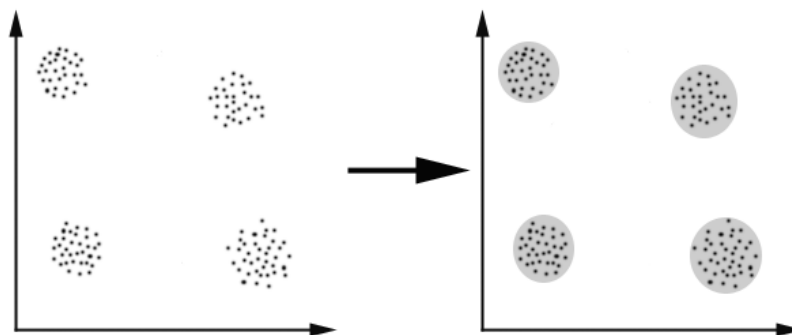


# Clustering

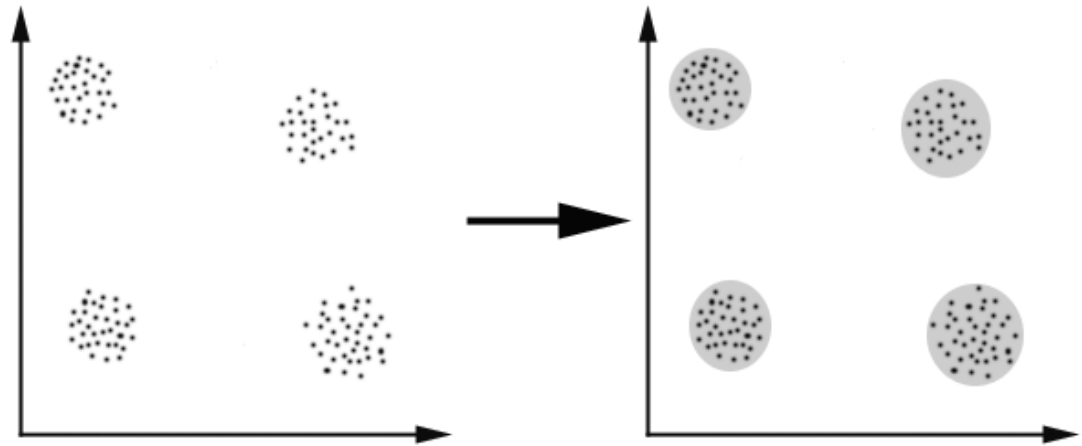
André E. Lazzaretti

UTFPR/CPGEI



# Objetivos

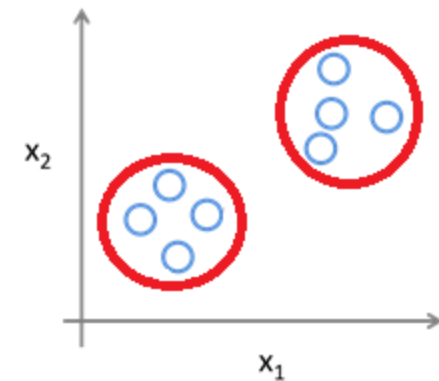
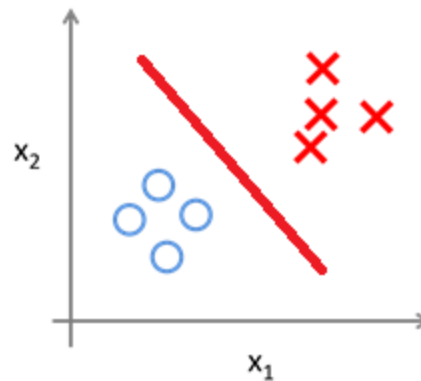
Clustering:



Supervised Learning

Unsupervised Learning

Supervised  
x Unsupervised:



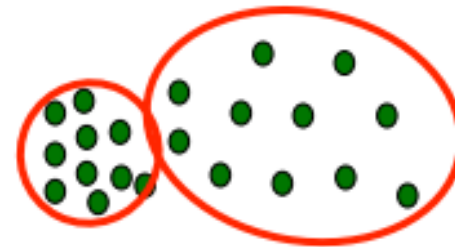
# Definição

- Agrupamento de um conjunto de indivíduos em uma população para representar uma certa estrutura nos dados. Agrupamento de Localização, Forma e Densidade.

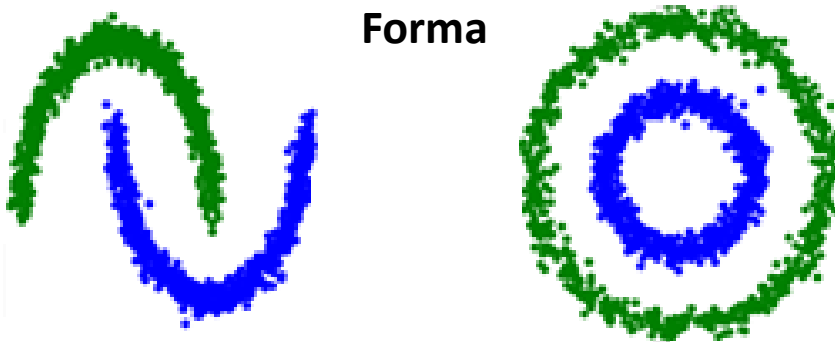
Localização



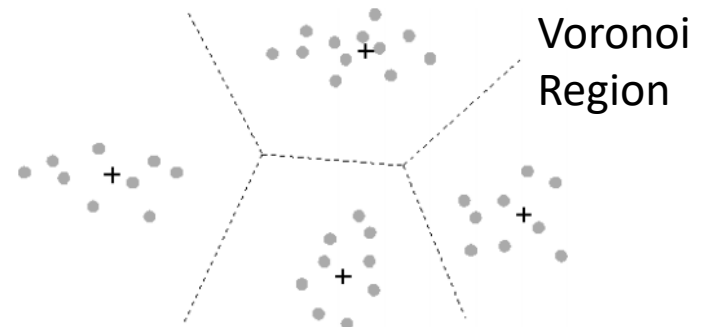
Densidade



Forma



Regiões de Proximidade:



# Métodos

- **Hierarchical methods:** Matriz de dissimilaridade (similaridade);
- **Mixture models:** função densidade de probabilidade;
- **Sum-of-squares methods:** minimização do erro quadrático médio;
- **Spectral clustering:** grafos, mapeamento, similaridade;
- **Cluster validity:** seleção dos modelos.

# Hierárquicos

- Definir uma medida de distância (dissimilaridade) entre os clusters
- Inicialização: todo exemplo é um cluster
- Processo Iterativo:
  - Calcula a distância entre todos os clusters
  - Combina os clusters mais próximos
- Dendrograma

# Hierárquico – Single-Link

	1	2	3	4	5	6
1	0	4	13	24	12	8
2		0	10	22	11	10
3			0	7	3	9
4				0	6	18
5					0	8.5
6						0



	1	2	(3, 5)	4	6
1	0	4	12	24	8
2		0	10	22	10
(3, 5)			0	6	8.5
4				0	18
6					0

At each stage of the algorithm, the closest two groups are fused to form a new group where the distance between two groups,  $A$  and  $B$ , is the distance between their closest members:

$$d_{AB} = \min_{i \in A, j \in B} d_{ij}$$

$$d_{1, (3, 5)} = \min\{d_{13}, d_{15}\} = 12$$

$$d_{2, (3, 5)} = \min\{d_{23}, d_{25}\} = 10$$

$$d_{4, (3, 5)} = 6, d_{6, (3, 5)} = 8.5$$

# Hierárquico – Single-Link

	(1, 2)	(3, 5)	4	6
(1, 2)	0	10	22	8
(3, 5)		0	6	8.5
4			0	18
6				0



	(1, 2)	(3, 4, 5)	6
(1, 2)	0	10	8
(3, 4, 5)		0	8.5
6			0



	(1, 2, 6)	(3, 4, 5)
(1, 2, 6)	0	8.5
(3, 4, 5)		0

$$d_{(1,2)(3,5)} = \min\{d_{13}, d_{23}, d_{15}, d_{25}\} = 10$$

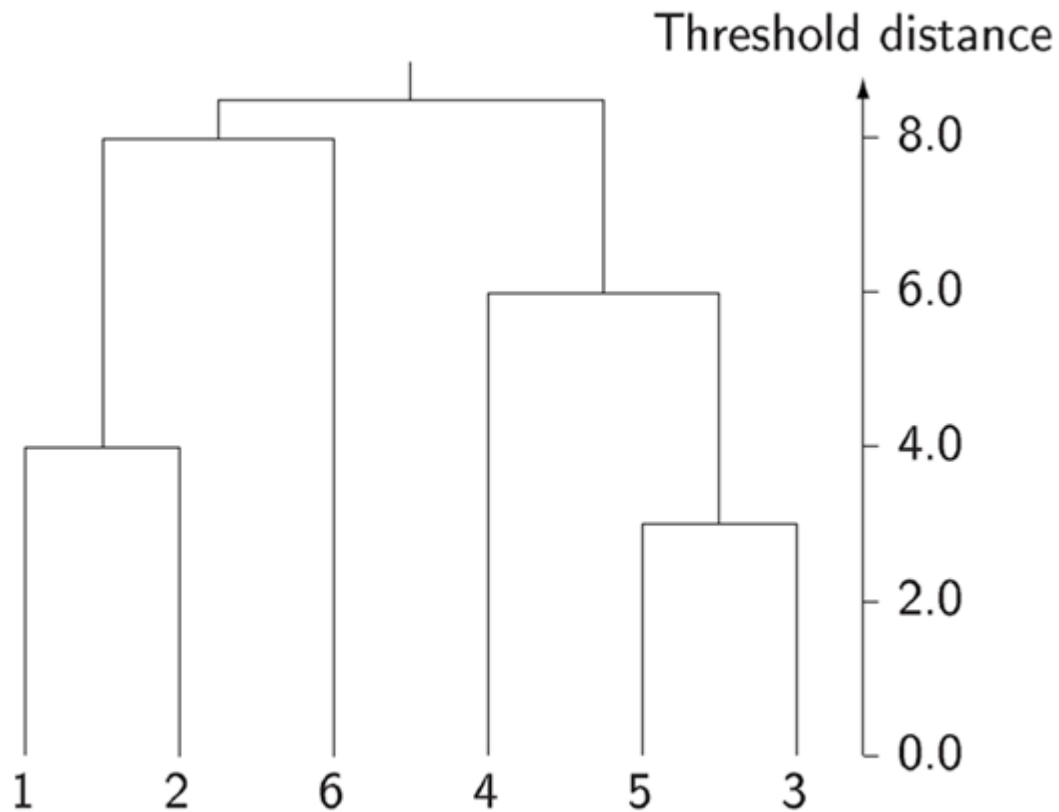
$$d_{(1,2)4} = \min\{d_{14}, d_{24}\} = 22$$

$$d_{(1,2)6} = \min\{d_{16}, d_{26}\} = 8$$

•  
•  
•

# Hierárquico – Single-Link

**Dendograma:**



	1	2	3	4	5	6
1	0	4	13	24	12	8
2		0	10	22	11	10
3			0	7	3	9
4				0	6	18
5					0	8.5
6						0

	1	2	(3, 5)	4	6
1	0	4	12	24	8
2		0	10	22	10
(3, 5)			0	6	8.5
4				0	18
6					0

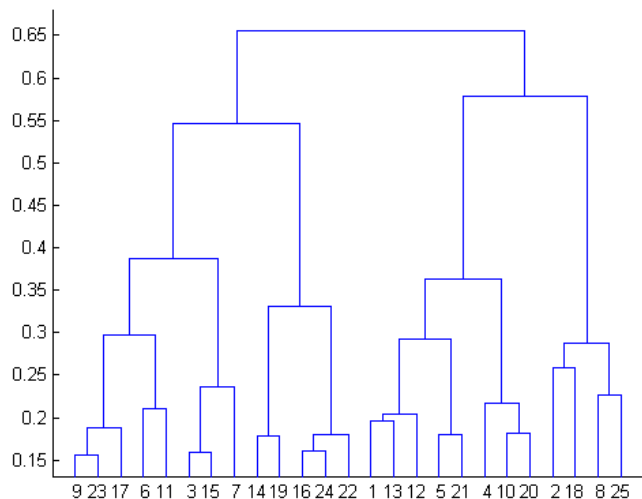
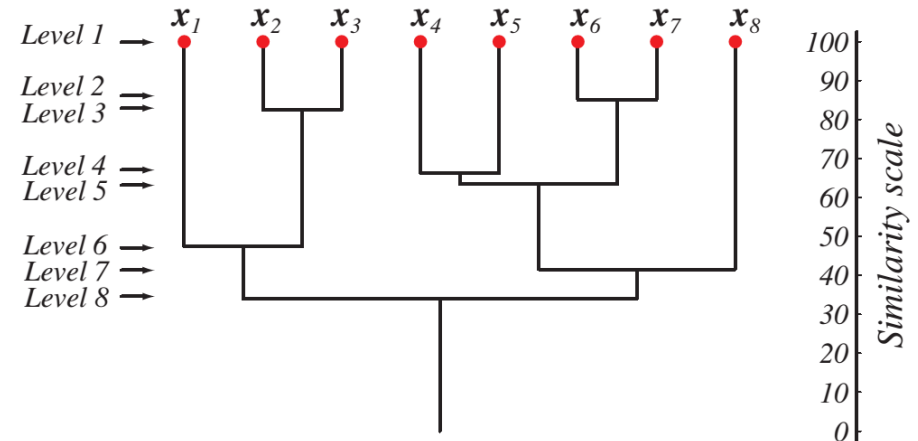
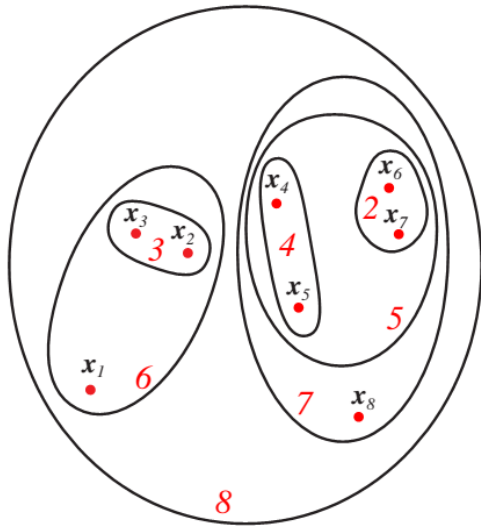
	(1, 2)	(3, 5)	4	6
(1, 2)	0	10	22	8
(3, 5)		0	6	8.5
4			0	18
6				0

	(1, 2)	(3, 4, 5)	6
(1, 2)	0	10	8
(3, 4, 5)		0	8.5
6			0

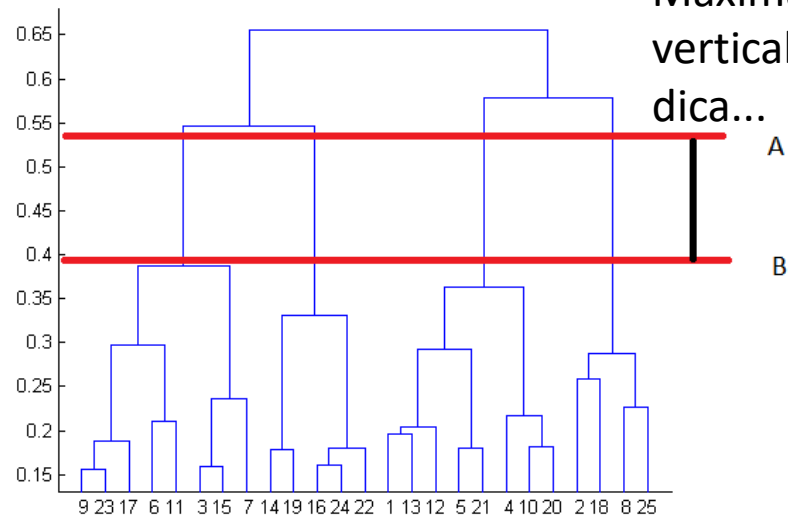
	(1, 2, 6)	(3, 4, 5)
(1, 2, 6)	0	8.5
(3, 4, 5)		0



# Hierárquico – Single-Link



Onde “cortar”?



