



ser
educacional

gente criando o futuro



Inteligência Analítica

unidade 3

ANDRÉ TIBA (andre.tiba@sereeducacional.com)

Sumário do curso



- Unidade 1 – Introdução à Estatística
- Unidade 2 – Introdução à Mineração de Dados
- **Unidade 3 – Introdução à Modelos de Agrupamento e Predição**
- Unidade 4 – Aplicação e Persistência do Conhecimento



Objetivos da unidade 3

- Entender os principais modelos de agrupamento.
- Implicações e extração de conhecimentos a partir de modelos de agrupamento.
- Conhecer modelos preditivos e como eles auxiliam na tomada de decisão.



Sumário da unidade 3

- 1) Modelos de Agrupamento
- 2) Análise de Modelos de Agrupamento
- 3) Construção de Modelos de Predição

1) Modelos de Agrupamento



- Também chamado de Clusterização (clustering)
- Baseia-se na ideia de agrupar dados com **mesmas** similaridades ou características.

1) Modelos de Agrupamento



- Modelos Hierárquicos
 - Aglomerativos
 - Divisivos
- Modelos Particionais
 - Exclusivos
 - Não Exclusivos

1) Modelos de Agrupamentos Hierárquicos

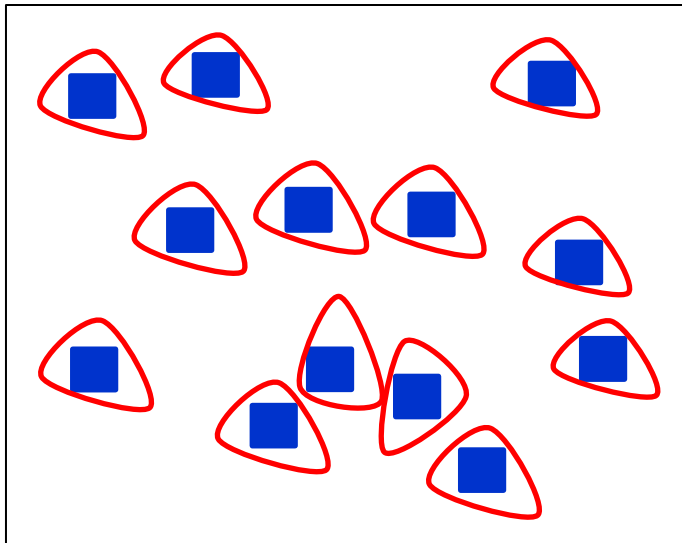


- Modelo Hierárquico Aglomerativo
- Em geral, **definimos** a quantidade de agrupamentos que desejamos para o final do processo.
- Baseado na distância entre os padrões:
 - quanto mais parecidos forem dois padrões, menor é a distância entre eles.

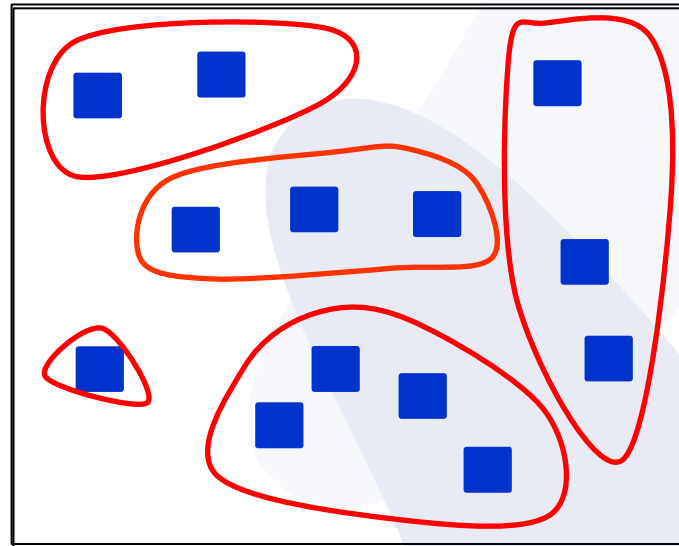
1) Modelos de Agrupamentos Hierárquicos



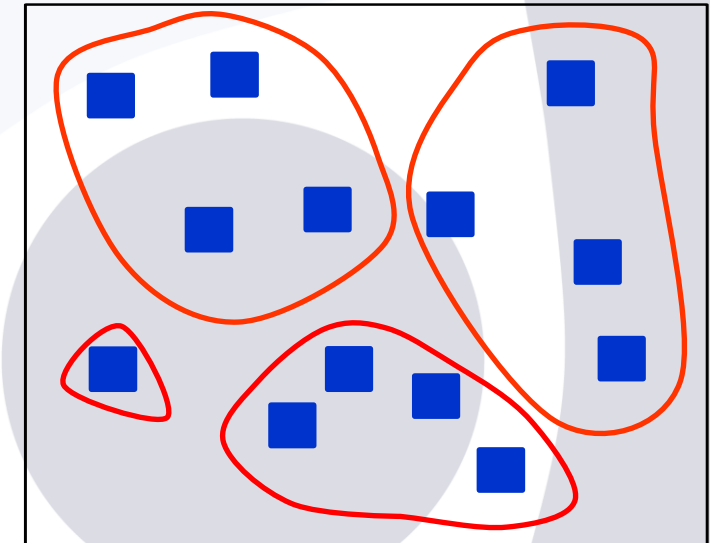
- Modelo Hierárquico Aglomerativo



início: cada padrão forma um grupo distinto (13 grupos).



padrões começam a aglomerar por similaridade, reduzindo a quantidade de grupos.



O processo de aglomeração e redução de grupos se encerra após alcançar um critério pré estabelecido.

1) Modelos de Agrupamentos Hierárquicos

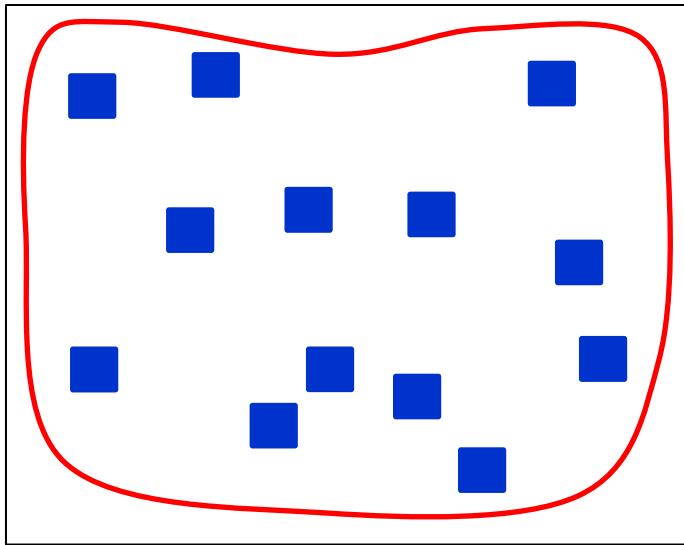


- Modelo Hierárquico Divisivo
- Em geral, **definimos** a quantidade de agrupamentos que desejamos para o final do processo.
- Baseado na distância entre os padrões:
 - quanto mais parecidos forem dois padrões, menor é a distância entre eles.

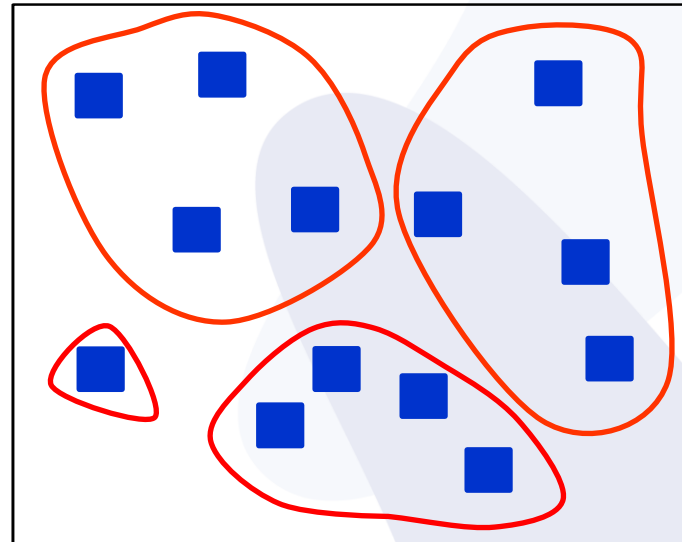
1) Modelos de Agrupamentos Hierárquicos



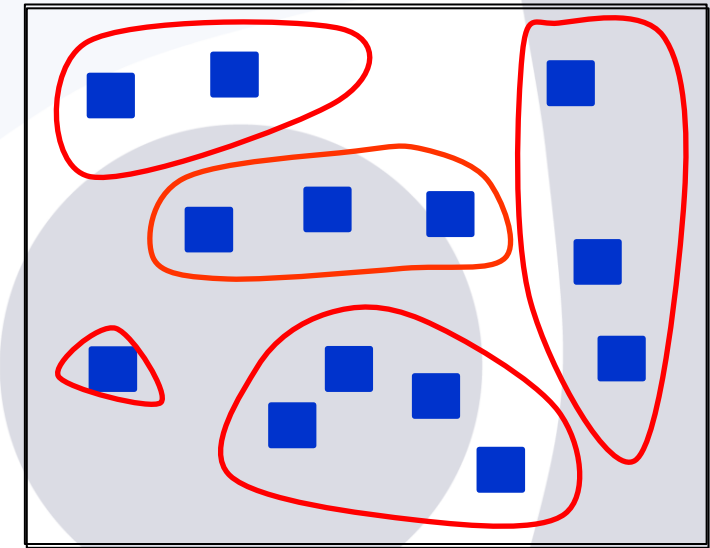
- Modelo Hierárquico Divisivo



início: todos os padrões formam um único grupo.



padrões começam se dividir por similaridade, aumentando a quantidade de grupos.

