Tema 11

Classificação

Não Supervisionada

Professora:
Ariane Machado Lima

SISTEMAS DE INFORMAÇÃO

EACH

Vídeo 1

Introdução

Classificação ou aprendizado não supervisionado

- Dois paradigmas:
  - Aprendizado supervisionado: *aprendizado por exemplos*
  - Aprendizado não supervisionado: *aprendizado por observações*
- Objetivo: classificar elementos onde:
  - Não se sabe a classe dos elementos (não há uma amostra de treinamento)
  - Não se conhece sobre o processo de geração dos padrões (das classes)
  - Às vezes não se sabe nem quantas classes estão envolvidas
- Normalmente a única informação são os vetores de características dos elementos

Agrupamentos (*Clustering*)

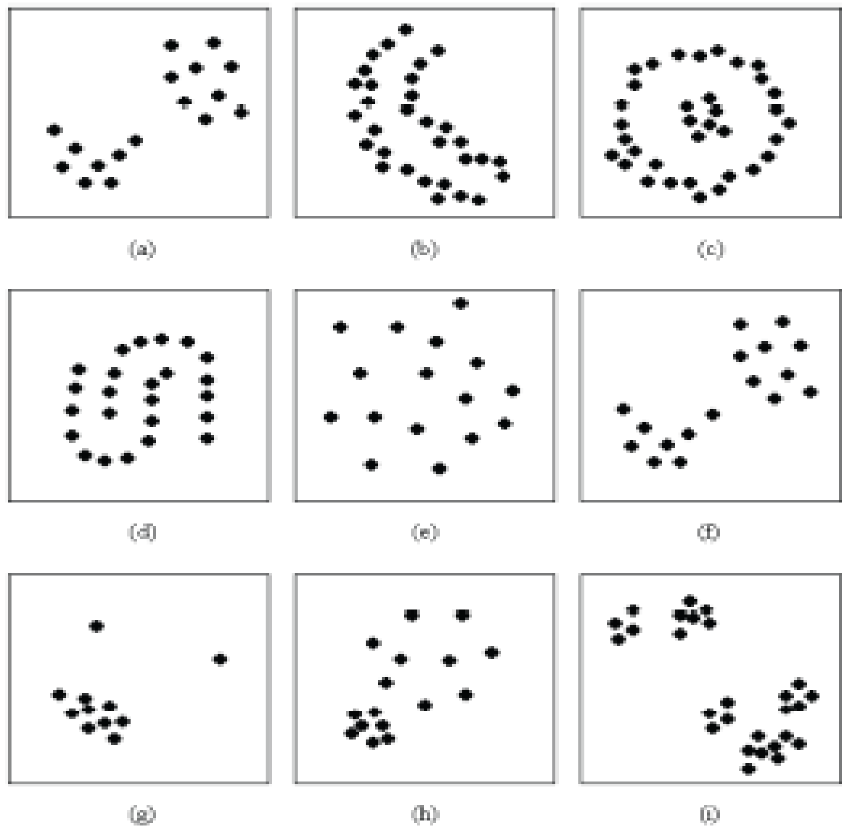
- Realização de partições do espaço com base em um critério

O que é dado?

Os vetores de características dos elementos a serem classificados

Podem apresentar diferentes distribuições no espaço

Exemplos - como você separaria?



[COSTA& CESAR, 2009]

Exemplos

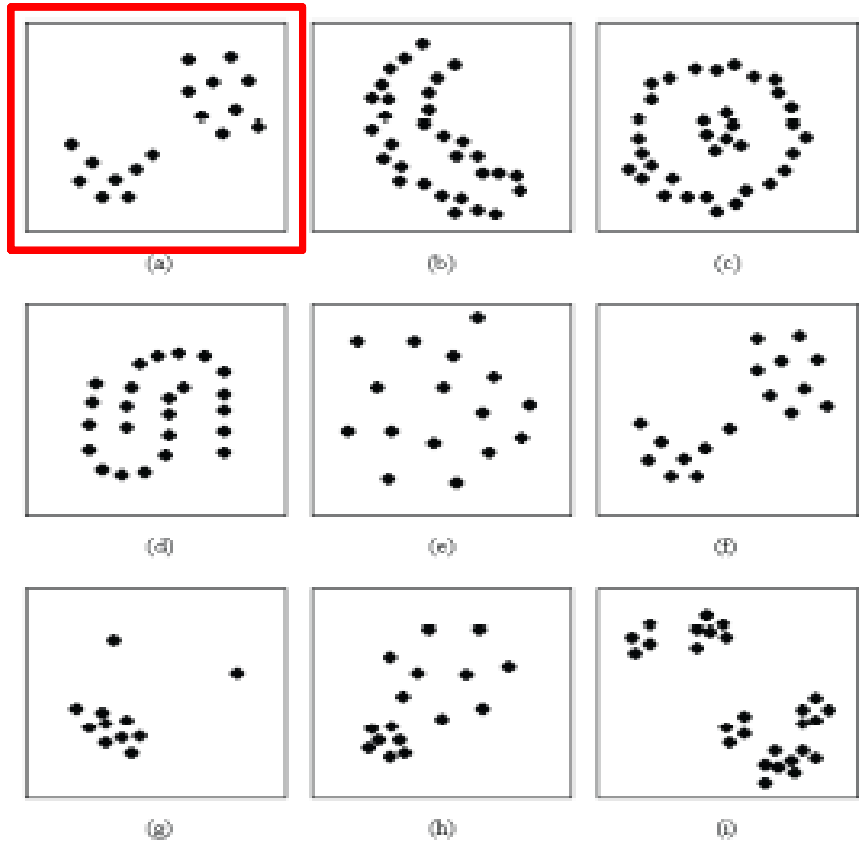
[COSTA& CESAR, 2009]

SISTEMAS DE INFORMAÇÃO

EACH

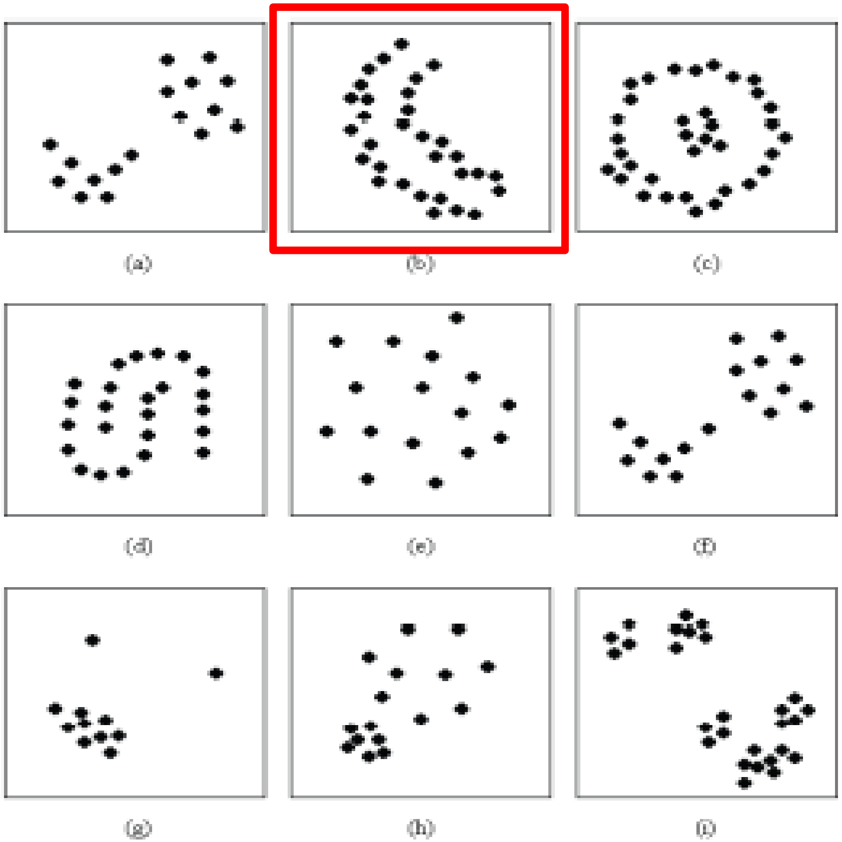
Exemplos

2 grupos
linearmente
separáveis



[COSTA& CESAR, 2009]

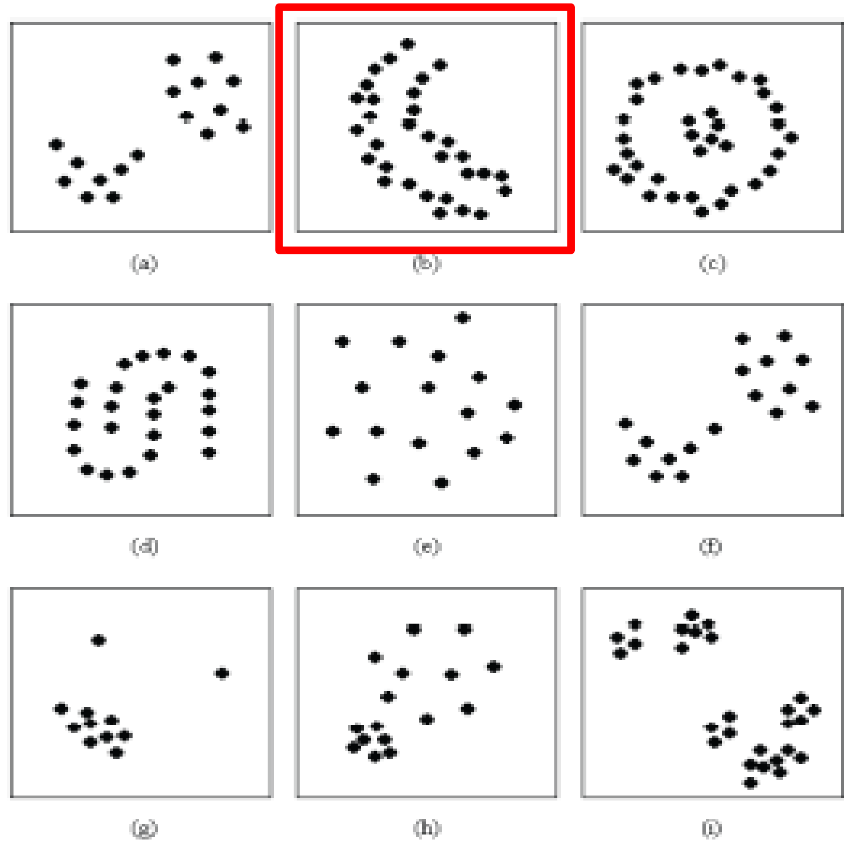
Exemplos



[COSTA& CESAR, 2009]

Exemplos

2 grupos
separação
não linear



[COSTA& CESAR, 2009]

Exemplos

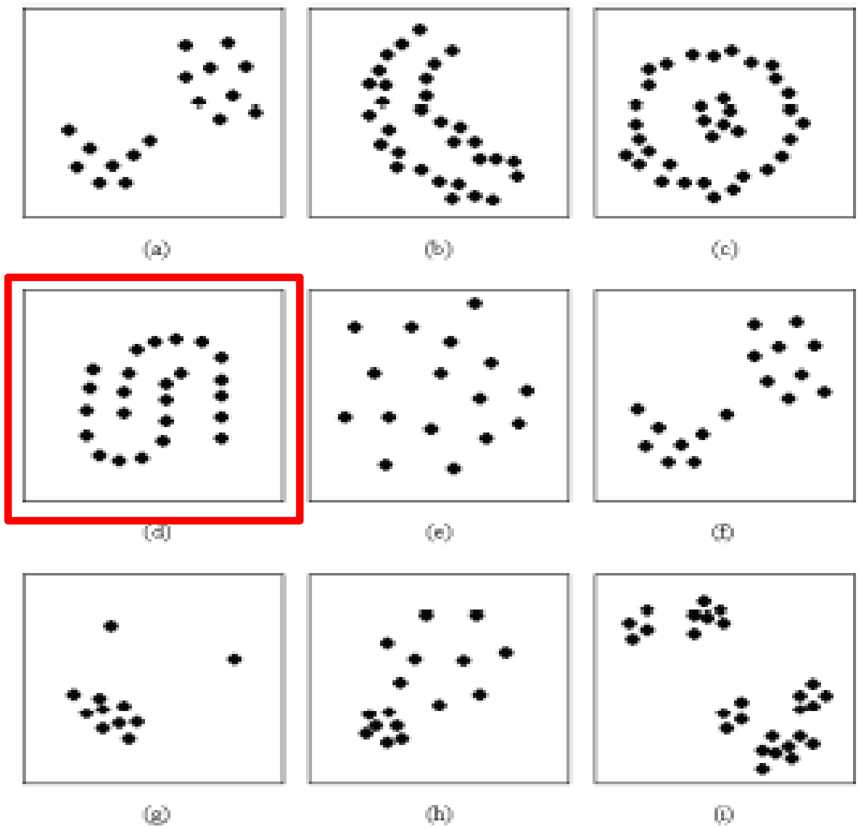
[COSTA& CESAR, 2009]

Exemplos

2 grupos
separação
não linear

[COSTA& CESAR, 2009]

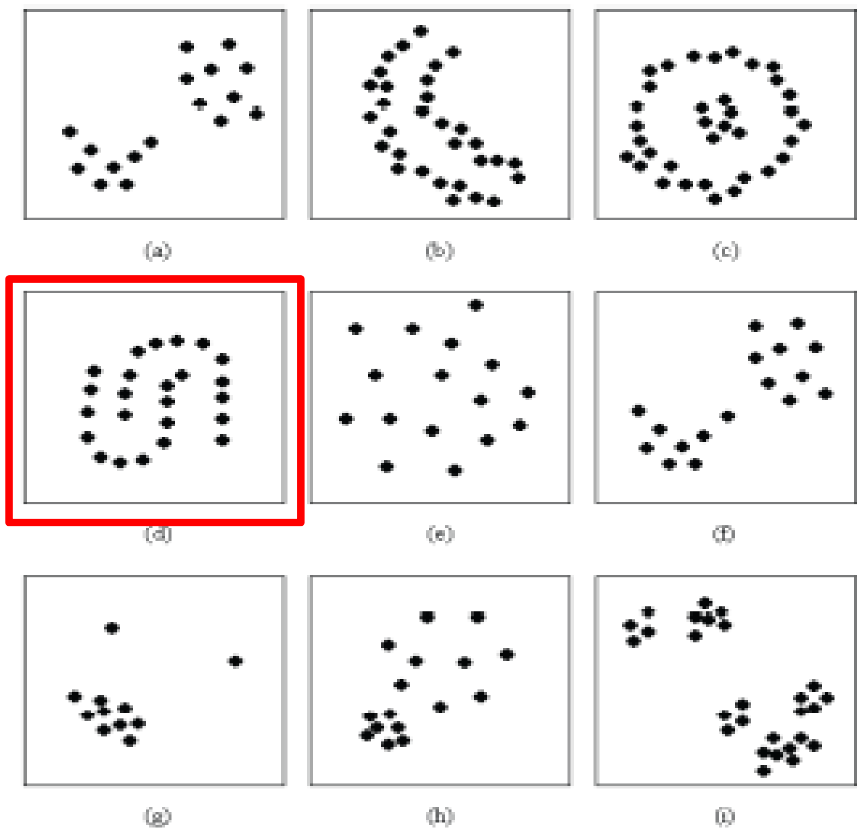
Exemplos



[COSTA& CESAR, 2009]

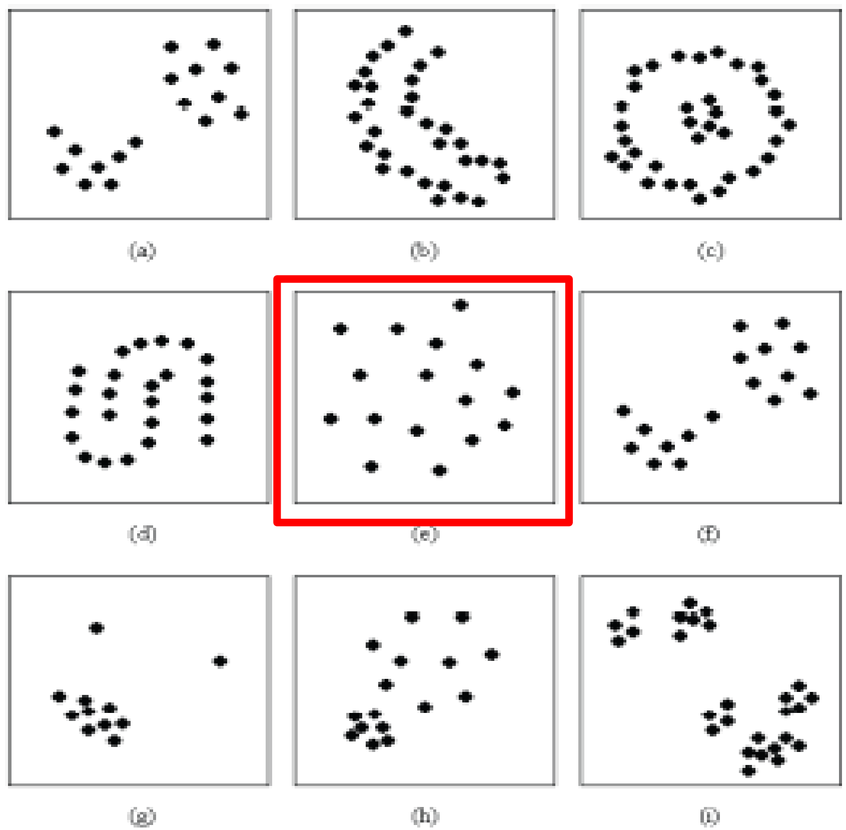
Exemplos

2 grupos
separação
não linear



[COSTA& CESAR, 2009]

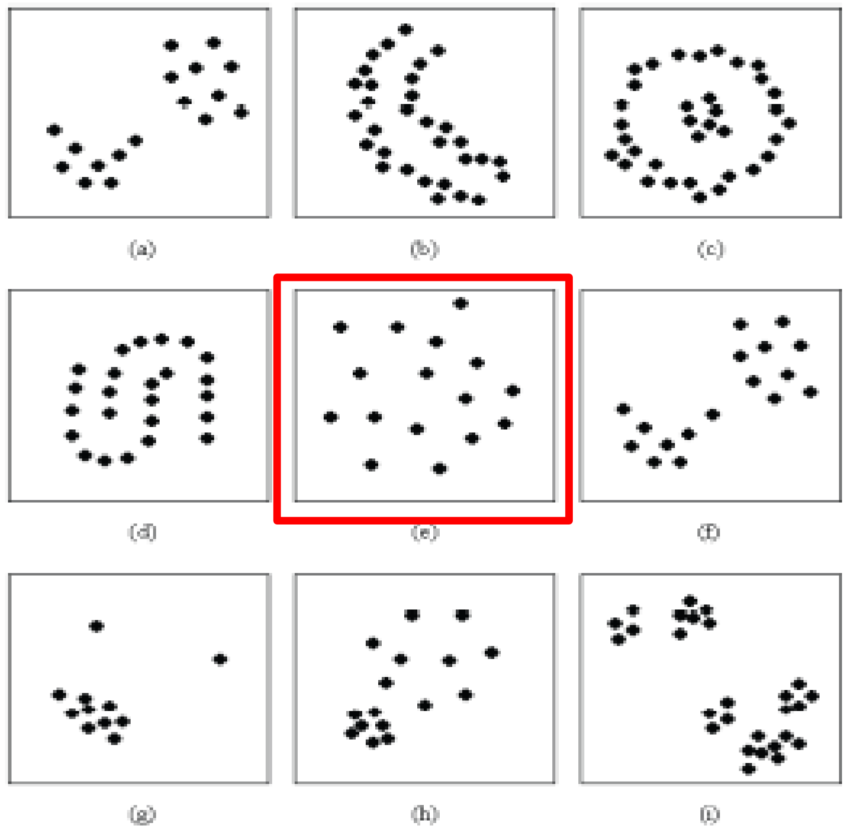
Exemplos



[COSTA& CESAR, 2009]

Exemplos

Quantos grupos?
Distribuição
uniforme



[COSTA& CESAR, 2009]

Exemplos

(variação da figura a)

[COSTA& CESAR, 2009]

Exemplos

(variação da figura a)
Dependendo do
deslocamento de
um ponto, podemos
imaginar 1 ou 2
grupos

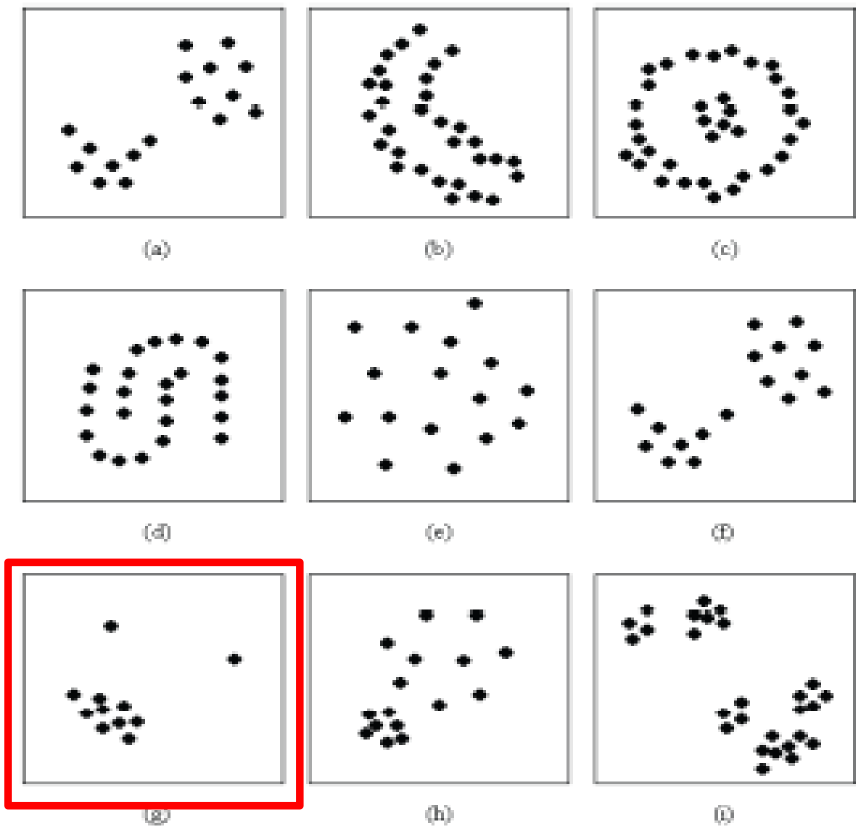
Ponto de ruído ?