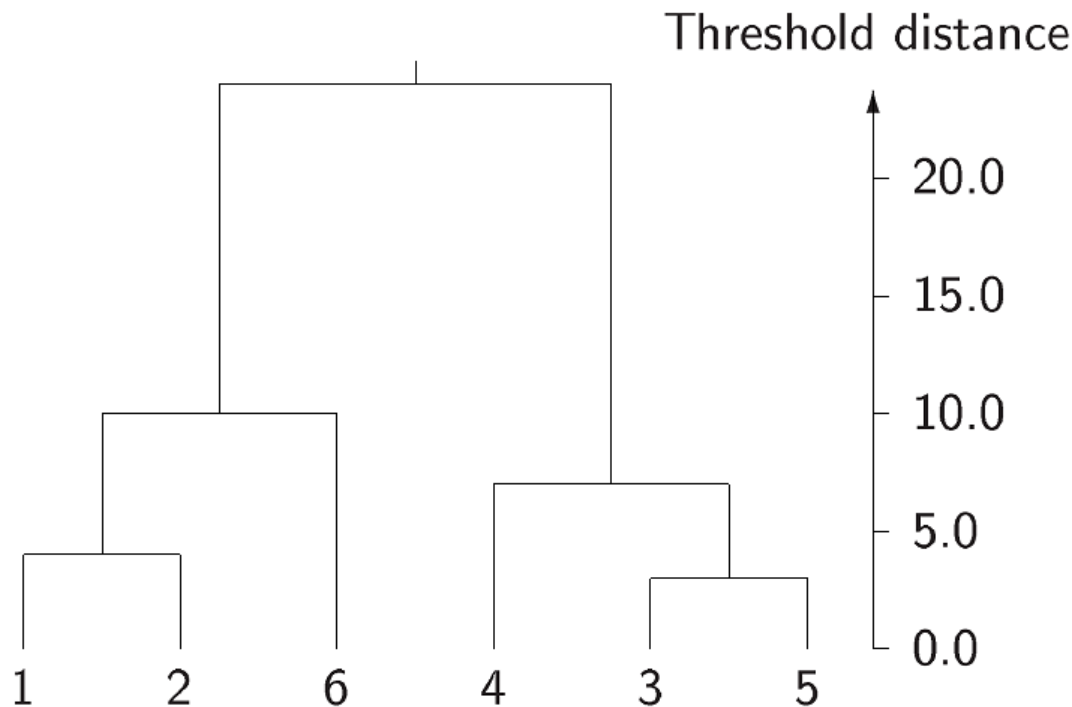
Máxima distância vertical é uma dica...

A

B

# Hierárquico – Complete-Link

**Dendograma:**



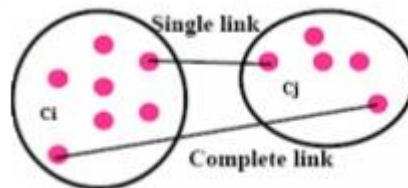
	1	2	3	4	5	6
1	0	4	13	24	12	8
2		0	10	22	11	10
3			0	7	3	9
4				0	6	18
5					0	8.5
6						0



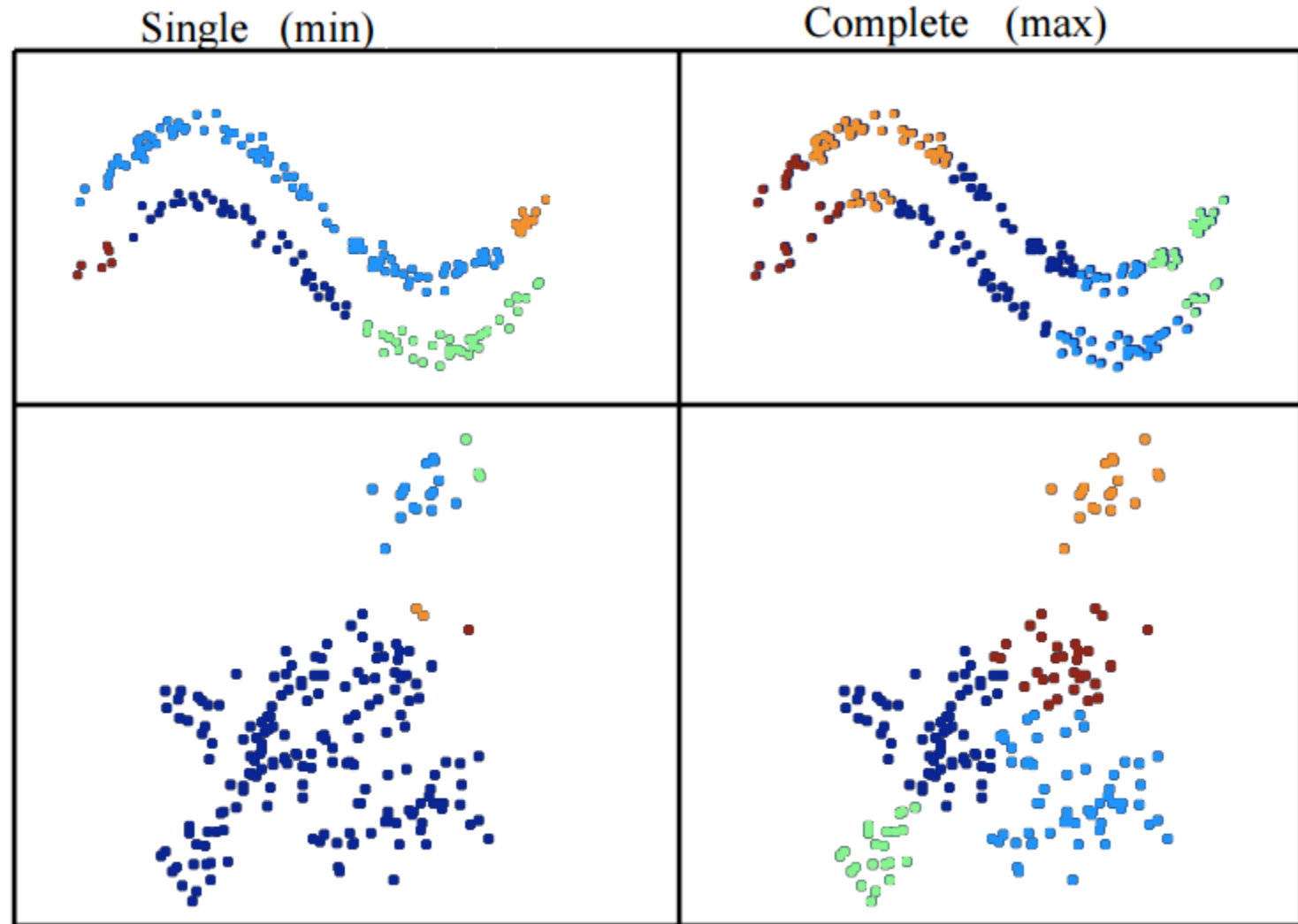
$$d_{AB} = \max_{i \in A, j \in B} d_{ij}$$

