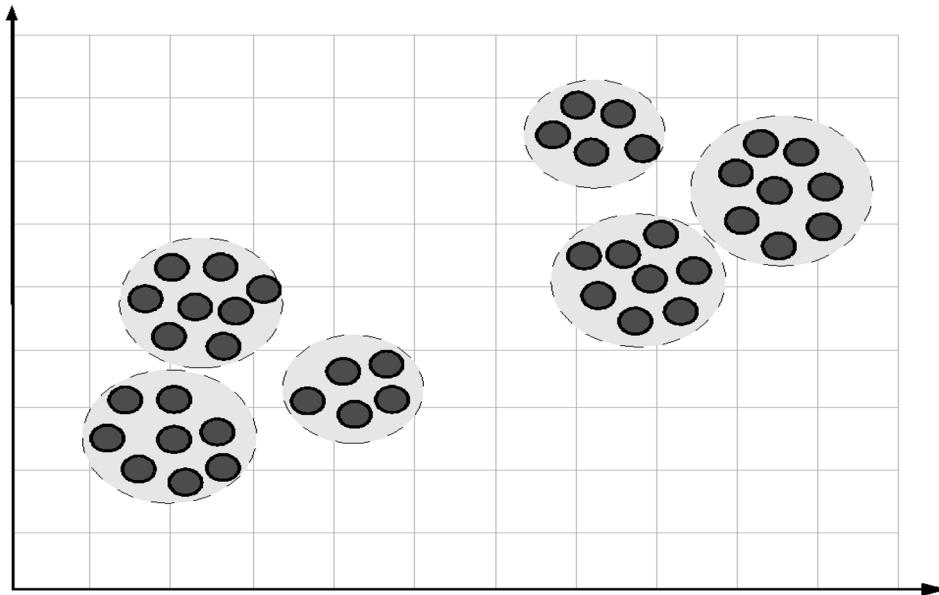
- Algoritmos podem encontrar estruturas em diferentes níveis de refinamento
  - Números de clusters diferentes ou de densidades diferentes
  - Dependendo de suas configurações de parâmetros
    - Importância de ajuste de parâmetros

# Nível de Refinamento



Estrutura **compacta**  
com dois clusters

# Nível de Refinamento

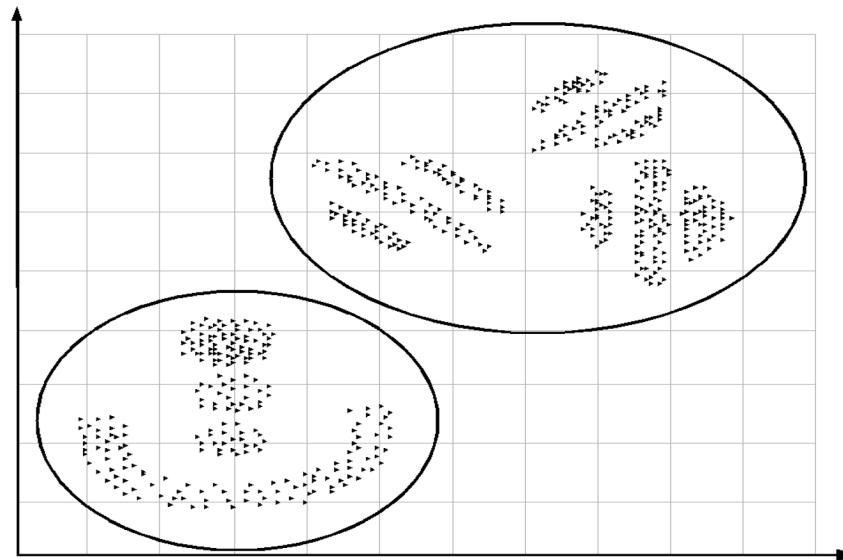


Estrutura **compacta**  
com seis clusters

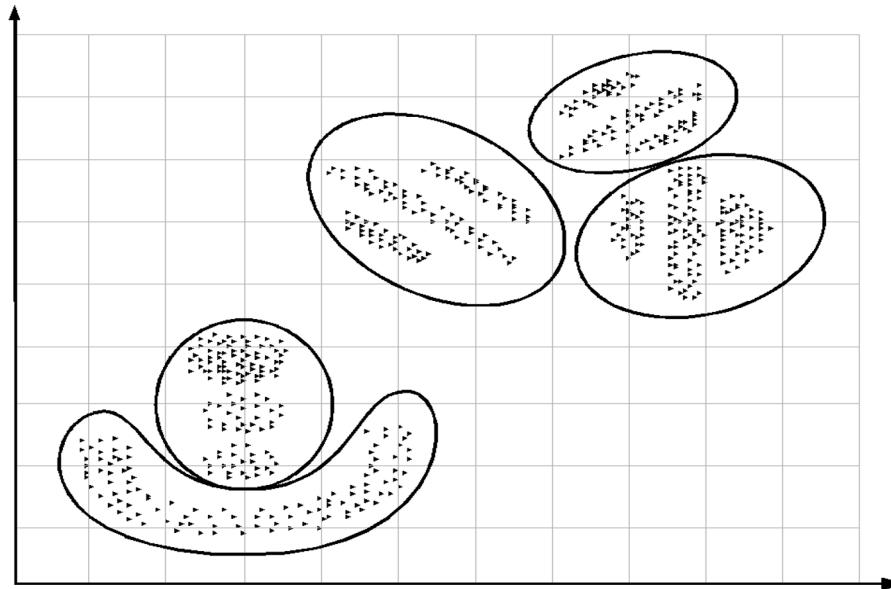
# Estruturas

- Um mesmo conjunto de dados pode ter mais de uma estrutura relevante
  - Cada uma representando uma diferente interpretação dos dados
  - Cada estrutura pode ser compatível com um critério de agrupamento diferente, estar em um nível diferente e/ou ser heterogênea

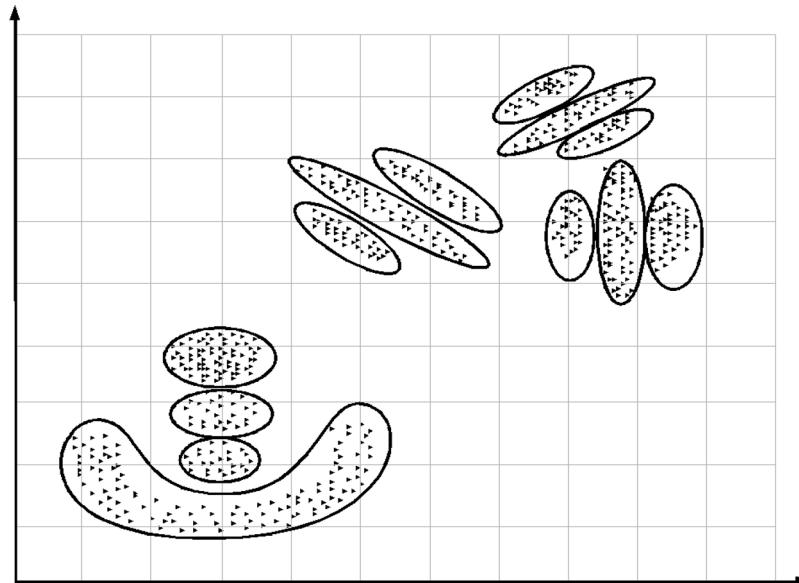
# Estruturas



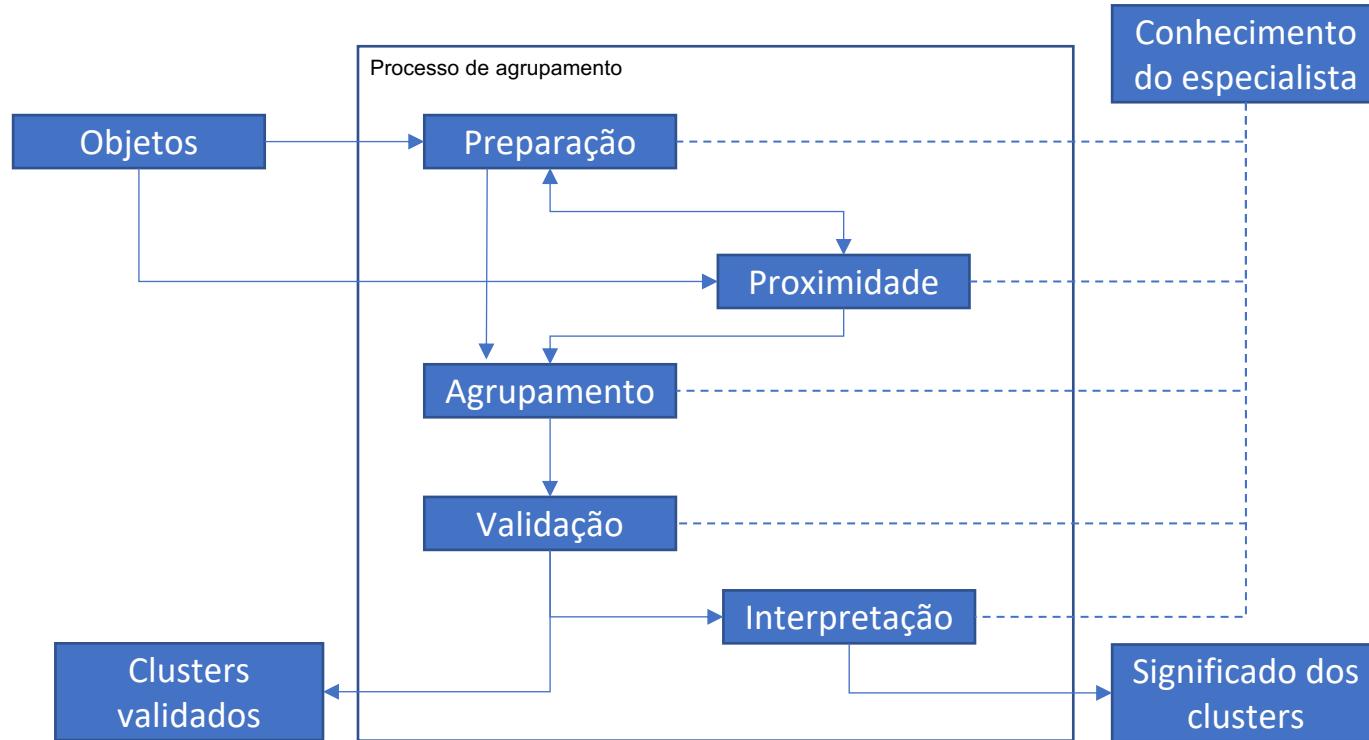
# Estruturas



# Estruturas



# Etapas da análise de agrupamento



# Preparação dos Dados

- Representação e pré-processamento

- Representação:

- Geralmente a atributo valor
    - Ou relação de proximidade entre objetos (matrizes e grafos de similaridade/dissimilaridade)

- Pré-processamentos podem incluir:

- Normalizações
    - Conversões de tipos
    - Redução de atributos
    - Extração de características



Sem uso de  
informação de classe

# Preparação dos Dados

- Matriz de similaridade/dissimilaridade:
  - Representa a similaridade/dissimilaridade entre cada par de objetos
  - Matriz  $S_{n \times n}$ 
    - Cada  $s_{ij}$ : distância ou similaridade entre objetos  $x_i$  e  $x_j$

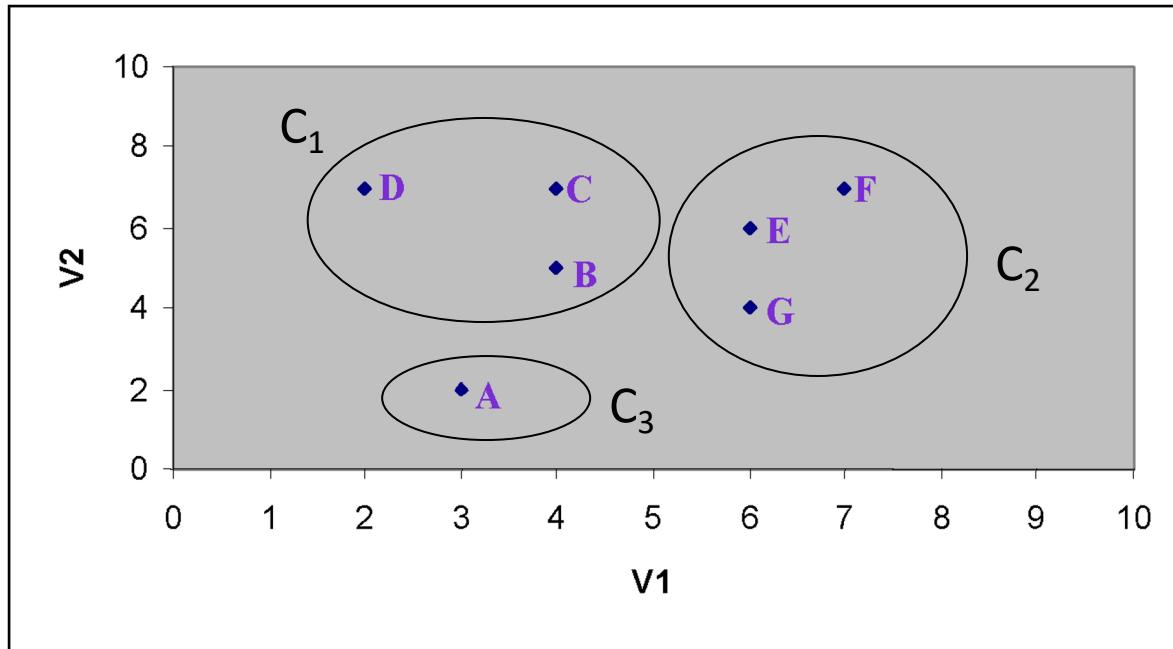
# Proximidade

- Medidas de distância e similaridade vistas em aula anterior
  - Distância Malahanobis, Euclideana, etc
  - Similaridade cosseno, de Pearson, etc

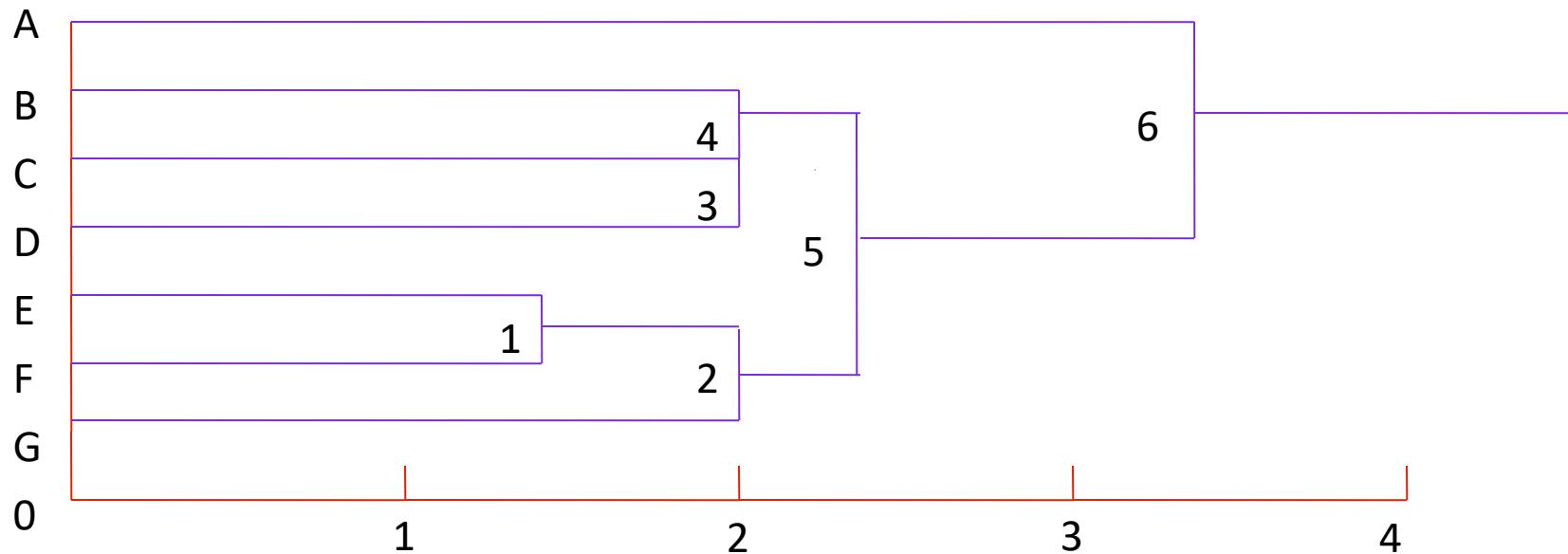
# Agrupamento

- Etapa central
  - Um ou mais algoritmos de agrupamento são aplicados aos dados
  - Tipos de estruturas que podem ser encontrados:
    - Particional
    - Hierárquica
    - Fuzzy

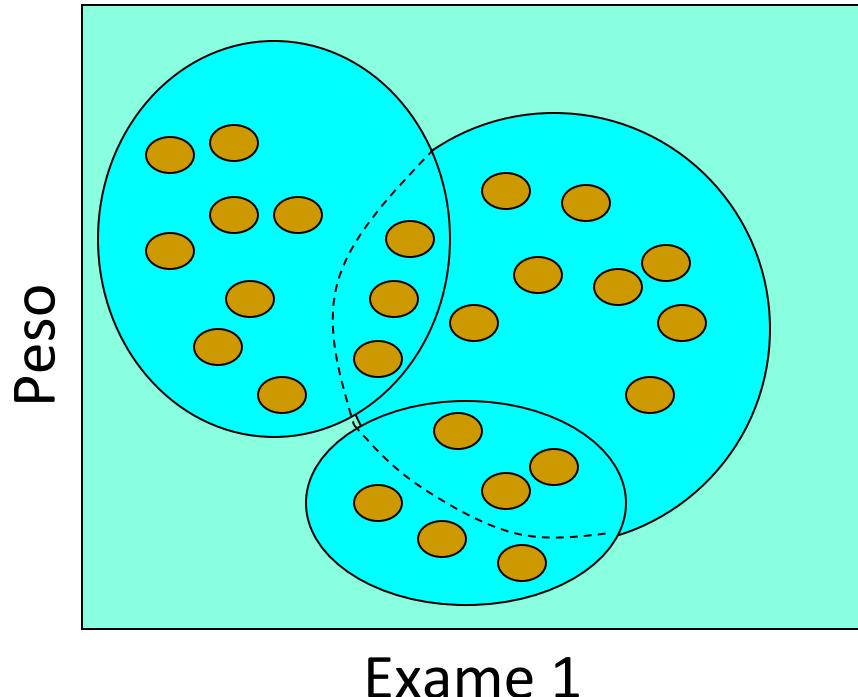
# Exemplo: Agrupamento Particional



# Agrupamento Hierárquico: Exemplo



# Agrupamento fuzzy



# Validação

- Avalia resultado de agrupamento
  - Determinar se clusters são significativos
  - Também pode ajudar na definição de parâmetros do algoritmo

# Interpretação

- Processo de examinar os clusters e rotulá-los
  - Descrevendo a natureza de cada um
  - Também é forma de validação dos clusters

# Algoritmos de agrupamento

# Algoritmos de Agrupamento

- Existe uma grande variedade
  - Diferentes critérios de agrupamento
- Classificação de acordo com método para definir os clusters:
  - Hierárquicos
  - Particionais
  - Baseados em grid
  - Baseados em densidade