[COSTA& CESAR, 2009]

SISTEMAS DE INFORMAÇÃO

EACH

Exemplos



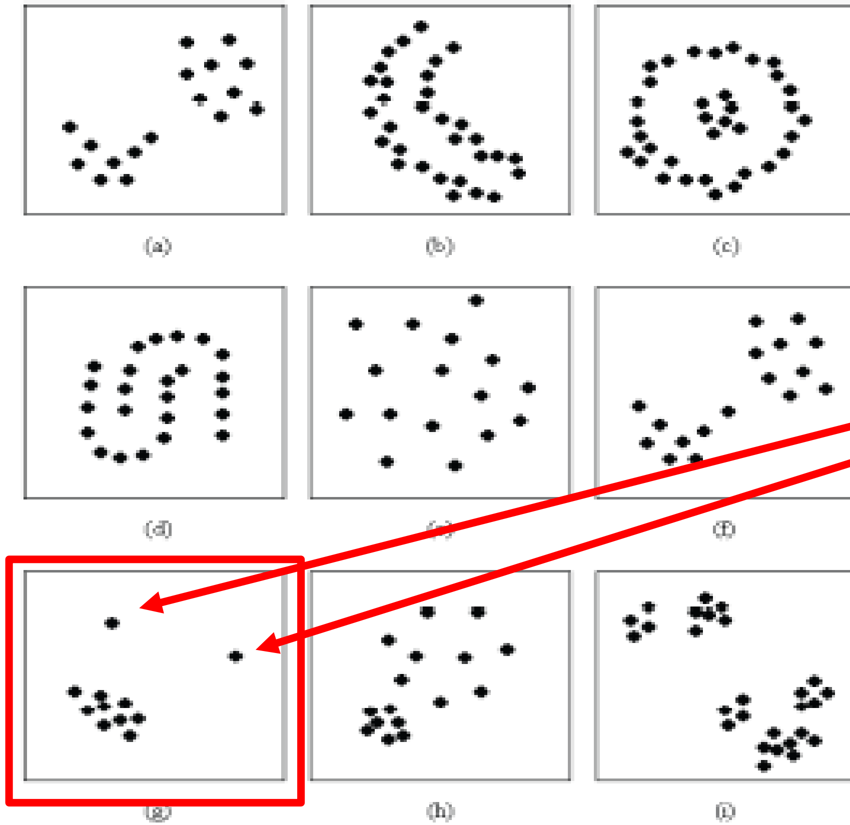
[COSTA& CESAR, 2009]

SISTEMAS DE INFORMAÇÃO

EACH

Exemplos

1, 2 ou 3 grupos?



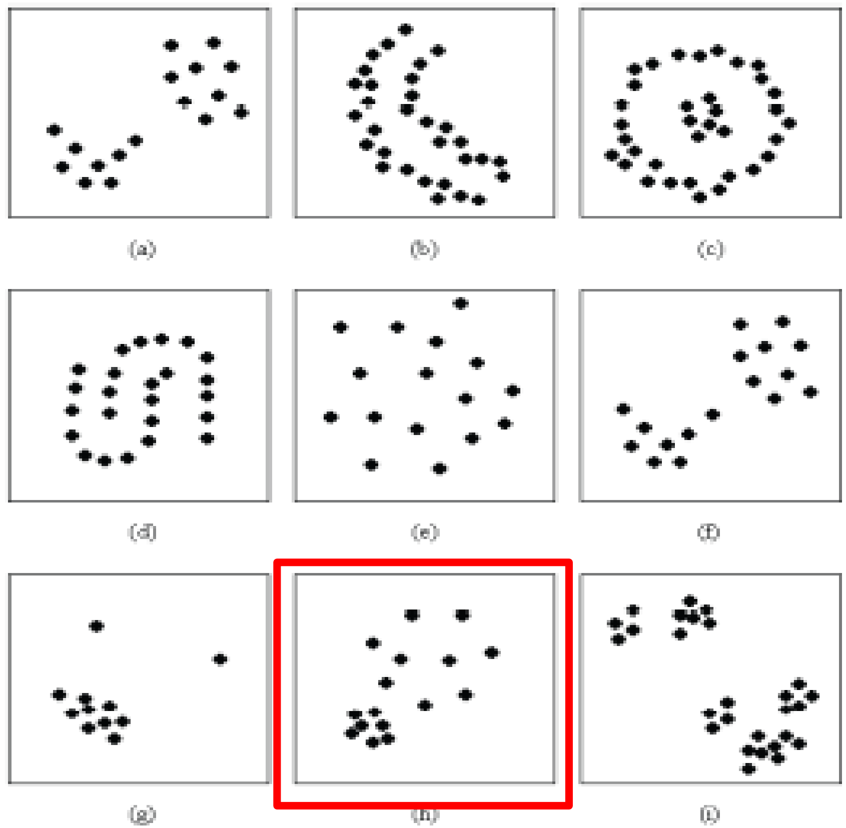
Outliers?

[COSTA& CESAR, 2009]

Exemplos

[COSTA& CESAR, 2009]

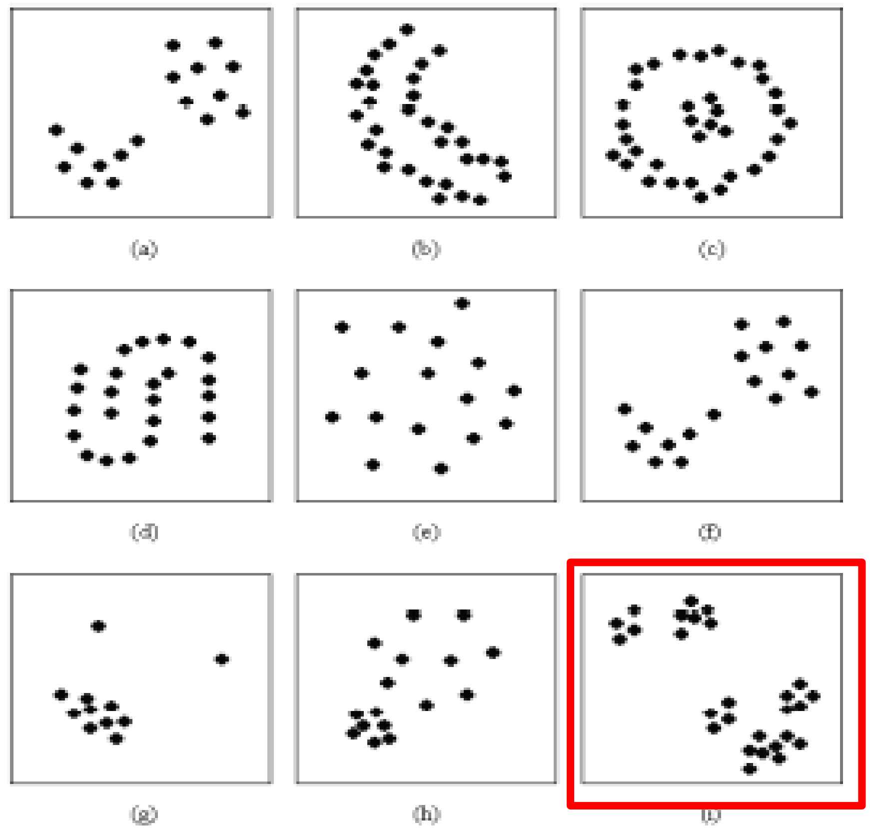
Exemplos



2 grupos...
… ou 1, se usar uma escala diferente (logaritmica)

[COSTA& CESAR, 2009]

Exemplos

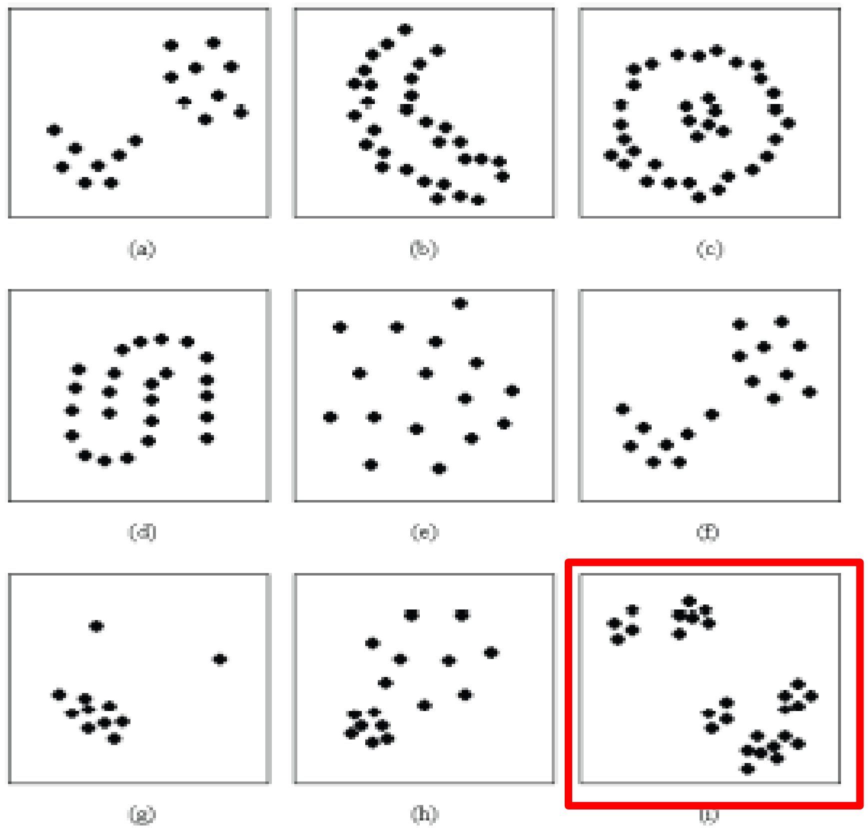


[COSTA& CESAR, 2009]

SISTEMAS DE INFORMAÇÃO

EACH

Exemplos



2 ou 5 grupos?
clusters de
clusters

[COSTA& CESAR, 2009]

Esteja consciente de que...

- O agrupamento é feito usando características e critérios que podem não ser adequados para a classificação verdadeira

- Qualquer critério vai impor uma estrutura sobre os dados, que pode não ser a real

- POR ISSO: qualquer informação extra que você tenha é VALIOSA (por ex: nr de classes)

Qual critério utilizar?

Considerar vários critérios, e analisar os resultados.

Em geral, os critérios seguem a ideia:

Critério de Similaridade: agrupar elementos de tal forma que

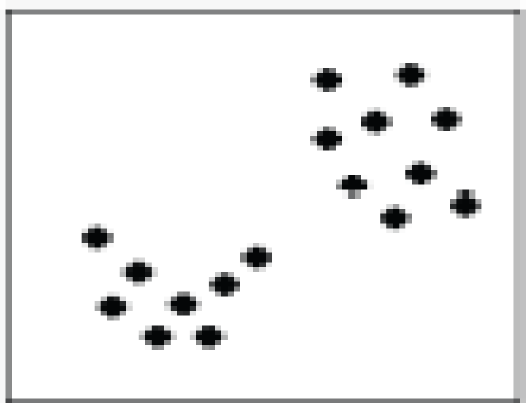
- elementos da mesma classe sejam o mais similares possível entre si
- elementos de classes distintas sejam o mais diferentes possível entre si

Cabe então definir o que é a similaridade utilizada

Critério de similaridade

Uma das formas de formas de formalizar o critério de similaridade é através da dispersão dos dados

- Cada grupo deve conter elementos com uma dispersão mais baixa (baixa dispersão intraclasse)
- Elementos de grupos distintos devem apresentar maior dispersão (alta dispersão interclasse)

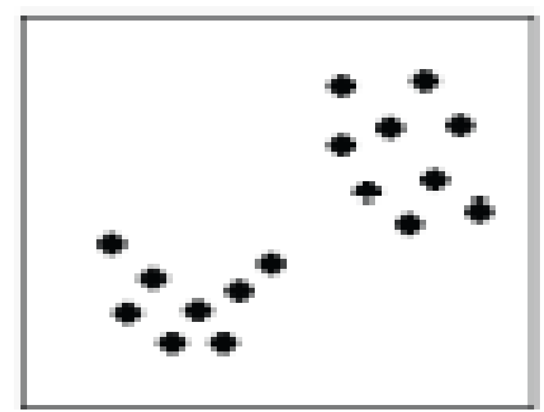


[COSTA& CESAR, 2009]

Critério de similaridade

Uma das formas de formas de formalizar o critério de similaridade é através da dispersão dos dados

- Cada grupo deve conter elementos com uma dispersão mais baixa (baixa dispersão intraclasse)
- Elementos de grupos distintos devem apresentar maior dispersão (alta dispersão interclasse)
- → Temos que nos preocupar em otimizar essas duas medidas?



[COSTA& CESAR, 2009]

Matrizes de Dispersão (*scatter matrices*)

- Suponha que conheçamos as classes...
- Dada uma matriz dataset NxM (elementos nas linhas, características nas colunas)
- K classes
- Cada classe C_i com N_i elementos
- Cada elemento f_j é um vetor Mx1
- $\vec{M}$ é o vetor médio Mx1 global (valor médio de cada caraterística)
- $\vec{\mu}_i$ é o vetor médio Mx1 da classe C_i

Exemplo

K = 3 classes:

Object #	Class	Feature 1	Feature 2
1	C_3	9.2	33.2
2	C_2	5.3	21.4
3	C_3	8.8	31.9
4	C_1	2.9	12.7
5	C_3	9.0	32.4
6	C_1	1.5	12.0
7	C_1	1.2	11.5

Matriz de dados:

$$F = \begin{bmatrix} 9.2 & 33.2 \\ 5.3 & 21.4 \\ 8.8 & 31.9 \\ 2.9 & 12.7 \\ 9.0 & 32.4 \\ 1.5 & 12.0 \\ 1.2 & 11.5 \end{bmatrix}$$

Exemplo

$$F = \begin{bmatrix} 9.2 & 33.2 \\ 5.3 & 21.4 \\ 8.8 & 31.9 \\ 2.9 & 12.7 \\ 9.0 & 32.4 \\ 1.5 & 12.0 \\ 1.2 & 11.5 \end{bmatrix}$$

$$\vec{M} = \begin{bmatrix} 5.4143 \\ 22.1571 \end{bmatrix}$$

$$\vec{f_1} = \begin{bmatrix} 9.2 \\ 33.2 \end{bmatrix}; \quad \vec{f_2} = \begin{bmatrix} 5.3 \\ 21.4 \end{bmatrix}; \quad \vec{f_3} = \begin{bmatrix} 8.8 \\ 31.9 \end{bmatrix}; \quad \vec{f_4} = \begin{bmatrix} 2.9 \\ 12.7 \end{bmatrix};$$

$$\vec{f_5} = \begin{bmatrix} 9.0 \\ 32.4 \end{bmatrix}; \quad \vec{f_6} = \begin{bmatrix} 1.5 \\ 12.0 \end{bmatrix}; \quad \vec{f_7} = \begin{bmatrix} 1.2 \\ 11.5 \end{bmatrix}$$

Exemplo

Matrizes de dados para as classes 1, 2 e 3:

$$F_1 = \begin{bmatrix} 2.9 & 12.7 \\ 1.5 & 12.0 \\ 1.2 & 11.5 \end{bmatrix}; \qquad F_2 = \begin{bmatrix} 5.3 & 21.4 \end{bmatrix}; \qquad F_3 = \begin{bmatrix} 9.2 & 33.2 \\ 8.8 & 31.9 \\ 9.0 & 32.4 \end{bmatrix}$$

$$\vec{\mu}_1 = \begin{bmatrix} 1.8667 \\ 12.0667 \end{bmatrix}; \qquad \vec{\mu}_2 = \begin{bmatrix} 5.3 \\ 21.4 \end{bmatrix}; \qquad \vec{\mu}_3 = \begin{bmatrix} 9.0 \\ 32.5 \end{bmatrix}$$

Matrizes de dispersão

- Matriz de dispersão total:

$$S = \sum_{i=1}^{N} \left(\vec{f_i} - \vec{\mathbf{M}}\right)\left(\vec{f_i} - \vec{\mathbf{M}}\right)^T$$

Matrizes de dispersão

- Matriz de dispersão total:

$$S = \sum_{i=1}^{N} \left(\vec{f_i} - \vec{\mathbf{M}}\right)\left(\vec{f_i} - \vec{\mathbf{M}}\right)^T$$

No nosso exemplo:

$$= \begin{bmatrix} 9.2 - 5.4143 \\ 33.20 - 22.1571 \end{bmatrix} \begin{bmatrix} 9.2 - 5.4143 & 33.20 - 22.1571 \end{bmatrix} + \cdots +$$

$$+ \begin{bmatrix} 1.2 - 5.4143 \\ 11.5 - 22.1571 \end{bmatrix} \begin{bmatrix} 1.2 - 5.4143 & 11.5 - 22.1571 \end{bmatrix}$$

$$= \begin{bmatrix} 78.0686 & 220.0543 \\ 220.0543 & 628.5371 \end{bmatrix}.$$

Matrizes de dispersão

- Matriz de dispersão total:

$$S = \sum_{i=1}^{N} \left(\vec{f_i} - \vec{\mathbf{M}}\right)\left(\vec{f_i} - \vec{\mathbf{M}}\right)^T$$

- Matriz de dispersão da classe C_i :

$$S_i = \sum_{i \in C_i} \left(\vec{f_i} - \vec{\mu}_i\right)\left(\vec{f_i} - \mu_i\right)^T$$

Matrizes de dispersão

- Matriz de dispersão total:

$$S = \sum_{i=1}^{N} \left(\vec{f_i} - \vec{\mathrm{M}}\right)\left(\vec{f_i} - \vec{\mathrm{M}}\right)^T$$

- Matriz de dispersão da classe C_i :

$$S_i = \sum_{i \in C_i} \left(\vec{f_i} - \vec{\mu}_i\right)\left(\vec{f_i} - \mu_i\right)^T$$

- Matriz de dispersão intraclasse:

$$S_{\text{intra}} = \sum_{i=1}^{K} S_i$$

Matrizes de dispersão

- Matriz de dispersão total:

$$S = \sum_{i=1}^{N} \left(\vec{f_i} - \vec{\mathbf{M}}\right)\left(\vec{f_i} - \vec{\mathbf{M}}\right)^T$$

- Matriz de dispersão da classe C_i :

$$S_i = \sum_{i \in C_i} \left(\vec{f_i} - \vec{\mu}_i\right)\left(\vec{f_i} - \mu_i\right)^T$$

- Matriz de dispersão intraclasse:

$$S_{\text{intra}} = \sum_{i=1}^{K} S_i$$

- Matriz de dispersão interclasse:

$$S_{\text{inter}} = \sum_{i=1}^{K} N_i \left(\vec{\mu}_i - \vec{\mathbf{M}}\right)\left(\vec{\mu}_i - \vec{\mathbf{M}}\right)^T$$

Matrizes de dispersão

- $S = S_{intra} + S_{inter}$

As matrizes são simétricas e a posição (i,i) mostra o quadrado da dispersão da característica i

- Trace(S): soma da diagonal principal de S

trace (S) = trace(S_{intra}) + trace(S_{inter})

Consequência: pode-se atentar a apenas diminuir S_{intra} ou aumentar S_{inter}

SISTEMAS DE INFORMAÇÃO

EACH

Matrizes de dispersão

- $S = S_{intra} + S_{inter}$

As matrizes são simétricas e a posição (i,i) mostra o quadrado da dispersão da característica i

- Trace(S): soma da diagonal principal de S

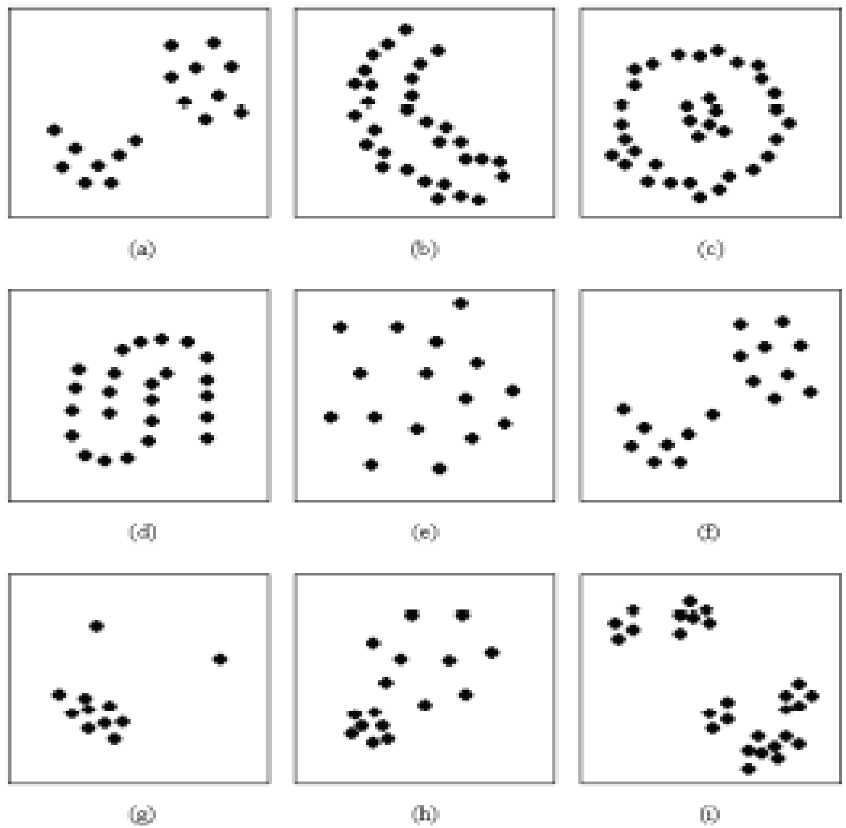
trace (S) = trace(S_{intra}) + trace(S_{inter})

Consequência: pode-se atentar a apenas diminuir S_{intra} ou aumentar S_{inter}

SISTEMAS DE INFORMAÇÃO

EACH

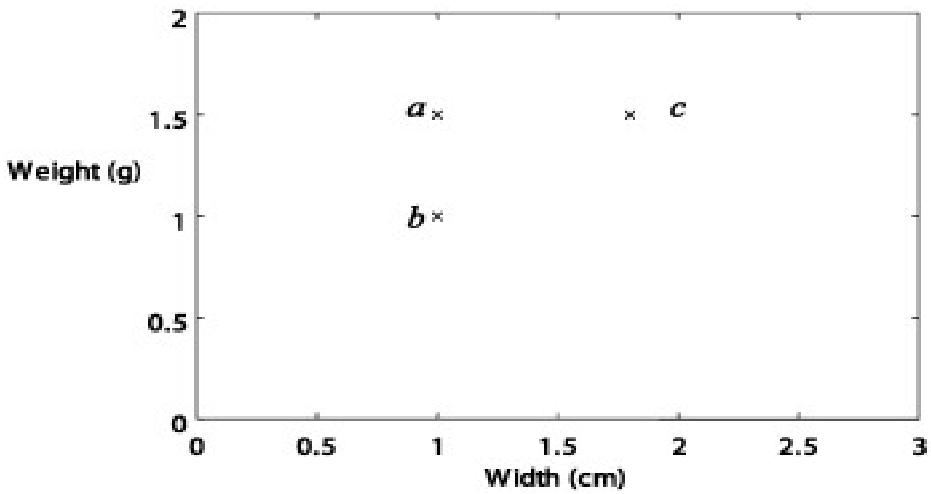
Voltando à nossa intuição...
Dispersão? Distância?