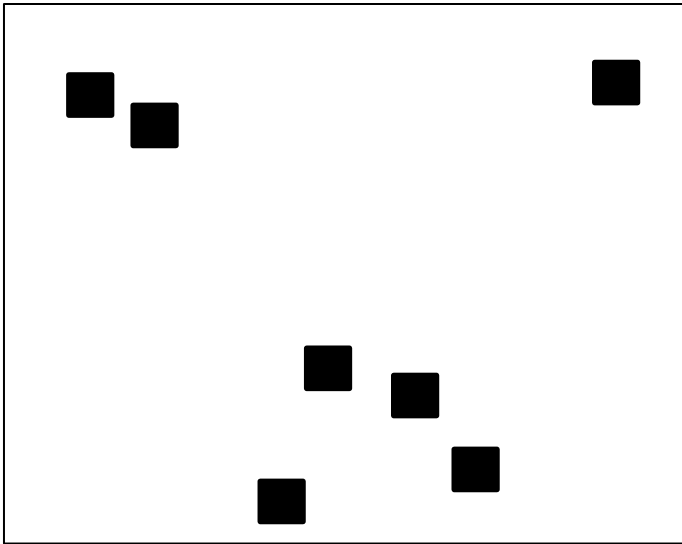


O processo de divisão e aumento de grupos se encerra após alcançar um critério pré estabelecido.

1) Modelos de Agrupamentos Hierárquicos



- Modelo Hierárquico Aglomerativo: dendograma

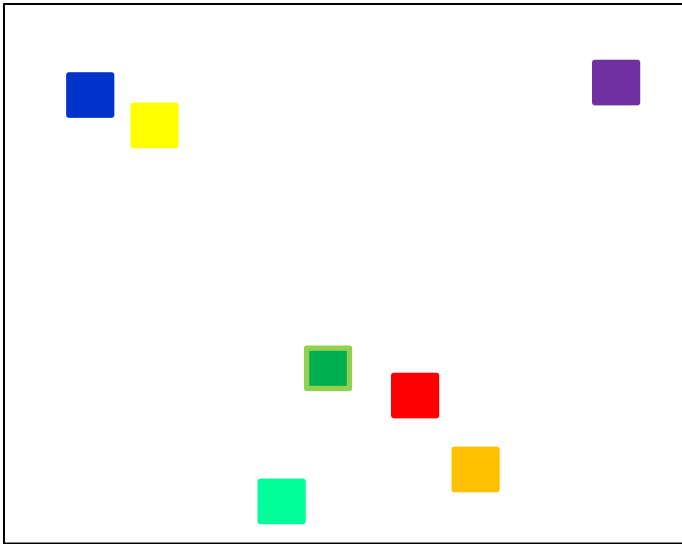


início: conjunto de padrões

1) Modelos de Agrupamentos Hierárquicos



- Modelo Hierárquico Aglomerativo: dendograma

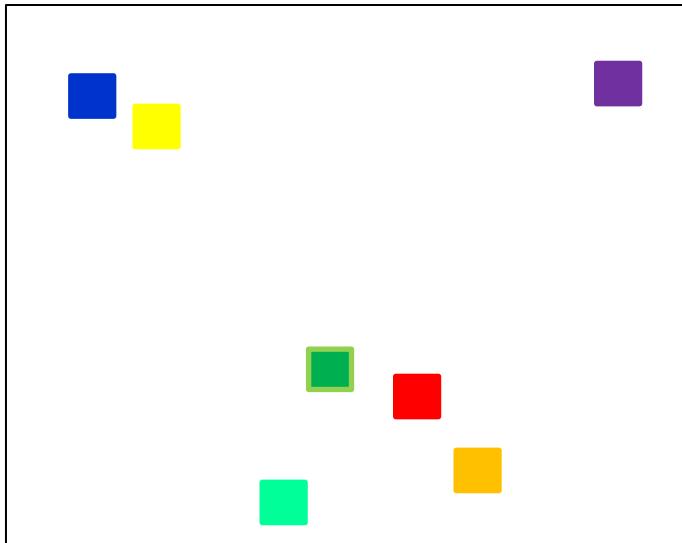


os padrões foram pintados apenas para diferenciá-los.



1) Modelos de Agrupamentos Hierárquicos



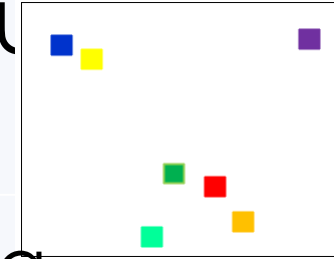
- Modelo Hierárquico Aglomerativo: dendograma



Distância entre padrões



							
	0						
	0,91	0					
	0,97	0,21	0				
	0,94	0,64	0,56	0			
	0,91	0,16	0,19	0,60	0		
	0,92	0,17	0,22	0,72	0,18	0	
	0,89	0,58	0,61	0,11	0,56	0,61	0

1) Modelos de Agrupamentos Hierárquicos



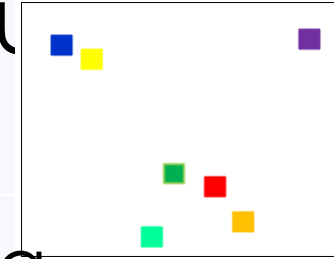
- Modelo Hierárquico Aglomerativo: dendograma

Distância entre padrões

							
	0						
	0,91	0					
	0,97	0,21	0				
	0,96	0,64	0,56	0			
	0,91	0,16	0,19	0,60	0		
	0,92	0,17	0,22	0,72	0,18	0	
	0,90	0,58	0,61	0,11	0,56	0,61	0












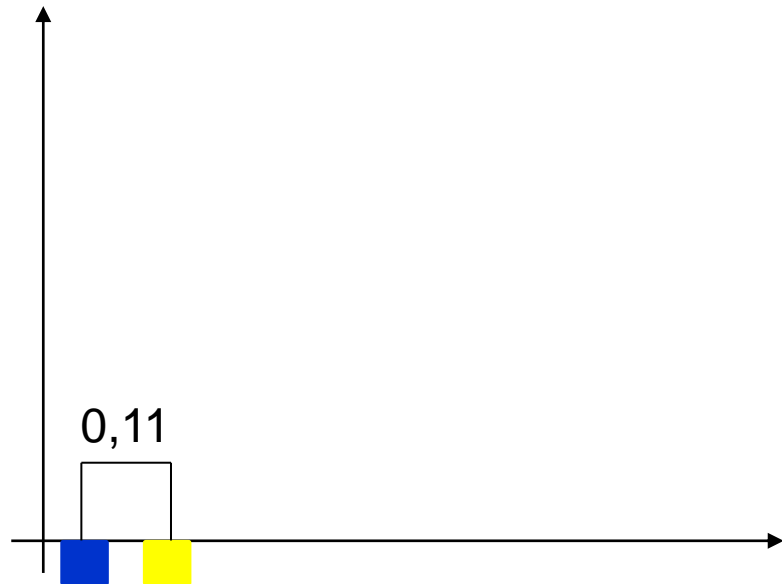
1) Modelos de Agrupamentos Hierárquicos



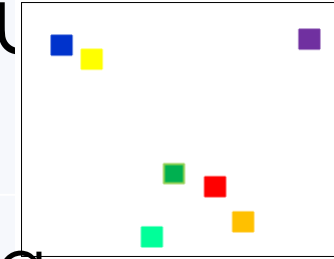
- Modelo Hierárquico Aglomerativo: dendograma

Distância entre padrões

							
	0						
	0,91	0					
	0,97	0,21	0				
	0,96	0,64	0,56	0			
	0,91	0,16	0,19	0,60	0		
	0,92	0,17	0,22	0,72	0,18	0	
	0,90	0,58	0,61	0,11	0,56	0,61	0

















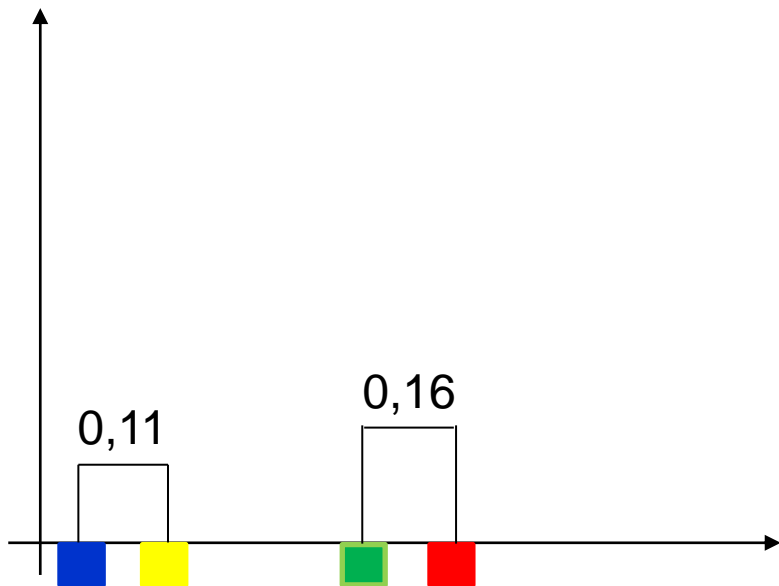
1) Modelos de Agrupamentos Hierárquicos



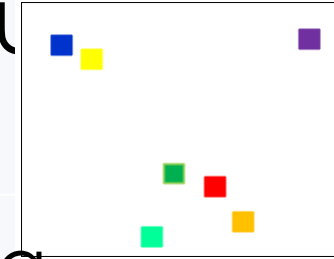
- Modelo Hierárquico Aglomerativo: dendograma