	1	2	(3, 5)	4	6
1	0	4	13	24	8
2		0	11	22	10
(3, 5)			0	7	9
4				0	18
6					0

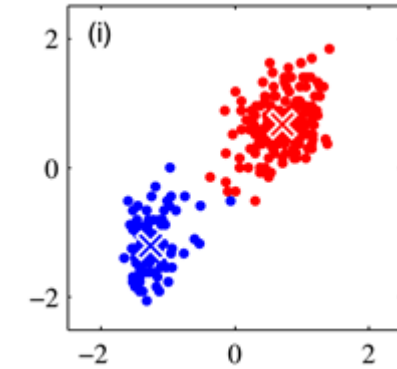
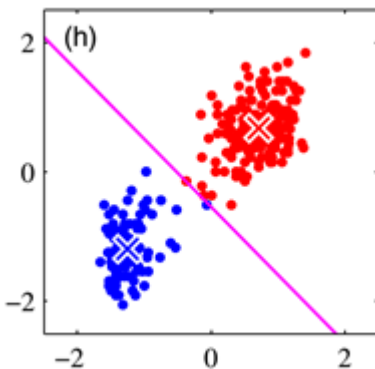
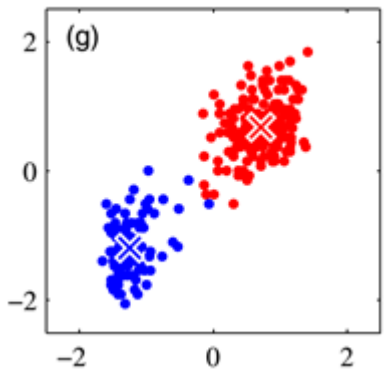
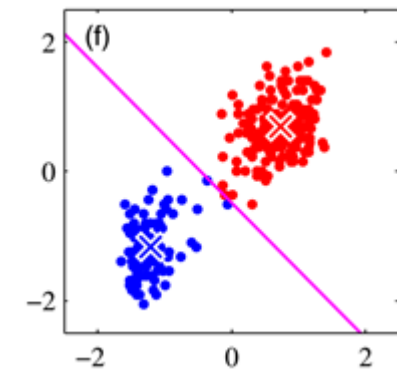
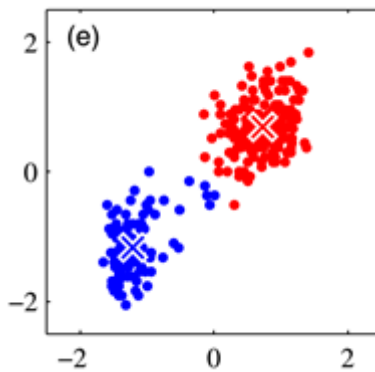
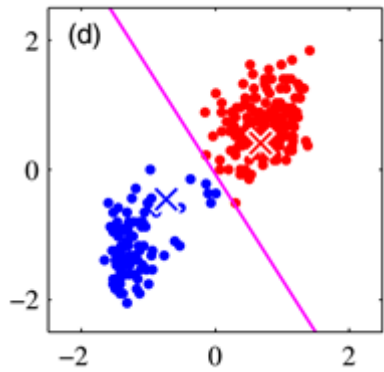
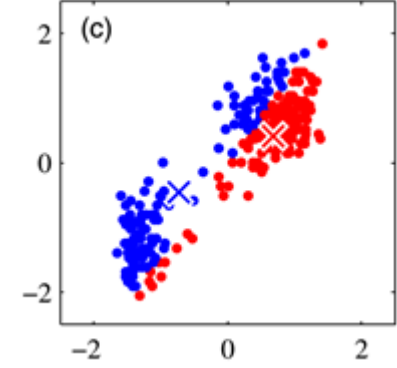
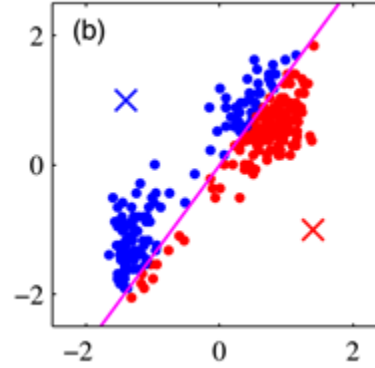
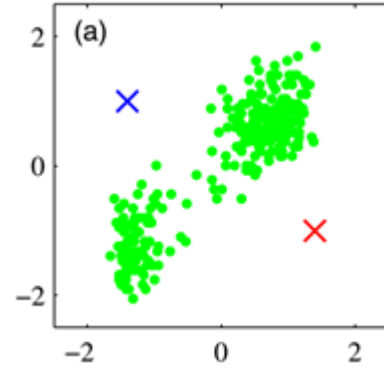
**Diferença em relação  
ao single-link:**



# Hierárquico – Complete x Single



# Sum-of-Squares: K-Means



# Sum-of-Squares: K-Means

- Objetivo:

$$J = \sum_{n=1}^N \sum_{k=1}^K r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2 \longrightarrow J(r_{nk}, \boldsymbol{\mu}_k)$$

- Alterna duas etapas (EM):

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg \min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

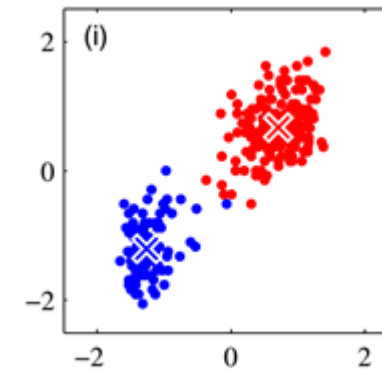
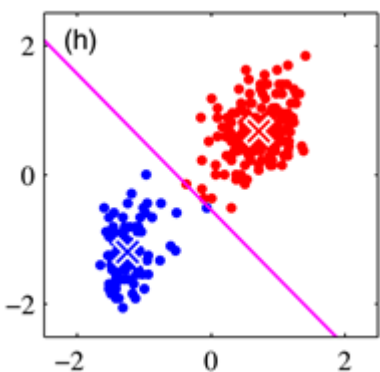
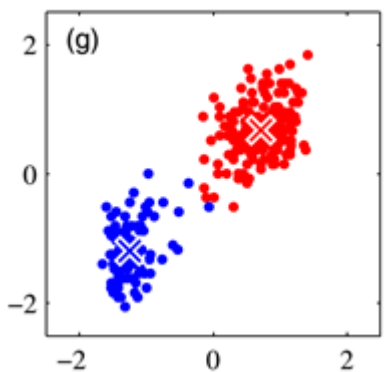
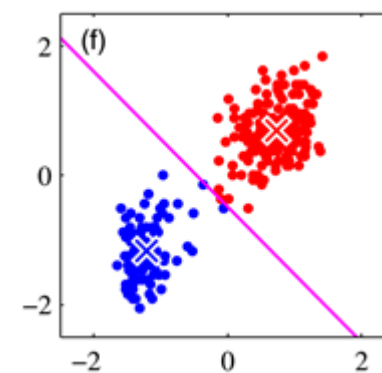
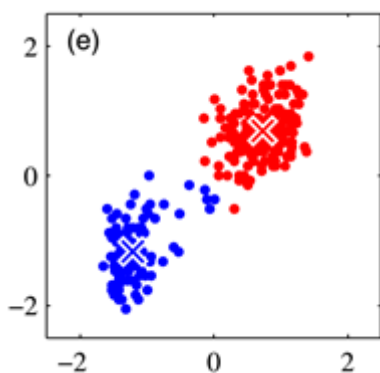
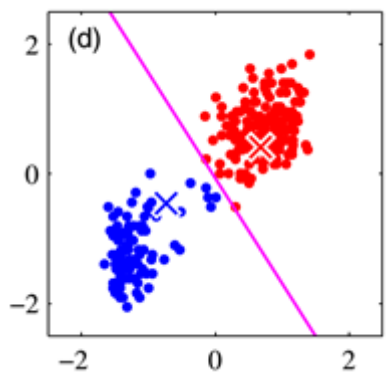
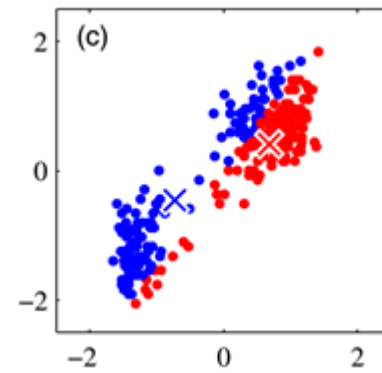
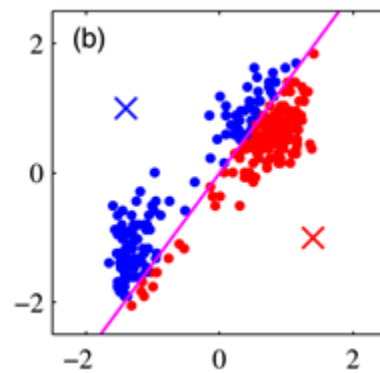
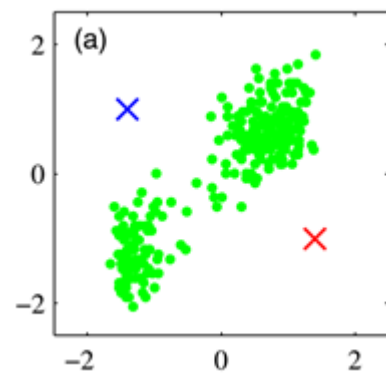
Cluster mais próximo  
(fixa  $\boldsymbol{\mu}_k$ )

$$2 \sum_{n=1}^N r_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k) = 0 \longrightarrow \boldsymbol{\mu}_k = \frac{\sum_n r_{nk} \mathbf{x}_n}{\sum_n r_{nk}}$$

Atualiza o Centróide  
(fixa  $r_{nk}$ )

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg \min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk} \mathbf{x}_n}{\sum_n r_{nk}}$$