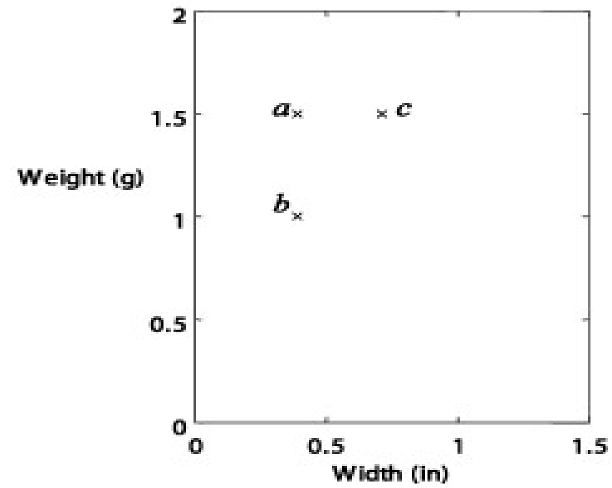


[COSTA& CESAR, 2009]

Necessidade de normalização

- A maioria das características são dimensionais (possuem unidade métrica)
- 3 objetos (por exemplo, folhas)
- Unidades de medidas afetam as distâncias entre os objetos no espaço de características





Pré-processamento

- Ainda há o problema da diferença entre as escalas de valores das características
- Para diminuir esse problema, é comum a normalização por z-score:

$$z^i_f = \frac{x^i_f - \bar{x}_f}{s_f}$$

Média

Desvio padrão

Pré-processamento

- Ainda há o problema da diferença entre as escalas de valores das características
- Para diminuir esse problema, é comum a normalização por z-score
- Há indicações do uso de um z-score baseado no desvio absoluto médio d_f, a fim de diminuir o efeito de outliers:

$$d_f = 1/n(|x^1_f - \overline{x_f}| + \dots + |x^n_f - \overline{x_f}|)$$

$$z^i_f = \frac{x^i_f - \overline{x}_f}{d_f}$$

Pré-processamento

- Ainda há o problema da diferença entre as escalas de valores das características
- Para diminuir esse problema, é comum a normalização por z-score
- Há indicações do uso de um z-score baseado no desvio absoluto médio d_f, a fim de diminuir o efeito de outliers
- Ainda há o problema da dimensionalidade (distâncias crescem nas dimensões, tempo de processamento, ...)
  - PCA (redução de dimensionalidade)

Fim do vídeo 1

Introdução

Vídeo 2

Agrupamento Hierárquico

Técnicas de agrupamentos

- Hierárquico: agrupamento progressivo de elementos (formação de subclasses)

- Particional: grupos obtidos a partir de uma partição do espaço de características com respeito a um número fixo de grupos

- Outros

Técnicas de agrupamentos

- Hierárquico: agrupamento progressivo de elementos (formação de subclasses)

- Particional: grupos obtidos a partir de uma partição do espaço de características com respeito a um número fixo de grupos

- Outros

Agrupamento Hierárquico

- Agrupamentos progressivos de N objetos em classes, de acordo com alguma medida de dissimilaridade (ex: distância)
- Objetos mais próximos são agrupados em subgrupos antes de objetos mais distantes
- No final
  - todos os objetos pertencem a um único e grande grupo
  - Você define no final quantos grupos vai considerar (número variável de subgrupos - classes)

Agrupamento Hierárquico

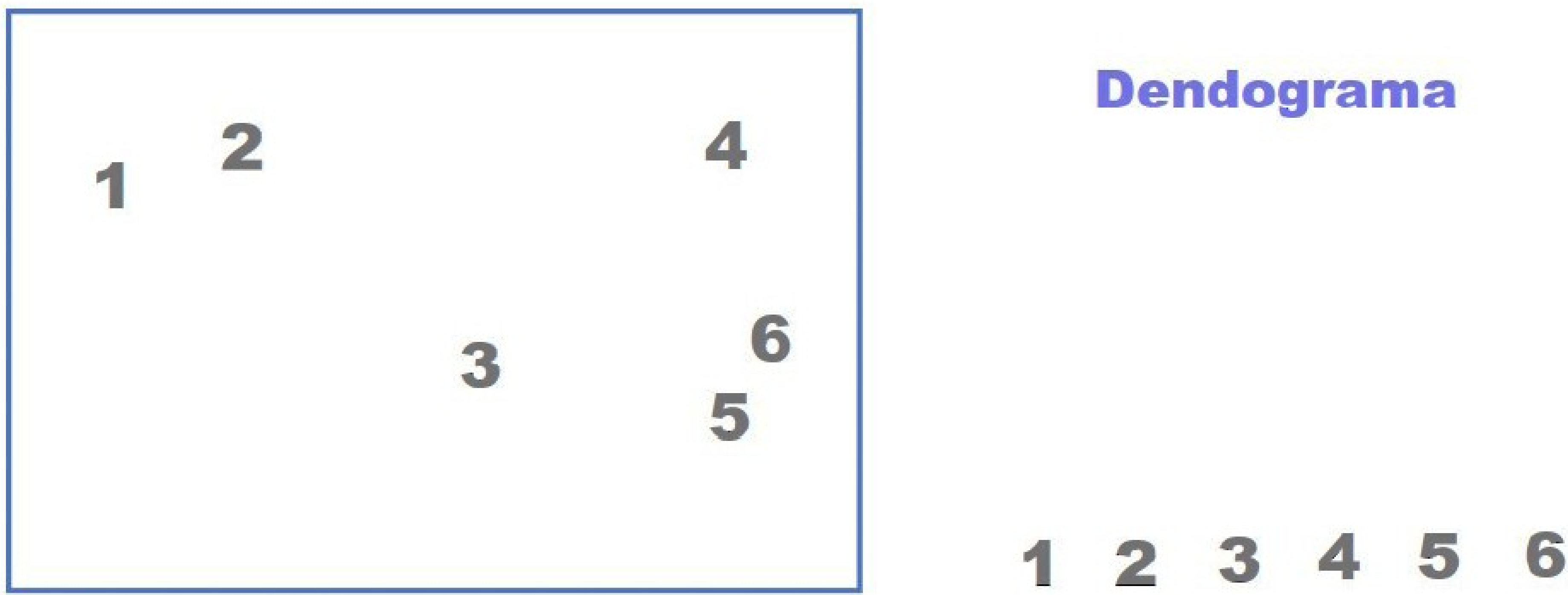
- Abordagem aglomerativa: parte de elementos individuais e vai agrupando subgrupos
  - Vamos ver aqui

- Abordagem divisiva: parte do grande grupo (contendo todos os elementos) e vai dividindo em subgrupos

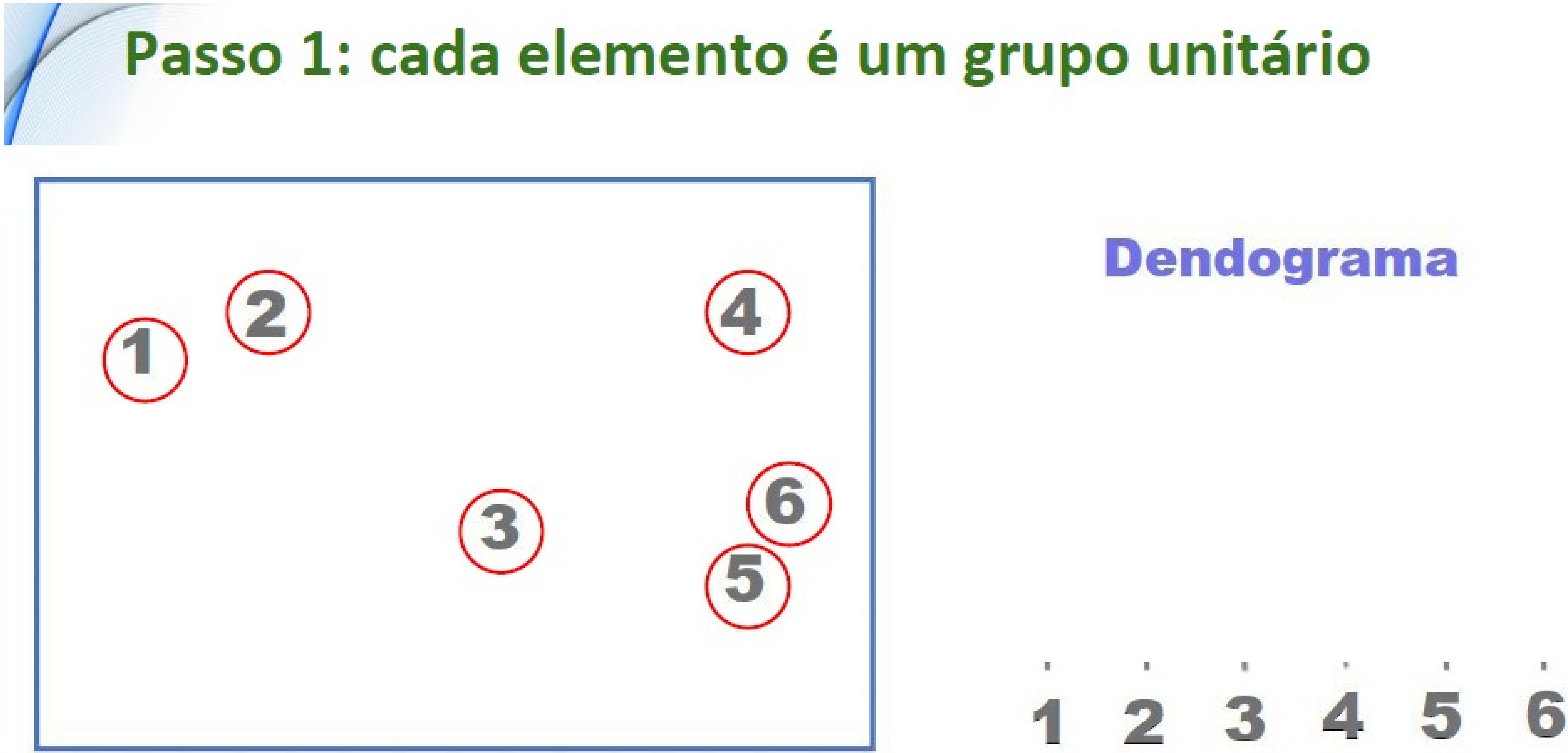
Agrupamento Hierárquico Aglomerativo

- Inicialmente cada objeto é um grupo
- Em cada iteração:
  - Matriz de distâncias entre os grupos correntes
  - Une os dois grupos mais próximos

Agrupamento Hierárquico Aglomerativo

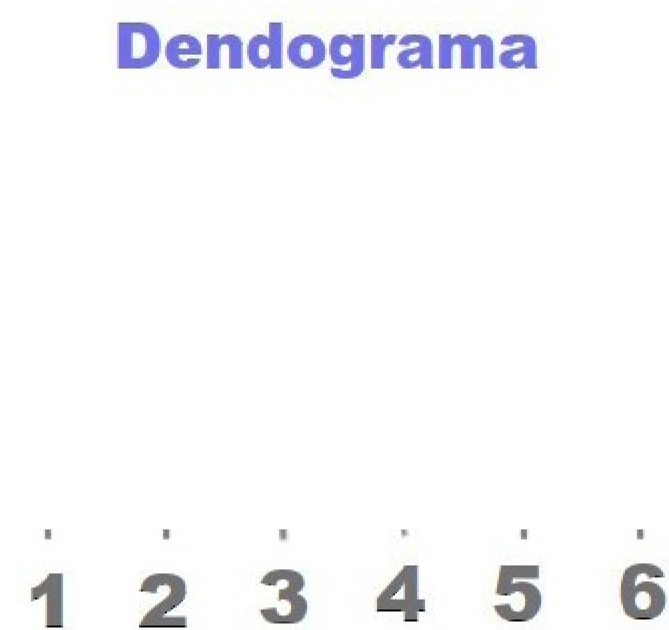


Agrupamento Hierárquico Aglomerativo



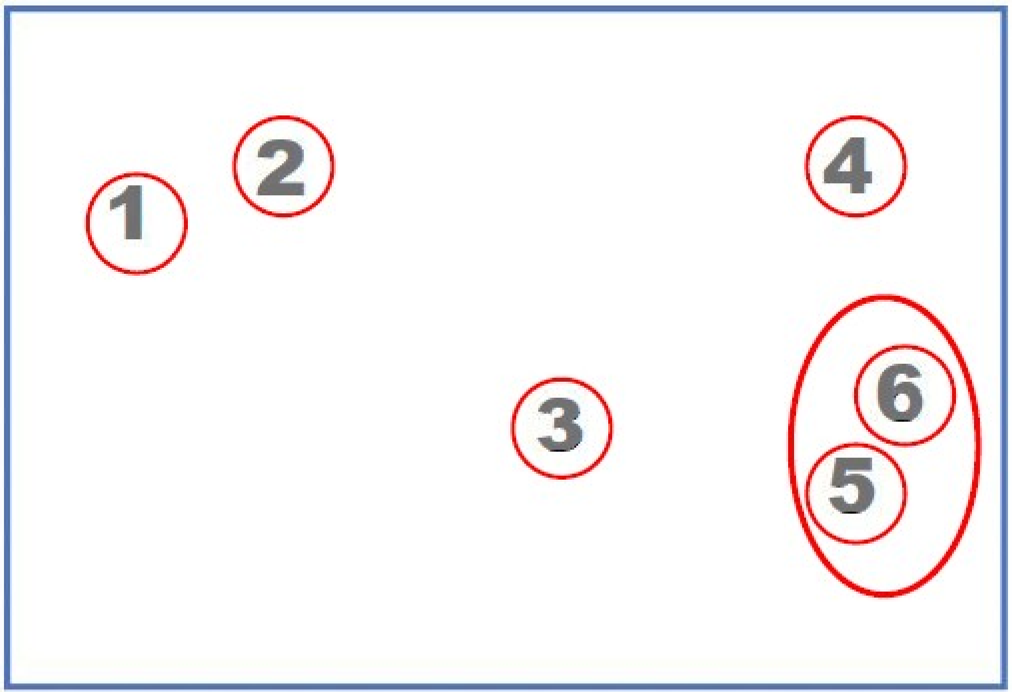
Agrupamento Hierárquico Aglomerativo

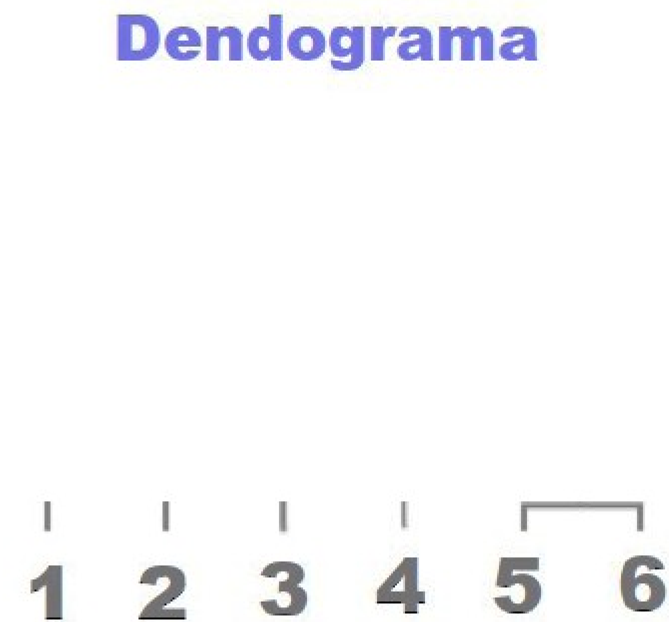
Passos seguintes: a cada passo juntar os dois grupos mais próximos, até que sobre apenas um



Agrupamento Hierárquico Aglomerativo

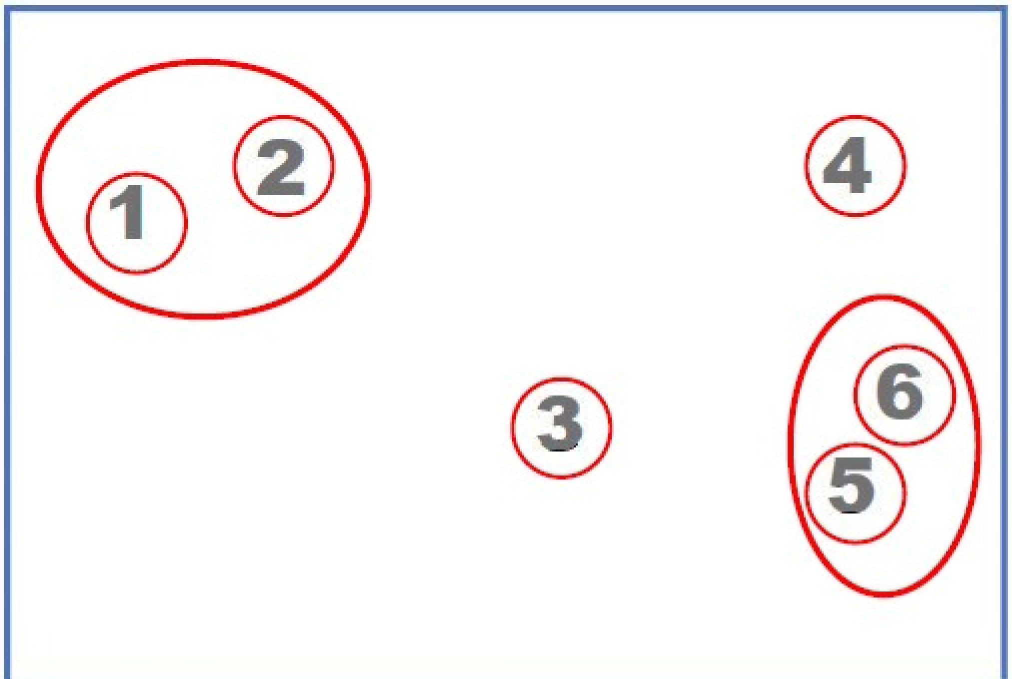
Passos seguintes: a cada passo juntar os dois grupos mais próximos, até que sobre apenas um





Agrupamento Hierárquico Aglomerativo

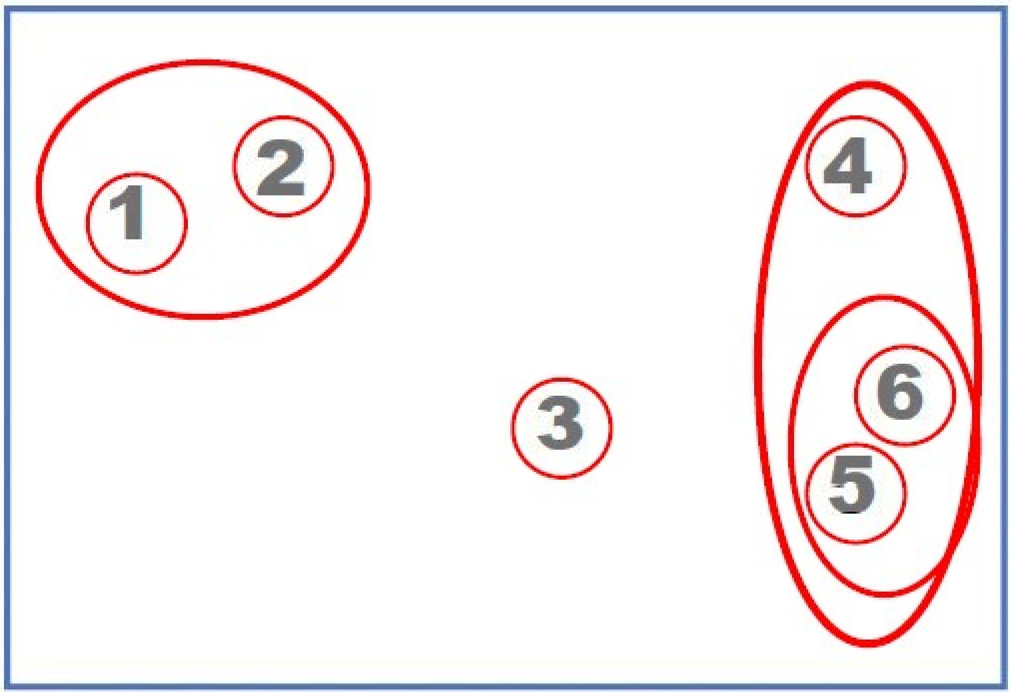
Passos seguintes: a cada passo juntar os dois grupos mais próximos, até que sobre apenas um



