

# Aula 8 - Reconhecimento e Interpretação



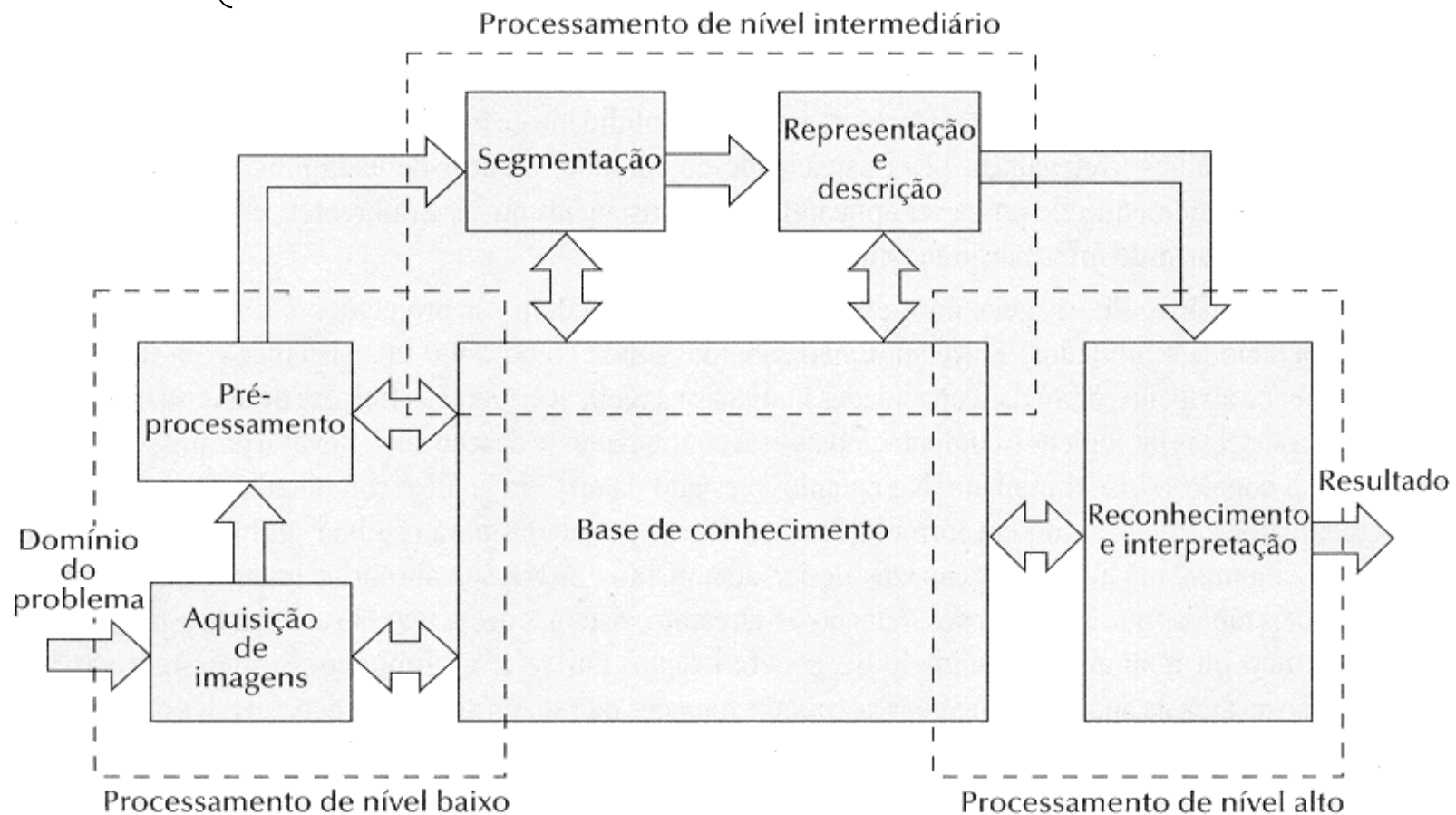
Prof. Adilson Gonzaga

# Elementos de Visão Computacional:

## Visão Computacional

Processamento de Baixo Nível  
Processamento de Nível Intermediário  
Processamento de Alto Nível