# Fuzzy c-Means

- Objetivo:

$$J_r = \sum_{i=1}^n \sum_{j=1}^g y_{ji}^r \| \mathbf{x}_i - \mathbf{m}_j \|^2 \quad \text{s.t.} \quad \sum_{j=1}^g y_{ij} = 1 \quad \text{e} \quad y_{ji} \geq 0$$

- Alterna duas etapas (EM):

$$y_{ji} = \frac{1}{\sum_{k=1}^g \left( \frac{d_{ij}}{d_{ik}} \right)^{\frac{2}{r-1}}}$$

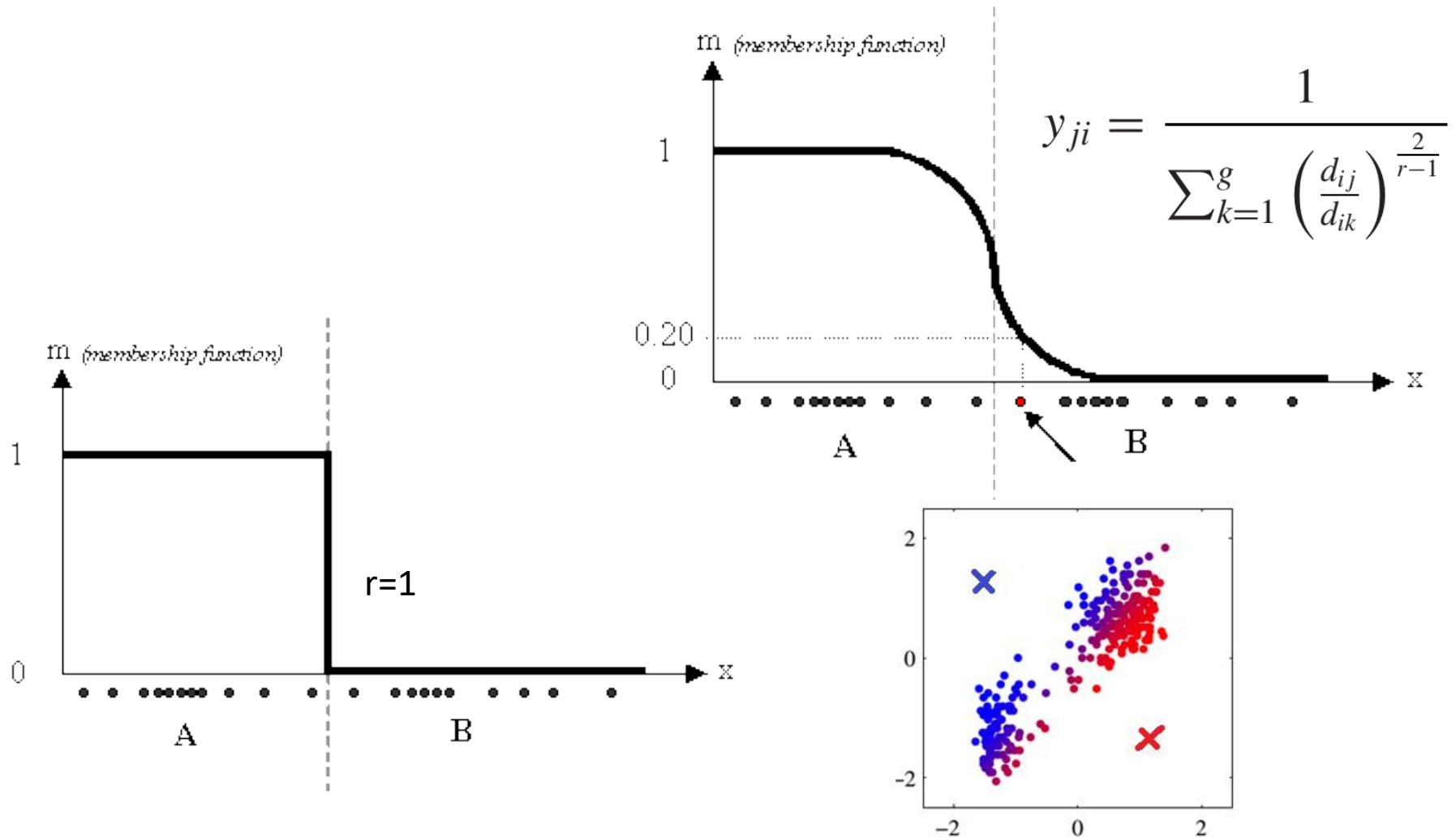
Atualizar graus de pertinência

$$\mathbf{m}_j = \frac{\sum_{i=1}^n y_{ji}^r \mathbf{x}_i}{\sum_{i=1}^n y_{ji}^r}$$

Atualiza o Centróide

# Fuzzy c-Means

- K-Means x C-Means:



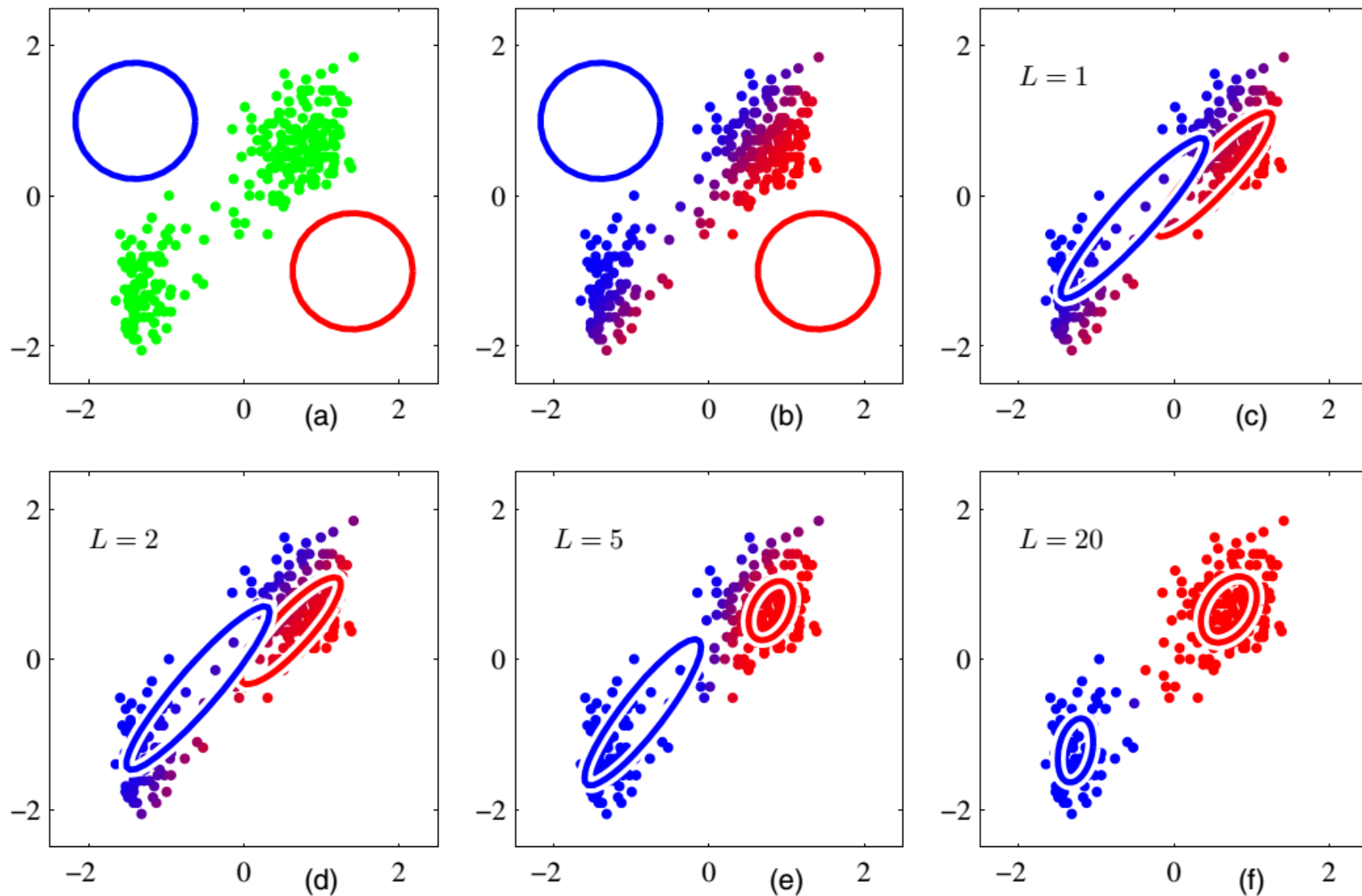


# Mixture of Densities: Maximização da Expectativa (EM)

**E-Step:**  $\left\{ \begin{aligned} \gamma(z_{nk}) &= \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \end{aligned} \right.$