# Algoritmos Hierárquicos

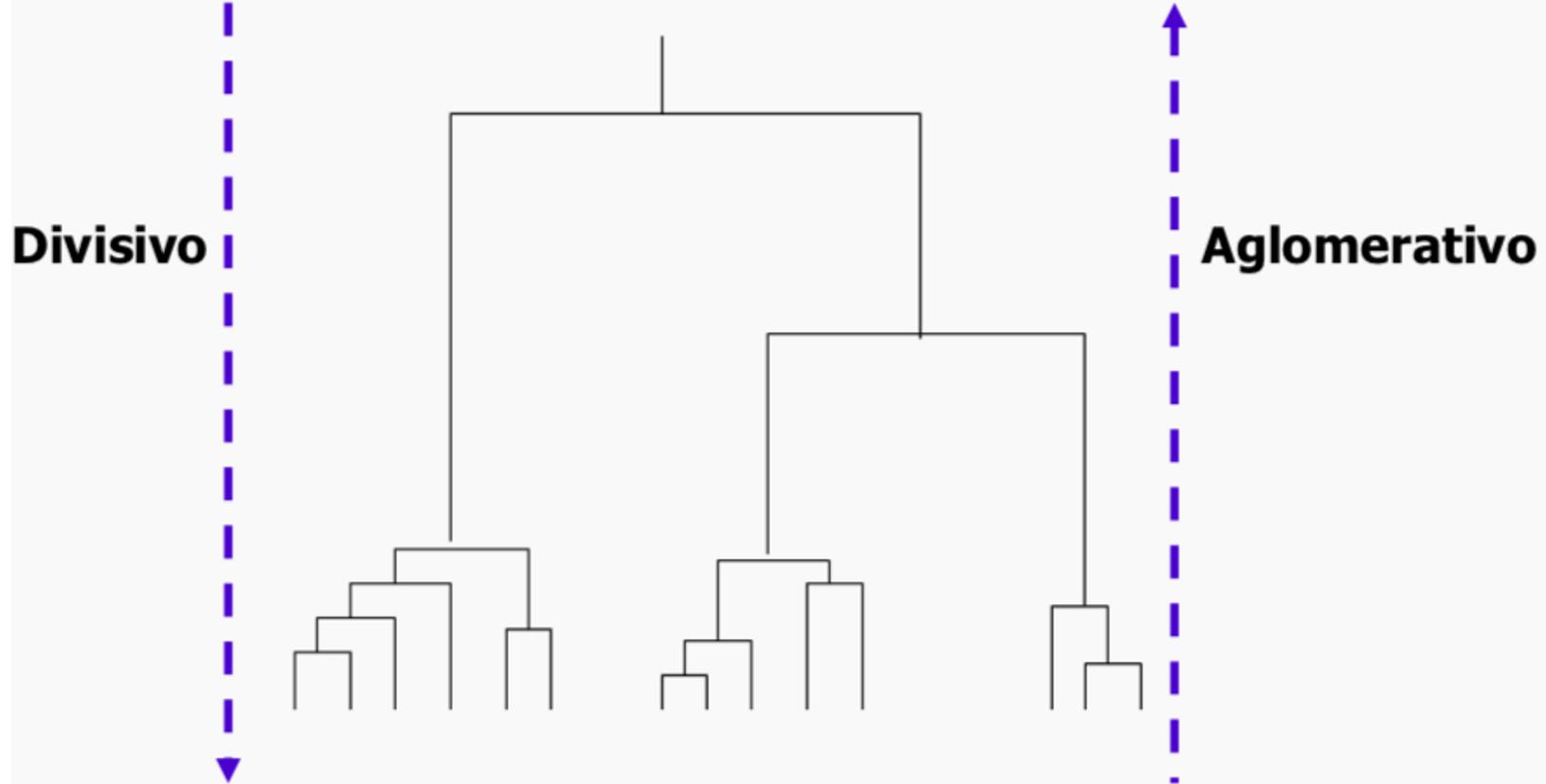
- Gera sequência de partições aninhadas
  - A partir de matriz de proximidade
  - Duas abordagens:

## Aglomerativa

- Começa com  $n$  clusters com um único objeto cada
- Agrupa clusters sucessivamente

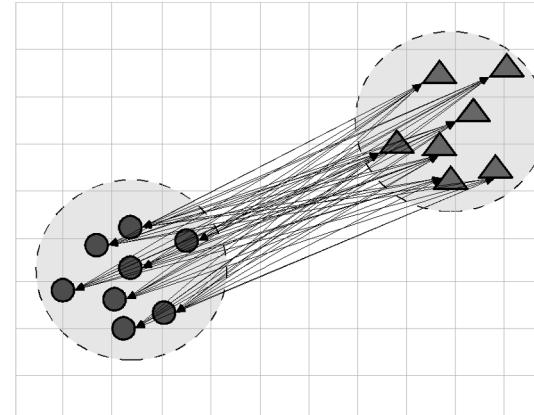
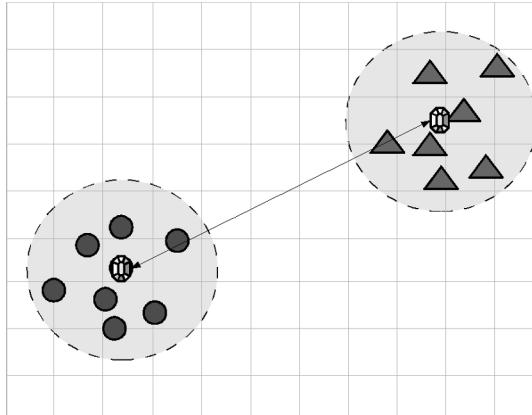
## Divisiva

- Começa com um único cluster contendo todos objetos
- Divide clusters sucessivamente



# Algoritmos Hierárquicos

- Mais clássicos: usam métricas de integração (linkage metrics)
  - Medidas de distância entre clusters distintos
    - Pode ser entre centroides dos clusters
    - Pode ser entre pares de exemplos dos clusters



# Métricas de Integração

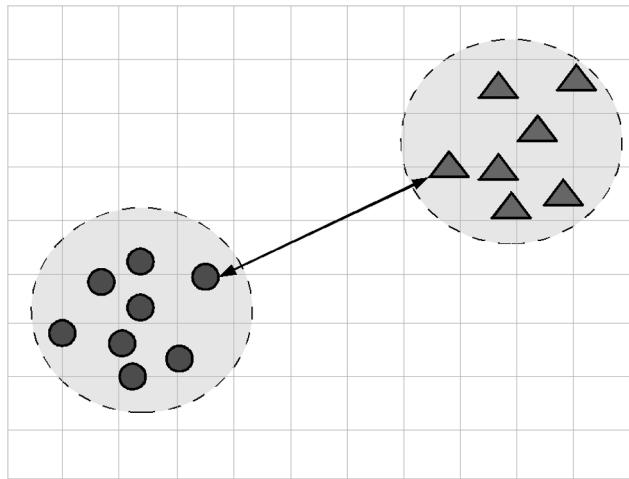
Mais comuns: dados dois clusters C1 e C2

- Ligação mínima
  - Distância/similaridade entre dois objetos mais próximos dos diferentes clusters
- Ligação média
  - Distância/similaridade média entre todos os objetos dos dois clusters
- Ligação máxima
  - Distância/similaridade entre os objetos mais distantes dos dois clusters

# Ligação Mínima

- Usada pelo algoritmo hierárquico single-link

$$d(\mathcal{C}_1, \mathcal{C}_2) = \min_{x_i \in \mathcal{C}_1, x_j \in \mathcal{C}_2} d(x_i, x_j)$$

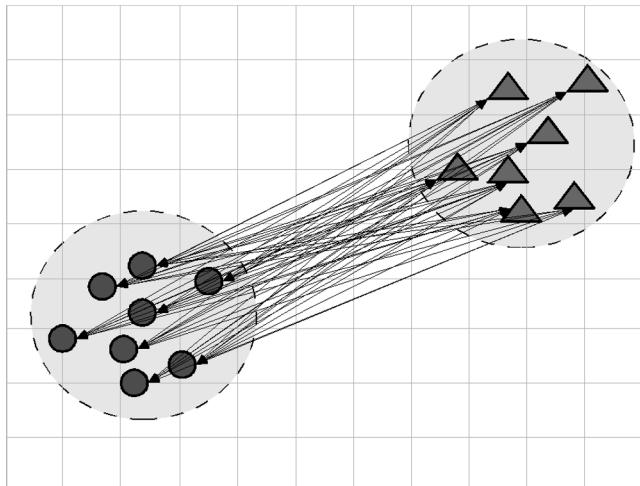


- Indicada para formas não elípticas
- Bastante sensível a ruídos e *outliers*
- Favorece clusters finos e alongados

# Ligaçāo Média

- Usada pelo algoritmo hierárquico average-link

$$d(\mathcal{C}_1, \mathcal{C}_2) = (1/n_1 n_2) \sum_{x_i \in \mathcal{C}_1, x_j \in \mathcal{C}_2} d(x_i, x_j)$$

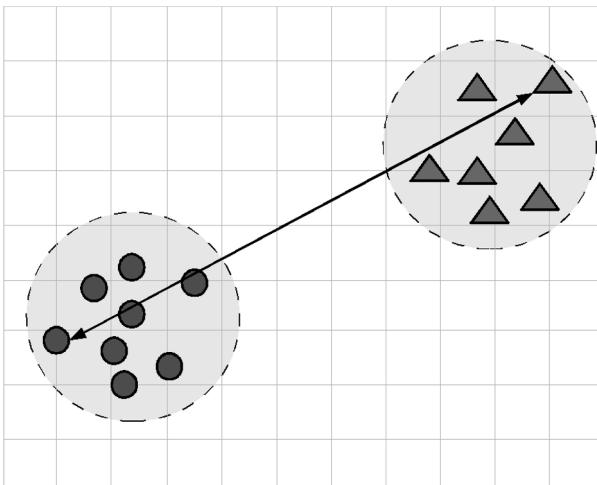


- Menos dependente de valores extremos
- Tende a combinar grupos de pequena variāncia
- Tende a produzir grupos com aproximadamente mesma variāncia

# Ligação Máxima

- Usada pelo algoritmo hierárquico complete-link

$$d(\mathcal{C}_1, \mathcal{C}_2) = \max_{x_i \in \mathcal{C}_1, x_j \in \mathcal{C}_2} d(x_i, x_j)$$



- Em geral favorece a obtenção de clusters esféricos
- Tende a quebrar grupos grandes

# Exemplo

- Seja o conjunto de dados:

	X1	X2	X3	X4	X5
C 1	7,000	10,000	9,000	7,000	10,000
C 2	9,000	9,000	8,000	9,000	9,000
C 3	5,000	5,000	6,000	7,000	7,000
C 4	6,000	6,000	3,000	3,000	4,000
C 5	1,000	2,000	2,000	1,000	2,000
C 6	4,000	3,000	2,000	3,000	3,000
C 7	2,000	4,000	5,000	2,000	5,000



Agrupar hierarquicamente

# Exemplo: método aglomerativo single-link

- Matriz de similaridade: usando Pearson

	X1	X2	X3	X4	X5
C_1	7,000	10,000	9,000	7,000	10,000
C_2	9,000	9,000	8,000	9,000	9,000
C_3	5,000	5,000	6,000	7,000	7,000
C_4	6,000	6,000	3,000	3,000	4,000
C_5	1,000	2,000	2,000	1,000	2,000
C_6	4,000	3,000	2,000	3,000	3,000
C_7	2,000	4,000	5,000	2,000	5,000

	Cliente_1	Cliente_2	Cliente_3	Cliente_4	Cliente_5	Cliente_6	Cliente_7
Cliente_1	1,000						
Cliente_2	-0,147	1,000					
Cliente_3	0,000	0,000	1,000				
Cliente_4	0,087	0,516	-0,824	1,000			
Cliente_5	0,963	-0,408	0,000	-0,060	1,000		
Cliente_6	-0,466	0,791	-0,354	0,699	-0,645	1,000	
Cliente_7	0,891	-0,516	0,165	-0,239	0,963	-0,699	1,000

# Ligação Simples: Exemplo

	C_1	C_2	C_3	C_4	C_5	C_6	C_7
C_1	1,000				0,963		
C_2	-0,147	1,000			-0,408		
C_3	0,000	0,000	1,000		0,000		
C_4	0,087	0,516	-0,824	1,000	-0,060		
C_5	0,963	-0,408	0,000	-0,060	1,000		
C_6	-0,466	0,791	-0,354	0,699	-0,645	1,000	
C_7	0,891	-0,516	0,165	-0,239	0,963	-0,699	1,000

Note que o C5 se “junta” a C1  
prevalecendo o maior valor  
(maior correlação) em cada linha:  
a linha/coluna C5 “desaparece”

	(C_1,C_5)	C_2	C_3	C_4	C_6	C_7
(C_1,C_5)	1,000					0,963
C_2	-0,147	1,000				-0,516
C_3	0,000	0,000	1,000			0,165
C_4	0,087	0,516	-0,824	1,000		-0,239
C_6	-0,466	0,791	-0,354	0,699	1,000	-0,699
C_7	0,963	0,516	0,165	-0,239	-0,699	1,000

# Ligação Simples: Exemplo

	(C_1,C_5)	C_2	C_3	C_4	C_6	C_7
(C_1,C_5)	1,000					0,963
C_2	-0,147	1,000				-0,516
C_3	0,000	0,000	1,000			0,165
C_4	0,087	0,516	-0,824	1,000		-0,239
C_6	-0,466	0,791	-0,354	0,699	1,000	-0,699
C_7	0,963	-0,516	0,165	-0,239	-0,699	1,000

Note que o C7 se “junta” a C1\_C5  
prevalecendo o maior valor  
(maior correlação) em cada linha:  
a linha/coluna C7 “desaparece”

	(C_1,C_5,C_7)	C_2	C_3	C_4	C_6
(C_1,C_5,C_7)	1,000	-0,147	0,165	0,087	-0,466
C_2	-0,147	1,000	-0,824	0,516	0,791
C_3	0,165	0,000	1,000	-0,824	-0,354
C_4	0,087	0,516	-0,824	1,000	0,699
C_6	-0,466	0,791	-0,354	0,699	1,000

# Ligação Simples: Exemplo

	(C_1,C_5,C_7)	C_2	C_3	C_4	C_6
(C_1,C_5,C_7)	1,000	-0,147	0,165	0,087	-0,466
C_2	-0,147	1,000	-0,824	0,516	0,791
C_3	0,165	0,000	1,000	-0,824	-0,354
C_4	0,087	0,516	-0,824	1,000	0,699
C_6	-0,466	0,791	-0,354	0,699	1,000

O C6 se “junta” a C2  
prevalecendo o maior valor  
(maior correlação) em cada linha:  
a linha/coluna C6 “desaparece”

	(C_1,C_5,C_7)	(C_2,C_6,C_4)	C_3
(C_1,C_5,C_7)	1,000	0,087	0,165
(C_2,C_6,C_4)	0,087	1,000	0,699
C_3	0,165	0,699	1,000

Na mesma linha C4 se junta a  
C2\_C6, etc

# Ligaçāo Simples: Exemplo

	(C_1,C_5,C_7)	(C_2,C_6,C_4)	C_3
(C_1,C_5,C_7)	1,000	0,087	0,165
(C_2,C_6,C_4)	0,087	1,000	0,699
C_3	0,165	0,699	1,000

	(C_1,C_5,C_7)	(C_2,C_6,C_4,C_3)
(C_1,C_5,C_7)	1,000	0,087
(C_2,C_6,C_4,C_3)	0,087	1,000

# Ligaçāo Simples: Exemplo

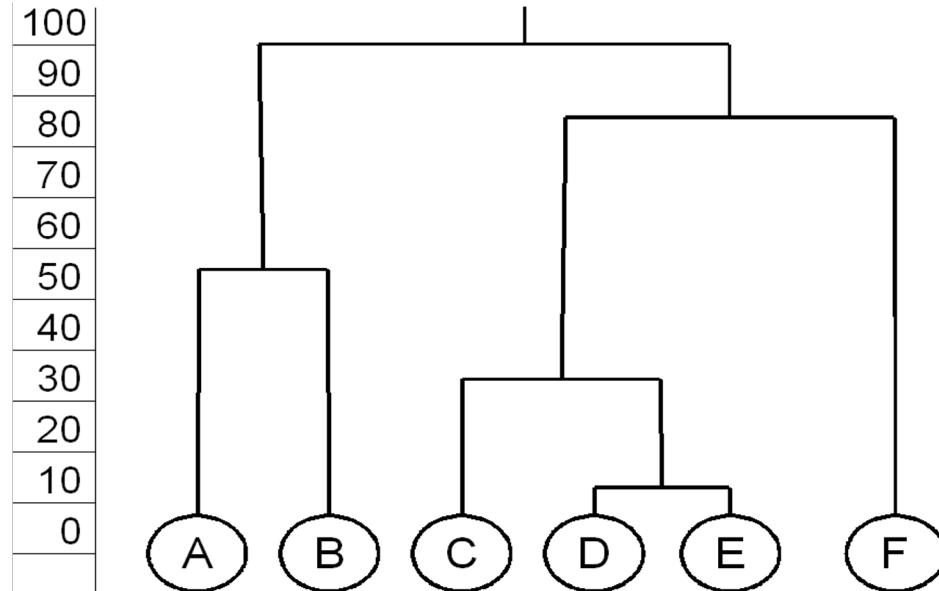
	(C_1,C_5,C_7)	(C_2,C_6,C_4,C_3)
(C_1,C_5,C_7)	1,000	0,087
(C_2,C_6,C_4,C_3)	0,087	1,000

(C\_1,C\_5,C\_7,C\_2,C\_6,C\_4,C\_3)