Dendograma

Agrupamento Hierárquico Aglomerativo

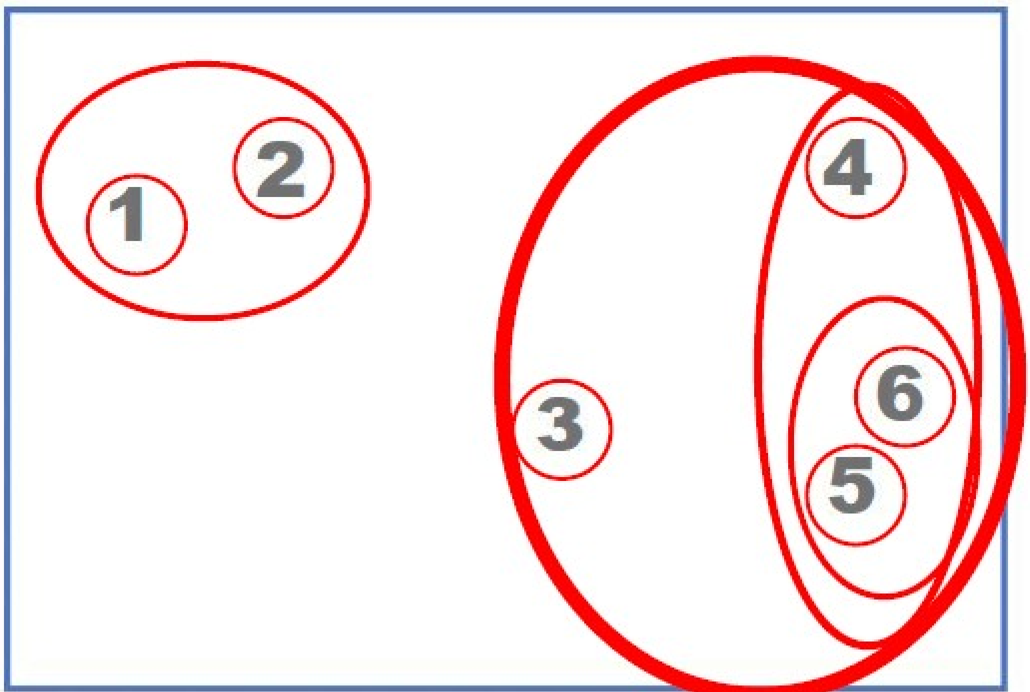
Passos seguintes: a cada passo juntar os dois grupos mais próximos, até que sobre apenas um



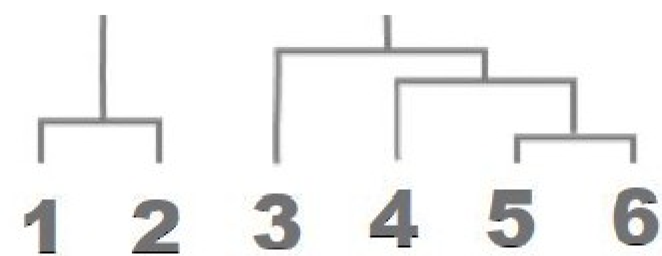
Dendograma

Agrupamento Hierárquico Aglomerativo

Passos seguintes: a cada passo juntar os dois grupos mais próximos, até que sobre apenas um



Dendograma

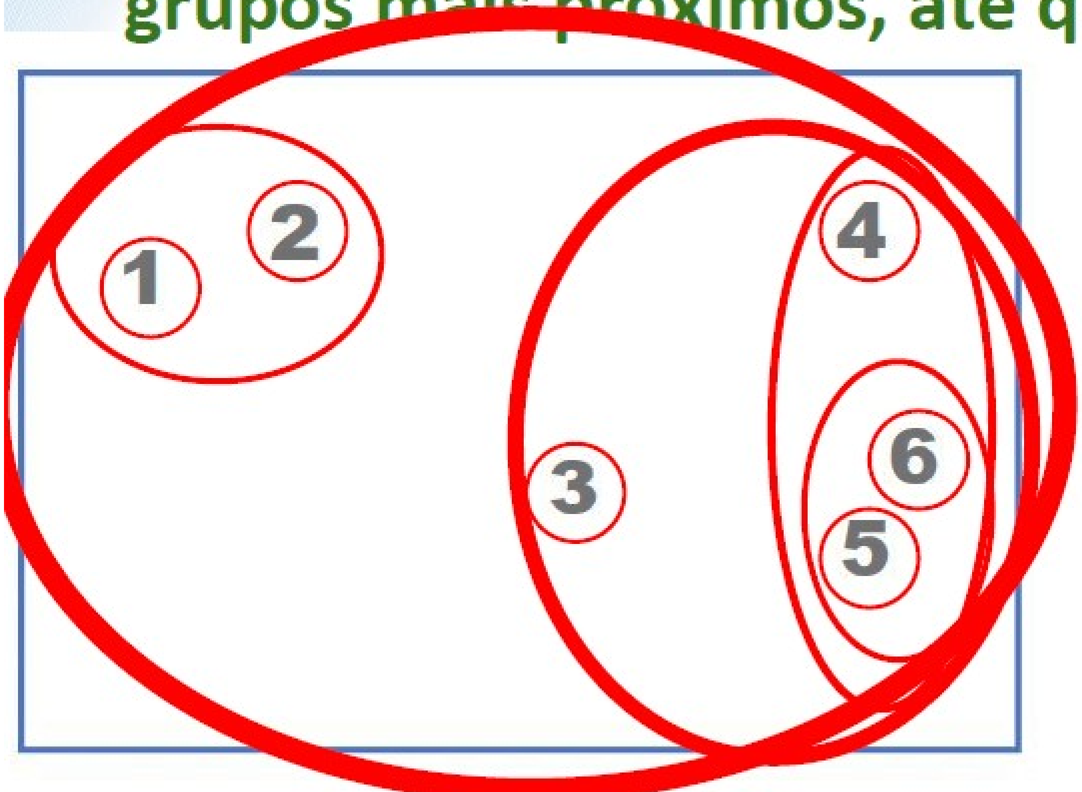


SISTEMAS DE INFORMAÇÃO

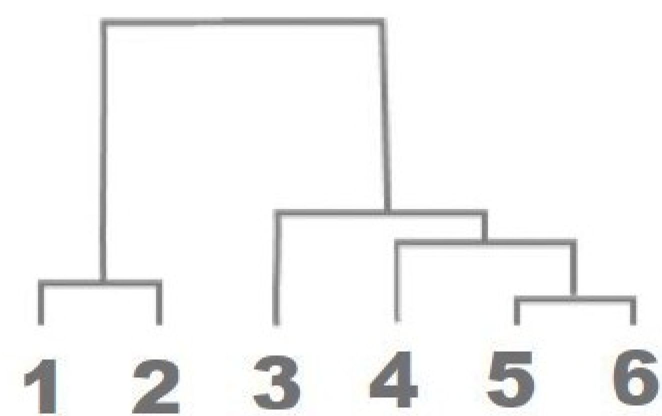
EACH

Agrupamento Hierárquico Aglomerativo

Passos seguintes: a cada passo juntar os dois grupos mais próximos, até que sobre apenas um

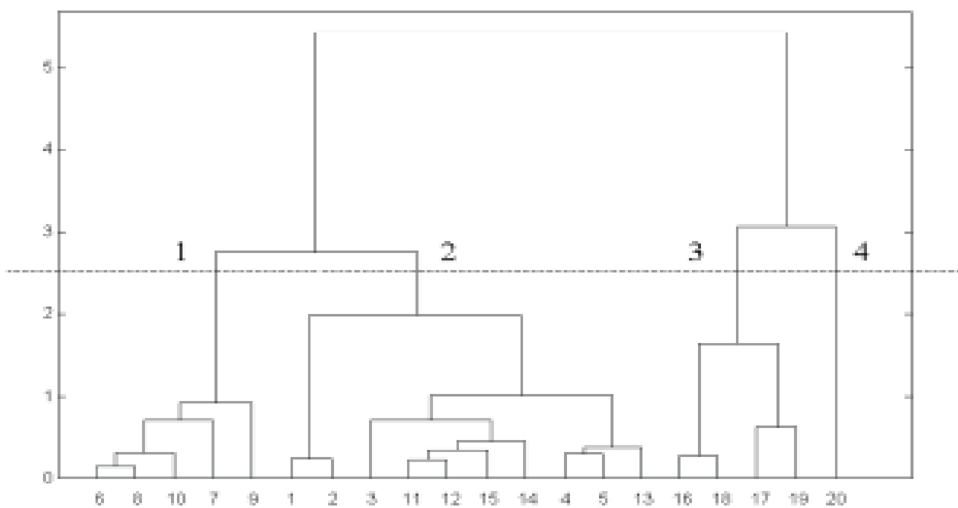


Dendograma



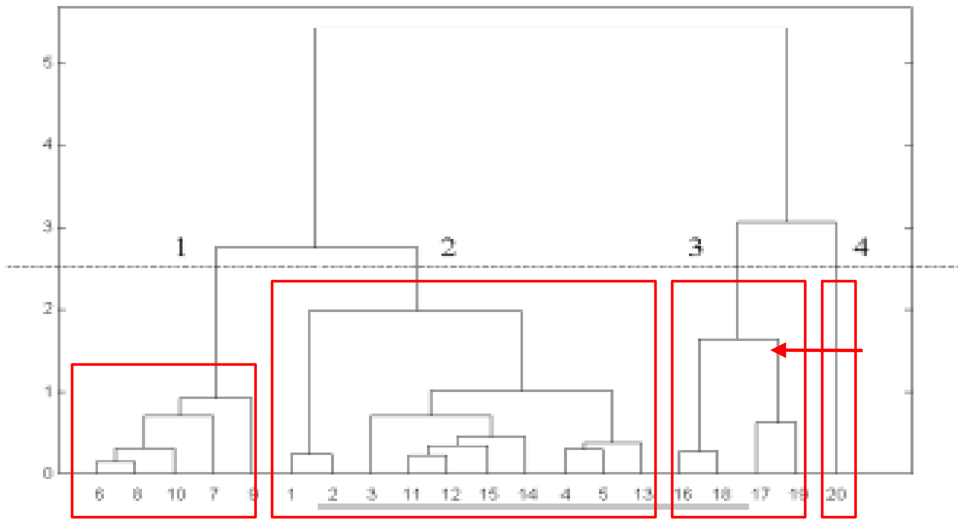
Agrupamento Hierárquico Aglomerativo

- Inicialmente cada objeto é um grupo
- Em cada iteração:
  - Matriz de distâncias entre os grupos correntes
  - Une os dois grupos mais próximos
- Dendograma: vários possíveis números de classes



Agrupamento Hierárquico Aglomerativo

- Inicialmente cada objeto é um grupo
- Em cada iteração:
  - Matriz de distâncias entre os grupos correntes
  - Une os dois grupos mais próximos
- Dendograma: vários possíveis números de classes



4 classes?

Agrupamento Hierárquico

- Como definir distâncias entre grupos?
- Diferentes formas de se medir essa distância implicam em diferentes tipos de agrupamentos hierárquicos
  - Single linkage
  - Complete linkage
  - Group average
  - Centroid
  - Ward’s linkage (NÃO usa distância “geométrica”)

Ward's linkage

- Ao invés de unir os dois grupos mais próximos une os dois grupos que menos aumentam a dispersão intraclasse

$x_{l,j,k}$: valor para a variável p na observação j pertencente ao cluster l

SS_l: soma dos erros quadrados dentro do cluster l

$$SS_l = \sum_{k=1}^{n_l}\sum_{j=1}^{p}\left(x_{l,k,j} - \bar{x}_{l,\bullet,j}\right)^2, \quad \bar{x}_{l,\bullet,j} = \frac{1}{n_l}\sum_{k=1}^{n_l} x_{l,k,j}$$

$SST_{l,i}$: soma total dos erros quadrados (agrupando os clusters l e i

$$SS_{l,i} = \sum_{k=1}^{n_l}\sum_{j=1}^{p}\left(x_{l,k,j} - \bar{x}_j\right)^2 + \sum_{k=1}^{n_i}\sum_{j=1}^{p}\left(x_{i,k,j} - \bar{x}_j\right)^2,$$

$$\bar{x}_j = \frac{1}{n_l + n_i}\left(\sum_{k=1}^{n_l} x_{l,k,j} + \sum_{k=1}^{n_i} x_{i,k,j}\right)$$

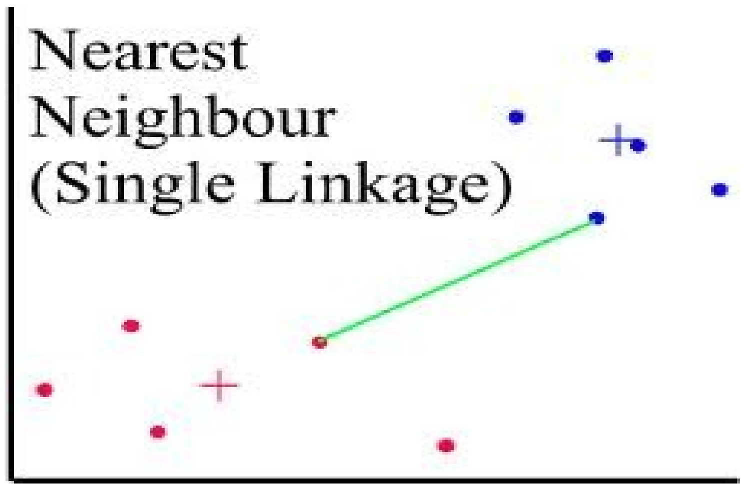
$$d(C_l, C_i) = SS_{l,i} - (SS_l + SS_i)$$

$$= \frac{n_l n_i}{n_l + n_i}\sum_{j=1}^{p}\left(\bar{x}_{l,\bullet,j} - \bar{x}_{i,\bullet,j}\right)^2$$

http://www.each.usp.br/lauretto/cursoR2017/04-AnaliseCluster.pdf

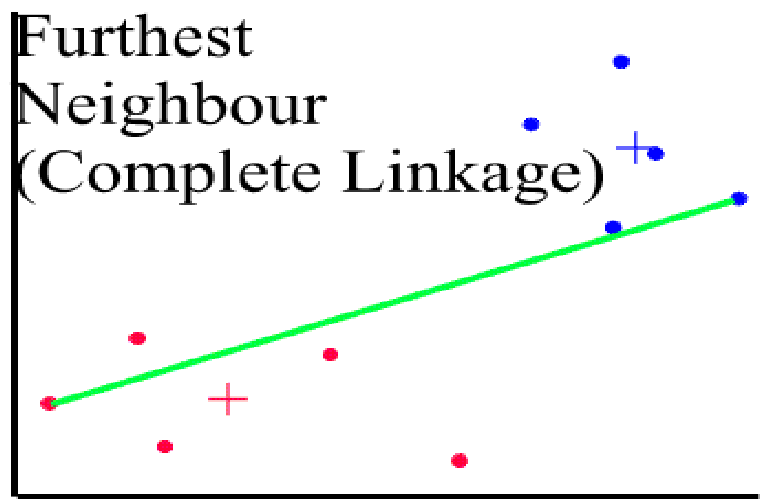
Single Linkage

- $dist(A,B) = \min dist(a,b),\ a \in A,\ b \in B$
- Distância mínima entre qualquer ponto da classe A e qualquer ponto da classe B



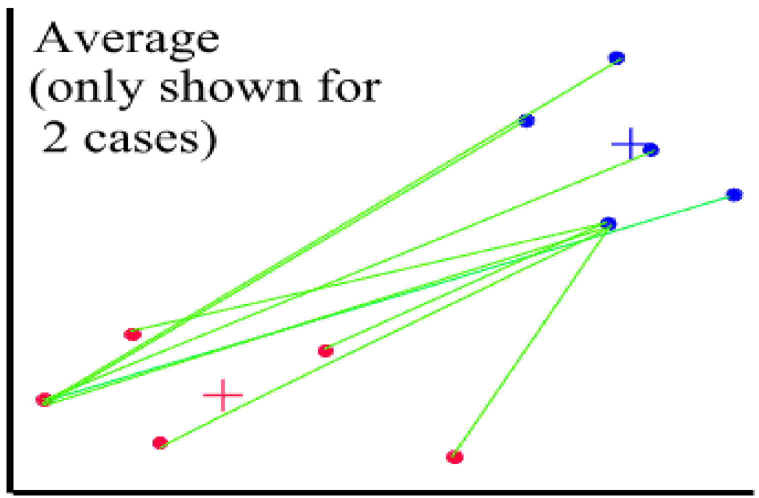
Complete Linkage

- dist(A,B) = max dist(a,b), a ∈ A, b ∈ B
- Distância máxima entre qualquer ponto da classe A e qualquer ponto da classe B



Average Linkage

- $dist(A,B) = 1/(N_A N_B) \sum dist(a,b), a \in A, b \in B$
- Distância média de todos os pares de pontos das classes A e B



SISTEMAS DE INFORMAÇÃO

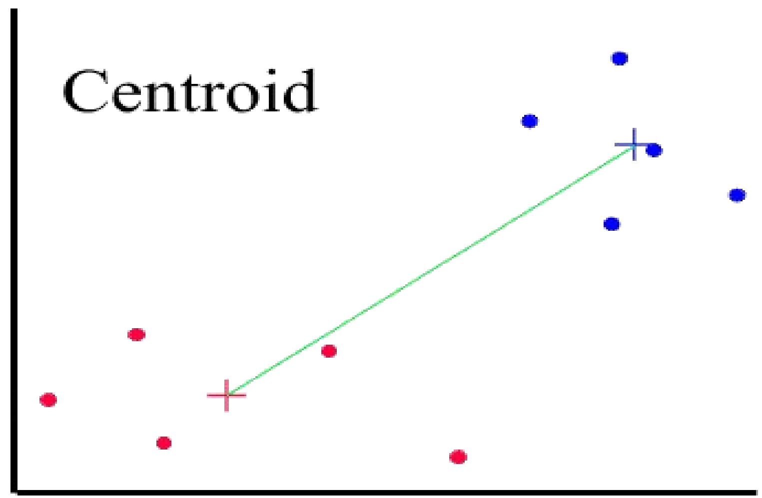
EACH

Centróide

- dist(A,B) = dist (C^A,C^B)

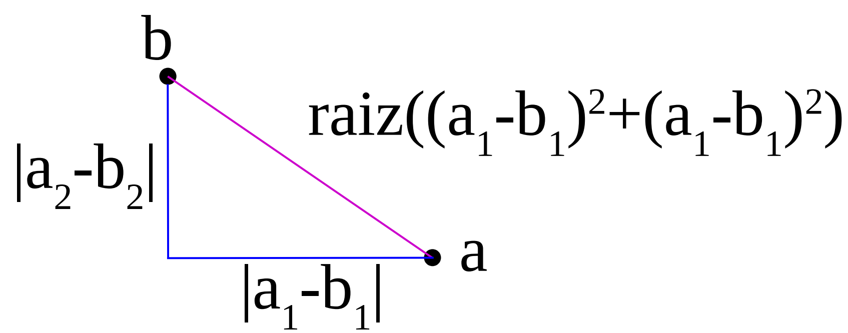
  C^A[i] é a média dos componentes i dos pontos do grupo A

Distância entre os centros de massa (centróides) da classe A (C^A) e da classe B (C^B)



Distâncias entre pontos

- E ainda é preciso definir a medida de distância entre dois pontos (**a**,**b**):
  - Distância euclidiana ou norma L2 = raiz($\Sigma_{i=1..n}(a_i\text{-}b_i)^2$)
  - Distância Manhattan, city block ou norma L1 = $\Sigma_{i=1..n} |a_i\text{-}b_i|$



Distâncias entre pontos

- E ainda é preciso definir a medida de distância entre dois pontos (**a**,**b**):
  - Distância euclidiana ou norma L2 = $raiz(\Sigma_{i=1..n}(a_i-b_i)^2)$
  - Distância Manhattan, city block ou norma L1 = $\Sigma_{i=1..n} |a_i-b_i|$
  - Distância Minkovski ou norma Lp = $(\Sigma_{i=1..n}|a_i-b_i|^p)^{1/p}$
  - Distâncias ponderadas, ex: $(\Sigma_{i=1..n} w_i|a_i-b_i|^p)^{1/p}$
  - Distância chessboard = $\max_i (|a_i-b_i|)$
  - Distância Mahalanobis = $((\mathbf{a}-\mathbf{b})^T S^{-1} (\mathbf{a}-\mathbf{b}))^{1/2}$

    S : matriz de covariância do dataset
    - é uma distância "normalizada"- atenua a diferença entre as escalas
    - Quando S é diagonal, é chamada

    distância euclidiana normalizada: $raiz(\Sigma_{i=1..n}((a_i-b_i)^2/s_i^2)$, s_i é o desvio padrão da característica i no dataset

    Lembram da ideia de executar o PCA antes?