Distância entre padrões

							
	0						
	0,91	0					
	0,97	0,21	0				
	0,96	0,64	0,56	0			
	0,91	0,16	0,19	0,60	0		
	0,92	0,17	0,22	0,72	0,18	0	
	0,90	0,58	0,61	0,11	0,56	0,61	0

















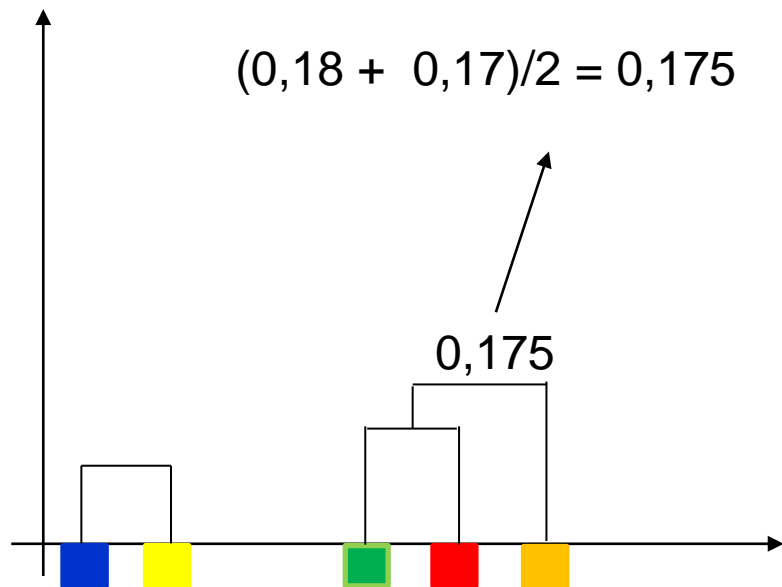
1) Modelos de Agrupamentos Hierárquicos



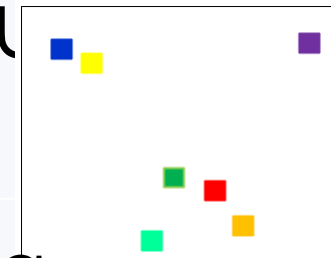
- Modelo Hierárquico Aglomerativo: dendograma

Distância entre padrões

							
	0						
	0,91	0					
	0,97	0,21	0				
	0,96	0,64	0,56	0			
	0,91	0,16	0,19	0,60	0		
	0,92	0,17	0,22	0,72	0,18	0	
	0,90	0,58	0,61	0,11	0,56	0,61	0






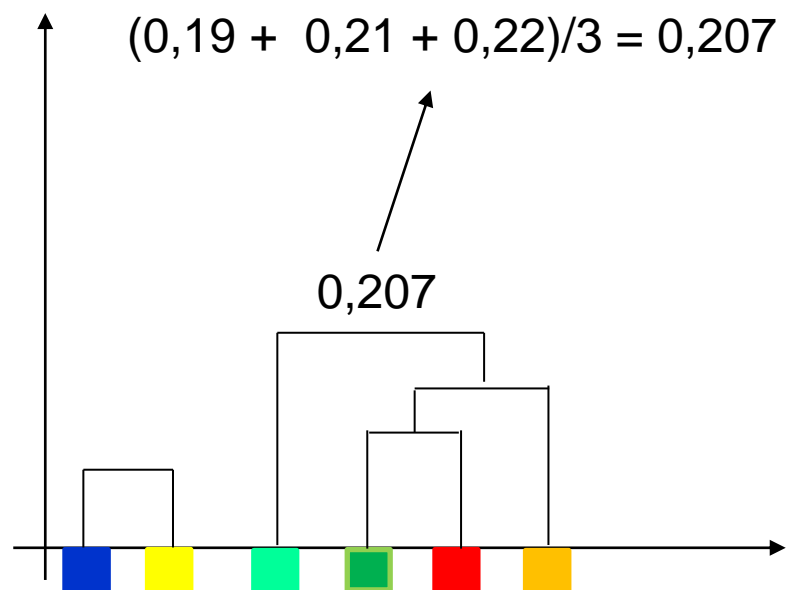
1) Modelos de Agrupamentos Hierárquicos



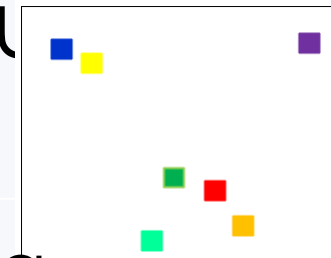
- Modelo Hierárquico Aglomerativo: dendograma

Distância entre padrões

							
	0						
	0,91	0					
	0,97	0,21	0				
	0,96	0,64	0,56	0			
	0,91	0,16	0,19	0,60	0		
	0,92	0,17	0,22	0,72	0,18	0	
	0,90	0,58	0,61	0,11	0,56	0,61	0

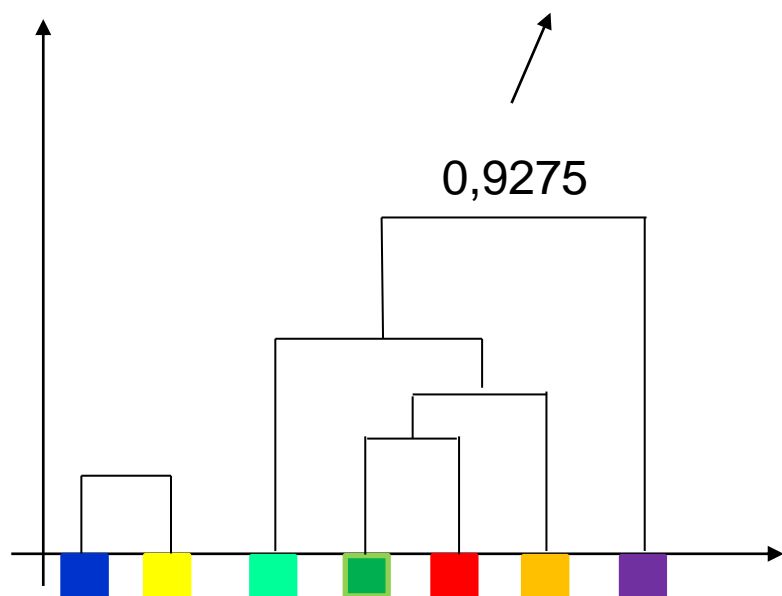


1) Modelos de Agrupamentos Hierárquicos

















- Modelo Hierárquico Aglomerativo: dendograma

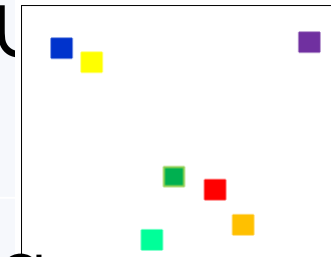
$$(0,91 + 0,97 + 0,91 + 0,92)/4 = 0,9275$$



Distância entre padrões

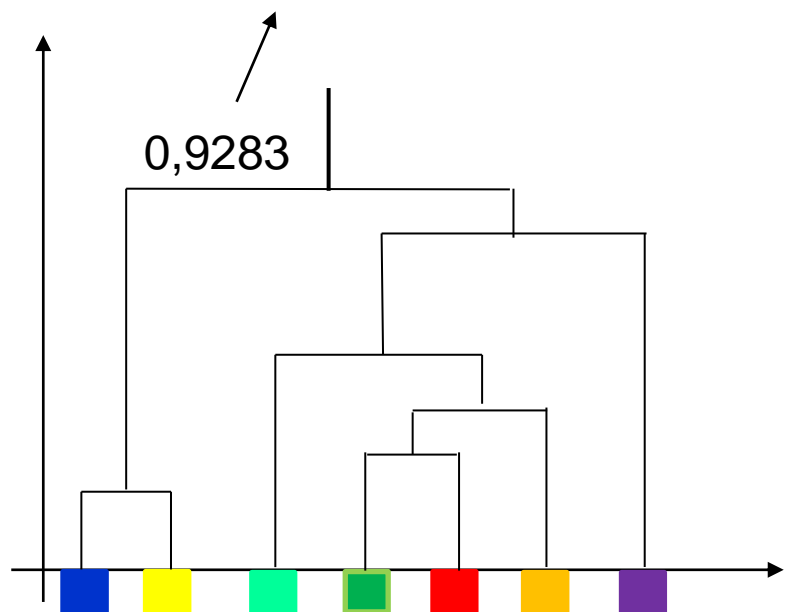
							
	0						
	0,91	0					
	0,97	0,21	0				
	0,96	0,64	0,56	0			
	0,91	0,16	0,19	0,60	0		
	0,92	0,17	0,22	0,72	0,18	0	
	0,90	0,58	0,61	0,11	0,56	0,61	0