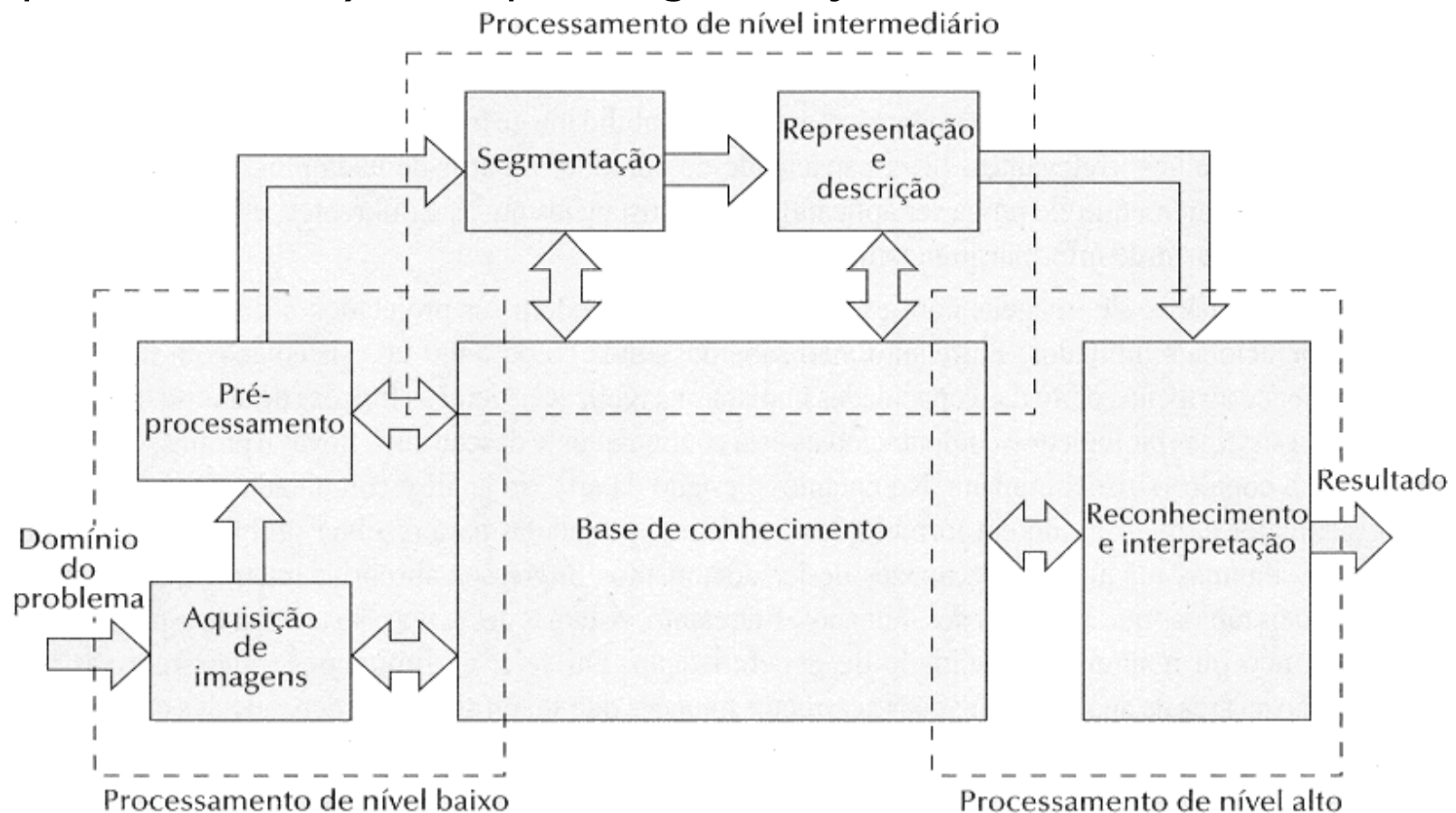
Processamento de Imagens



## Elementos de Visão Computacional:

As linhas tracejadas mostram que a divisão não é rígida:

Ex: **Limiarização** - pode ser usada tanto para melhoramento da Imagem (pré-processamento) como para segmentação.



## Padrões:

- ❑ Um **Padrão** é uma descrição quantitativa ou estrutural de um objeto ou alguma outra entidade de interesse em uma Imagem.
- ❑ Um Padrão é formado por um ou mais **descritores**. Os descritores são também chamados de características.
- ❑ O ato de gerar os descritores que caracterizam um objeto ou partes de uma imagem é chamado de **Extração de Características**.
- ❑ Uma **Classe de Padrões** é uma família de padrões que compartilham algumas propriedades comuns e são denotadas como  $w_1, w_2, w_3, \dots, w_M$  onde  $M$  é o número de classes.