Agrupamento Hierárquico
Algoritmo

Construa uma matriz de distâncias D NxN

para n de 1 até N-1

(a) Determine a distância mínima em D, e os grupos C_j e C_k (j<k) que determinam essa distância

(b) $C_{N+n} \leftarrow C_j \cup C_k$

(c) Atualize a matriz, agora sem a linha e coluna k, e considerando que a linha e coluna j correspondem agora ao grupo C_{N+n}

SISTEMAS DE INFORMAÇÃO

EACH

Exemplo – Single Linkage

N = 5 objetos

$$F = \begin{bmatrix} 1.2 & 2.0 \\ 3.0 & 3.7 \\ 1.5 & 2.7 \\ 2.3 & 2.0 \\ 3.1 & 3.3 \end{bmatrix}$$

$$D^{(1)} = \begin{array}{cc} \begin{bmatrix} 0 & & & & \\ 2.4759 & 0 & & & \\ 0.7616 & 1.8028 & 0 & & \\ 1.1000 & 1.8385 & 1.0630 & 0 & \\ 2.3022 & 0.4123 & 1.7088 & 1.5264 & 0 \end{bmatrix} & \begin{array}{c} C_1 \\ C_2 \\ C_3 \\ C_4 \\ C_5 \end{array} \end{array}$$

n = 1, e dmin = 0.4123 (j =2 e k = 5)

$C_6 \leftarrow C_2 \cup C_5$

Exemplo – Single Linkage (cont.)

$$D^{(2)} = \begin{array}{c} \begin{bmatrix} 0 & & & \\ 2.3022 & 0 & & \\ 0.7616 & 1.7088 & 0 & \\ 1.1000 & 1.5264 & 1.0630 & 0 \end{bmatrix} \end{array} \begin{array}{c} C_1 \\ C_2C_5 = C_6 \\ C_3 \\ C_4 \end{array}$$

n = 2, e dmin = 0.7616 (j =1 e k = 3)

$C_7 \leftarrow C_1 \cup C_3$

Exemplo – Single Linkage (cont.)

$$D^{(2)} = \begin{array}{c} \begin{bmatrix} 0 & & \\ 1.7088 & 0 & \\ 1.0630 & 1.5264 & 0 \end{bmatrix} \end{array} \begin{array}{c} C_1C_3 = C_7 \\ C_2C_5 = C_6 \\ C_4 \end{array}$$

n = 3, e dmin = 1.0630 (j =1 e k = 3)

$C_8 \leftarrow C_7 \cup C_4$

Exemplo – Single Linkage (cont.)

$$D^{(2)} = \begin{array}{c} \begin{bmatrix} 0 & \\ 1.5264 & 0 \end{bmatrix} \end{array} \begin{array}{l} C_1C_3C_4 = C_8 \\ C_2C_5 = C_6 \end{array}$$

n = 4, e dmin = 1.5264 (j =1 e k = 2)

$C_9 \leftarrow C_8 \cup C_6$

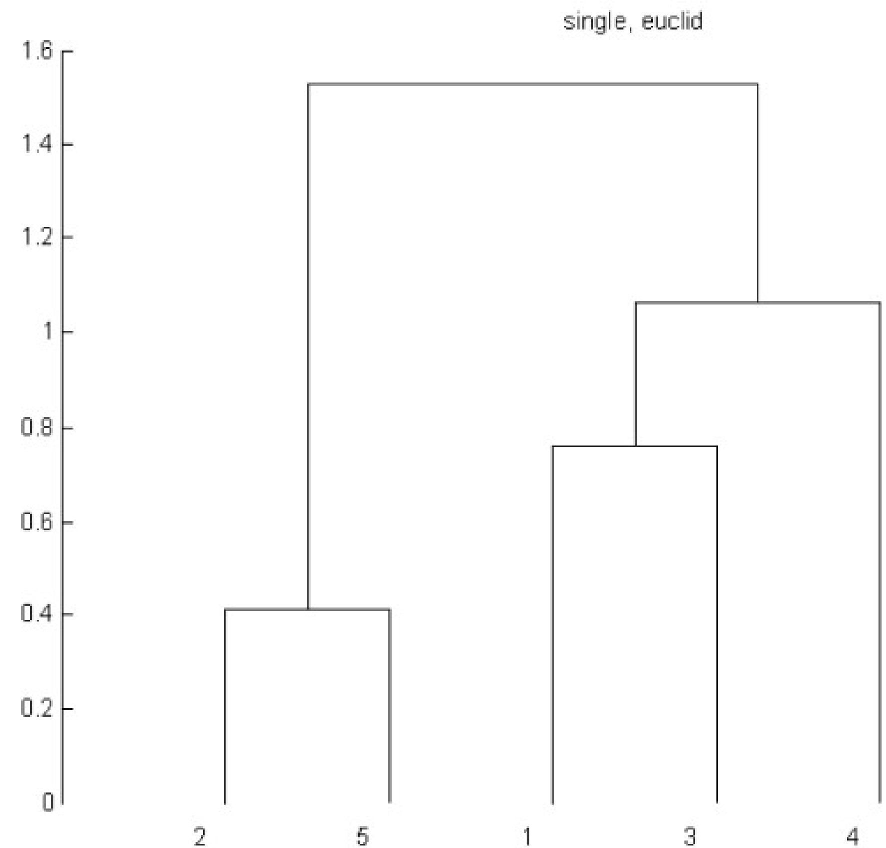
Exemplo – Single Linkage (cont.)

n = 1, e dmin = 0.4123 (j =2 e k = 5)
$C_6 \leftarrow C_2 \cup C_5$

n = 2, e dmin = 0.7616 (j =1 e k = 3)
$C_7 \leftarrow C_1 \cup C_3$

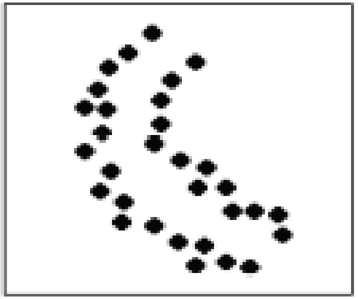
n = 3, e dmin = 1.0630 (j =1 e k = 3)
$C_8 \leftarrow C_7 \cup C_4$

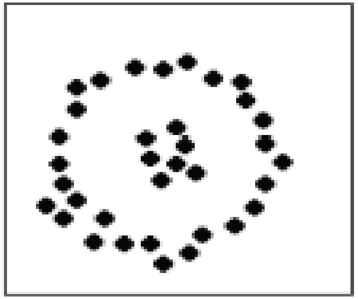
n = 4, e dmin = 1.5264 (j =1 e k = 2)
$C_9 \leftarrow C_8 \cup C_6$



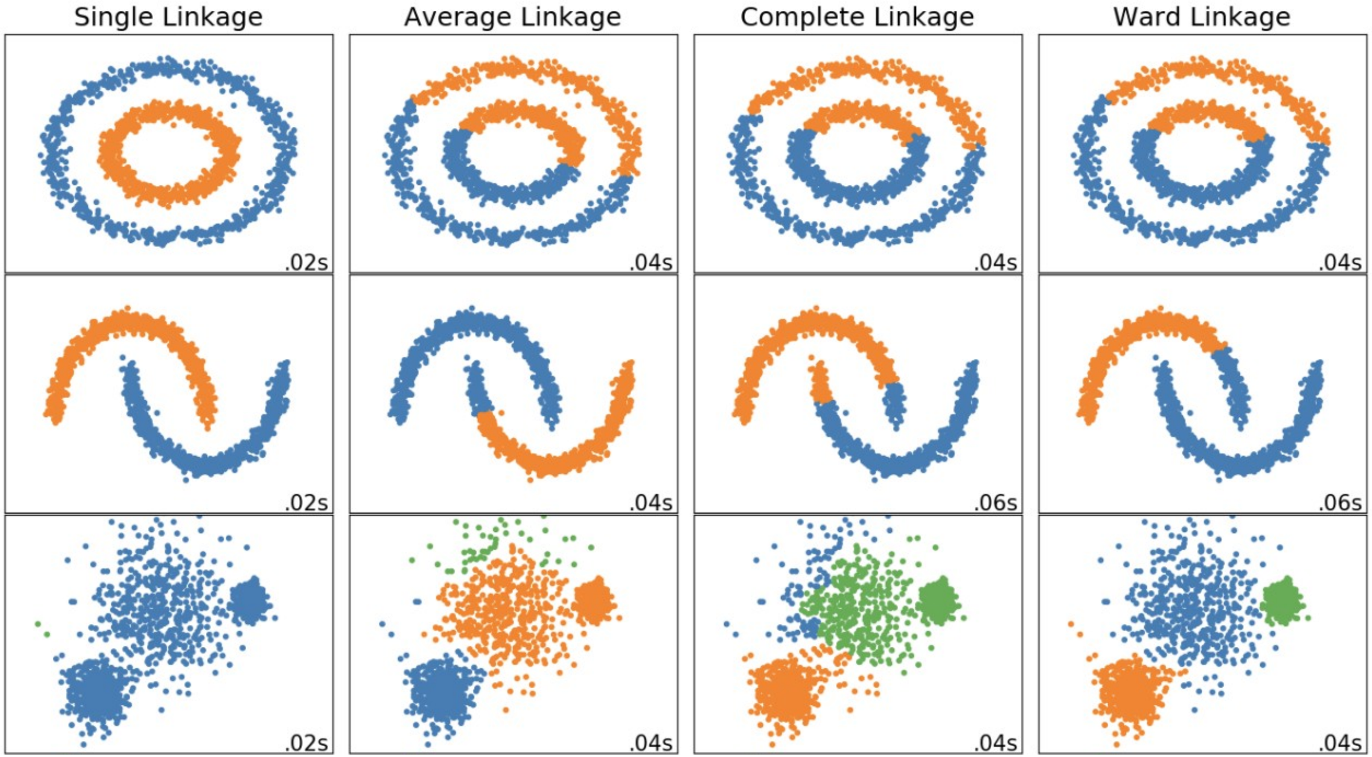
Comparação entre métodos hierárquicos

- Single Linkage:
  - Tendência de *chaining*, o que implica na união de grupos bem separados mas conectados por alguns poucos pontos
  - Não adequado para dados gaussianos (BAYNE, 1980)
  - Menos afetado por *outliers*
  - Um dos poucos que funcionam bem para dados não elipsoides
  - Desempenho pobre no caso geral





Comparação entre métodos hierárquicos



https://towardsdatascience.com/machine-learning-algorithms-part-12-hierarchical-agglomerative-clustering-example-in-python-1e18e0075019

- Centroid linkage:
  - Sugerido apenas para distância Euclidiana
  - Adequado para tratar grupos de diferentes tamanhos

Comparação entre métodos hierárquicos

- Conclusão:
  - Nenhum é o melhor para todos os casos
  - Solução: testar vários métodos e escolher aquele mais compatível com o esperado
  - Outros comentários no final do último vídeo

Determinando a relevância e o número de grupos

- Agrupamento hierárquico permite a escolha de K grupos, $1 \leq K \leq N$
- Qual K escolher?

Determinando a relevância e o número de grupos

- Agrupamento hierárquico permite a escolha de K grupos, $1 \leq K \leq N$
- Qual K escolher?

- Nenhuma regra de ouro

- Mas há dicas para tentar ter uma idéia da relevância de certos grupos...

Determinando a relevância e o número de grupos

- Dica 1: olhar para os grupos com maior tempo de vida – intervalo de distância entre a criação de um grupo e sua união com outro grupo
- Ex: indicação de 2 grupos

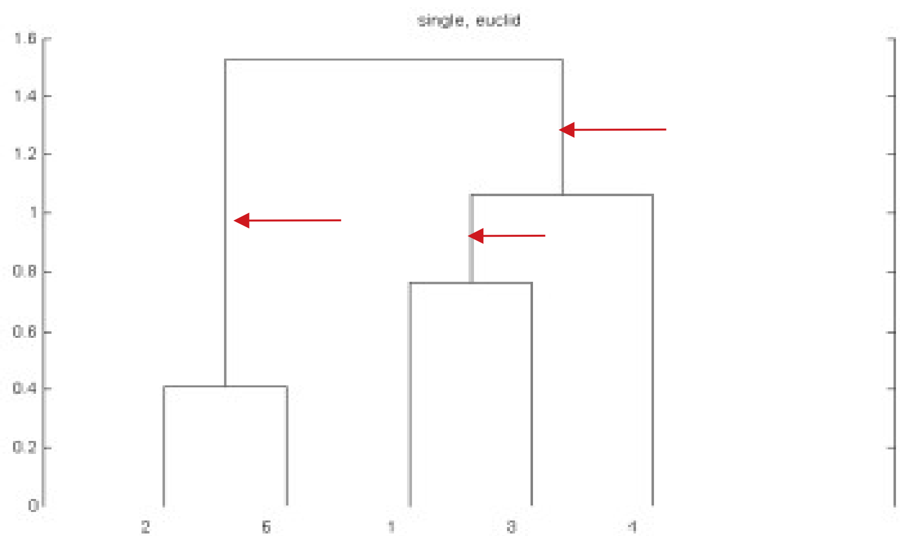


Figure 8.34: *The obtained dendrogram.*

Determinando a relevância e o número de grupos

- Dica 2: olhar para o maior salto entre as distâncias em que houve união de algum grupo (se acima de um limiar). Ex:

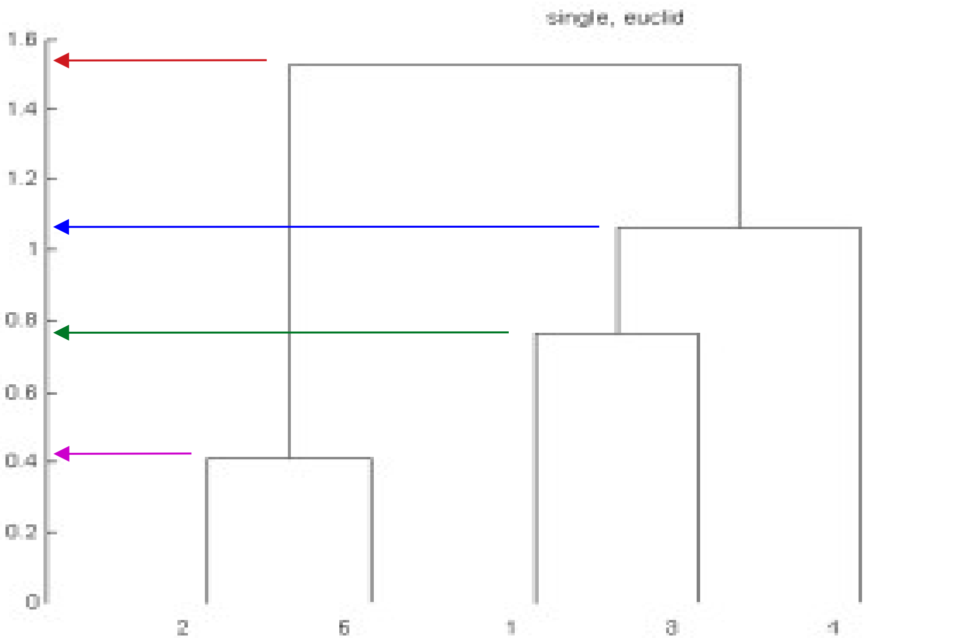


Figure 8.34: *The obtained dendrogram.*

Number of clusters	Distance	Distance jump
4	0.4123	0.3493
3	0.7616	0.3014
2	1.0630	$\boxed{0.4643}$
1	1.5264	—

Fim do vídeo 2

Agrupamento Hierárquico

Vídeo 3

Agrupamento Particional

Técnicas de agrupamentos

- Hierárquico: agrupamento progressivo de elementos (formação de subclasses)

- Particional: grupos obtidos a partir de uma partição do espaço de características com respeito a um número fixo de grupos

- Outros

Agrupamento Particional – algoritmo simples

- Um algoritmo possível baseado nas dispersões (para um dado número fixo de classes):
- Ideia: um bom agrupamento deveria exibir
  - Baixa dispersão intraclasse
  - Alta dispersão interclasse
- Função custo: trace(S_{intra})

Agrupamento Particional – algoritmo simples

Entrada: dataset (matriz dos dados - vetores de características)

Associe aleatoriamente uma classe a cada objeto

Enquanto não satisfizer o critério de parada

    Selecione aleatoriamente um objeto

    Mude a classe desse objeto (aleatoriamente, mas sem deixar classes vazias)

    Se trace(S_{intra}^{nova}) > trace($S_{intra}^{anterior}$)

        Volte o objeto à sua classe anterior

Agrupamento Particional – algoritmo simples

Entrada: dataset (matriz dos dados - vetores de características)

Associe aleatoriamente uma classe a cada objeto

Enquanto não satisfizer o critério de parada (ex: quando os grupos estabilizarem, por exemplo, quando o número de interações sem alteração de classificação for acima de um limiar)

    Selecione aleatoriamente um objeto

    Mude a classe desse objeto (aleatoriamente, mas sem deixar classes vazias)

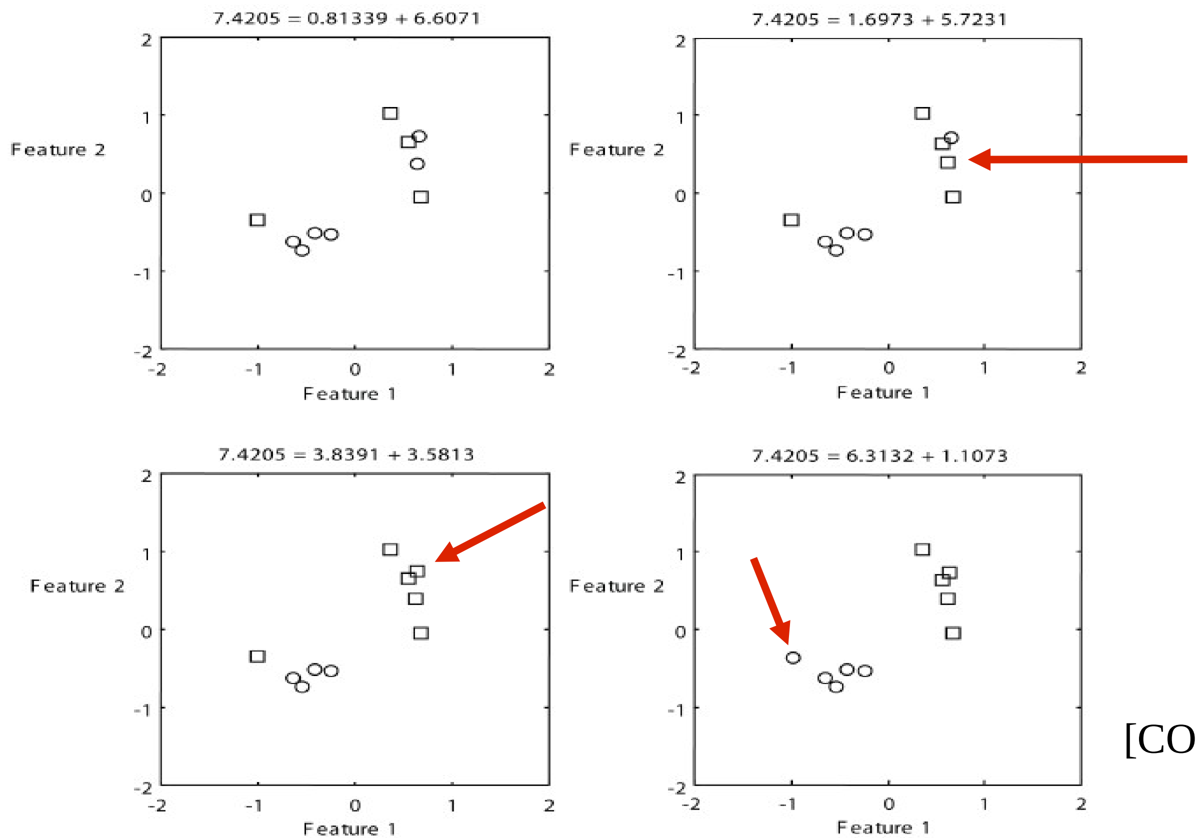
    Se trace(S_{intra}^{nova}) > trace($S_{intra}^{anterior}$)

        Volte o objeto à sua classe anterior

Agrupamento Particional – algoritmo simples

- Note a importância do número de grupos ser fixo! Caso contrário haveria uma tendência a aumentar o número de grupos (grupos menores tendem a ter menores dispersões intraclasse)

Agrupamento Particional Simples – Ex:(apenas sitações intermediárias de decréscimo de S_{intra})



[COSTA& CESAR, 2009]

Agrupamento Particional – algoritmo simples

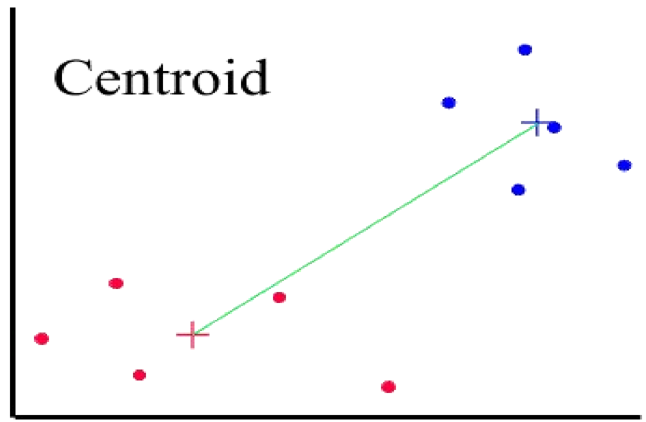
- Equivalente a métodos baseados em erro quadrático
- Rápida convergência mas...
- Sofre do problema de mínimos locais

Agrupamento Particional
K-Médias (k-means)

- Algoritmo mais famoso dessa categoria
- Também sofre de mínimos locais
- Número de classes pré-definido (K)
- Distância entre pontos (objetos) → Matriz de Distâncias ao invés de Matriz de Dispersão como medida de dissimilaridade
- K pontos iniciais para representar cada classe (sementes)
  - Não necessariamente pontos de objetos
  - Oportunidade para conhecer algum conhecimento a priori (senão, seleção aleatória)

Agrupamento Particional K-Médias

- Centróide: centro de massa



- Cada centróide define uma área de influência
  - Pontos mais próximos dele do que de qualquer outro centróide

Agrupamento Particional
K-Médias

Entrada: k, dataset (matriz dos dados – vetores de caract.)