1) Modelos de Agrupamentos Hierárquicos

















- Modelo Hierárquico Aglomerativo: dendograma

$$(0,91 + 0,97 + 0,91 + 0,92 + 0,90 + 0,96)/6 = 0,9283$$



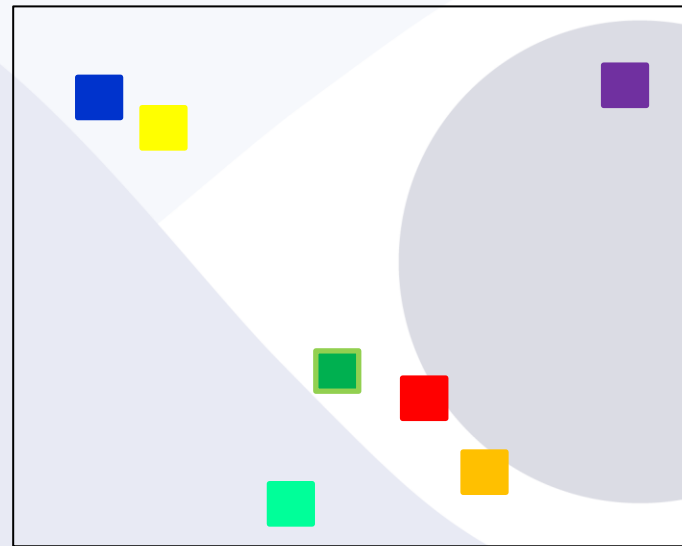
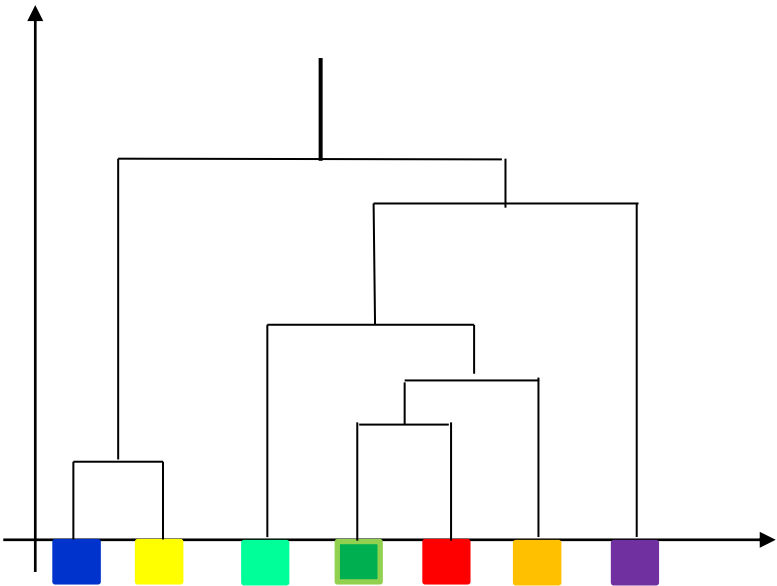
Distância entre padrões

							
	0						
	0,91	0					
	0,97	0,21	0				
	0,96	0,64	0,56	0			
	0,91	0,16	0,19	0,60	0		
	0,92	0,17	0,22	0,72	0,18	0	
	0,90	0,58	0,61	0,11	0,56	0,61	0

1) Modelos de Agrupamentos Hierárquicos



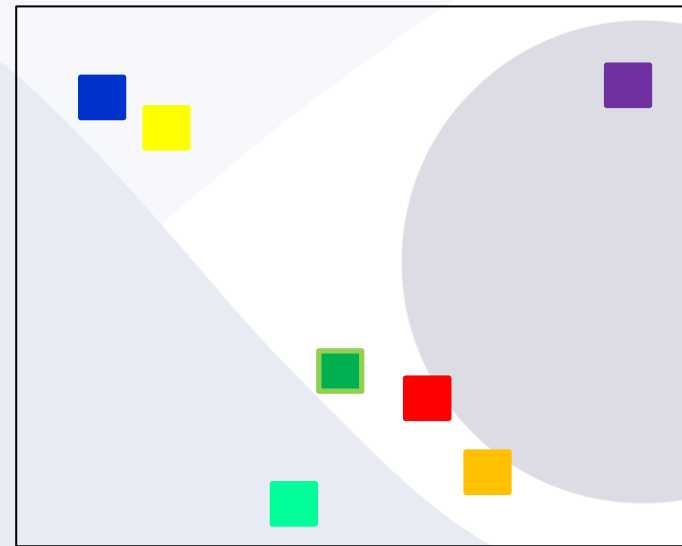
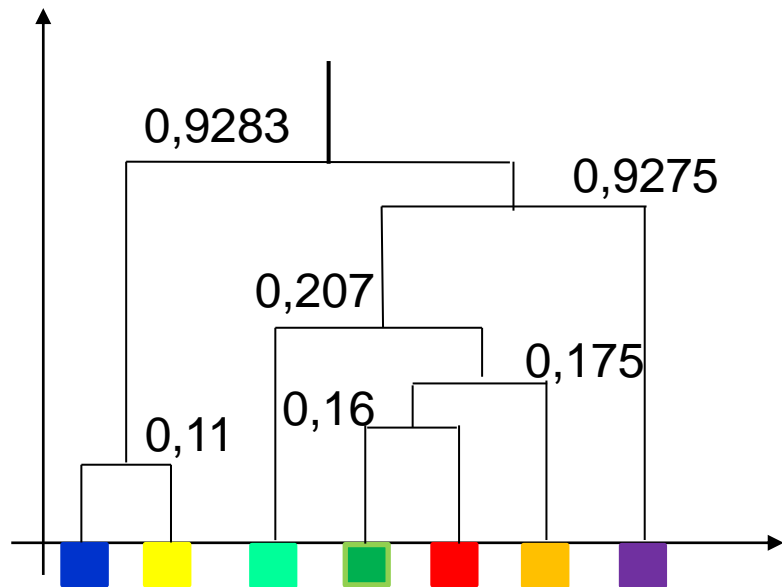
- Modelo Hierárquico Aglomerativo: dendograma