## Arranjos de Padrões:

- Vetores (descrições quantitativas)
- Cadeias e Árvores (descrições estruturais)

## Vetores de Características:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$x_i$  é o  $i$ -ésimo descritor

$n$  é o número de descritores ou características

## Exemplo\_1:

Descrever três tipos de flores:  
( Iris setosa, virginica e versicolor)



3 classes  
 $W_1, W_2, W_3$

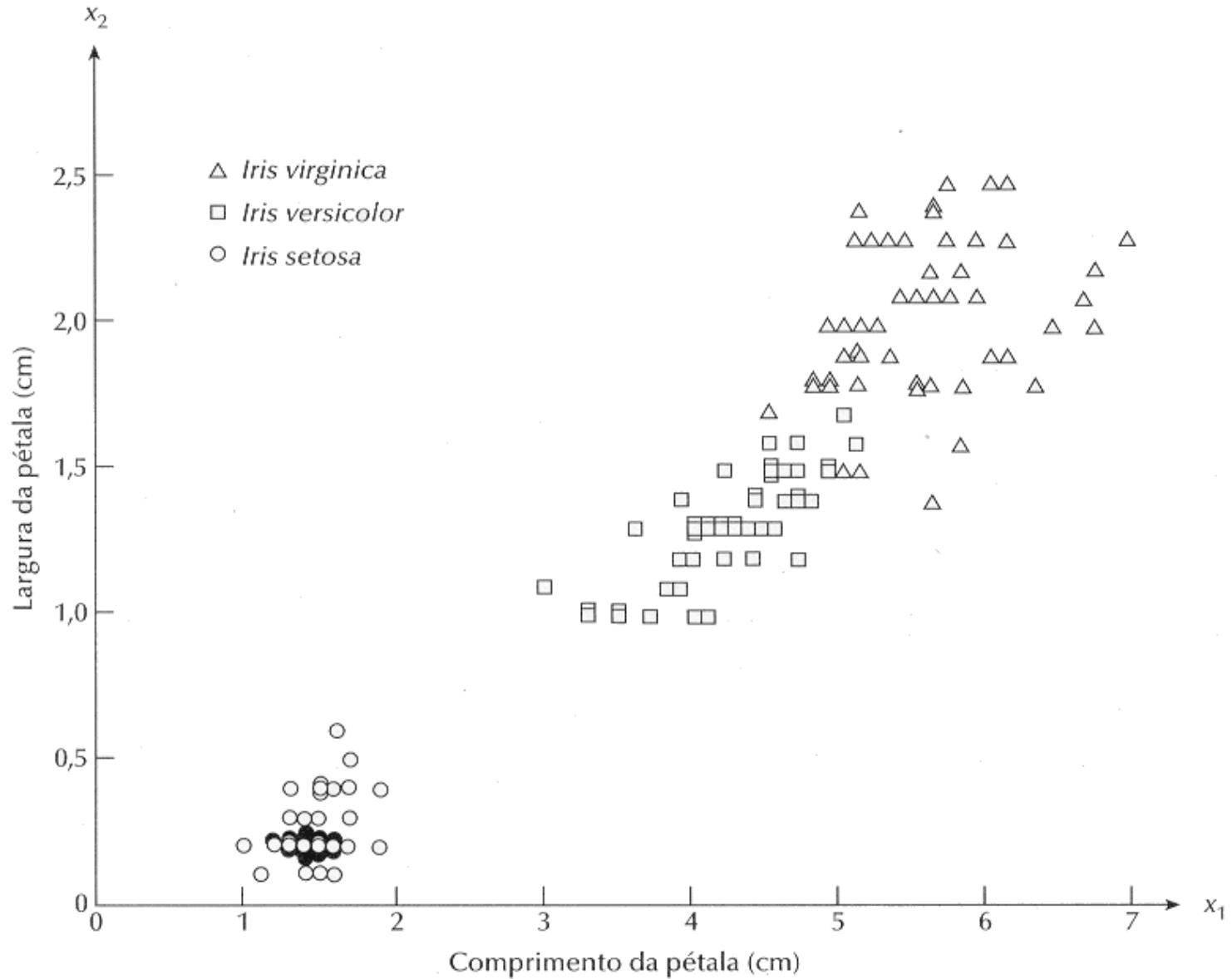


Características a serem utilizadas:  
(**largura e comprimento** de suas pétalas)

□ Uma vez que um conjunto de medidas tenha sido selecionado, um vetor de características torna-se a representação completa de cada amostra física.

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Cada flor do conjunto de amostras de flores, é um ponto no espaço euclidiano bi-dimensional.