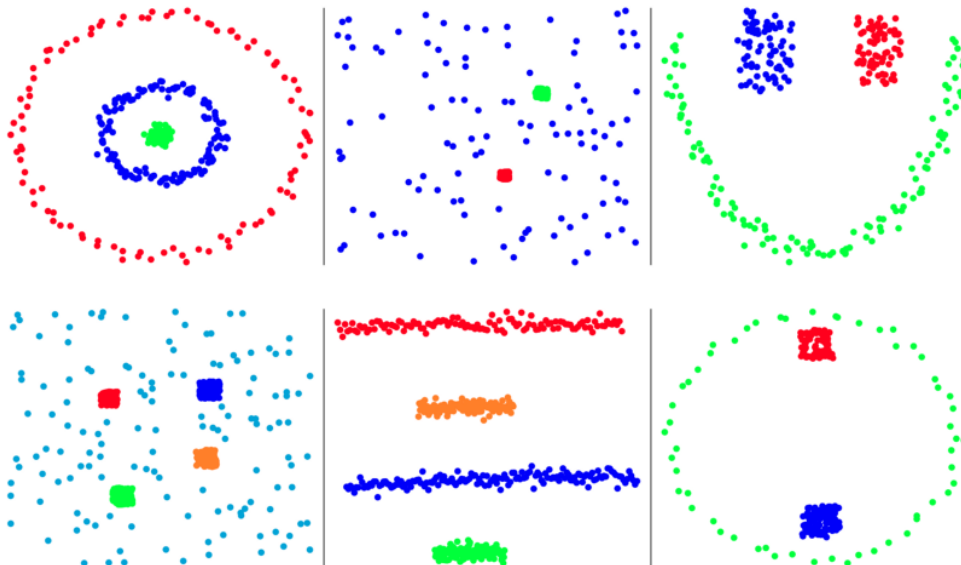
**M-Step:**  $\left\{ \begin{aligned} \boldsymbol{\mu}_k^{\text{new}} &= \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}_n \\ \boldsymbol{\Sigma}_k^{\text{new}} &= \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}) (\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}})^T \\ \pi_k^{\text{new}} &= \frac{N_k}{N} \\ N_k &= \sum_{n=1}^N \gamma(z_{nk}) \end{aligned} \right.$

# Maximização da Expectativa (EM)



# Spectral Clustering

- Os modelos vistos até aqui assumem uma estrutura pré-definida para os grupos:
  - *K-Means*: esférico (distância Euclidiana)
  - Hierárquico: esférico (distância Euclidiana)
  - MoG: esférico ou elíptico (matriz de covariância)
- E nos casos a seguir?



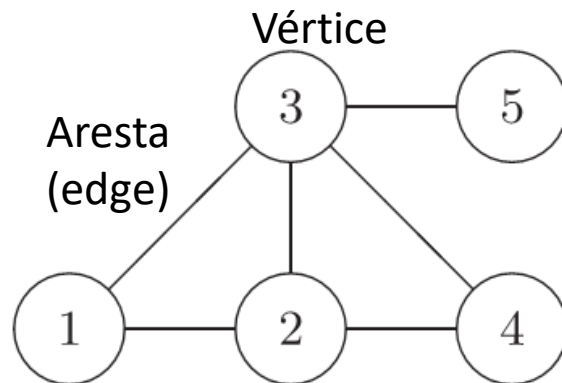
## Spectral Clustering!

- 1) Constrói um grafo baseado na similaridade (vizinhança) dos pontos.
- 2) Realiza mapeamento dos pontos para um espaço onde a representação dos grupos fica evidenciada (*spectral embedding*).
- 3) Realiza o agrupamento neste novo espaço (p.ex. *k-means*).

# Spectral Clustering

- Teoria de Grafos:

$$G = \{V, E\}$$



Matriz de Adjacência (forma mais simples):

$A_{ij} = 1$  if node  $i$  and node  $j$  are connected

$A_{ij} = 0$  if node  $i$  and node  $j$  are not connected

	1	2	3	4	5
1	0	1	1	0	0
2	1	0	1	1	0
3	1	1	0	1	1
4	0	1	1	0	0
5	0	0	1	0	0

A **teoria dos grafos** é um ramo da matemática que estuda as relações entre os objetos de um determinado conjunto.

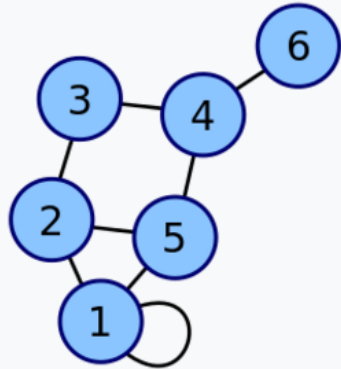


# Spectral Clustering

- A matriz de adjacência possui relação com outras duas matrizes:  $\mathbf{L} = \mathbf{D} - \mathbf{A}$ .
- Sendo  $\mathbf{D}$ , uma diagonal cuja diagonal corresponde a (na sua forma mais simples):

$$d_{i,j} := \begin{cases} \deg(v_i) & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

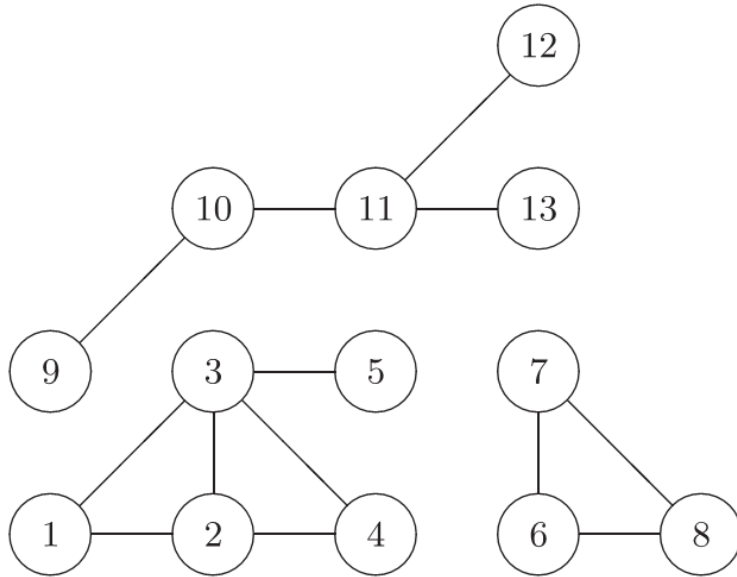
$\deg(v_i)$  corresponde a um grau (número) de conexões de um determinado vértice.

Vertex labeled graph	Degree matrix
	$\begin{pmatrix} 4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$

- E  $\mathbf{L}$  a matriz Laplaciana.

# Spectral Clustering

- A matriz  $L$  possui propriedades que facilitam o processo de agrupamento: decomposição em autovalores e autovetores ( $L\mathbf{v} = \lambda\mathbf{v}$ ) - *spectrum of a graph*.