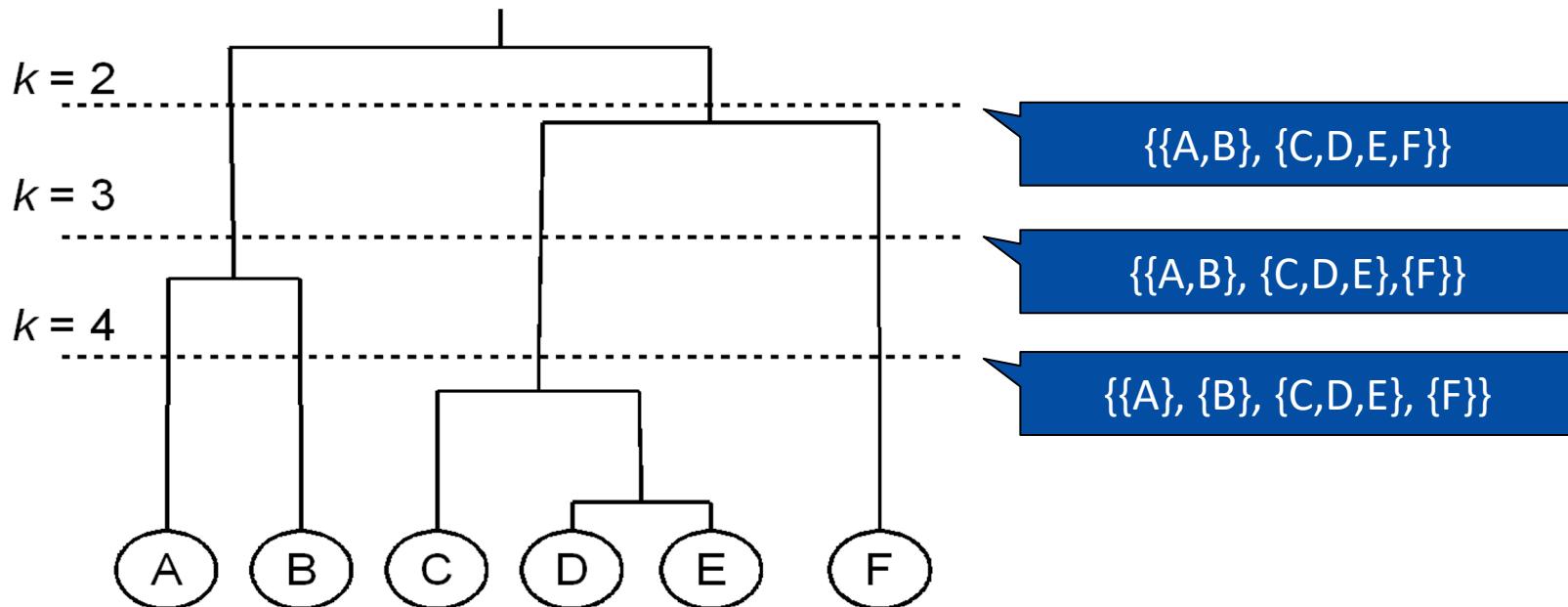
# Dendrograma

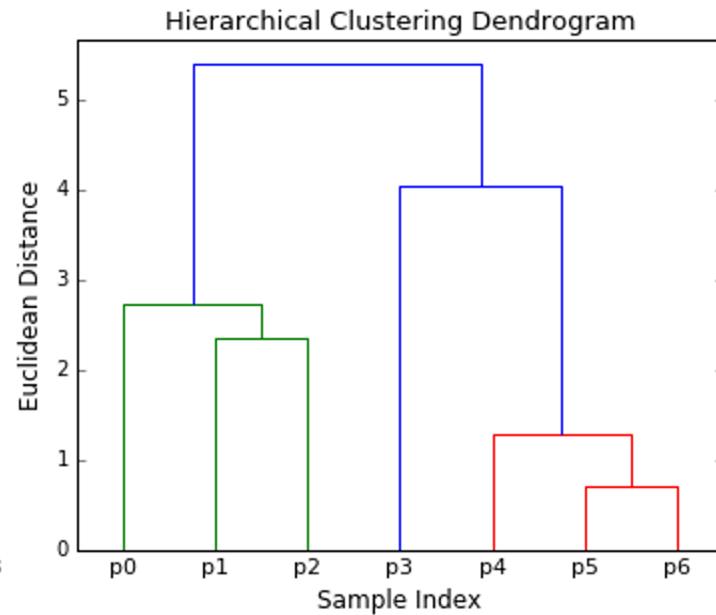
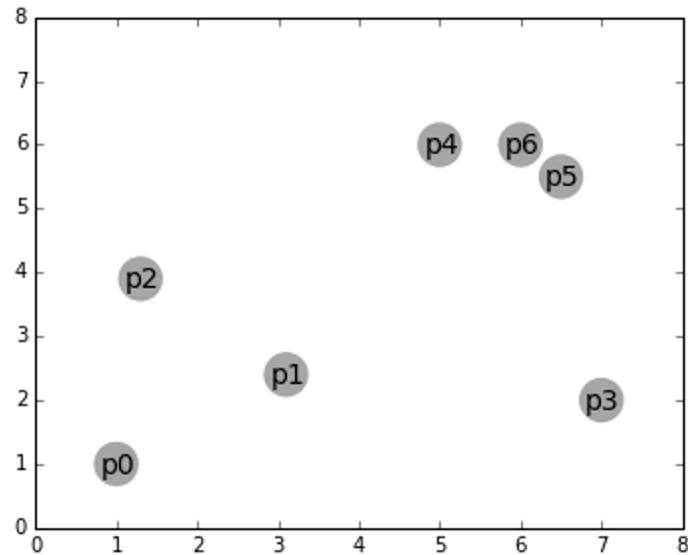
- Soluções de algoritmo hierárquico geralmente são representadas por dendrogramas
  - Árvore binária representando uma hierarquia de partições
  - O corte de um dendograma em qualquer nível produz uma simples partição

# Dendrograma



# Dendrograma

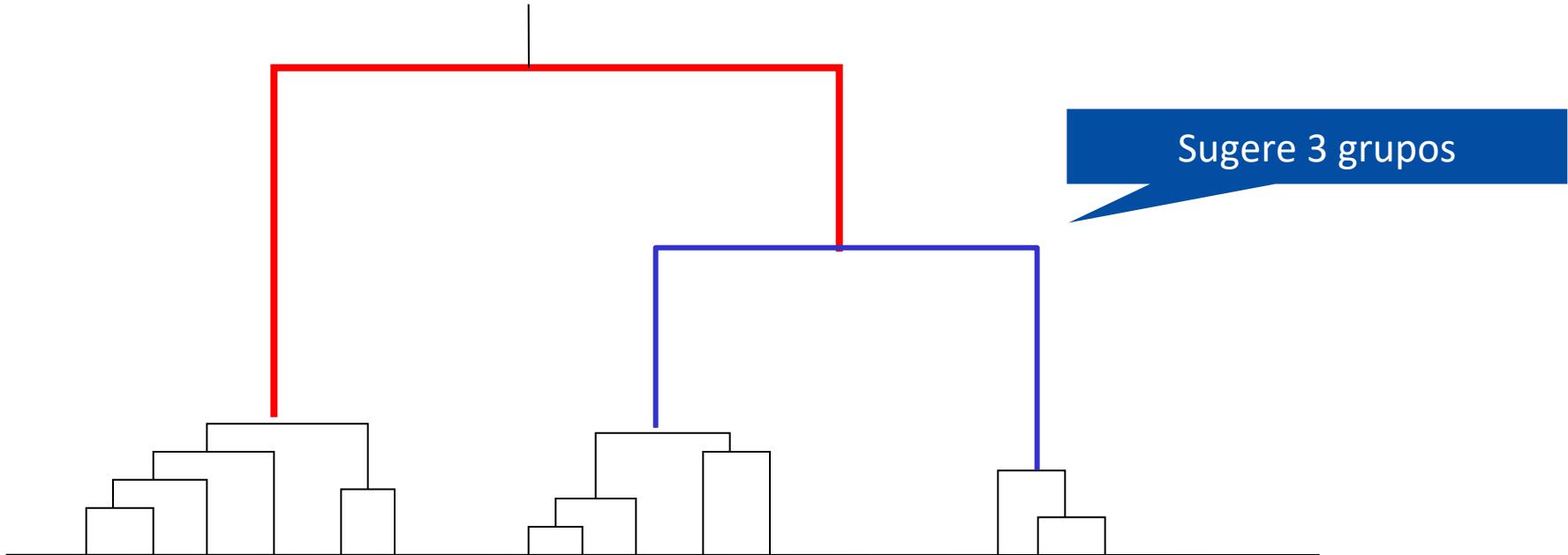




# Algoritmos Hierárquicos

- Como escolher uma partição?
  - Partição com k clusters
    - Selecionando partição com k clusters na seqüência de agrupamentos da hierarquia
  - Partição que melhor se encaixa nos dados
    - Procurar no dendograma grandes mudanças em níveis adjacentes
      - Nesse caso, uma mudança de j para j-1 grupos pode indicar que j é o melhor número de grupos
      - Existem outros procedimentos, alguns mais objetivos

# Algoritmos Hierárquicos



# Algoritmos Hierárquicos

- Aspectos positivos:
  - Flexibilidade com respeito ao nível de granularidade
    - Não requerem a princípio especificação do número de clusters
  - Fácil utilização
  - Determinístico
  - Geração de taxonomias é natural em alguns domínios

# Algoritmos Hierárquicos

- Aspectos negativos:
  - Critério de terminação vago
  - Não lidam bem com ruídos e outliers
  - Estrutura somente de árvore
  - Objetos agrupados segundo decisões locais, que não são reavaliadas

# k-means

- Realiza agrupamento particional
  - Não há hierarquias, os dados são particionados
  - O número de grupos ( $k$ ) tem que ser definido a priori

# k-means

- Passo 1: Os primeiros k centros são escolhidos aleatoriamente
- Passo 2: Cada objeto é atribuído ao grupo associado com o centro mais próximo
- Passo 3: Computa-se um novo centro para cada grupo (centroide)
- Passo 4: Repetir Passo 2 (com os novos centros) e Passo 3 até que não haja mudança nos centros

# k-Médias: Exemplo

- Exemplo do uso do k-médias, com Correlação de Pearson como medida de proximidade

	X1	X2	X3	X4	X5
Cliente_1	7,000	10,000	9,000	7,000	10,000
Cliente_2	9,000	9,000	8,000	9,000	9,000
Cliente_3	5,000	5,000	6,000	7,000	7,000
Cliente_4	6,000	6,000	3,000	3,000	4,000
Cliente_5	1,000	2,000	2,000	1,000	2,000
Cliente_6	4,000	3,000	2,000	3,000	3,000
Cliente_7	2,000	4,000	5,000	2,000	5,000

- Passo 1: Os primeiros k centros dos aglomerados são escolhidos aleatoriamente.
  - Por exemplo, escolheu-se o CLIENTE 2 e o CLIENTE 5, para K = 2.

# k-Médias: Exemplo

- Passo 2: Cada objeto é atribuído ao grupo associado com o centro mais próximo

	Cliente_1	Cliente_2	Cliente_3	Cliente_4	Cliente_5	Cliente_6	Cliente_7
Cliente_1	1,000						
Cliente_2	-0,147	1,000	0,000	0,516	-0,408	0,791	-0,516
Cliente_3	0,000	0,000	1,000				
Cliente_4	0,087	0,516	-0,824	1,000			
Cliente_5	0,963	-0,408	0,000	-0,060	1,000	-0,645	0,963
Cliente_6	-0,466	0,791	-0,354	0,699	-0,645	1,000	
Cliente_7	0,891	-0,516	0,165	-0,239	0,963	-0,699	1,000

Esta tabela mostra a correlação entre as instâncias (pearson)

- O cliente 1 se aproxima mais do centro5:  $(0,963 > -0,147)$ .
- O cliente 3 se aproxima dos dois  $(0,0)$ : atribui-se aleatoriamente para algum deles.
- E assim por diante ...