1) Modelos de Agrupamentos Hierárquicos



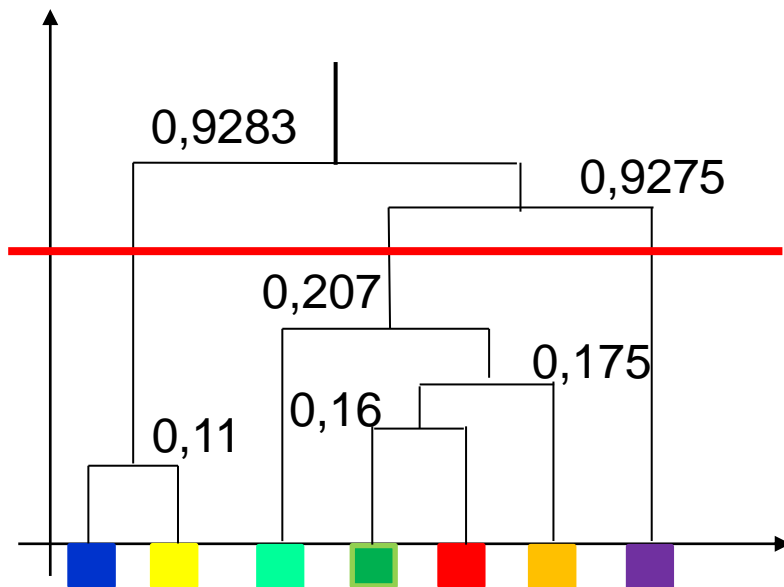
- Modelo Hierárquico Aglomerativo



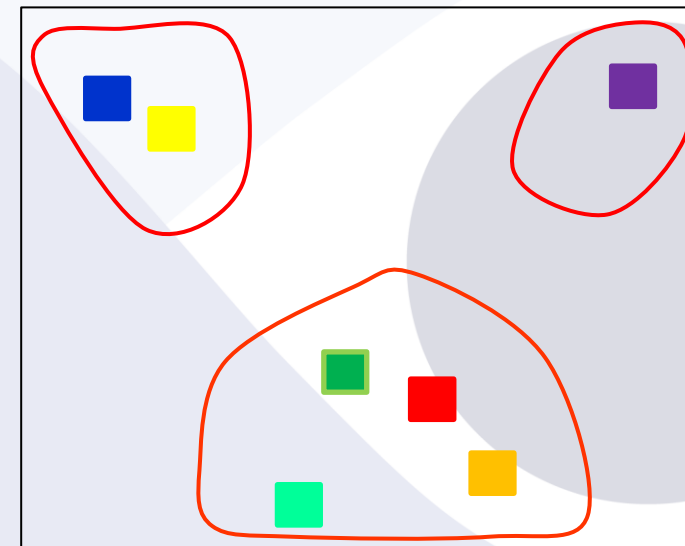
1) Modelos de Agrupamentos Hierárquicos



- Modelo Hierárquico Aglomerativo: dendograma



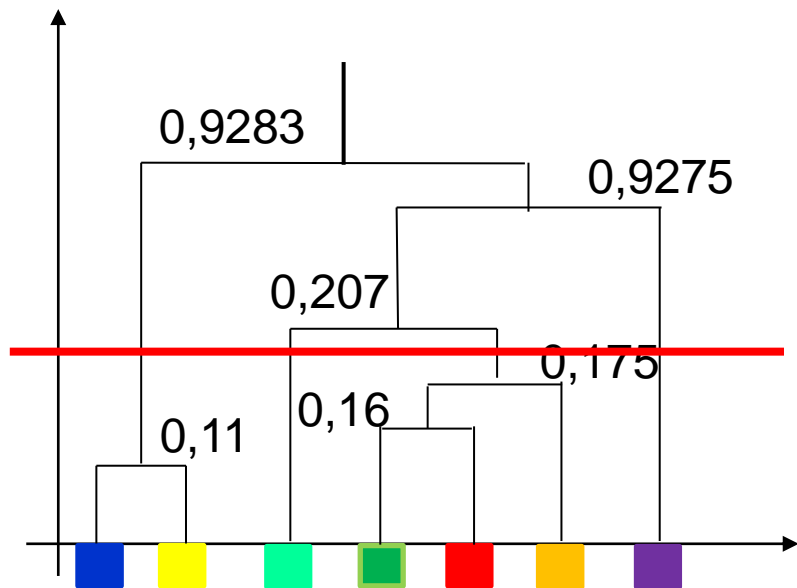
três grupos



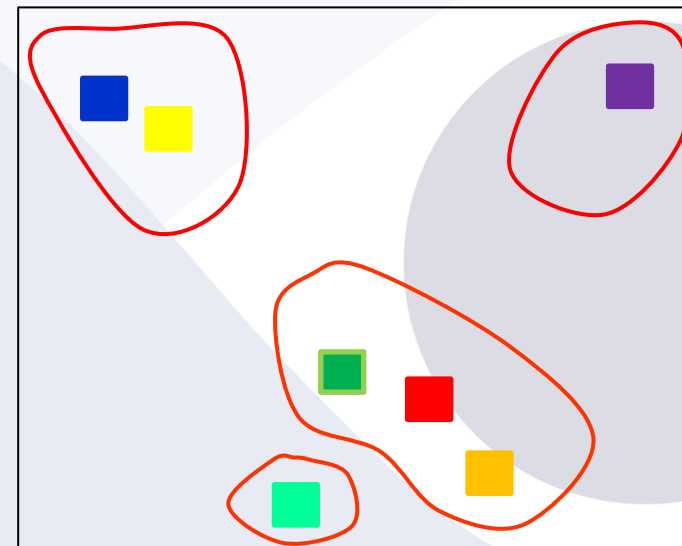
1) Modelos de Agrupamentos Hierárquicos



- Modelo Hierárquico Aglomerativo: dendograma



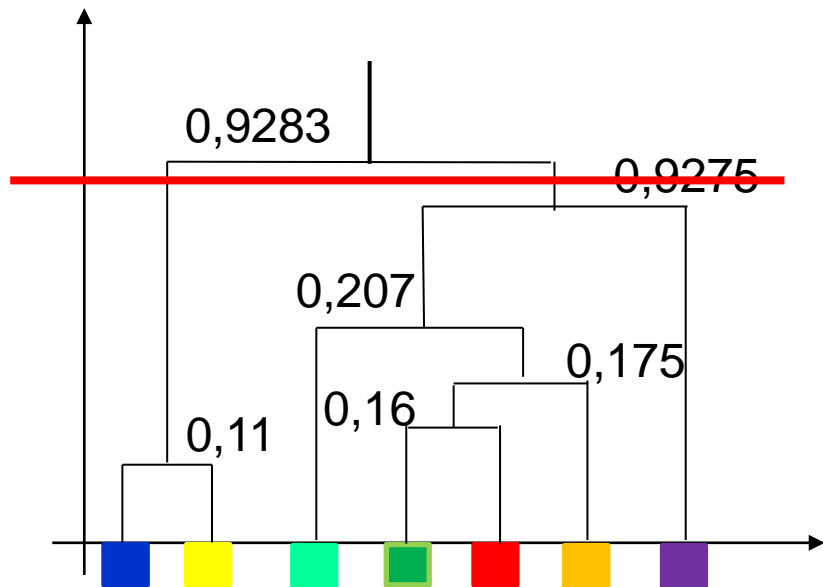
quatro grupos



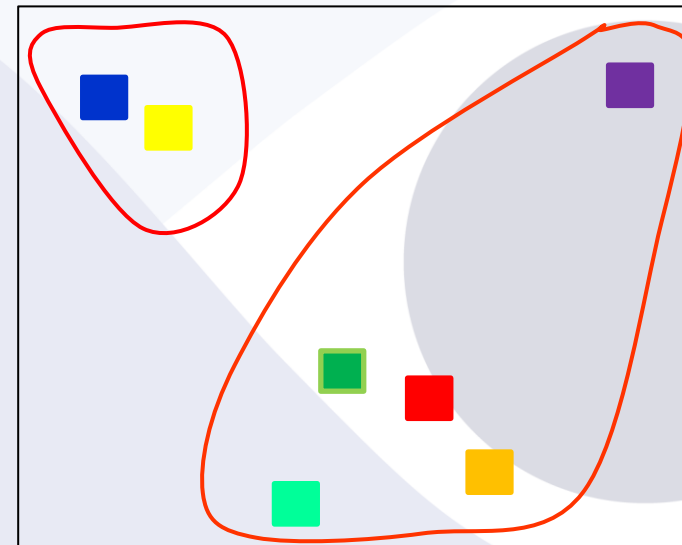
1) Modelos de Agrupamentos Hierárquicos



- Modelo Hierárquico Aglomerativo: dendograma



dois grupos



1) Modelos de Agrupamentos Hierárquicos



- Modelo Hierárquico Aglomerativo: dendograma

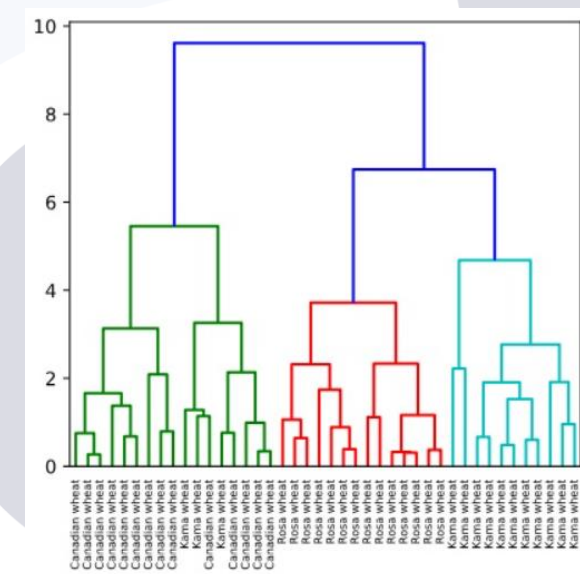


h é a medida similaridade entre dois grupos

esta representação forma uma árvore!



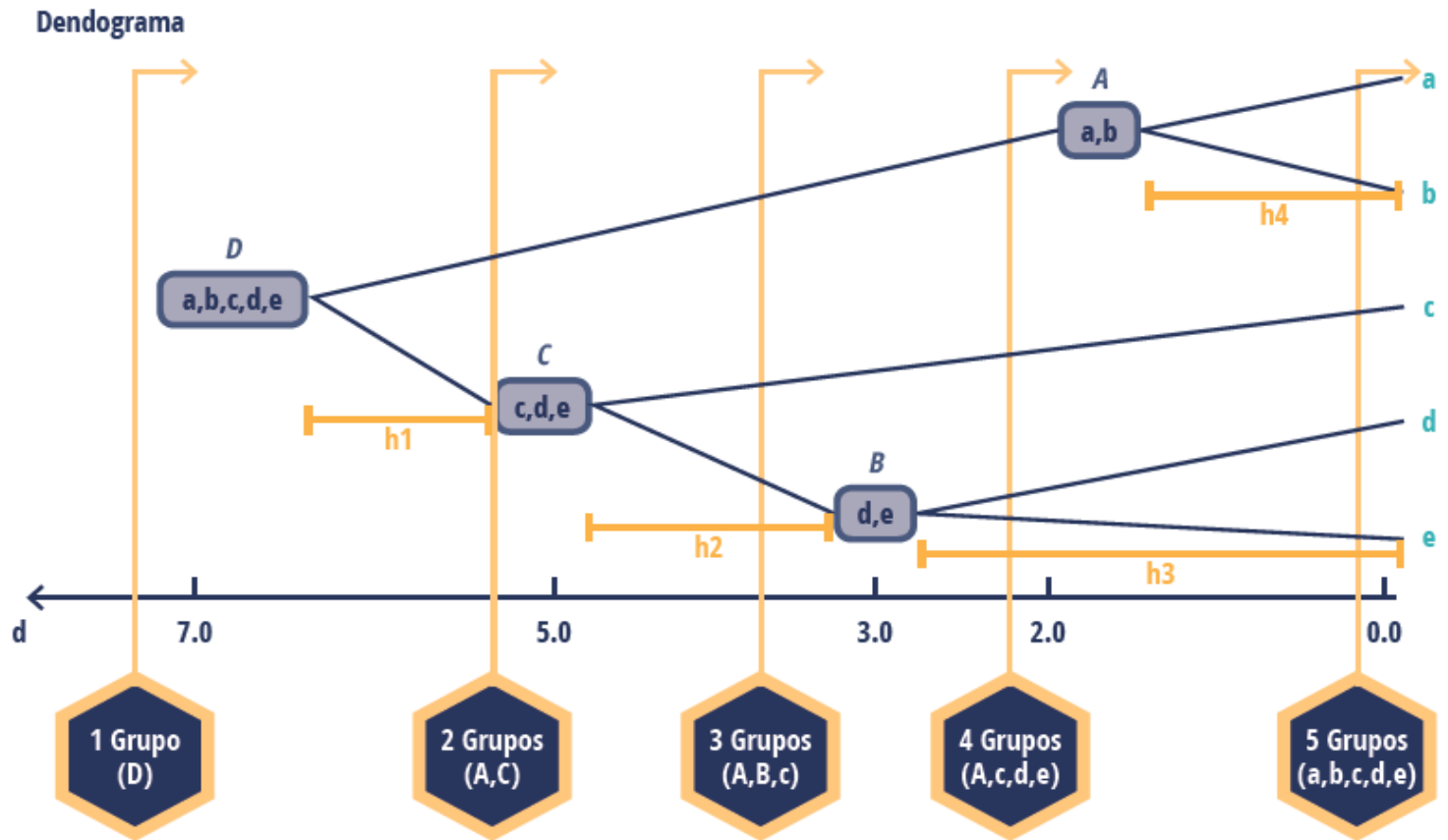
-
- Dendrogram**
Complete Linkage, Euclidean Distance
- Distance
- Observations



1) Modelos de Agrupamento Hierárquicos



- dendograma



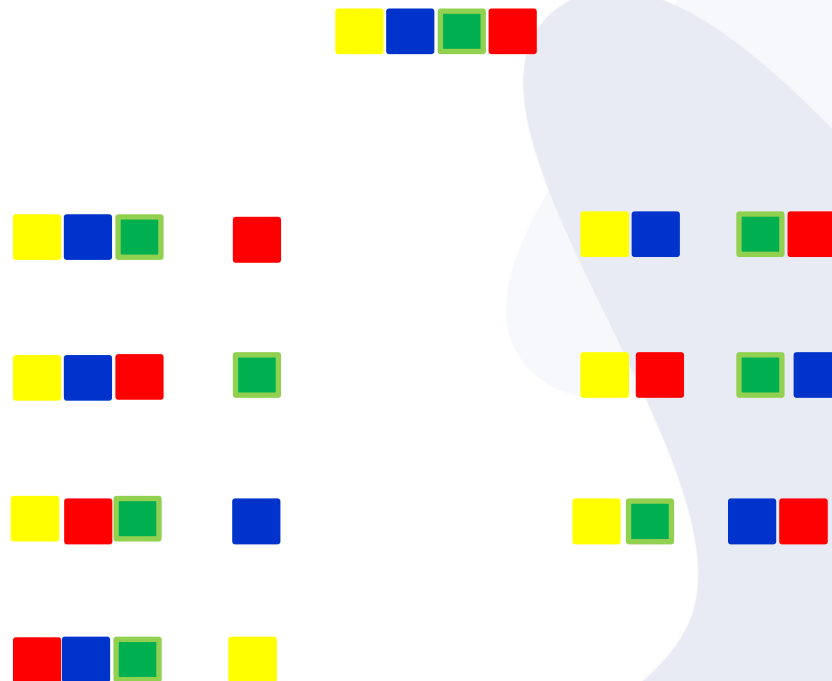
h é a medida similaridade entre dois grupos

esta representação forma uma árvore!