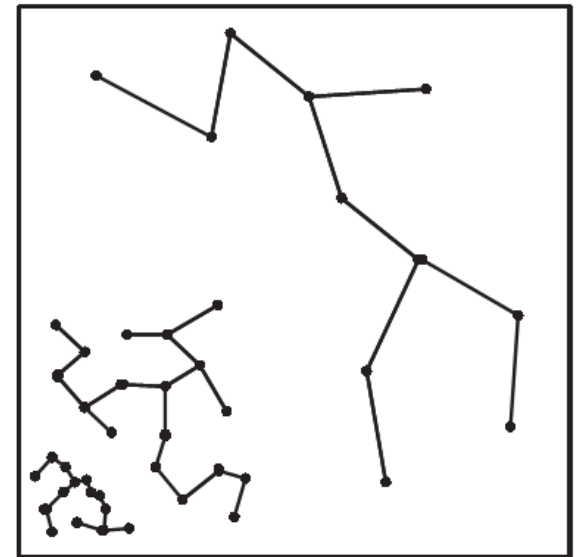
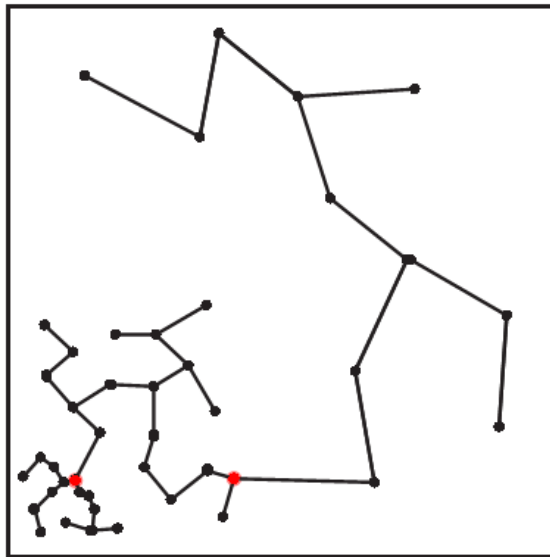
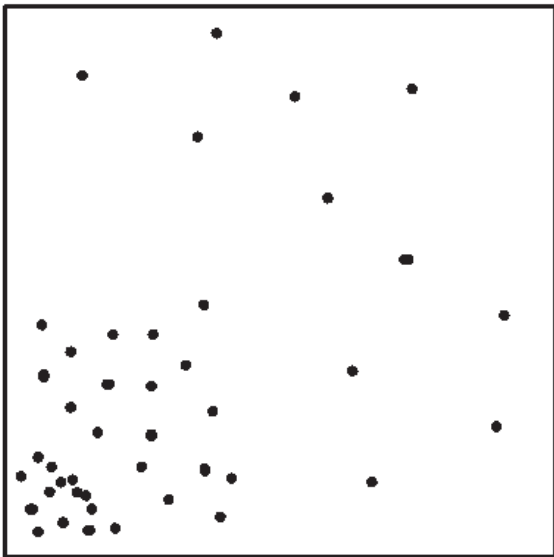
Eigenvectors  $v_1, v_2, v_3$  corresponding to the eigenvalue  $\lambda = 0$

	$v_1$	$v_2$	$v_3$
1	0	0	1
2	0	0	1
3	0	0	1
4	0	0	1
5	0	0	1
6	0	1	0
7	0	1	0
8	0	1	0
9	1	0	0
10	1	0	0
11	1	0	0
12	1	0	0
13	1	0	0

A multiplicidade do autovalor com valor 0 é igual ao número de componentes conectados no grafo. Os autovetores correspondentes fornecem indicadores para mostrar a qual componente pertence um determinado nó.

# Spectral Clustering

- Na prática, ainda é necessário especificar a matriz de adjacência para representar a conectividade:



# Spectral Clustering

- Como definir a matriz de adjacência?
- Medida de similaridade:  $A_{ij} = s(\mathbf{x}_i, \mathbf{x}_j)$
- Conexão:

Fully connected graph

$$s(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2}\right)$$

$\epsilon$ -neighbourhood

$$s(\mathbf{x}, \mathbf{y}) = \begin{cases} 1 & \|\mathbf{x} - \mathbf{y}\| < \epsilon \\ 0 & \|\mathbf{x} - \mathbf{y}\| \geq \epsilon \end{cases}$$

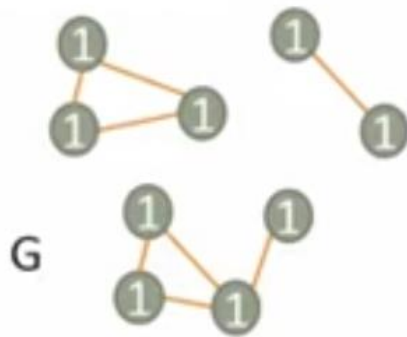
- Matriz Laplaciana (**L**) e a Matriz de Adjacência:  
**L** = **D** – **A**, sendo **D** (definição mais geral):

$$d_i = \sum_{j=1}^n A_{ij}$$



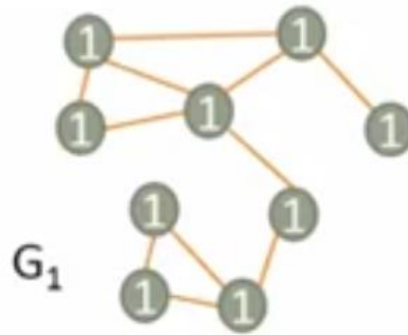
# Spectral Clustering

- Na prática, nem todos os autovalores serão nulos, porém, caso a matriz de adjacência represente a estrutura do cluster, pode-se analisar os autovalores de magnitudes mais próximas de zero:



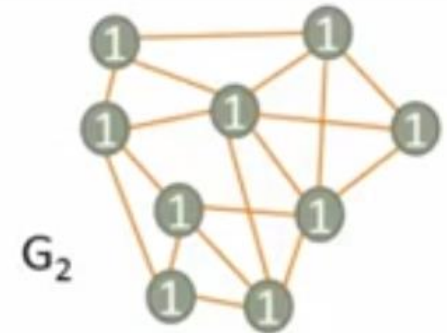
$L_G$  has  $0 = \lambda_1 = \lambda_2 = \lambda_3$

Demais  
Não-nulos



Both  $L_{G_1}$  and  $L_{G_2}$  have  $0 = \lambda_1$  and  $\lambda_2 > 0$ .  $\lambda_2(L_{G_1}) < \lambda_2(L_{G_2})$

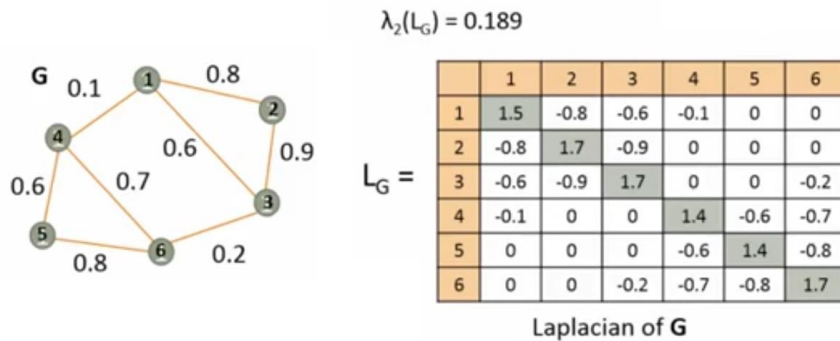
Um único  
componente



“densidade”  
de conexões

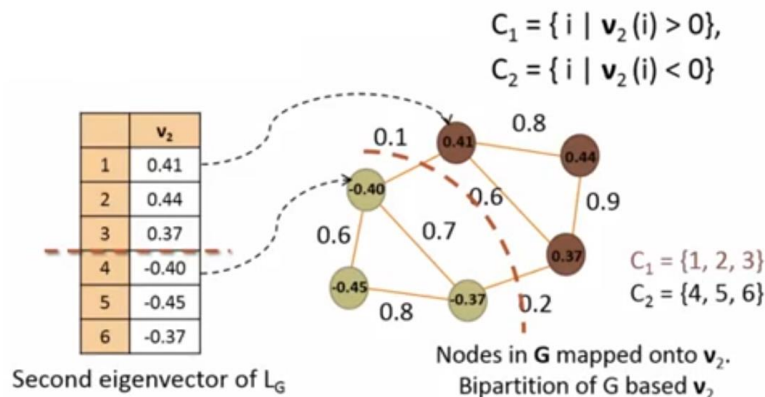
# Spectral Clustering

- Se for um problema de dois clusters, pode-se usar o autovetor  $\mathbf{v}_2$ , correspondente à  $\lambda_2$ :



Autovetor  $\mathbf{v}_2$ , fornece um indicativo mais claro do agrupamento. Está associado ao segundo menor autovalor.