Agrupamento Particional
K-Médias

1. Você define k grupos (ex: k = 2)

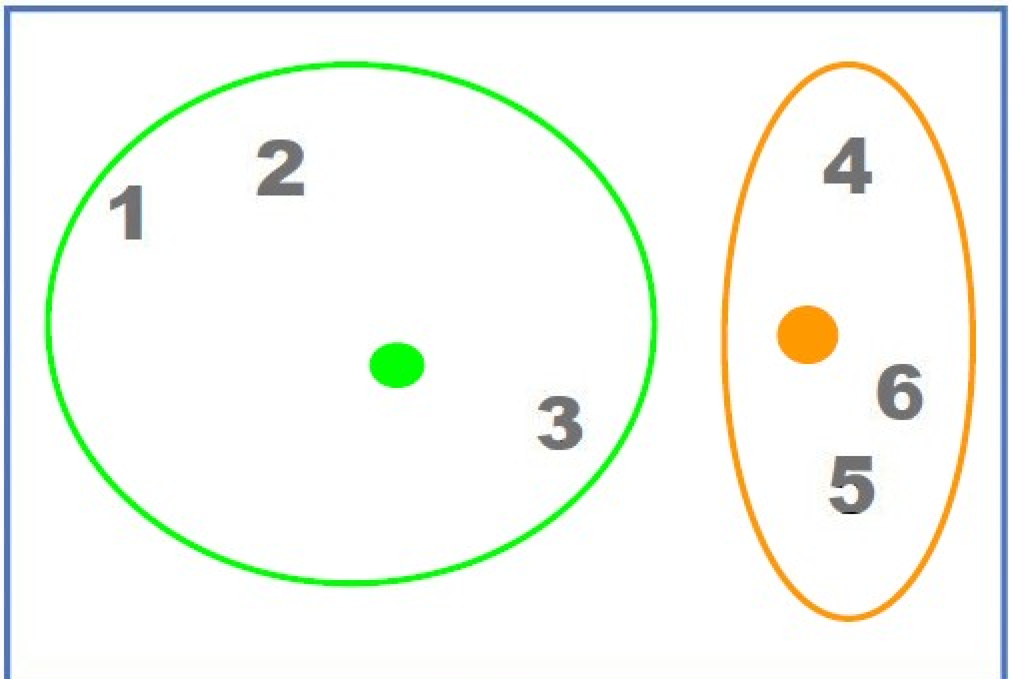
Agrupamento Particional
K-Médias

1. Você define k grupos (ex: k = 2)
2. Escolha aleatoriamente dois pontos representantes dessa classe (sementes) quaisquer (se tiver algum palpite acerca dos grupos use-o!)

Agrupamento Particional
K-Médias

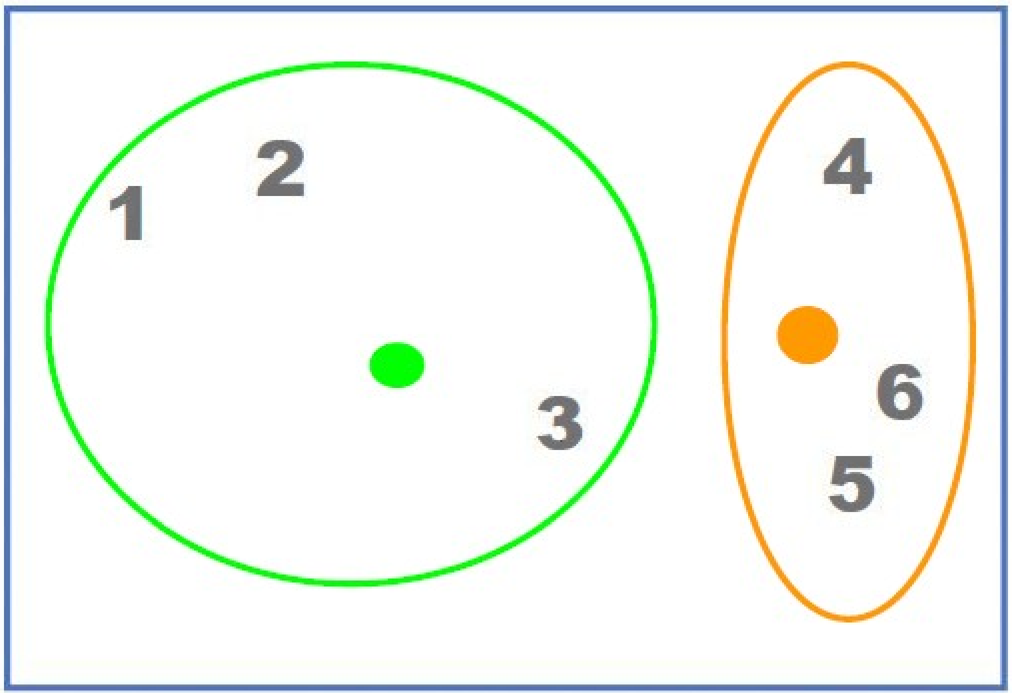
3. Classifique todos os seus dados de acordo com a semente mais próxima

Agrupamento Particional
K-Médias



3. Classifique todos os seus dados de acordo com a semente mais próxima

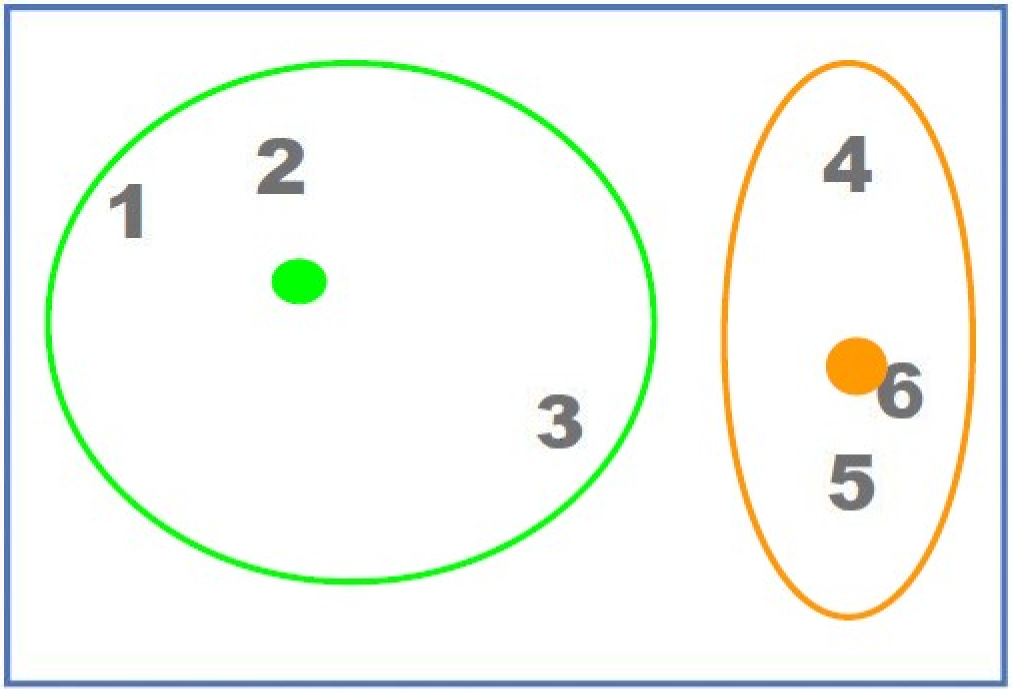
Agrupamento Particional
K-Médias



3. Classifique todos os seus dados de acordo com a semente mais próxima

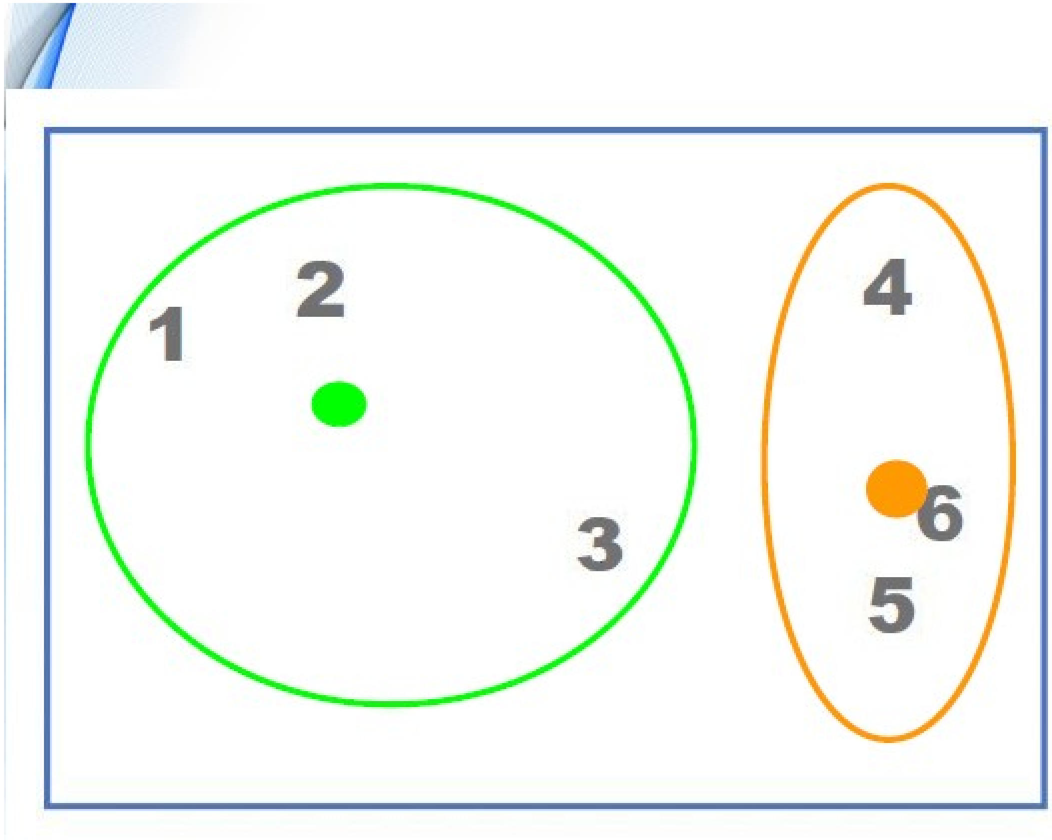
4. Defina a nova semente no centro (de massa)

Agrupamento Particional
K-Médias



3. Classifique todos os seus dados de acordo com a semente mais próxima
4. Defina a nova semente no centro (de massa)

Agrupamento Particional
K-Médias

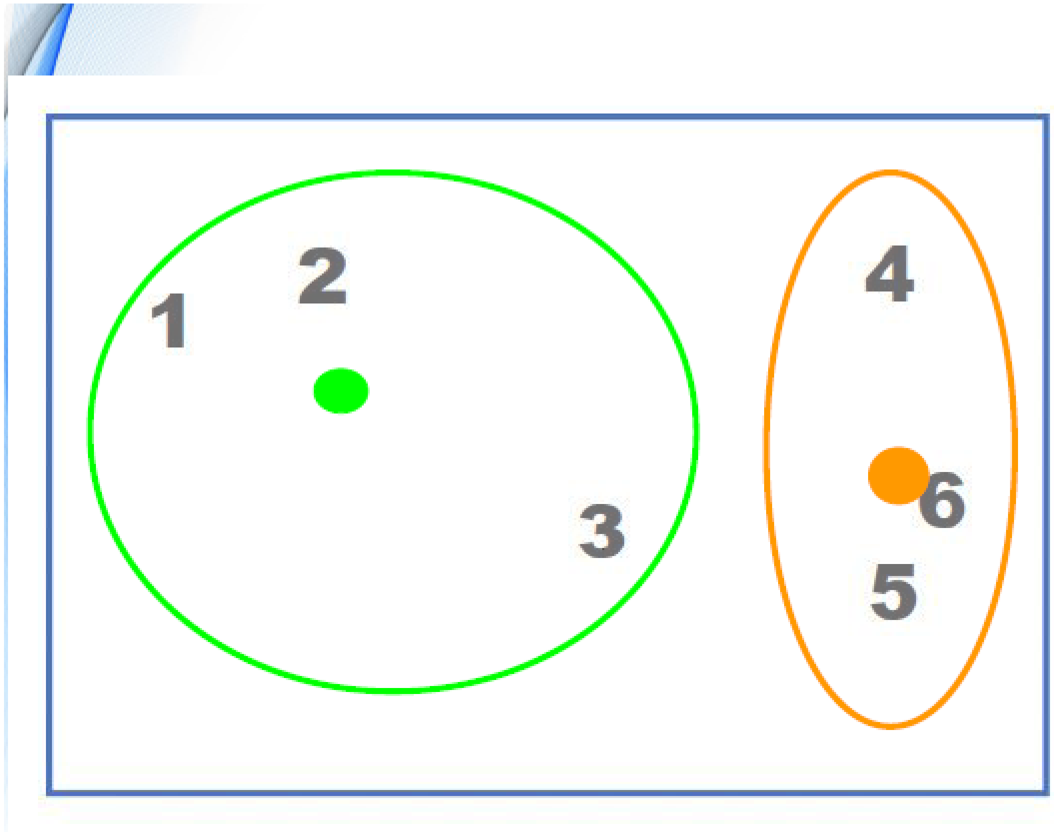


3. Classifique todos os seus dados de acordo com a semente mais próxima

4. Defina a nova semente no centro (de massa)

5. Repita 3 e 4 até que a classificação dos pontos não mude mais

Agrupamento Particional
K-Médias

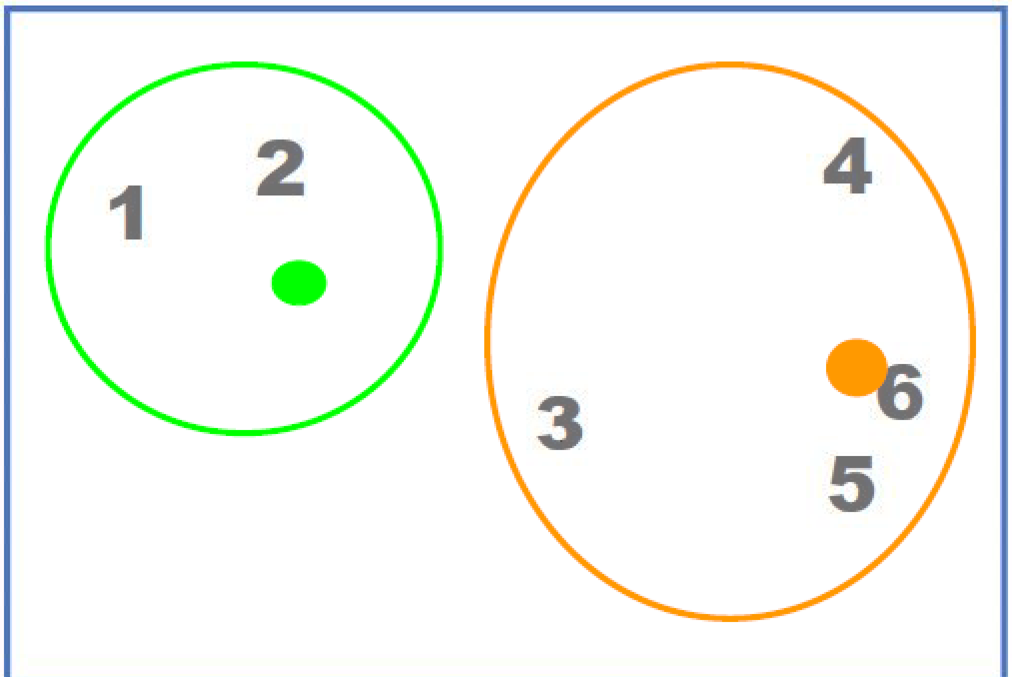


3. Classifique todos os seus dados de acordo com a semente mais próxima

4. Defina a nova semente no centro (de massa)

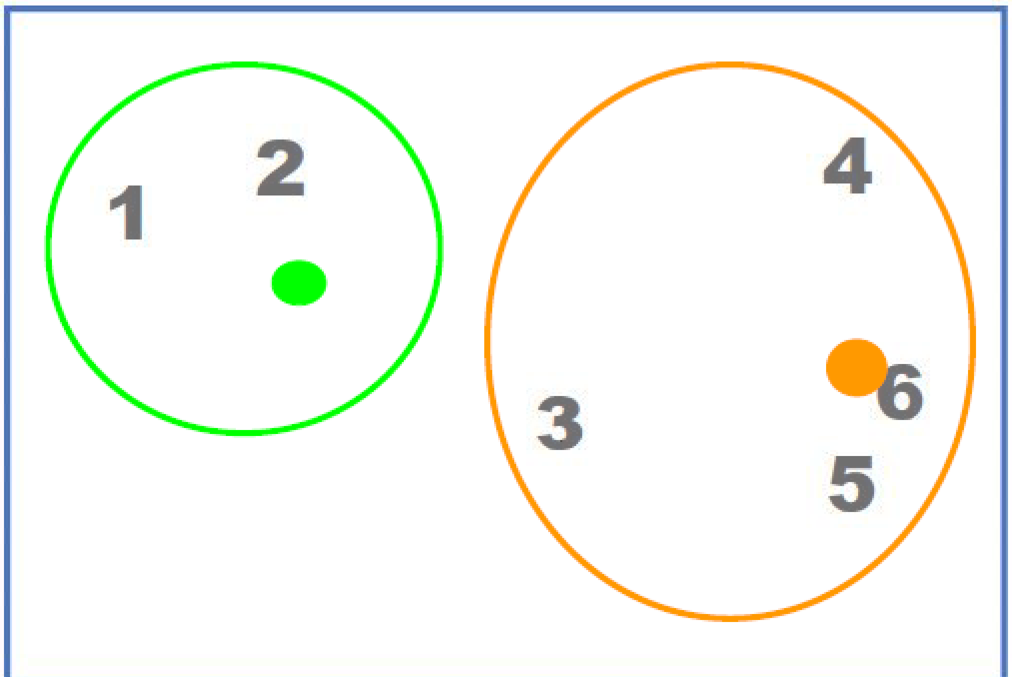
5. Repita 3 e 4 até que a classificação dos pontos não mude mais

Agrupamento Particional
K-Médias



3. Classifique todos os seus dados de acordo com a semente mais próxima
4. Defina a nova semente no centro (de massa)
5. Repita 3 e 4 até que a classificação dos pontos não mude mais

Agrupamento Particional
K-Médias

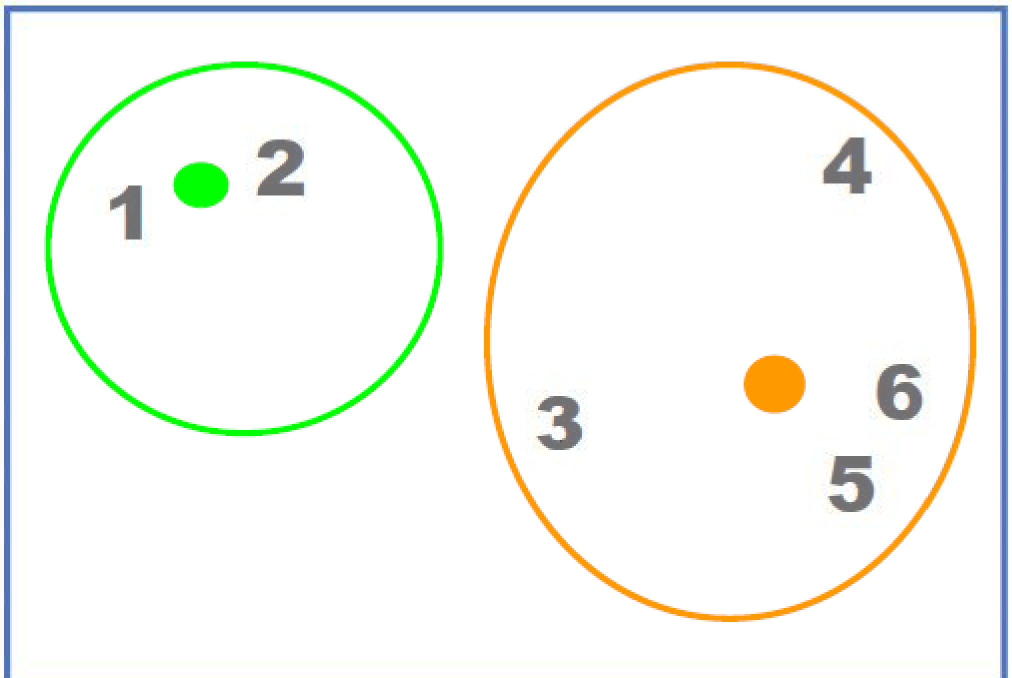


3. Classifique todos os seus dados de acordo com a semente mais próxima

4. Defina a nova semente no centro (de massa)

5. Repita 3 e 4 até que a classificação dos pontos não mude mais

Agrupamento Particional
K-Médias

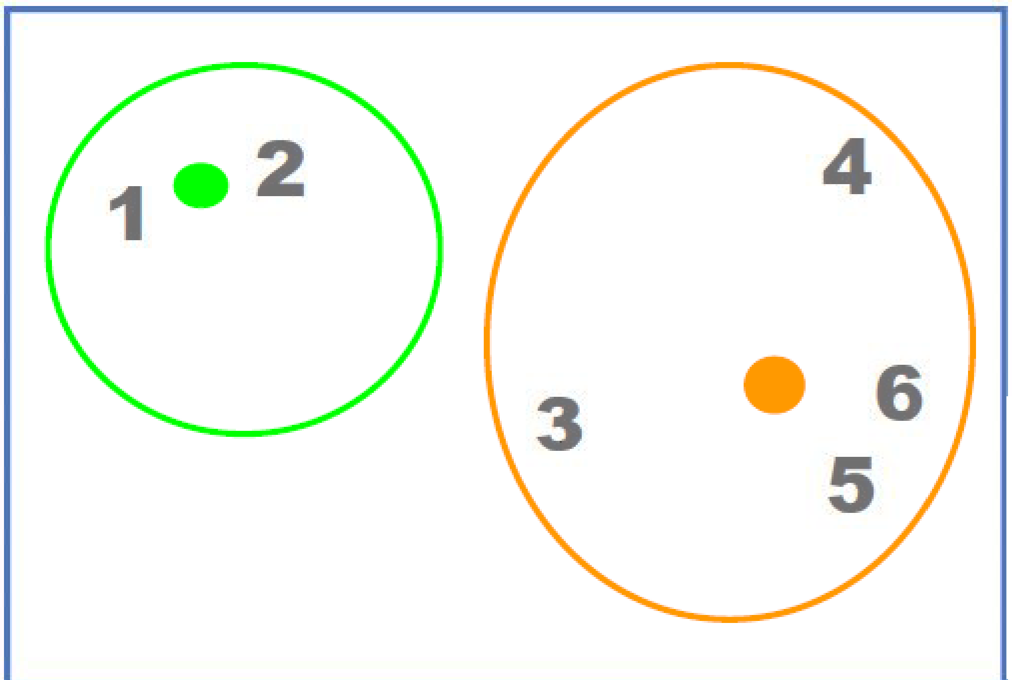


3. Classifique todos os seus dados de acordo com a semente mais próxima

4. Defina a nova semente no centro (de massa)

5. Repita 3 e 4 até que a classificação dos pontos não mude mais

Agrupamento Particional
K-Médias



3. Classifique todos os seus dados de acordo com a semente mais próxima

4. Defina a nova semente no centro (de massa)

5. Repita 3 e 4 até que a classificação dos pontos não mude mais

Agrupamento Particional
K-Médias

3. Classifique todos os seus dados de acordo com a semente mais próxima
4. Defina a nova semente no centro (de massa)
5. Repita 3 e 4 até que a classificação dos pontos não mude mais

Agrupamento Particional
K-Médias

Entrada: k, dataset D (matriz dos dados – vetores de caract.)