# k-Médias: Exemplo

- Passo3: Compute um novo centro para cada grupo  
(média dos valores de todos os objetos - centróide)

	X1	X2	X3	X4	X5
Cliente_2	9,000	9,000	8,000	9,000	9,000
Cliente_3	5,000	5,000	6,000	7,000	7,000
Cliente_4	6,000	6,000	3,000	3,000	4,000
Cliente_6	4,000	3,000	2,000	3,000	3,000
Centro_1	6,000	5,750	4,750	5,500	5,750

	X1	X2	X3	X4	X5
Cliente_1	7,000	10,000	9,000	7,000	10,000
Cliente_5	1,000	2,000	2,000	1,000	2,000
Cliente_7	2,000	4,000	5,000	2,000	5,000
Centro_2	3,333	5,333	5,333	3,333	5,667

Calcula-se os novos centros pelos valores médios dos atributos de todos os exemplos que ficaram em cada grupo

# k-Médias: Exemplo

	Cliente_1	Cliente_2	Cliente_3	Cliente_4	Cliente_5	Cliente_6	Cliente_7	Centro_1	Centro_2
Cliente_1	1	-0,1474	0	0,087	0,9631	-0,4663	0,8913	-0,1371	0,9723
Cliente_2	-0,1474	1	0	0,516	-0,4082	0,7906	-0,516	0,93	-0,3498
Cliente_3	0	0	1	-0,8242	0	-0,3536	0,1648	-0,2599	0,068
Cliente_4	0,087	0,516	-0,8242	1	-0,0602	0,6994	-0,2391	0,737	-0,0698
Cliente_5	0,9631	-0,4082	0	-0,0602	1	-0,6455	0,9631	-0,3797	0,9926
Cliente_6	-0,4663	0,7906	-0,3536	0,6994	-0,6455	1	-0,6994	0,919	-0,6011
Cliente_7	0,8913	-0,516	0,1648	-0,2391	0,9631	-0,6994	1	-0,4799	0,9723
Centro_1	-0,1371	0,93	-0,2599	0,737	-0,3797	0,919	-0,4799	1	-0,322
Centro_2	0,9723	-0,3498	0,068	-0,0698	0,9926	-0,6011	0,9723	-0,322	1

- Calcula-se então a correlação em relação aos novos centros (vide as duas últimas linhas)
- cliente\_1 fica no grupo do centro\_2 ( $0,9723 > -0,1371$ )  
cliente\_2 fica no grupo do centro\_1 ( $0,93 > -0,3498$ )  
.... E assim por diante ...
- Formados: (2, 4, 6) e (1, 3, 5, 7)

# k-Médias: Exemplo

	X1	X2	X3	X4	X5
Cliente_2	9,000	9,000	8,000	9,000	9,000
Cliente_4	6,000	6,000	3,000	3,000	4,000
Cliente_6	4,000	3,000	2,000	3,000	3,000
Centro_1	6,333	6,000	4,333	5,000	5,333

	X1	X2	X3	X4	X5
Cliente_1	7,000	10,000	9,000	7,000	10,000
Cliente_3	5,000	5,000	6,000	7,000	7,000
Cliente_5	1,000	2,000	2,000	1,000	2,000
Cliente_7	2,000	4,000	5,000	2,000	5,000
Centro_2	3,750	5,250	5,500	4,250	6,000

Calcula-se os novos centros pelos valores médios dos atributos de todos os exemplos que ficaram em cada grupo

# k-Médias: Exemplo

	Cliente_1	Cliente_2	Cliente_3	Cliente_4	Cliente_5	Cliente_6	Cliente_7	Centro_1	Centro_2
Cliente_1	1	-0,1474	0	0,087	0,9631	-0,4663	0,8913	-0,1106	0,9175
Cliente_2	-0,1474	1	0	0,516	-0,4082	0,7906	-0,516	0,75	-0,3323
Cliente_3	0	0	1	-0,8242	0	-0,3536	-0,6281	0,3377	
Cliente_4	0,087	0,516	-0,8242	1	-0,0602	0,6994	-0,2391	0,9389	-0,2939
Cliente_5	0,9631	-0,4082	0	-0,0602	1	-0,6455	0,9631	-0,3062	0,9372
Cliente_6	-0,4663	0,7906	-0,3536	0,6994	-0,6455	1	-0,6994	0,8883	-0,6686
Cliente_7	0,8913	-0,516	0,1648	-0,2391	0,9631	-0,6994	1	-0,4564	0,962
Centro_1	-0,1106	0,75	-0,6281	0,9389	-0,3062	0,8883	-0,4564	1	-0,4473
Centro_2	0,9175	-0,3323	0,3377	-0,2939	0,9372	-0,6686	0,962	-0,4473	1

Calcula-se então a correlação em relação aos novos centros  
(vide as duas últimas linhas)

cli1 fica no grupo do centro2 ( $0,9175 > -0,1106$ )

cli2 fica no grupo do centro1 ( $0,75 > -0,3323$ )

.... E assim por diante ...

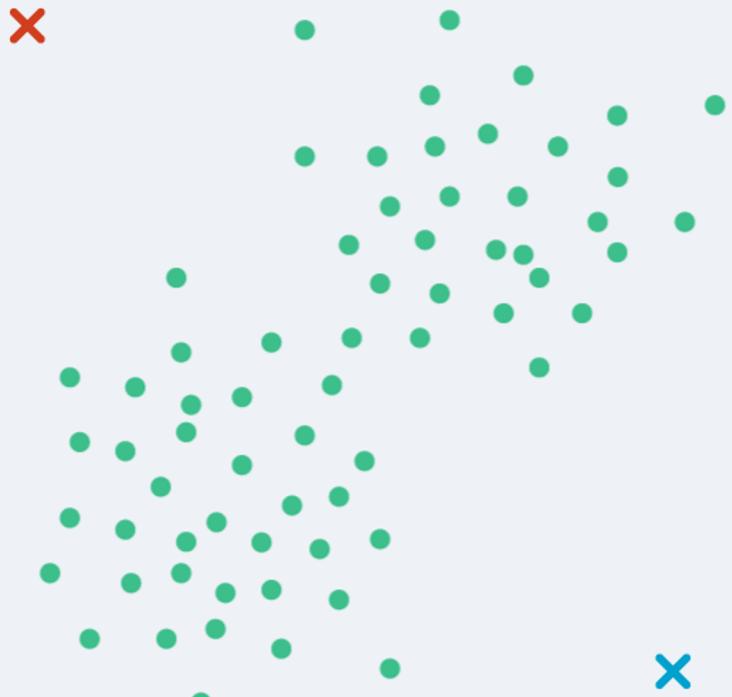
Formados: (2, 4, 6) e (1, 3, 5, 7)

Os grupos são os mesmos da iteração anterior: logo, o algoritmo convergiu.

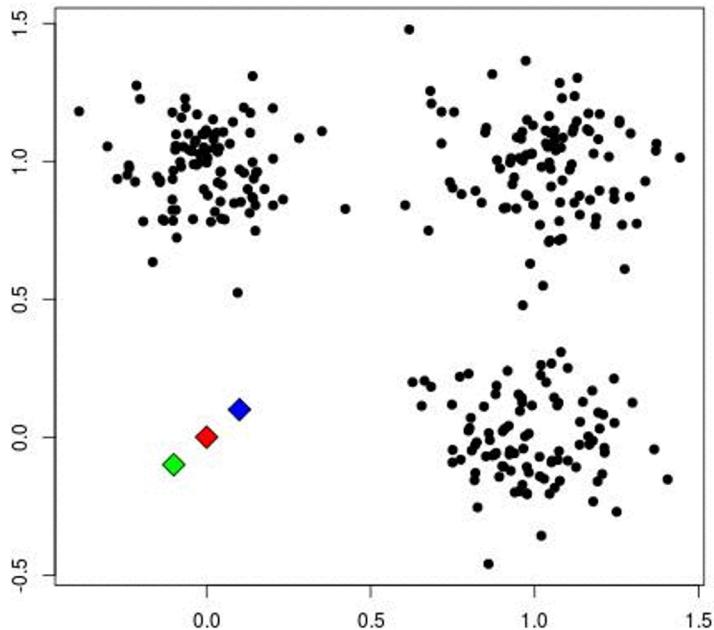
# k-Médias: Exemplo

	X1	X2	X3	X4	X5
Cliente_2	9,000	9,000	8,000	9,000	9,000
Cliente_4	6,000	6,000	3,000	3,000	4,000
Cliente_6	4,000	3,000	2,000	3,000	3,000
Centro_1	6,333	6,000	4,333	5,000	5,333

	X1	X2	X3	X4	X5
Cliente_1	7,000	10,000	9,000	7,000	10,000
Cliente_3	5,000	5,000	6,000	7,000	7,000
Cliente_5	1,000	2,000	2,000	1,000	2,000
Cliente_7	2,000	4,000	5,000	2,000	5,000
Centro_2	3,750	5,250	5,500	4,250	6,000



Start!



# k-means

- Aspectos positivos:
  - Eficiente
    - $O(n)$
  - Como usa critério de compactação, é indicado para encontrar grupos hiper esféricos

# k-means

- Aspectos negativos:
  - Podem convergir para ótimos locais
  - Sensível à inicialização
  - Clusters em geral desbalanceados
  - Como determinar o valor de k



# Validação

# Validação de agrupamento

- Dois objetivos:
  - Avaliação e comparação de algoritmos
  - Validação das estruturas encontradas
    - Determinar se estrutura é válida, se não ocorreu por acaso

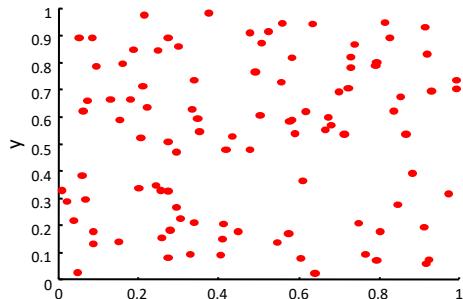
Garantir a corretude, validade e reproduzibilidade dos experimentos e das conclusões obtidas

# Validação de agrupamento

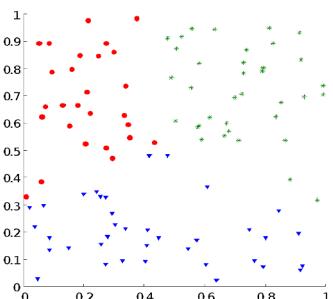
- Importante:
  - Não existe um resultado correto esperado
    - É uma tarefa não supervisionada
    - Em alguns estudos usa-se dados para os quais se conhecem uma ou mais estruturas
      - E algoritmos são avaliados em habilidade de encontrá-las
  - Nem sempre existe uma solução única
    - Podem existir várias estruturas em um mesmo conjunto de dados