1) Modelos de Agrupamento Hierárquicos: Aglomerativo x Divisivo

- Modelos Aglomerativos precisam de menos computação pois os Modelos Divisivos testam todas as divisões possíveis de partição em grupos. **Impraticável!**



Para um conjunto com 4 padrões,
existem 7 combinações possíveis

1) Modelos de Agrupamento Hierárquicos: Aglomerativo x Divisivo



- Por outro lado, os modelos divisivos **erram menos** que os aglomerativos, portanto são mais confiáveis.

1) Modelos de Agrupamento



- Modelos Hierárquicos
 - Aglomerativos
 - Divisivos
- Modelos Particionais
 - Exclusivos
 - Não Exclusivos



1) Modelos de Agrupamentos Particionais

- Baseados na **minimização** de uma função de custo.
 - Grupos são formados de maneira a minimizar a função de custo.
- Observe que nesta abordagem, um padrão pode **mudar de grupo** ao longo do processo (o que não ocorre em modelos hierárquicos).
- É necessário escolher o número de grupos a priori (algo que nem sempre é conhecido).



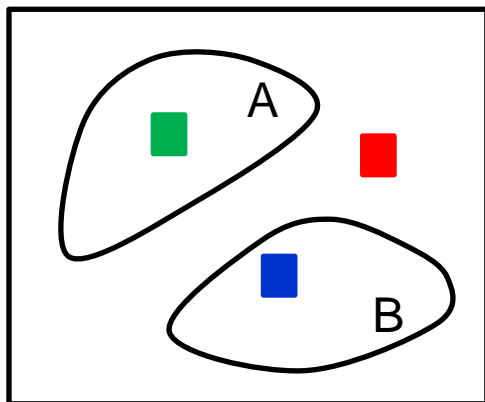
1) Modelos de Agrupamentos Particionais

- Modelos Exclusivos
 - Define a quantidade de grupos (ou classes), e um padrão analisado, **pertence a um grupo** e não pertence aos demais grupos.
- Modelos Não Exclusivos
 - Define a quantidade de grupos (classes), e um padrão analisado, possui **grau de pertinência** para **CADA** grupo.
 - Ex: lógica fuzzy

1) Modelos de Agrupamentos Particionais



Modelo Exclusivo



■ $\in A$ e ■ $\notin B$

■ $\notin A$ e ■ $\in B$

■ $\notin A$ e ■ $\notin B$

μ_A é o grau de pertinência do conjunto A

■
$$\begin{cases} \mu_A = 1 \\ \mu_B = 0 \end{cases}$$

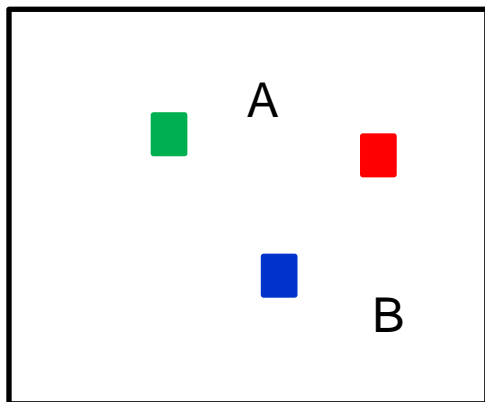
■
$$\begin{cases} \mu_A = 0 \\ \mu_B = 1 \end{cases}$$

■
$$\begin{cases} \mu_A = 0 \\ \mu_B = 0 \end{cases}$$

1) Modelos de Agrupamentos Particionais



Modelo não Exclusivo
ex: Lógica Fuzzy



μ_A é o grau de pertinência do conjunto A

■ $\begin{cases} \mu_A = 0.82 \\ \mu_B = 0.11 \end{cases}$

■ $\begin{cases} \mu_A = 0.22 \\ \mu_B = 0.94 \end{cases}$

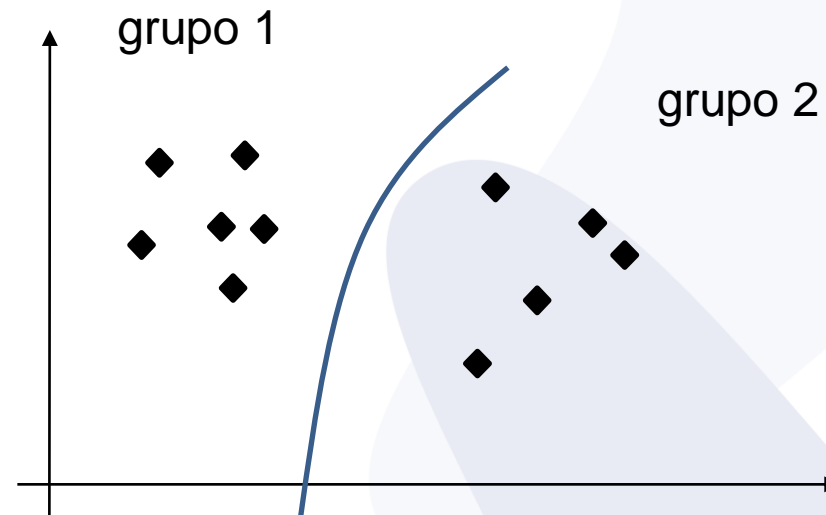
■ $\begin{cases} \mu_A = 0.47 \\ \mu_B = 0.45 \end{cases}$

1) Modelos de Agrupamentos Particionais



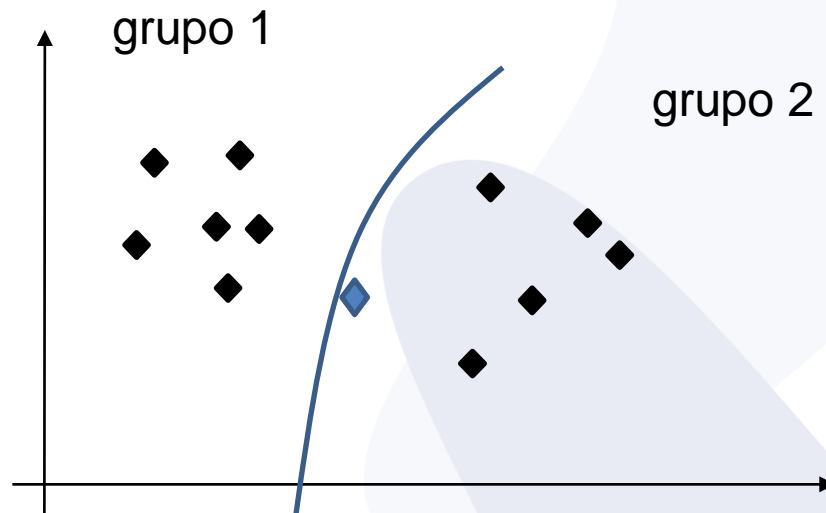
- Modelos não Exclusivos
 - Possibilita tratar ambiguidade através de um grau de pertinência.
 - Modelada de forma não estruturada.

1) Modelos de Agrupamentos Particionais – modelos não Exclusivos



Um conjunto de padrões e uma curva decorrente de um processo de aprendizagem, que separa dois grupos de padrões.

1) Modelos de Agrupamentos Particionais – modelos não Exclusivos

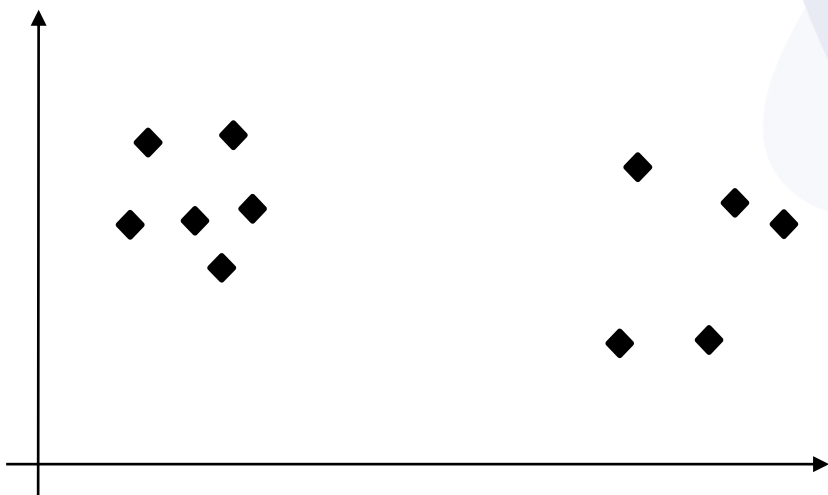


- O novo padrão **azul** será agrupado no grupo 2.
- Mas veja que ele está MUITO próximo da curva de decisão.
- Se estivesse só um pouco mais a esquerda seria agrupado no grupo 1.
- Com lógica fuzzy, $\mu_1 = 0.47$, $\mu_2 = 0.53$ teremos graus de pertinência para cada grupo.