Escolha os K pontos protótipos iniciais (sementes) e guarde-os na lista W

Enquanto não satisfizer o critério de parada

  Calcule todas as distâncias entre cada objeto e os pontos protótipos P_i (matriz D KxN)

  Use a matriz D para identificar os objetos mais próximos de cada protótipo P_i (guarde-os na lista L_i)

  Obtenha como novos protótipos os centróides de cada L_i

Agrupamento Particional
K-Médias

- Critério de parada:
    - Ex: Deslocamento de cada centróide é menor que um limiar

Agrupamento Particional
K-Médias

- Critério de parada:
  - Ex: Deslocamento de cada centróide é menor que um limiar
- No final, os objetos mais próximos de um centróide pertencem à sua classe

Agrupamento Particional
K-Médias

- No final, os objetos mais próximos de um centróide P_i pertencem à sua classe C_i
- Distância utilizada: euclidiana
  - Isso equivale a usar uma função custo baseada em erro quadrático

$$E = \arg\min_{C} \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - P_i\|^2$$

SISTEMAS DE INFORMAÇÃO

EACH

Agrupamento Particional
K-Médias

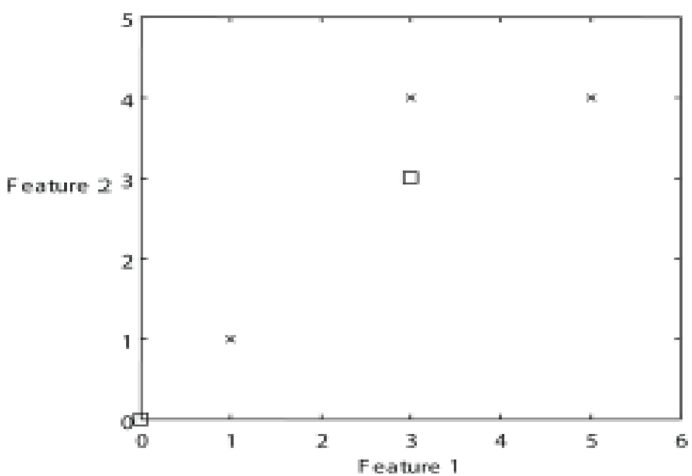
- Note que um grupo pode ficar vazio!
  - Possíveis soluções:
    - Reconhecer que há uma classe a menos
    - Escolher outra semente para aquela classe e re-executar o algoritmo (a partir do loop enquanto)

Agrupamento Particional
K-Médias - Exemplo

Object	Feature 1	Feature 2
X_1	1	1
X_2	3	4
X_3	5	4

Protótipos iniciais:

$P_1 = (0,0)$ e $P_2 = (3,3)$



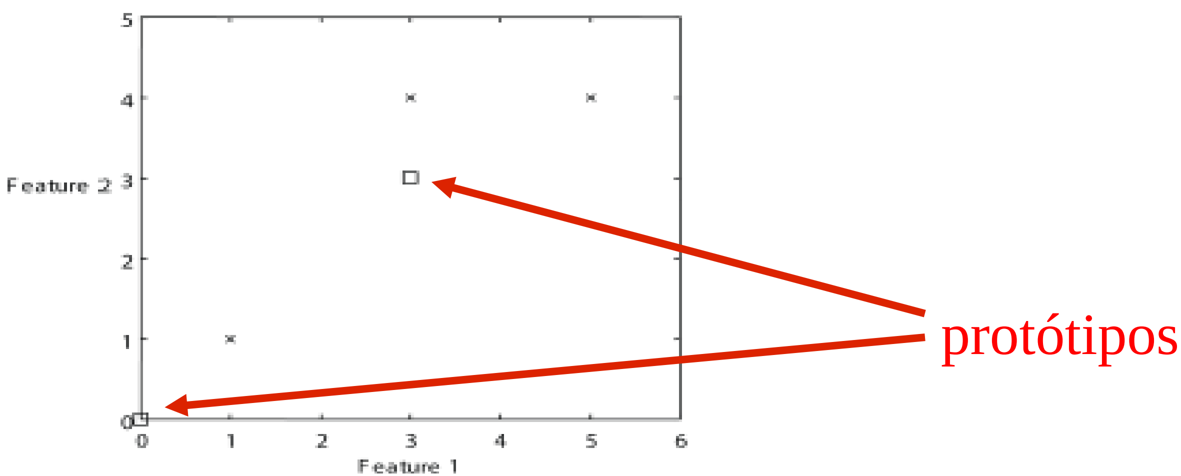
Critério de Parada: maior deslocamento de centróide m < 0,25

Agrupamento Particional
K-Médias - Exemplo

Object	Feature 1	Feature 2
X_1	1	1
X_2	3	4
X_3	5	4

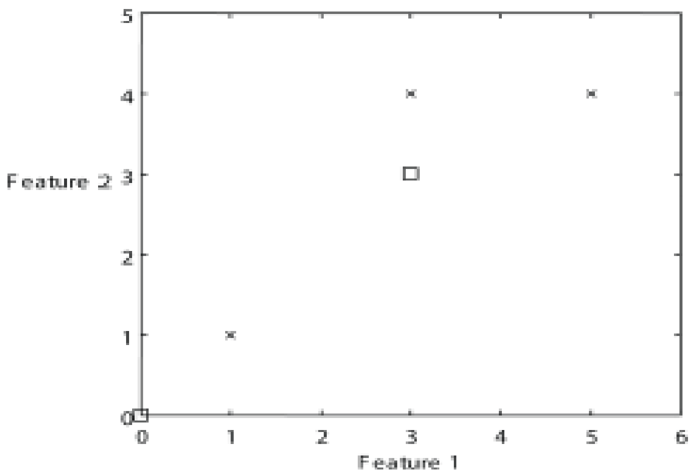
Protótipos iniciais:

$P_1 = (0,0)$ e $P_2 = (3,3)$



Agrupamento Particional
K-Médias - Exemplo

Object	Feature 1	Feature 2
X_1	1	1
X_2	3	4
X_3	5	4



Protótipos iniciais:

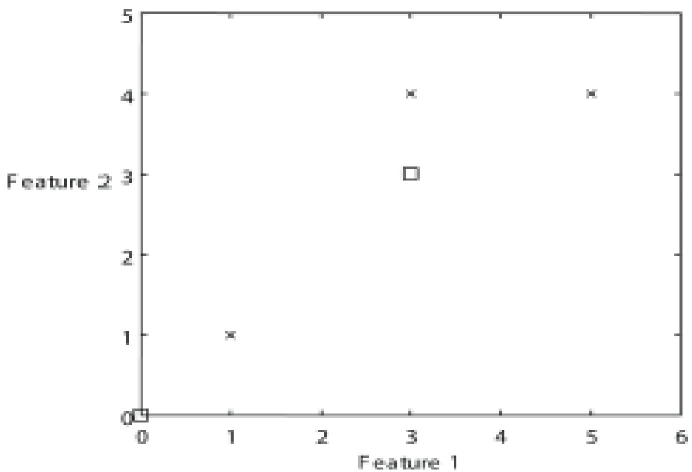
$P_1 = (0,0)$ e $P_2 = (3,3)$

$$D = \begin{bmatrix} \sqrt{2} & 5 & \sqrt{41} \\ 2\sqrt{2} & 1 & \sqrt{5} \end{bmatrix}$$

$$L_1 = (X_1) \quad \text{and} \quad L_2 = (X_2, X_3)$$

Agrupamento Particional
K-Médias - Exemplo

Object	Feature 1	Feature 2
X_1	1	1
X_2	3	4
X_3	5	4



Protótipos iniciais:

$P_1 = (0,0)$ e $P_2 = (3,3)$

$$D = \begin{bmatrix} \sqrt{2} & 5 & \sqrt{41} \\ 2\sqrt{2} & 1 & \sqrt{5} \end{bmatrix}$$

$$L_1 = (X_1) \quad \text{and} \quad L_2 = (X_2, X_3)$$

Protótipos novos:

P'_1 = média $\{X_1\}$ = (1,1) e

P'_2 = média $\{X_2,X_3\}$ = (4,4)

Agrupamento Particional K-Médias - Exemplo

Protótipos iniciais:

$P_1 = (0,0)$ e $P_2 = (3,3)$

Protótipos novos:

$P'_1 = \text{média } \{X_1\} = (1,1)$ e

$P'_2 = \text{média } \{X_2,X_3\} = (4,4)$

$$m = \max \{\|P'_1 - P_1\|, \| P'_2 - P_2 \|\}$$

$$= \max\{\text{raiz}(2), \text{raiz}(2) \} > 0.25$$

CONTINUA!

SISTEMAS DE INFORMAÇÃO

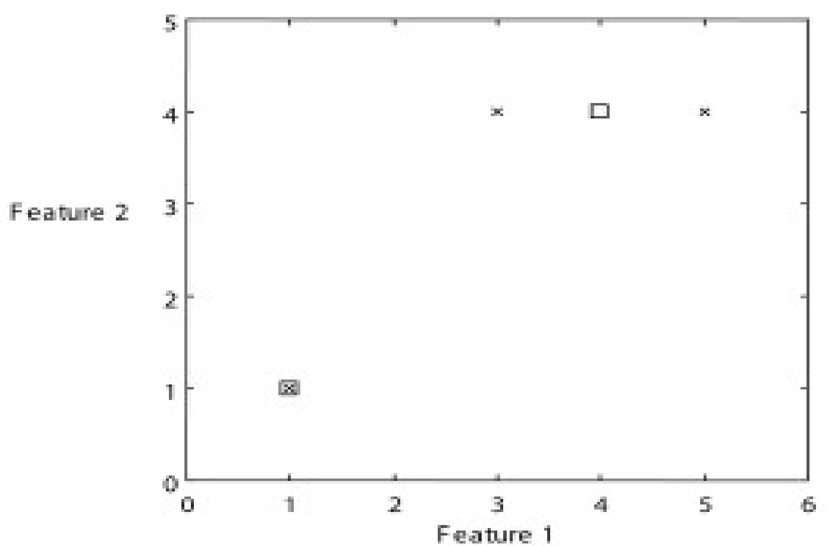
EACH

Agrupamento Particional
K-Médias - Exemplo

Object	Feature 1	Feature 2
X_1	1	1
X_2	3	4
X_3	5	4

Protótipos:

$P_1 = (1,1)$ e $P_2 = (4,4)$

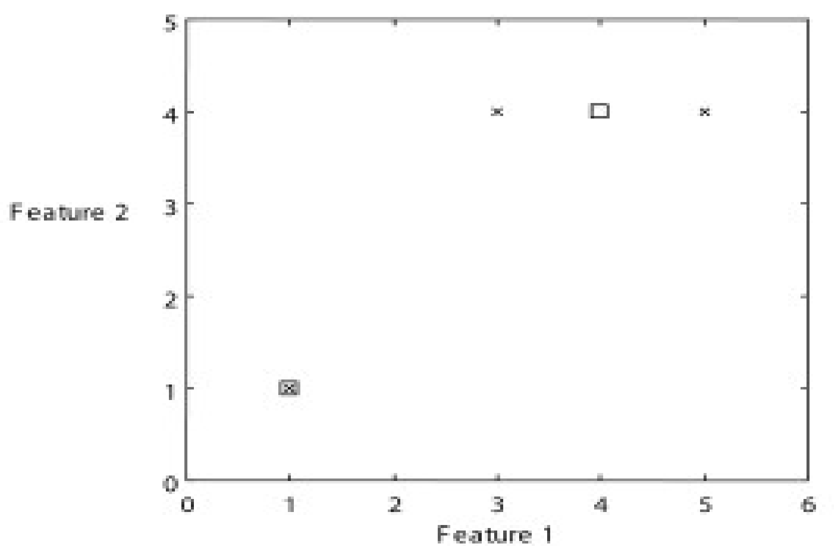


SISTEMAS DE INFORMAÇÃO

EACH

Agrupamento Particional
K-Médias - Exemplo

Object	Feature 1	Feature 2
X_1	1	1
X_2	3	4
X_3	5	4



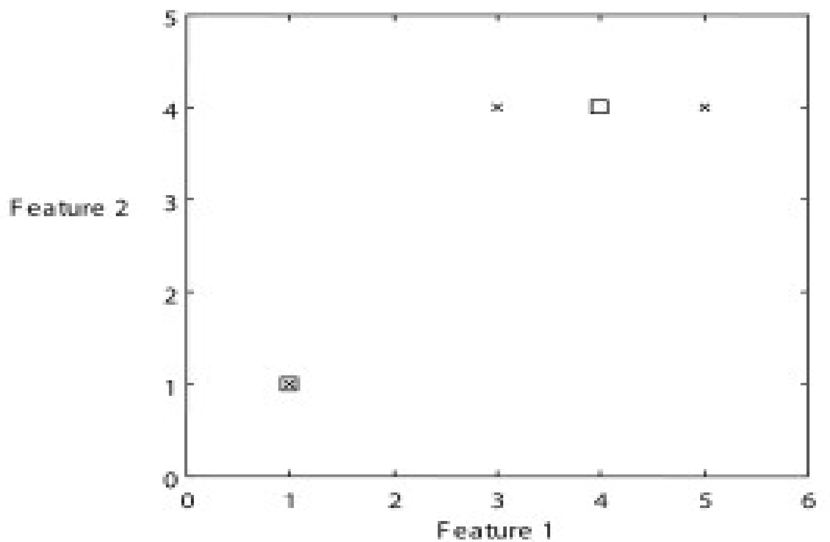
Protótipos:

$P_1 = (1,1)$ e $P_2 = (4,4)$

$$D = \begin{bmatrix} 0 & \sqrt{13} & 5 \\ \sqrt{18} & 1 & 1 \end{bmatrix}$$

Agrupamento Particional
K-Médias - Exemplo

Object	Feature 1	Feature 2
X_1	1	1
X_2	3	4
X_3	5	4



Protótipos:

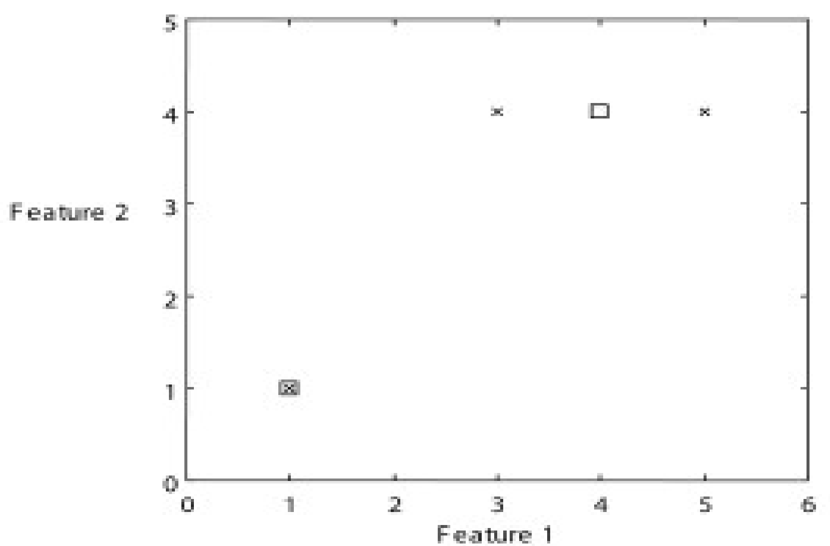
$P_1 = (1,1)$ e $P_2 = (4,4)$

$$D = \begin{bmatrix} 0 & \sqrt{13} & 5 \\ \sqrt{18} & 1 & 1 \end{bmatrix}$$

$$L_1 = (X_1) \quad \text{and} \quad L_2 = (X_2, X_3)$$

Agrupamento Particional
K-Médias - Exemplo

Object	Feature 1	Feature 2
X_1	1	1
X_2	3	4
X_3	5	4



Protótipos:

$P_1 = (1,1)$ e $P_2 = (4,4)$

$$D = \begin{bmatrix} 0 & \sqrt{13} & 5 \\ \sqrt{18} & 1 & 1 \end{bmatrix}$$

$$L_1 = (X_1) \quad \text{and} \quad L_2 = (X_2, X_3)$$

Protótipos novos:

P'_1 = média $\{X_1\} = (1,1)$ e

P'_2 = média $\{X_2,X_3\} = (4,4)$

Agrupamento Particional
K-Médias - Exemplo

Protótipos anteriores:

$P_1 = (1,1)$ e $P_2 = (4,4)$

Protótipos novos:

$P'_1 = \text{média } \{X_1\} = (1,1)$ e

$P'_2 = \text{média } \{X_2,X_3\} = (4,4)$

$m = \max \{\|P'_1 - P_1\|, \| P'_2 - P_2 \|\}$

$= \max\{0, 0\} < 0.25$

TERMINA!

SISTEMAS DE INFORMAÇÃO

EACH

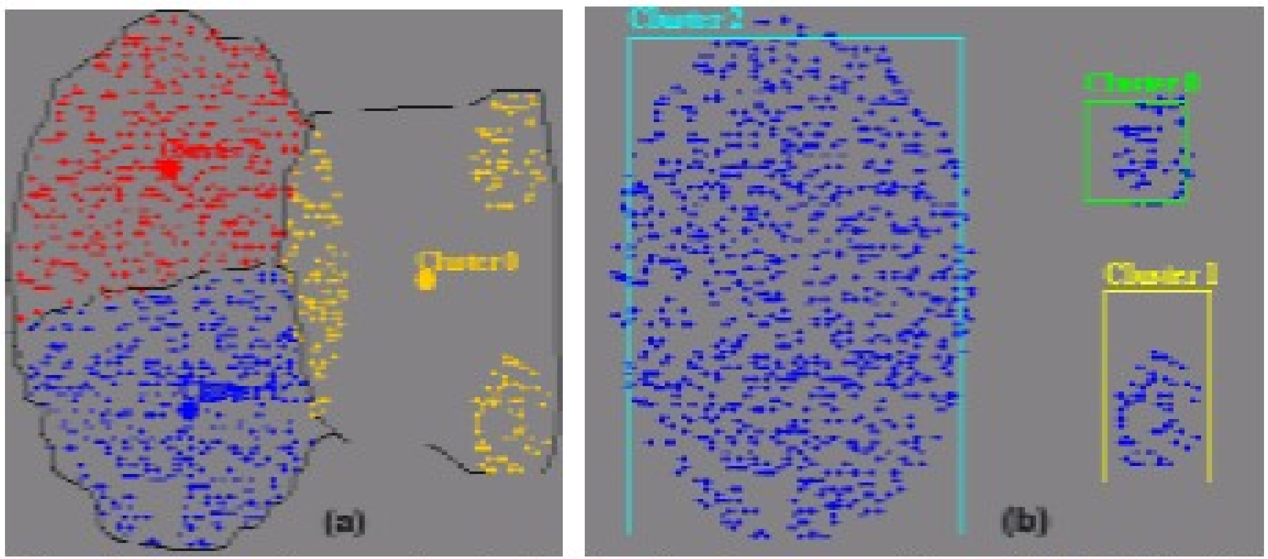
Agrupamento Particional
K-Médias

- A convergência para a menor dispersão não é garantida

- Alternativas:
  - várias rodadas (com diferentes sementes) e escolher a configuração com menor matriz de dispersão intraclasse
  - Juntar grupos de centróides próximos e subdividir grupos com alta dispersão intra-classe

Agrupamento Particional
K-Médias

- Sensibilidade à escolha inicial dos protótipos:



A Genetic Rule-Based Data Clustering Toolkit
I Sarafis, AMS Zalzala and P W Trinder

Agrupamento Particional
K-Médias

- Opções quanto à escolha inicial dos protótipos:
  - Totalmente aleatório
  - K pontos bem espaçados (ideia do kmeans++)
  - Com base em conhecimento a priori
  - M execuções (com protótipos diferentes): escolher o resultado com menor erro quadrático
  - Executar um agrupamento hierárquico, e escolher um protótipo para cada classe

Variações baseadas em K-Médias

- Escolha dos k protótipos iniciais
- Cálculo de dissimilaridade
- Estratégia para calcular o protótipo de cada classe

Variações baseadas em K-Médias

- Escolha dos k protótipos iniciais
  - Ex: k-means++
    - Cada objeto i tem um peso w_i (pesos inicialmente iguais)

    Para j = 1 até k

    sorteia protótipo Pj com base nos pesos w_i

    para cada objeto o_i que não é protótipo

    atualiza $w_i = (\min_l d(o_i, P_l))^2$ (l protótipos já definidos)

Variações baseadas em K-Médias

- **K-medoids** ou **PAM (partitioning around medoids)**:
  - protótipo deve ser um objeto da classe (para diminuir sensibilidade a outliers)