# Validação de agrupamento

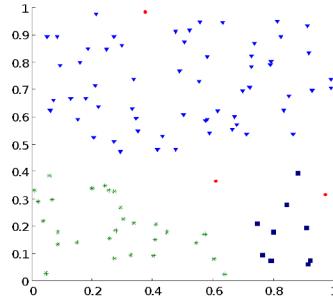
Pontos



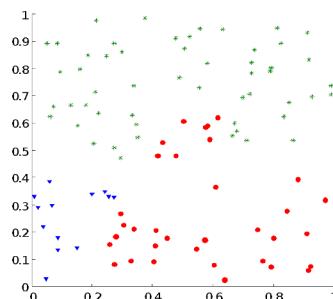
Aleatórios



K-médias



DBSCAN



Complete-  
Link

# Validação de agrupamento

- Baseada em índices estatísticos
  - Julgam o mérito das estruturas encontradas
  - Quantificando qualidade do agrupamento
  - Forma como são aplicados é dada por critério de validação
    - Critério expressa estratégia usada para validar estrutura
    - Índice é estatística pela qual validade é testada

# Validação de agrupamento

Três tipos de critérios:

- Critérios relativos:

- Comparam agrupamentos com respeito a algum aspecto (qual é o mais estável, por exemplo)
- Usados para comparar diversos algoritmos de agrupamento ou determinar parâmetros de algoritmos (ex. número de clusters)

- Critérios internos:

- Medem qualidade do agrupamento com base nos dados apenas

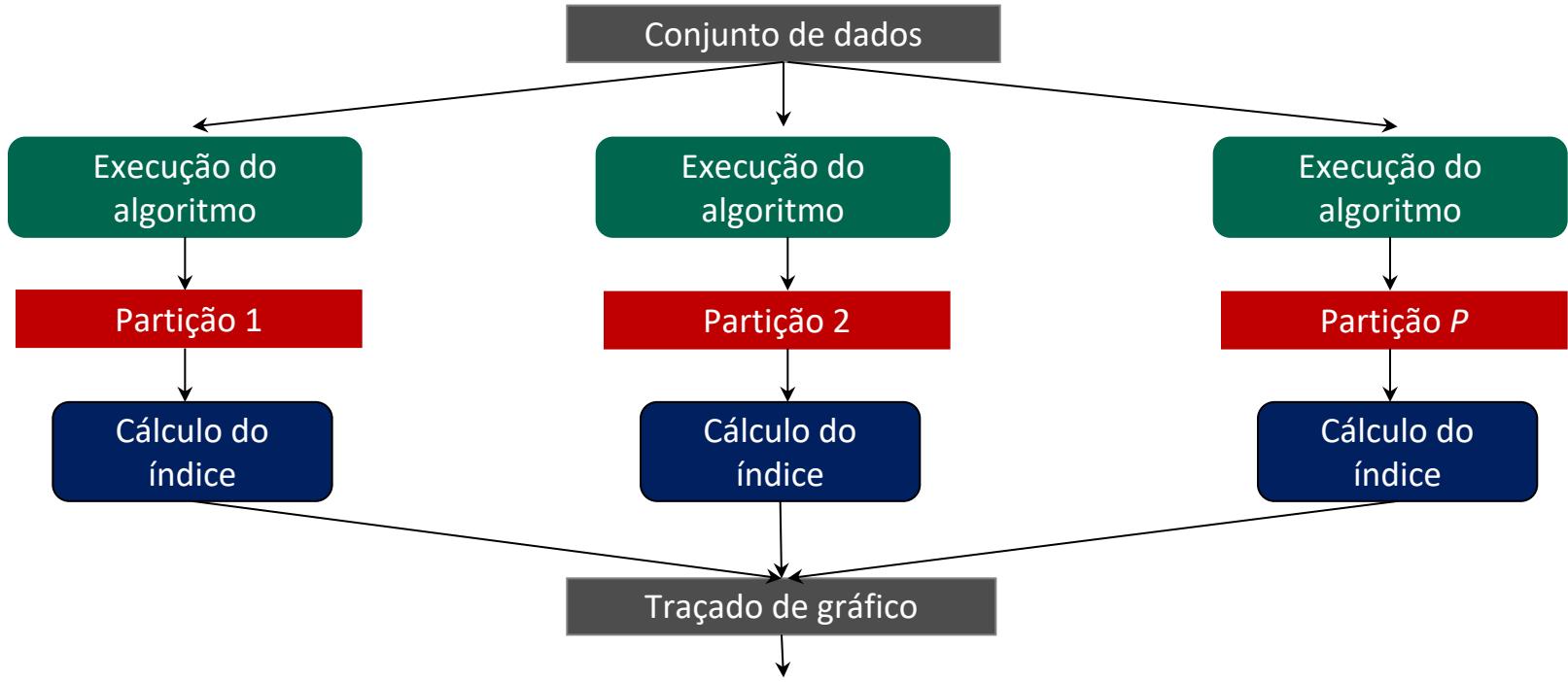
- Critérios externos:

- Avaliam agrupamento de acordo com uma estrutura estabelecida previamente

# Critérios relativos

- Forma mais comum de uso:
  - Calcula valor para vários agrupamentos sendo comparados
  - Melhor agrupamento: determinado pelo valor que se destaca na sequência
    - Ex. máximo, mínimo, inflexão em gráfico com valores

# Critérios relativos



Algoritmo ou valor mais apropriado para parâmetro(s)

# Índices relativos

- Variância intracluster:  $\text{var}(\pi)$ 
  - Mede compactação de clusters
  - Valores entre  $[0, \infty]$ 
    - Quanto menor o valor, melhor a partição

$$\text{var}(\pi) = \sqrt{\frac{1}{n} \sum_{C_k \in \pi} \sum_{x_i \in C_k} d(x_i, \bar{x}^{(k)})}$$

# Índices relativos

- Conectividade:  $\text{con}(\pi)$

- Ligada ao conceito de encadeamento

- Grau com que vizinhos são colocados no mesmo cluster

- Valores entre  $[0, v]$

- $v$  é o número de vizinhos mais próximos

- $nn_{ij}$  é o  $j$ -ésimo vizinho mais próximo a  $x_i$

- Quanto menor o valor, melhor a partição

$$\text{con}(\pi) = \sqrt{\sum_{x_i \in X} \sum_{j=1}^v f(x_i, nn_{ij})}$$

$$f(x_i, nn_{ij}) = \begin{cases} 1/j & \text{se } x_i \in C_k, nn_{ij} \in C_k \\ 0 & \text{caso contrário} \end{cases}$$

# Índices relativos

- Índices Dunn:  $D(\pi)$ 
  - Distância intracluster / dispersão do cluster
  - Razão da separação entre os clusters e dentro dos clusters
  - Identificação de clusters compactos e bem separados
    - Valores altos indicam presença desse tipo de cluster
    - Ponto máximo de gráfico contra  $k$  indica número de clusters
    - Complexo e sensível a ruído

$$D(\pi) = \min_{a=1,\dots,k} \left\{ \min_{b=a+1,\dots,k} \left\{ \frac{d(\mathcal{C}_a, \mathcal{C}_b)}{\max_{l=1,\dots,k} d(\mathcal{C}_l)} \right\} \right\}$$

$$d(\mathcal{C}_a, \mathcal{C}_b) = \min_{x_i \in \mathcal{C}_a, x_j \in \mathcal{C}_b} d(x_i, x_j)$$

$$d(\mathcal{C}_a) = \max_{x_i, x_j \in \mathcal{C}_a} d(x_i, x_j)$$

# Índices relativos

- Silhueta:  $\text{sil}(\pi)$ 
  - Baseia na proximidade entre objetos de um cluster e distância deles ao cluster mais próximo
  - Avalia adequação de objeto ao seu cluster, de cada cluster, de partição
  - Valores em [-1,1]
    - Melhor valor próximo de 1
  - Apropriadas para identificação de clusters compactos e bem separados (esféricos)
    - Favorece agrupamentos disjuntos

# Índices relativos

- Silhueta de um objeto:
  - Próximo de 1: objeto está bem situado em seu cluster
  - Próximo de -1: objeto deveria estar em outro cluster
    - $a(x_i, C_i)$ : distância média de  $x_i$  a todos objetos em  $C_i$
    - $b(x_i)$ : menor distância de  $x_i$  em relação aos demais clusters

$$sil(x_i) = \begin{cases} 1 - a(x_i, C_i) & se\ a(x_i, C_i) < b(x_i) \\ 0 & se\ a(x_i, C_i) = b(x_i) \\ \frac{b(x_i)}{a(x_i, C_i)} - 1 & se\ a(x_i, C_i) > b(x_i) \end{cases}$$

$$a(x_i, C_i) = \frac{1}{|C_k|} \sum_{x_i, x_j \in C_k, x_i \neq x_j} d(x_i, x_j)$$

$$b(x_i) = \min_{x_i \in C_i, C_i \neq C_j} a(x_i, C_j)$$

# Índices relativos

- Silhueta de um cluster:

$$sil(\mathcal{C}_k) = \frac{1}{|\mathcal{C}_k|} \sum_{x_i \in \mathcal{C}_k} sil(x_i)$$

- Largura média da silhueta:

$$sil(\pi) = \frac{1}{n} \sum_{i=1}^n sil(x_i)$$

Melhor k: resulta em maior  $sil(\pi)$

# Índices relativos

- Coeficiente de silhueta: SC
  - Máximo  $\text{sil}(\pi)$  para  $\pi$  com  $k = 2, 3, \dots, n-1$
  - Quantifica estrutura descoberta

## Interpretação SC:

- $SC \leq 0,25$  : não foi encontrada uma estrutura substancial
- $0,26 \leq SC \leq 0,5$  : estrutura é fraca e pode ser artificial
- $0,5 \leq SC \leq 0,7$  : estrutura razoável foi encontrada
- $0,7 \leq SC \leq 1$  : estrutura é forte