1) Modelos de Agrupamentos Particionais baseados em distância



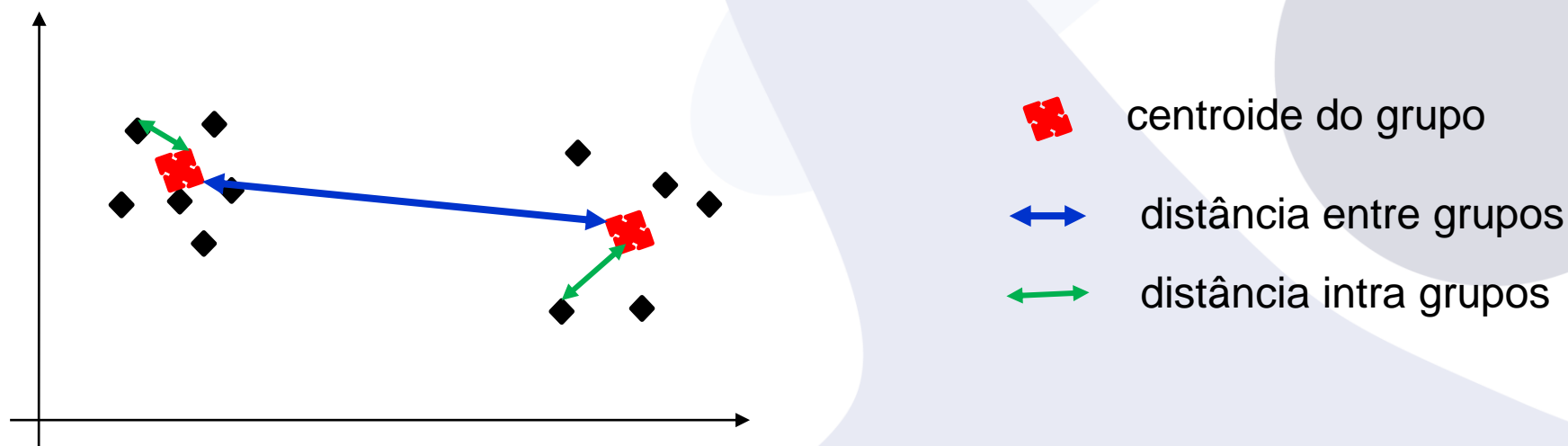
- Mede-se a distância entre os grupos
- Procura-se maximizar a distância entre grupos, ao mesmo tempo que deseja-se minimizar a distância intra grupos.



1) Modelos de Agrupamentos Particionais baseados em distância



- Mede-se a distância entre os grupos
- Procura-se maximizar a distância entre grupos, ao mesmo tempo que deseja-se minimizar a distância intra grupos.



1) Modelos de Agrupamentos Particionais baseados em distância



- **Similaridade** (lê-se como semelhança) entre padrões:
 - Padrões similares são semelhantes (similaridade alta) e portanto, a distância entre eles é pequena.
 - Padrões dissimilares são diferentes (similaridade baixa) e portanto, a distância entre eles é grande.
 - A similaridade é um número real no intervalo $[0,1]$
 - $0.78 \rightarrow$ alta similaridade
 - $0.12 \rightarrow$ baixa similaridade

1) Modelos de Agrupamentos Particionais baseados em distância



- **Similaridade** (lê-se como semelhança) entre padrões:
 - Padrões similares são semelhantes (similaridade alta) e portanto, a distância entre eles é pequena.
 - Padrões dissimilares são diferentes (similaridade baixa) e portanto, a distância entre eles é grande.
 - A similaridade é um número real no intervalo $[0,1]$
 - $0.78 \rightarrow$ alta similaridade
 - $0.12 \rightarrow$ baixa similaridade

1) Modelos de Agrupamentos Particionais baseados em distância



- Padrões **homogêneos (mesma forma)** possuem alta **Similaridade**.
- Padrões **heteromogêneos (diferente forma)** possuem alta **Dissimilaridade**.



1) Modelos de Agrupamentos Particionais baseados em distância

- Algoritmo K-means (de longe o mais usado):
 - Simples de ser implementado. É necessário definir a quantidade de grupos
 - Inicializa-se k centroides, calculando para cada padrão a distância até o centroide.
 - O padrão é estabelecido ao centroide do grupo mais próximo
 - Faz-se a correção das posições dos centroides e o processo iterativo recomeça.

1) Modelos de Agrupamentos: avaliação da qualidade do agrupamento

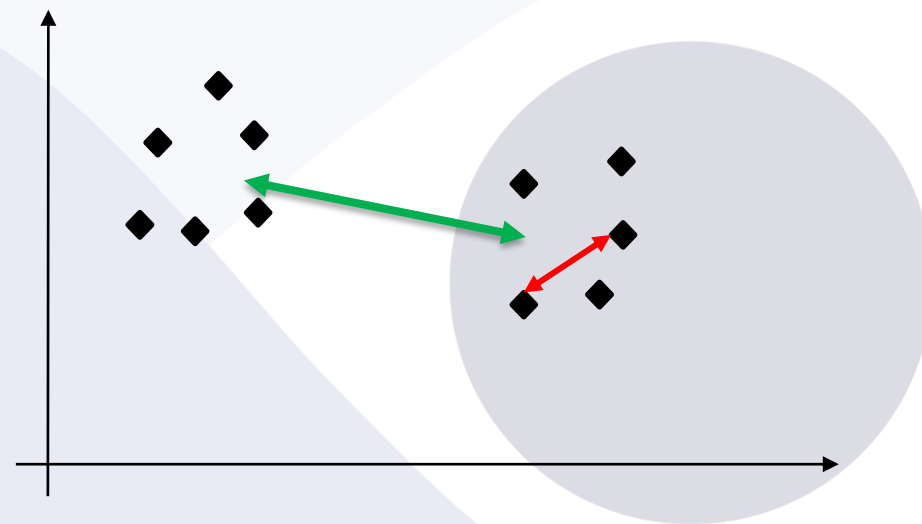
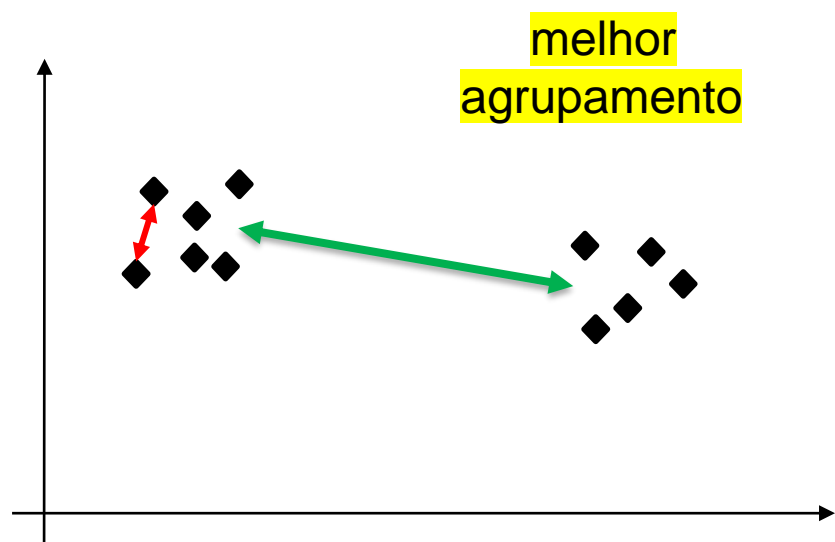


- Baseados na medida de variabilidade dentro grupo e entre grupos.

1) Modelos de Agrupamentos: avaliação da qualidade do agrupamento



- Baseados na medida de variabilidade dentro grupo e entre grupos.