Variações baseadas em K-Médias

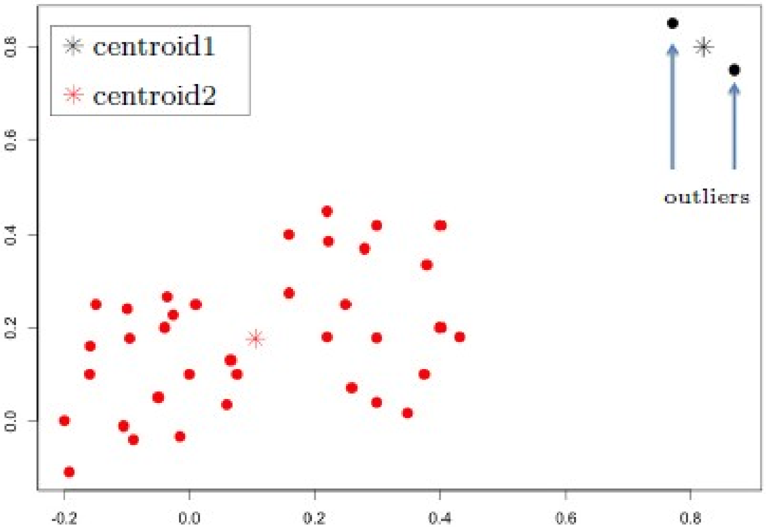
- **K-medoids** ou **PAM (partitioning around medoids)**:
  - protótipo deve ser um objeto da classe (para diminuir sensibilidade a outliers)
  - Função custo (erro absoluto): $E = \sum_{i=1}^{k} \sum_{x \in C_i} |x - P_i|$

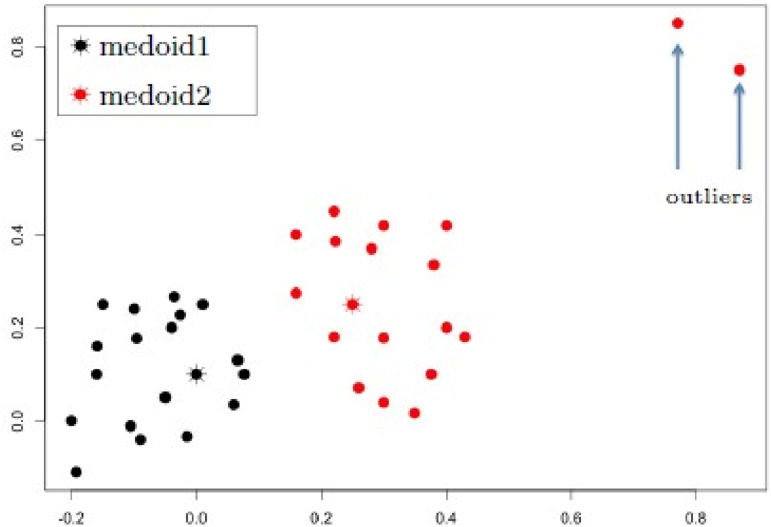
  (também diminui efeito de outliers)

  a cada iteração, verifica se é possível diminuir o custo total ao subtituir algum protótipo por um objeto aleatório o_{random}
  - Mais robusto na presença de outliers porém custoso que o k-means
  - CLARA e variantes: para datasets grandes (subamostragem)

Variações baseadas em K-Médias

- **K-medoids** ou **PAM (partitioning around medoids)**:





Generalized k-means based clustering for temporal data under time warp
Saeid SOHEILY-KHAH

  – Mais robusto na presença de outliers porém custoso que o k-means
  – CLARA e variantes: para datasets grandes (subamostragem)

Variações baseadas em K-Médias

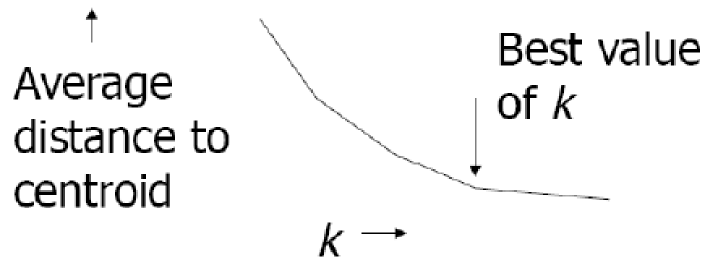
- K-mode para variáveis categóricas
- K-prototypes para datasets mistos (mistura de variáveis numéricas e categóricas)
- Fuzzy k-médias (cada objeto tem um probabilidade de pertencer a uma classe)

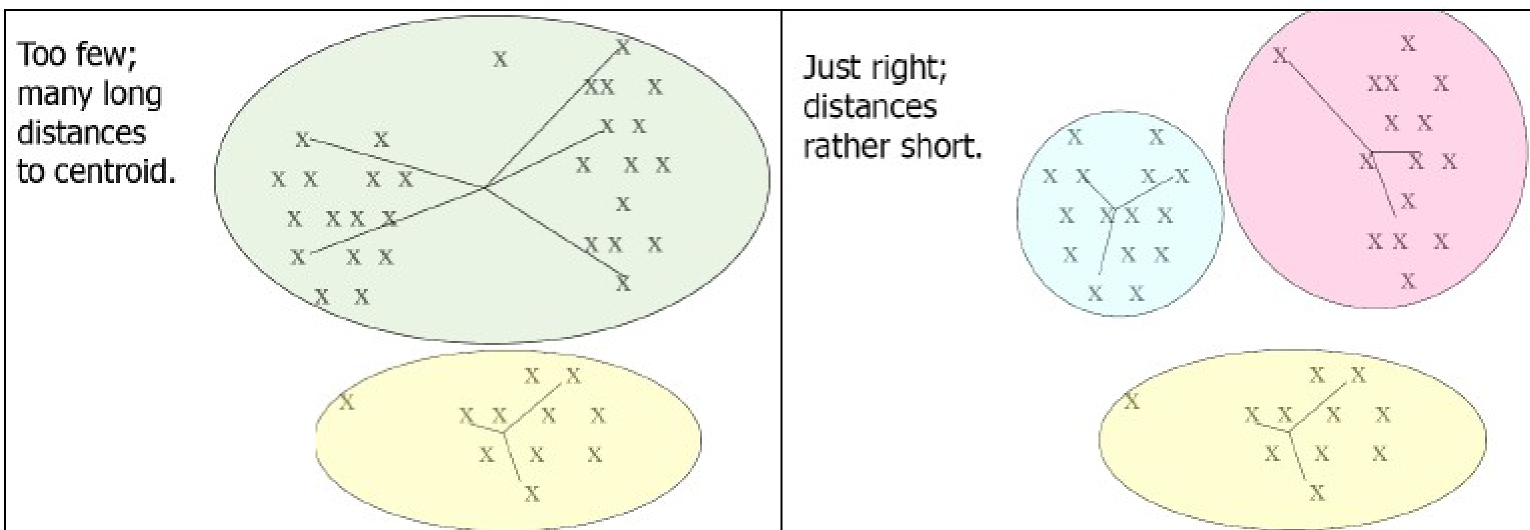
E se eu não sei o k?

- Pode ir direto para outros métodos (hierárquicos)
- Conseguir pistas do k:
  - Interpretação dos resultados de agrupamentos hierárquicos

E se eu não sei o k?

- Testar diferentes valores de k e calcular a distância média dos objetos aos seus respectivos centróides





https://web.stanford.edu/class/cs345a/slides/12-clustering.pdf

Comentários gerais

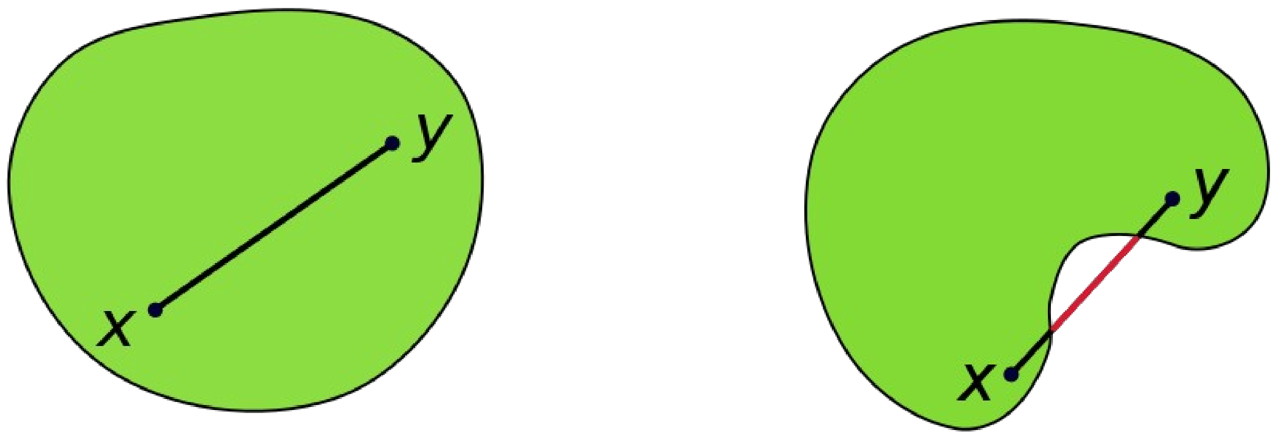
- Esses são algoritmos de realocação iterativa (um tipo de EM - Expectation-Maximization)
  - Achar a solução ótima (garantida) é NP-difícil

Comentários gerais

- Tendem a formar agrupamentos compactos (fronteiras convexas)

Conjunto convexo

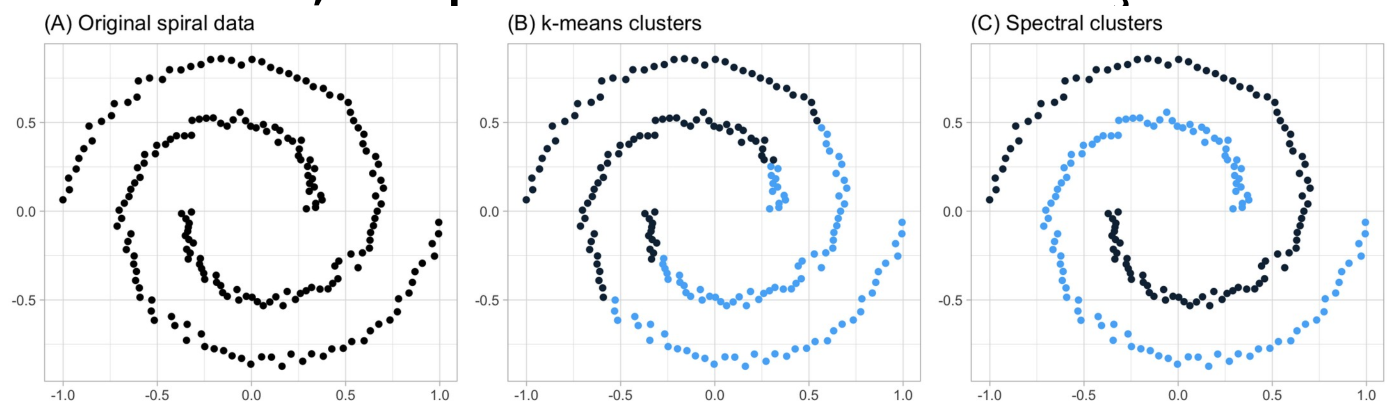
Conjunto não convexo



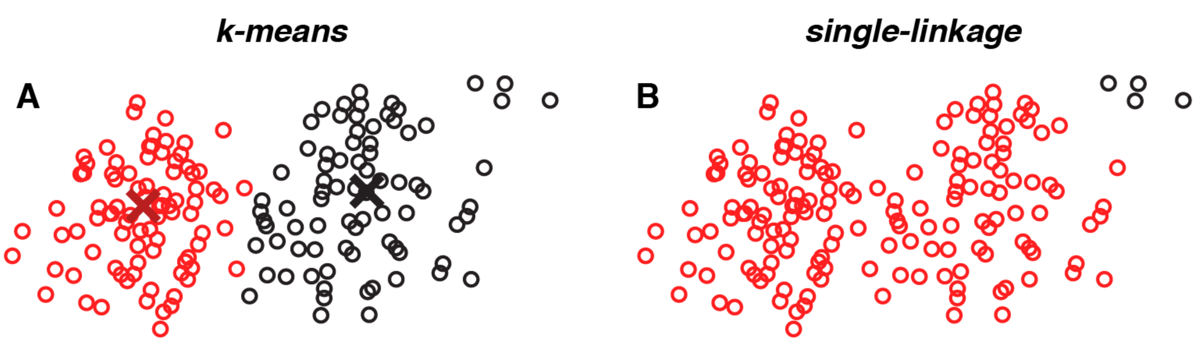
https://en.wikipedia.org/wiki/Convex_set

Comentários gerais

- Tendem a formar agrupamentos compactos (fronteiras convexas)
  - Pode ser bom ou não, depende da distribuição dos dados



https://bradleyboehmke.github.io/HOML/kmeans.html



← single-linkage

http://alexhwilliams.info/itsneuronalblog/2015/09/11/clustering1/

Fim do vídeo 3

Agrupamento Particional

Vídeo 4

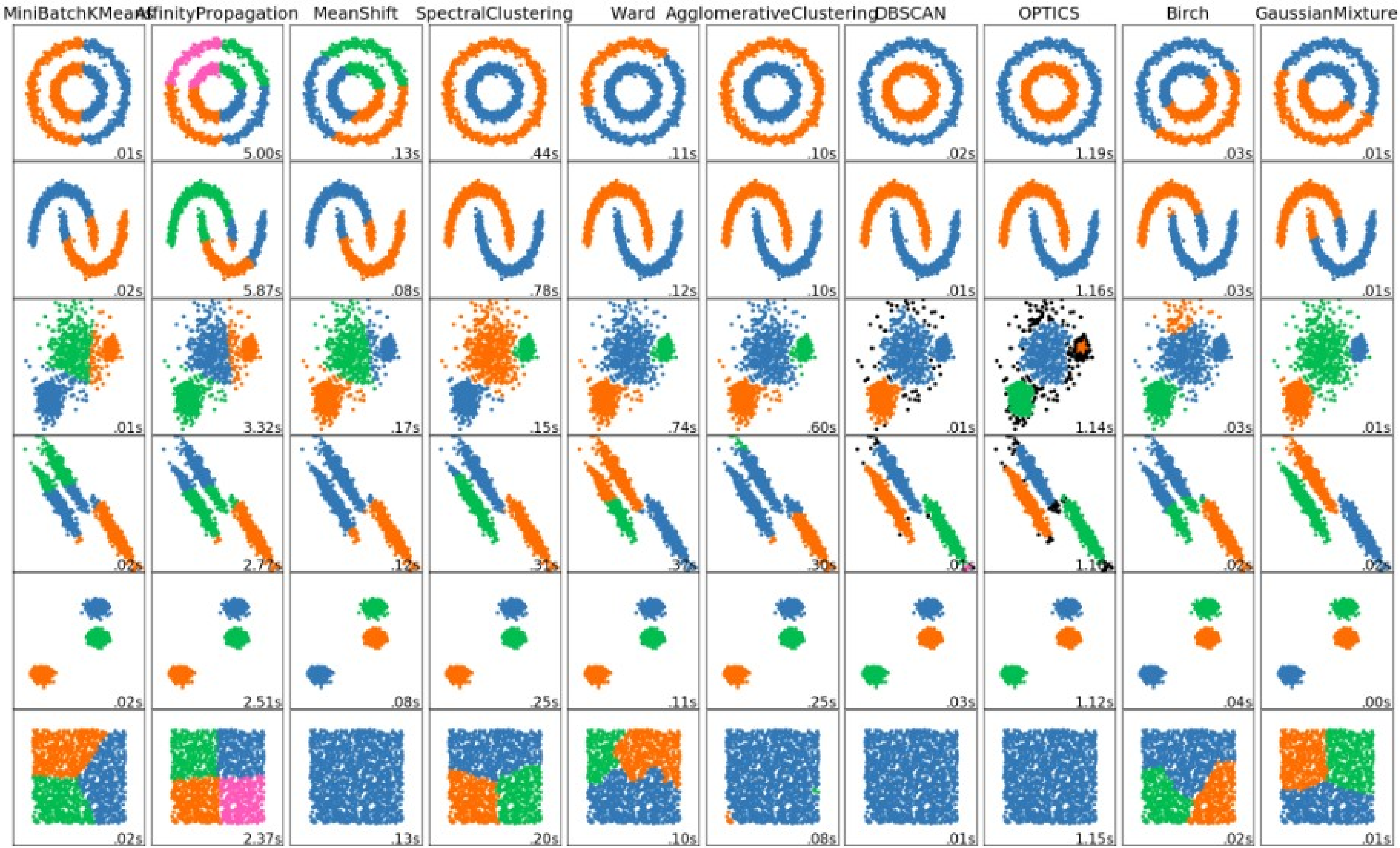
Outras técnicas, validação e comentários finais

SISTEMAS DE INFORMAÇÃO

EACH

Técnicas de agrupamentos

- **Hierárquico**: agrupamento progressivo de elementos (formação de subclasses) (ex: BIRCH, Chameleon, ROCK – para atributos categóricos)
- **Particional**: grupos obtidos a partir de uma partição do espaço de características com respeito a um número fixo de grupos
- **Baseados em densidades**: não assumem um formato dos grupos, mas que há grupos densos separados por regiões de baixa densidade (ex: DBSCAN, OPTICS, DENCLUE)
- **Baseados em modelos** matemáticos/estatísticos: otimizam o ajuste dos grupos aos modelos (ex: mclust (R))
- **Baseados em grafos** (agrupamento **spectral**): grafo de similaridade dos objetos; redução de dimensionalidade baseado nos autovalores da matriz de similaridade
- **Baseados em redes neurais**: procuram aprender um protótipo de cada classe – cada objeto é classificado de acordo com o protótipo mais próximo ou similar
- **Baseados em grid**: dividem o espaço em um grid (multiresolução) e aplica operações sobre ele (ex: STING, WaveCluster)
- **Baseados em restrições**: encontram grupos que satisfazem restrições definidas pelo usuário ou pela aplicação. Ex: tamanho min/max dos grupos, nr min/max de grupos, objetos com determinadas características devem ser agrupados em um mesmo grupo, diferentes pesos/distâncias para determinadas características
- **Outros**



https://scikit-learn.org/stable/modules/clustering.html#spectral-clustering

Tendência de agrupamento

- Na verdade, antes mesmo de executar um algoritmo de agrupamento, é importante avaliar a tendência de agrupamento dos dados (se são agrupáveis)
- Estatísticas para medir tendência
  - Ex: Hopkins - quanto mais próximo de 1 mais agrupável é o conjunto

Validação

- Como avaliar os resultados?

Validação

- Como avaliar os resultados?

- Uma simples e possível alternativa é replicação
  - Testa com vários subconjuntos
  - Testa com vários algoritmos de agrupamentos
  - Espera-se obter uma certa concordância de grupos

Validação

A estrutura C (partição dos pontos) resultante de um algoritmo pode ser avaliada por três tipos de critérios:

- **Critérios externos**: avaliação de aspectos anteriores à execução de agrupamentos. Ex: o que sabe-se sobre a estrutura do problema? C confirma essa estrutura?
- **Critérios internos**: compara medidas baseadas nos dados. Ex: comparar C com a matriz de proximidade dos objetos
- **Critérios relativos**: comparar diferentes agrupamentos (mesmo algoritmo com diferentes parâmetros ou diferentes algoritmos)

Pode-se ainda avaliar a validade de um grupo específico.

Tudo isso discutido no cap 16 de (THEODORIDIS & KOUTROUMBAS, 2003): uso de teste de hipóteses

Agrupamento

- É uma técnica de classificação
  - Supervisionada ou não supervisionada?

Agrupamento

- É uma técnica de classificação
  - Supervisionada ou não supervisionada?

Agrupamento

- É uma técnica de classificação
  - Supervisionada ou não supervisionada?
  - Paramétrica ou não paramétrica?

Agrupamento

- É uma técnica de classificação
  - Supervisionada ou não supervisionada?
  - Paramétrica ou não paramétrica?
    - Depende... Vimos nos vídeos 2 e 3 métodos não paramétricos, mas há métodos que se utilizam de famílias específicas de distribuições (métodos paramétrics)

Agrupamento

- É uma técnica de classificação
  - Supervisionada ou não supervisionada?
  - Paramétrica ou não paramétrica?
    - Depende... Vimos nos vídeos 2 e 3 métodos não paramétricos, mas há métodos que se utilizam de famílias específicas de distribuições (métodos paramétricos)

- Bastante utilizado para identificação de *outliers* (HAN & KAMBER, 2006), seção 7.11

Aprendizado semi-supervionado

- Mistura de dados rotulados e não rotulados
- Classificação semi-supervisionada: usar os dados rotulados para rotular os dados não rotulados, e então aplicar aprendizado supervisionado
- Clustering semi-supervisionado: usar os dados rotulados para auxiliar os agrupamentos
- Dissertação de mestrado https://teses.usp.br/teses/disponiveis/55/55134/tde-12102003-140536/publico/Dissertacao_MKS.pdf

Referências

- AHMAD, A.; KHAN, S. S. Survey of the state-of-the-art mixed data clustering algorithms. **IEEE Access** vol 7, p. 31883-31902, 2019
- BAYNE, C. K. *et al*. Monte Carlo comparisons of selected clustering procedures. **Pattern Recognition**, v. 12, p.51-62, 1980
- COSTA, L. F.; CESAR Jr, R. B**. Shape Classification and Analysis: Theory and Practice**. CRC Press, 2009, 2 ed. (Cap. 8.3)
- HAN, J; KAMBER, M. Data Mining: Concepts and Techniques. 2 ed. Elsevier. 2006. Cap 7
- JAIN, A. K.; DUIN, R. P. W.; MAO, J. Statistical Pattern Recognition: A Review. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 22 n. 1 p. 4-37, 2000 (seção 8).
- THEODORIDIS, S.; KOUTROUMBAS, K. Pattern Recognition, 2 ed. Elsevier. 2003. Cap 11 a 16.

Fim do vídeo 4

Outras técnicas, validação e comentários finais