# Índices relativos

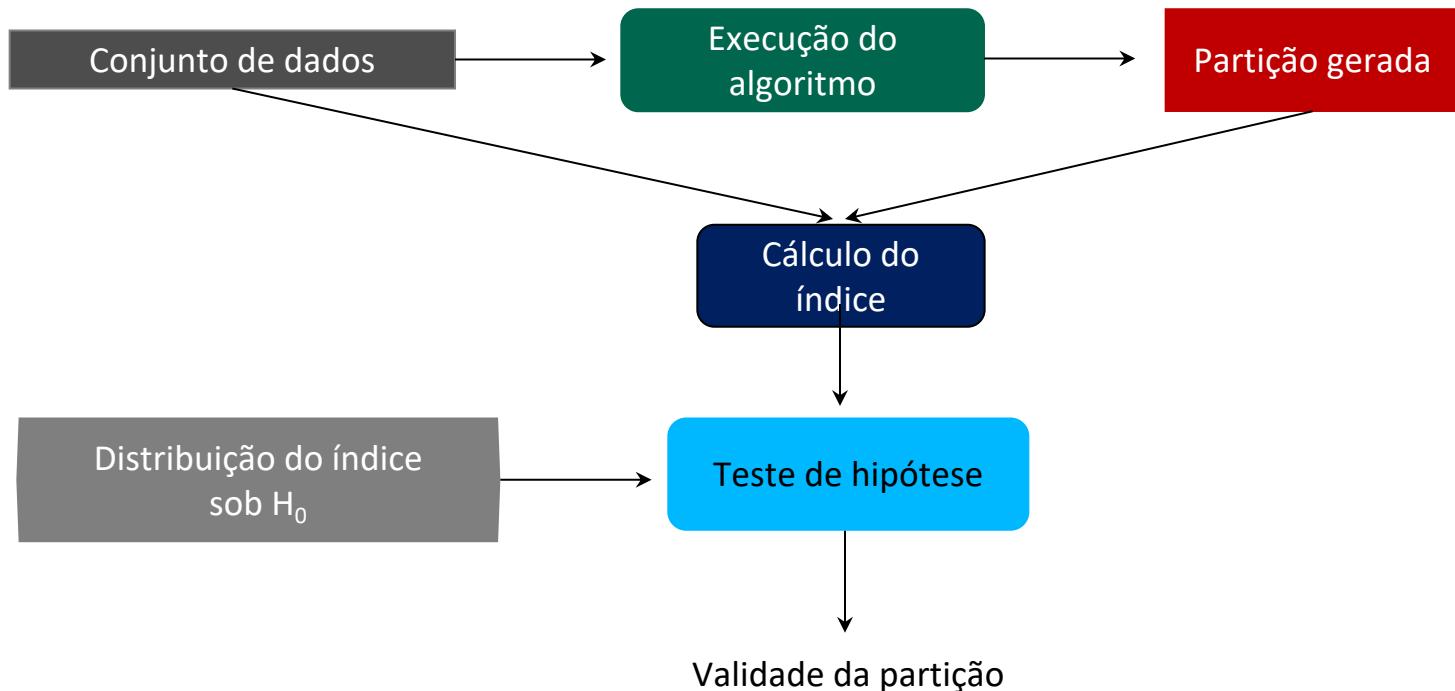
- Análise de replicação: semelhante a cross-validation em aprendizado supervisionado
  - Medir estabilidade de um algoritmo
  - Medida comparando partição do algoritmo a uma partição obtida em um subconjunto independente de dados
- Uso mais comum: determinar número adequado de clusters
  - Algoritmo é executado para entre  $k_{\min}$  e  $k_{\max}$  clusters
  - Valores de índices plotados em função de  $k$
  - Escolhe-se um ponto no gráfico

# Critérios externos e internos

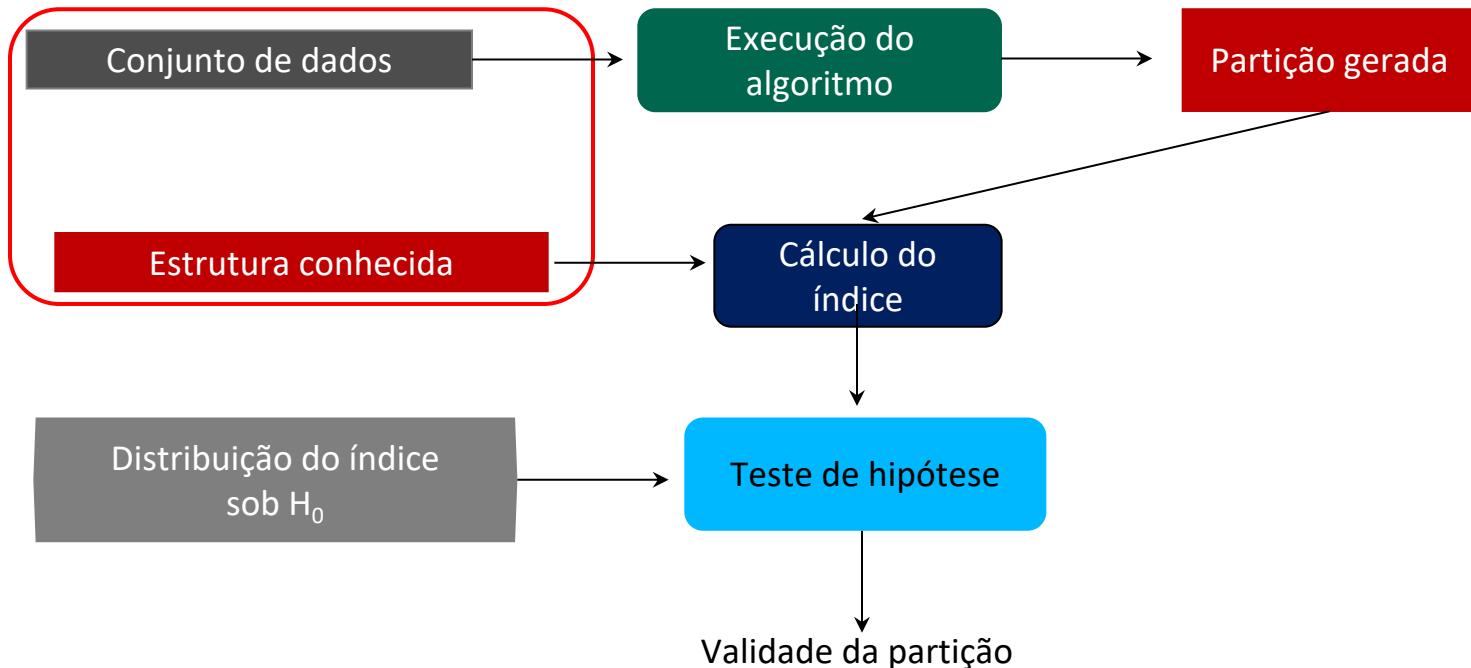
- Baseados em testes estatísticos
  - Medem quanto resultado confirma uma hipótese
    - Testes de hipóteses que determinam se estrutura obtida é adequada para os dados
      - Verificando se valor do índice é muito grande ou pequeno
    - Têm um maior custo computacional

**Hipótese nula  $H_0$ :** afirmação sobre aleatoriedade ou falta de estrutura nos dados

# Critérios internos



# Critérios externos



# Critérios internos e externos

- Problema: definir limiares estatísticos dos índices
  - Na prática, são definidos com ferramentas como a análise de Monte Carlo e bootstrapping
- Uso de reamostragem dos dados, com substituição
  - Amostras podem ser dos objetos ou dos atributos
  - Amostras são usadas para construir o modelo nulo

# Critérios internos

- Medem grau com que uma partição obtida representa uma estrutura presente nos dados
  - Baseia-se apenas na matriz de objetos ou de similaridade
  - Medem ajuste entre a partição gerada e os dados usados
  - Apresenta dificuldades por dependência dos dados
  - Podem ser usados como índices relativos
    - E índices anteriores também podem ser usados como internos

# Índices internos

- Gap:
  - Avalia dispersão intracluster em relação a esperança sob uma distribuição de referência nula

$$Gap(k) = E^*\{\log(W_k)\} - \log(W_k)$$

$$W_k = \sum_{r=1}^k \frac{1}{2n_r} D_r$$

$$D_r = \sum_{x_i, x_j \in C_r} d(x_i, x_j)$$

- Estimar número de clusters: máximo Gap
- Encontrar distribuição de referência:
  - Ex. Uniformemente para cada atributo
- Usa método Monte Carlo e faz B execuções
- Assume cluster se uniformes bem separados

# Critérios externos

- Medem quanto o agrupamento obtido confirma uma hipótese pré-especificada
  - Usa testes de hipóteses
  - Normalmente usada com método Monte Carlo
  - Sejam:
    - $\pi_e$  uma partição resultante do agrupamento
    - $\pi_r$  uma partição independente dos dados (com base em alguma estrutura real dos dados)

# Índices externos

- Rand: probabilidade de que dois objetos pertençam ao mesmo cluster ou a clusters diferentes em  $\pi_e$  e  $\pi_r$

$$R(\pi^e, \pi^r) = \frac{(a1 + a4)}{M}$$

- a1 = número de pares de objetos no mesmo cluster em  $\pi_e$  e  $\pi_r$
  - a4 = número de pares de objetos em clusters diferentes em ambos  $\pi_e$  e  $\pi_r$
  - M = n(n-1)/2
- Normalmente é normalizado entre 0 e 1: rand corrigido
    - 0 indica partição ao acaso
    - 1 indica partição que casa perfeitamente com a real

# Índices externos

- Jaccard: probabilidade de que dois objetos pertencentes ao mesmo cluster em uma partição também pertencem ao mesmo cluster na outra partição

$$J(\pi^e, \pi^r) = \frac{a1}{(a1 + a2 + a3)}$$

- a1 = número de pares de objetos no mesmo cluster em  $\pi_e$  e  $\pi_r$
- a2 = número de pares de objetos que estão no mesmo cluster em  $\pi_e$  e em clusters diferentes em  $\pi_r$
- a3 = número de pares de objetos em clusters diferentes em  $\pi_e$  e no mesmo cluster em  $\pi_r$

Slides construídos com base no material fornecido pela autora do livro ‘inteligência artificial: uma abordagem de aprendizado de máquina’ (Faceli, 2011).