↔ maximizar distância entre grupos
↔ minimizar distâncias dentro do grupo

2) Análise de Modelos de Agrupamento: conceitos



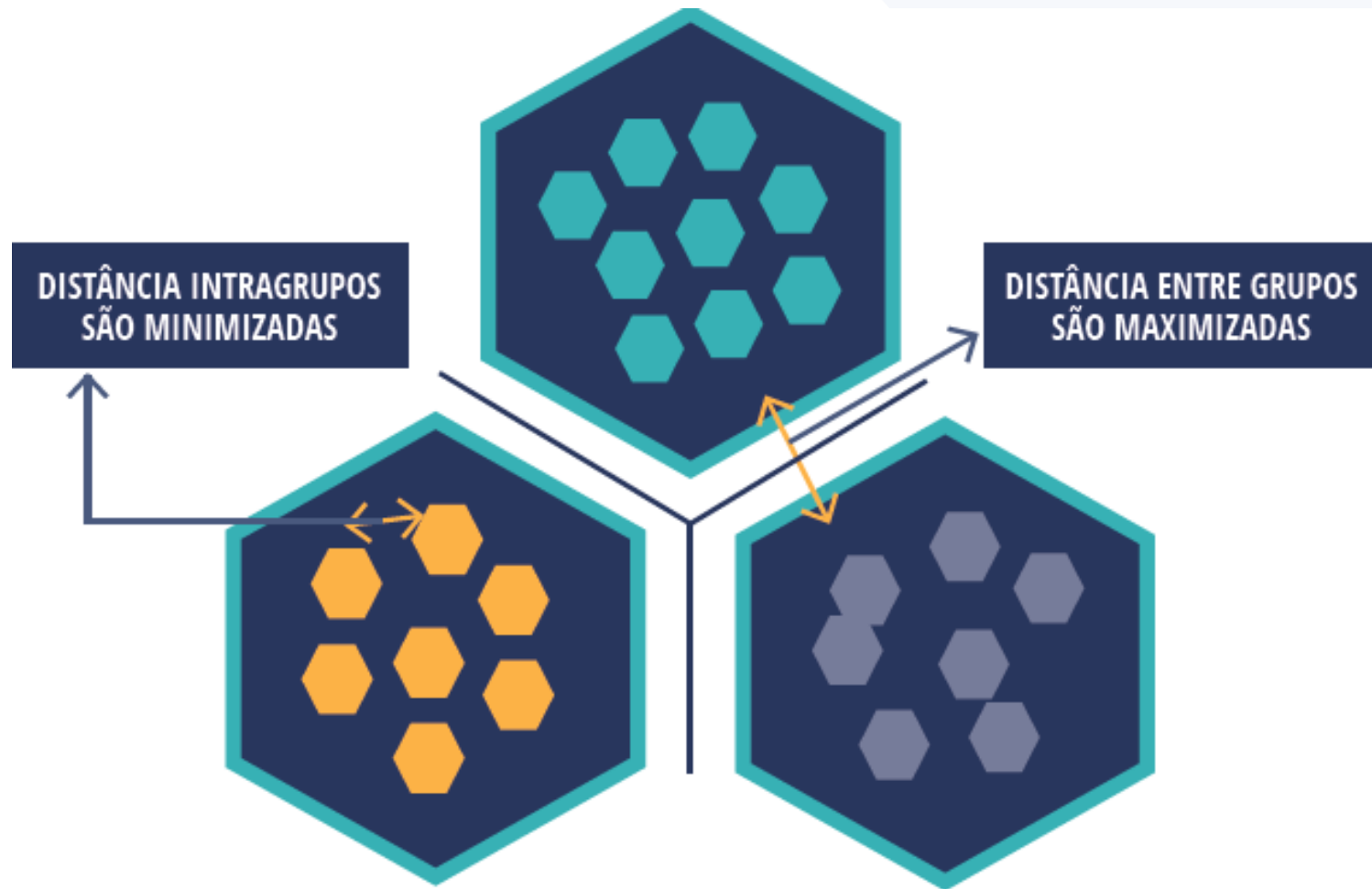
- Clustering (ou agrupamento): técnicas computacionais cujo objetivo é realizar a separação de dados em grupos.
- Utiliza os atributos dos dados para poder separar os grupos.
- Os dados podem ser numéricos e/ou categóricos.

2) Análise de Modelos de Agrupamento: conceitos



- Utiliza técnicas não supervisionadas
- Procura-se por homogeneidade (similaridade) entre componentes do grupo e heterogeneidade (dissimilaridade) em relação a outros grupos.

2) Análise de Modelos de Agrupamento: conceitos



2) Análise de Modelos de Agrupamento: Aplicações



- **Armazenamento de dados:** proporciona a redução de dimensionalidade dos dados.
- **Marketing:** ajuda na criação de perfis de consumidores, agrupando consumidores com mesmo perfil de compra.
- **Finanças:** entender perfis de clientes para disponibilidade de crédito, ofertas de serviços, etc ...

2) Análise de Modelos de Agrupamento: Limitações



- Não existe uma receita de bolo para encontrar **a melhor solução** para todos os problemas.
- Para base de dados de **cada problema**, teremos que procurar por uma boa solução, **experimentando diferentes técnicas de agrupamento** e ajustando os parâmetros. É muito empírico!

2) Análise de Modelos de Agrupamento: Medidas de Similaridade



- Conjunto de padrões em uma matriz: cada linha da matriz é um padrão, e as colunas são os atributos de cada padrão.

$$X = \begin{bmatrix} X_{11} & X_{12} & X_{13} & \cdots & X_{1p} \\ X_{21} & X_{22} & X_{23} & \cdots & X_{2p} \\ X_{31} & X_{32} & X_{33} & \cdots & X_{3p} \\ X_{41} & X_{42} & X_{43} & \cdots & X_{4p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & X_{n3} & \cdots & X_{np} \end{bmatrix}$$



2) Análise de Modelos de Agrupamento: Medidas de Similaridade

- Cálculo da dissimilaridade entre os padrões: quanto maior for a dissimilaridade mais diferentes são os padrões (medido para cada par de padrões).

$$\begin{bmatrix} 0 & d(1,2) & d(1,3) & \dots & d(1,n) \\ d(2,1) & 0 & d(2,3) & \dots & d(2,n) \\ d(3,1) & d(3,2) & 0 & \dots & d(3,n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d(n,1) & d(n,2) & d(n,3) & \dots & 0 \end{bmatrix} = D$$

3) Construção de Modelos de Predição



- O que é um modelo preditivo?
- Construimos um modelo de agrupamento, capaz de extrair informações e agrupar uma DETERMINADA base de dados (ou padrões).
- Porém, agora para um novo padrão (que não pertence a base de dados original). É necessário analisá-lo, ou seja, ver em qual dos grupos ele se encaixa melhor.
- Ao realizar este procedimento, dizemos que o modelo está fazendo uma previsão.

3) Construção de Modelos de Predição



- O que é um modelo preditivo?
- Em resumo: é um modelo capaz de tomar uma decisão/avaliação para um novo estímulo (novo padrão), diferente de todos os estímulos (conjunto de padrões usado no treinamento) usados no processo de aprendizagem do modelo.

3) Construção de Modelos de Predição: séries temporais



- **Séries Temporais:**

- É **uma sequencia** de medições extraídas ao longo do tempo.
 - Ex: temperatura média diária, por um ano
- Dados podem ser contínuos ou discretos
 - Contínuos: medições vindas de sistemas analógicos
 - Discretos: medições vindas de sistemas digitais

3) Construção de Modelos de Predição: séries temporais



- Decomposição da Série para gerar um conjunto de dados:
 - A série temporal é usada na montagem de um conjunto de dados.
 - Ex: temperatura média diária → deseja-se prever a temperatura média com base nos valores da temperatura média dos três dias anteriores (janela).

3) Construção de Modelos de Predição: séries temporais



Data	Temperatura (°C)
01/01	27.6
02/01	27.4
03/01	28.1
04/01	28.8
05/01	29.3
06/01	29.0
...	...

Conjunto de dados (entrada, saída)
para uma janela temporal de tamanho três

[27.6; 27.4; 28.1; **28.8**]

[27.4; 28.1; 28.8; **29.3**]

[28.1; 28.8; 29.3; **29.0**]

...

3) Construção de Modelos de Predição: séries temporais



- Três características presentes em Séries Temporais:
 - Tendência
 - Sazonalidade
 - Resíduo

3) Construção de Modelos de Predição: séries temporais



- Três características presentes em Séries Temporais:
 - Tendência
 - Sazonalidade
 - Resíduo

3) Construção de Modelos de Predição: séries temporais



- Tendência: avalia o comportamento de mudança nos dados → crescente ou decrescente
 - Linear → mudança com taxa constante
 - Exponencial → mudança com taxa progressiva (cresce)
 - Amortecida → taxa de mudança decresce

3) Construção de Modelos de Predição: séries temporais



- Sazonalidade: comportamento que se repete, em ciclos.
 - Ex: estações do ano → no verão as temperaturas médias tendem a serem altas, e no inverno tendem a serem baixas.

3) Construção de Modelos de Predição: séries temporais



- Resíduo: comportamento atípicos de imprevisibilidade.
 - Ex: suponha que estamos estudando o preço médio do sorvete. Sabemos que no inverno o preço deve ser menor, mas ocorreu que em determinado ano, houve grande escassez de leite, e o preço do sorvete disparou no inverno.

3) Construção de Modelos de Predição: séries temporais



- Quando estudamos uma série temporal, desejamos analisar alguns objetivos, tais como:
 - **Descrição:** verifica comportamento da série, a procura de tendências, sazonalidades e resíduos.
 - **Explicação:** entender a variação de uma série através de outra série.

3) Construção de Modelos de Predição: séries temporais



- Quando estudamos uma série temporal, desejamos analisar alguns objetivos, tais como:
 - **Predição:** descobrir quais os fatores que possibilitam realizar a previsão do valor futuro da série.
 - **Controle:** em determinados processos, o controle é fundamental para melhoria de qualidade.

3) Construção de Modelos de Predição: análise dos modelos de predição



- Modelos univariados
- Modelos multivariados

3) Construção de Modelos de Predição: análise dos modelos de predição



- Modelos univariados:
 - Série temporal de uma única variável (seus valores passados) é utilizada para prever o seu futuro (o valor da variável).
 - Ex: temos os índices de inflação mensal dos últimos dez anos e queremos prever a inflação do próximo mês. Podemos testar usando vários tamanhos de janelas.

3) Construção de Modelos de Predição: análise dos modelos de predição



- Modelos multivariados:
 - Série temporal de múltiplas variáveis usadas para prever o valor futuro de uma variável.
 - Ex: para prever a inflação do próximo mês, usaremos a série histórica da inflação mensal, do preço mensal do dólar, e do PIB mensal.

3) Construção de Modelos e análise dos modelos de

- Como um modelo deve ser construído:



3) Construção de Modelos de Predição: métodos estatísticos



- Modelos Autorregressivos Integrados de Média Móvel (**ARIMA**, Autoregressive Integrated Moving Average).
- Chamado apenas de Modelos ARIMA
- Baseiam-se nas autocorrelações presentes na série
- Modelos de Suavização Exponencial
 - buscam descrever a tendência e a sazonalidade na série.

OBRIGADO(A)



UNINASSAU.DIGITAL



UNINABUCO.DIGITAL



UNAMA.DIGITAL



UNG.DIGITAL
UNIVERSIDADE GUARULHOS



UNIVERITAS.DIGITAL



UNINORTE.DIGITAL



UNIFACIMED
.DIGITAL



UNIFAEI
CENTRO UNIVERSITÁRIO