- $L$  deve ser semi-positiva definida.
- Para tanto, é necessária a seguinte normalização:



- $\tilde{L} = D^{-1/2} A D^{-1/2}$

# Spectral Clustering

Given a set of points  $S = \{s_1, \dots, s_n\}$  in  $\mathbb{R}^l$  that we want to cluster into  $k$  subsets:

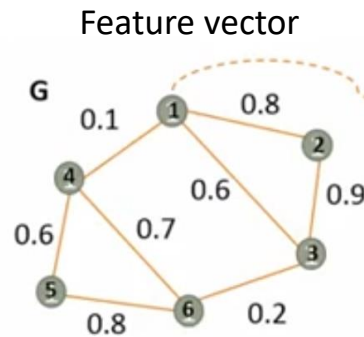
1. Form the affinity matrix  $A \in \mathbb{R}^{n \times n}$  defined by  $A_{ij} = \exp(-\|s_i - s_j\|^2 / 2\sigma^2)$  if  $i \neq j$ , and  $A_{ii} = 0$ .
2. Define  $D$  to be the diagonal matrix whose  $(i, i)$ -element is the sum of  $A$ 's  $i$ -th row, and construct the matrix  $L = D^{-1/2} A D^{-1/2}$ .
3. Find  $x_1, x_2, \dots, x_k$ , the ~~largest~~ <sup>first</sup>  $k$  eigenvectors of  $L$  (chosen to be orthogonal to each other in the case of repeated eigenvalues), and form the matrix  $X = [x_1 x_2 \dots x_k] \in \mathbb{R}^{n \times k}$  by stacking the eigenvectors in columns.
4. Form the matrix  $Y$  from  $X$  by renormalizing each of  $X$ 's rows to have unit length (i.e.  $Y_{ij} = X_{ij} / (\sum_j X_{ij}^2)^{1/2}$ ).
5. Treating each row of  $Y$  as a point in  $\mathbb{R}^k$ , cluster them into  $k$  clusters via K-means or any other algorithm (that attempts to minimize distortion).
6. Finally, assign the original point  $s_i$  to cluster  $j$  if and only if row  $i$  of the matrix  $Y$  was assigned to cluster  $j$ .

Para mais clusters (Ng, Jordan, Weiss):

**Exemplos Matlab!**

	1	2	3	4	5	6
1	1.0	-0.5	-0.4	-0.1	0	0
2	-0.5	1.0	-0.5	0	0	0
3	-0.4	-0.5	1.0	0	0	-0.1
4	-0.1	0	0	1.0	-0.4	-0.5
5	0	0	0	-0.4	1.0	-0.5
6	0	0	-0.1	-0.5	-0.5	1.0

$L_{\text{norm}}(G)$



	$v_1$	$v_2$	$v_3$
1	$v_1(1)$	$v_2(1)$	$v_3(1)$
2	$v_1(2)$	$v_2(2)$	$v_3(2)$
3	$v_1(3)$	$v_2(3)$	$v_3(3)$
4	$v_1(4)$	$v_2(4)$	$v_3(4)$
5	$v_1(5)$	$v_2(5)$	$v_3(5)$
6	$v_1(6)$	$v_2(6)$	$v_3(6)$

$U$  for  $k = 3$

# Métricas de Desempenho

- **Ideia geral:** como definir uma métrica de comparação de desempenho de diferentes resultados de agrupamento (assumindo que os dados possam ser agrupados)?
- Métricas de Avaliação:
  - Critério Externo – correspondência em relação a uma estrutura pré-definida (p.ex. *labels* de classes);
  - Critério Interno – compactação e separação de grupos.

# Validação Externa

- Supondo uma distribuição de clusters e uma distribuição pré-especificada:  $C=\{C_1,\dots,C_m\}$  e  $P=\{P_1,\dots,P_s\}$
- Para um dado par de vetores  $(\mathbf{x}_v, \mathbf{x}_u)$ :
  - SS: ambos os exemplos pertencem ao mesmo grupo  $P$  e cluster  $C$  - definido como  $a$
  - SD: ambos os exemplos pertencem ao mesmo cluster  $C$  e diferentes grupos  $P$  - definido como  $b$
  - DS: ambos os exemplos pertencem ao mesmo grupo  $P$  e diferentes cluster  $C$  - definido como  $c$
  - DD: ambos os exemplos pertencem a diferentes grupos  $P$  e cluster  $C$  – definido como  $d$
- $M$  é definido como número total de exemplos na base.

# Validação Externa

- Exemplo:

$$X = \{\mathbf{x}_i, i = 1, \dots, 6\}$$

$$\mathcal{C} = \{\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}, \{\mathbf{x}_4, \mathbf{x}_5\}, \{\mathbf{x}_6\}\}$$

$$\mathcal{P} = \{\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}, \{\mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}\}$$

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$
$\mathbf{x}_1$		SS	SS	DD	DD	DD
$\mathbf{x}_2$			SS	DD	DD	DD
$\mathbf{x}_3$				DD	DD	DD
$\mathbf{x}_4$					SS	DS
$\mathbf{x}_5$						DS
$\mathbf{x}_6$						

$$a = 4, b = 0, c = 2, \text{ and } d = 9.$$

SS      SD      DS      DD

- SS: ambos os exemplos pertencem ao mesmo grupo P e cluster C
- SD: ambos os exemplos pertencem ao mesmo cluster C e diferentes grupos P
- DS: ambos os exemplos pertencem ao mesmo grupo P e diferentes cluster C
- DD: ambos os exemplos pertencem a diferentes grupos P e cluster C

# Validação Externa

- Rand:  $R = (a + d)/M$
- Jaccard:  $J = a/(a + b + c)$
- Fowlkes e Mallows:

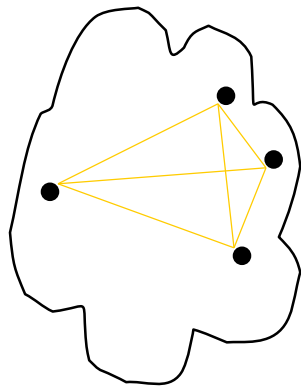
$$FM = a/\sqrt{m_1 m_2} = \sqrt{\frac{a}{a+b} \frac{a}{a+c}}$$

$$m_1 = a + b$$

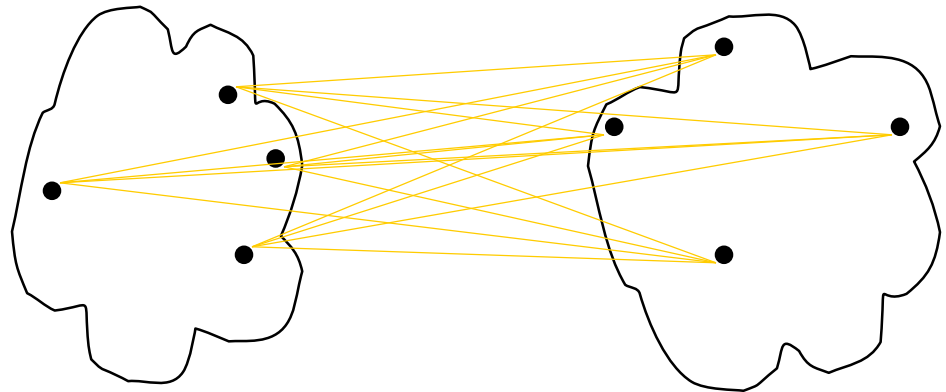
$$m_2 = a + c$$

# Validação Interna

- **Coesão:** relação entre os exemplos em um mesmo cluster. Pode ser medida pela distância entre os exemplos e o centro do cluster ou entre os exemplos de um mesmo cluster.
- **Separabilidade:** Mede a separação entre os diversos clusters de um agrupamento. Pode ser medida pela distância entre os centros dos clusters. Ou entre os pares de exemplos de diferentes clusters.



Coesão

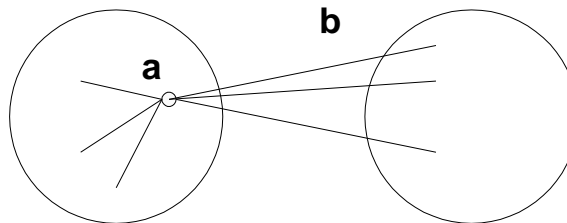


Separabilidade



# Coeficiente de Silhueta

- Combina ideias de coesão e separabilidade
- Para um exemplo  $i$ 
  - Calcula  $a$  = distância média entre  $i$  e os demais exemplos do mesmo cluster
  - Calcula  $b$  = min (distância média entre  $i$  e os demais exemplos dos demais clusters)
  - Calcula o índice:  $s = 1 - a/b$
  - Tipicamente entre 0 e 1.
  - Quanto mais próximo de 1, melhor.



# Referências

- Livro Andrew Webb (Statistical Pattern Recognition) – Capítulo 10
- Artigo Ben Hur – Support Vector Clustering
- Aula de Clustering do Prof. Omar Sobh (Univ. Illinois) – disponíveis no Youtube.