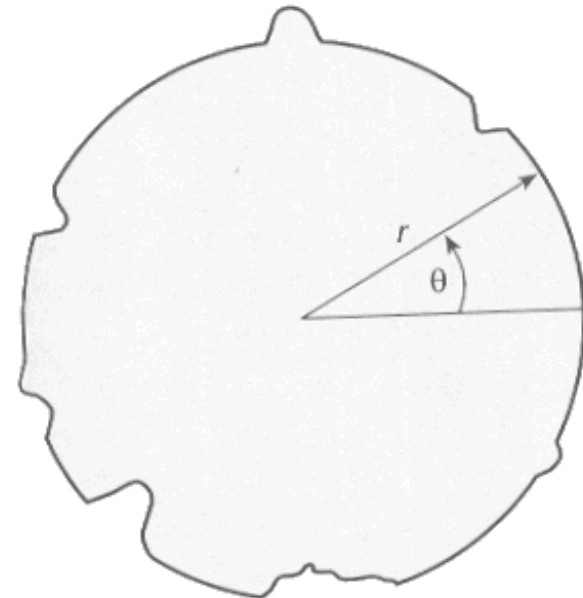


## Seleção das Características:

- ❑ No exemplo anterior, as características “comprimento e largura” das pétalas permitiram separar bem apenas a classe das “Iris Setosa”.
- ❑ A **Seleção das Características** que geram o Vetor de Características, possui uma influência profunda no desempenho de um sistema de Visão Computacional.

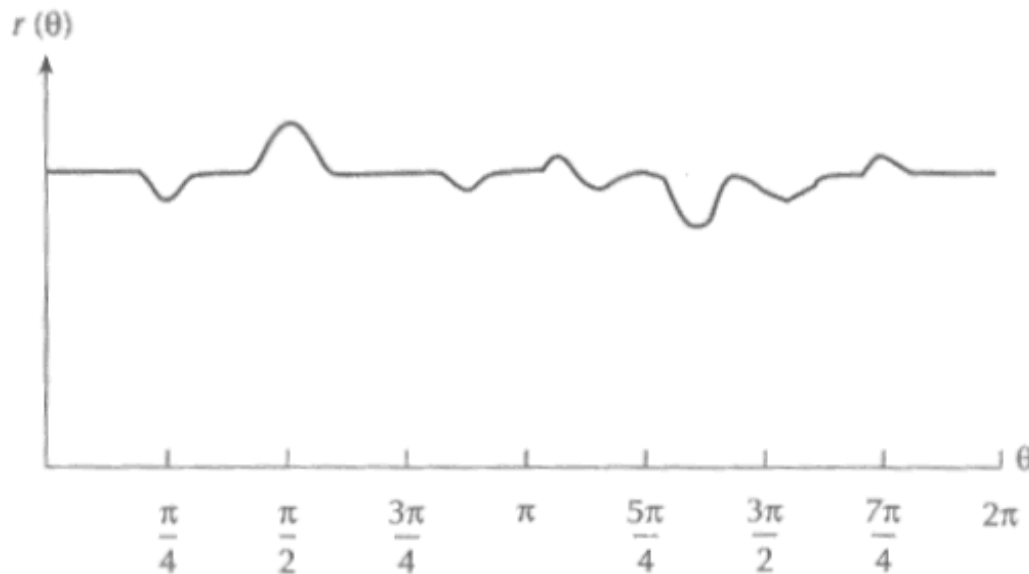
### Exemplo\_2:

Escolher as características para o Vetor de Características, visando classificar diversas formas ruidosas quase-circulares. (Peças com desgastes ou defeitos)





Uma solução seria utilizar a descrição por Assinatura:



A cada intervalo  $\Delta\theta$  dado por  $\theta_1, \theta_2, \dots, \theta_n$

Gerar os Vetores de Características:

$$x_1 = r(\theta_1)$$

$$x_2 = r(\theta_2)$$

$\vdots$

$$x_n = r(\theta_n)$$

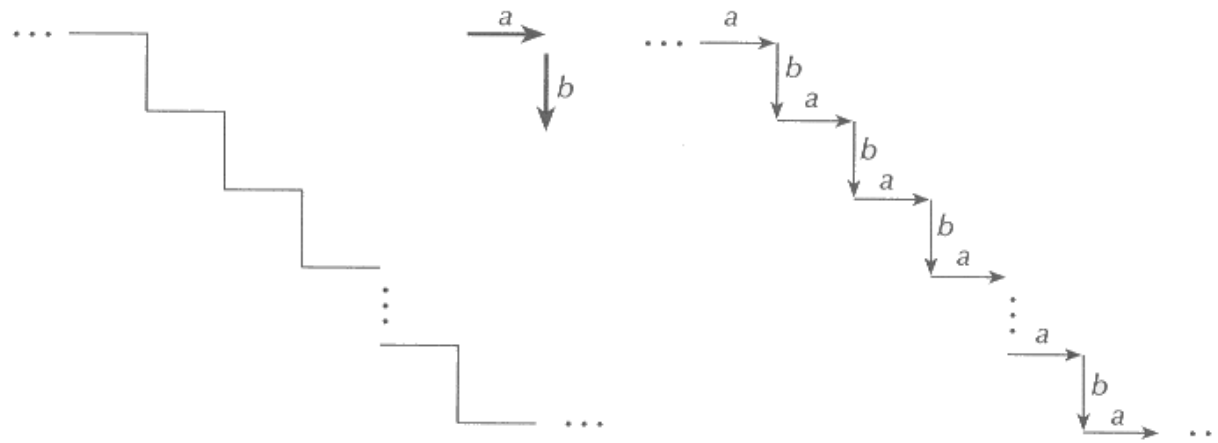
Esses vetores tornam-se pontos no espaço n-dimensional, e as Classes de Padrões "nuvens" de n dimensões.

- ❑ Outra solução seria calcular os Momentos de cada peça e gerar o Vetor de Características com os  $m$  primeiros momentos.
- ❑ As Características a serem selecionadas podem gerar o Vetor de Características através de diversas técnicas como por exemplo:
  - Momentos
  - Número do Formato
  - Descritores Topológicos
  - outros....
- ❑ Geralmente a utilização de características geradas por diferentes metodologias, tornam o Reconhecimento facilitado.

## Descrições Estruturais:

- ❑ Problemas de reconhecimento em que não apenas as medidas quantitativas sobre cada característica, mas também as relações espaciais entre as características determinam as classes, são geralmente melhor resolvidos por **abordagens estruturais**.
- ❑ Representações por **Cadeias** geram padrões cujas estruturas baseiam-se em conectividade de primitivas, usualmente associadas com a forma da fronteira.

Exemplo:



Uma escada pode ser descrita como uma cadeia de símbolos:

$w = \dots abababab\dots$

## Métodos de Decisão:

Funções de Decisão ou Funções Discriminantes.

Seja  $x = (x_1, x_2, \dots, x_n)^T$  um Vetor de Características n-dimensional e  $w_1, w_2, \dots, w_M$  M Classes de Padrões.

O Reconhecimento de Padrões consiste em encontrar as M funções de decisão  $d_1(x), d_2(x), \dots, d_M(x)$  tal que:

- Se o padrão  $x$  pertencer à classe  $w_i$ , então:

$$d_i(x) > d_j(x) \quad j = 1, 2, \dots, M; j \neq i$$

Ou seja:

$$x \in w_i \quad \text{se } d_i(x) \text{ é o maior valor}$$

## Fronteira de Decisão:

A Fronteira que separa duas classes  $w_i$  e  $w_j$  é dada pelos valores de  $x$  para os quais  $d_i(x) = d_j(x)$ , ou seja:

$$d_i(x) - d_j(x) = 0$$

Pode-se identificar a Fronteira de Decisão entre duas classes através da função:

$$d_{ij} = d_i(x) - d_j(x) = 0$$

Ou seja, se  $d_{ij}(x) > 0$  o padrão pertence à classe  $w_i$  e se  $d_{ij}(x) < 0$  o padrão pertence à classe  $w_j$

## Classificador de Distância Mínima:

- ❑ Uma Classe de Padrões pode ser representada por um vetor protótipo (ou médio).

$$m_j = \frac{1}{N_j} \sum_{x \in w_j} x \quad j = 1, 2, \dots, M$$

- ❑ Uma maneira de definir a pertinência de um Vetor de Características ( $x$ ) desconhecido, é atribuí-lo à classe de seu protótipo mais próximo.

Distância euclidiana:  $D_j(x) = \|x - m_j\| \quad j = 1, 2, \dots, M$

Onde:  $\|a\| = (a^T a)^{1/2}$  é a norma euclidiana.

$x \in w_i$  se  $D_i(x)$  for a menor distância

## Classificador de Distância Mínima:

Isso equivale a avaliar as funções:

$$d_j(x) = x^T m_j - \frac{1}{2} m_j^T m_j \quad j = 1, 2, \dots, M$$

e atribuir  $x$  à classe  $w_i$  se  $d_i(x)$  for o maior valor.

□ A **Fronteira de Decisão** entre as classes  $w_i$  e  $w_j$  para o Classificador de Distância Mínima é:

$$d_{ij} = d_i(x) - d_j(x) =$$

$$x^T (m_i - m_j) - \frac{1}{2} (m_i - m_j)^T (m_i - m_j) = 0$$

$n=2$  ----- uma reta

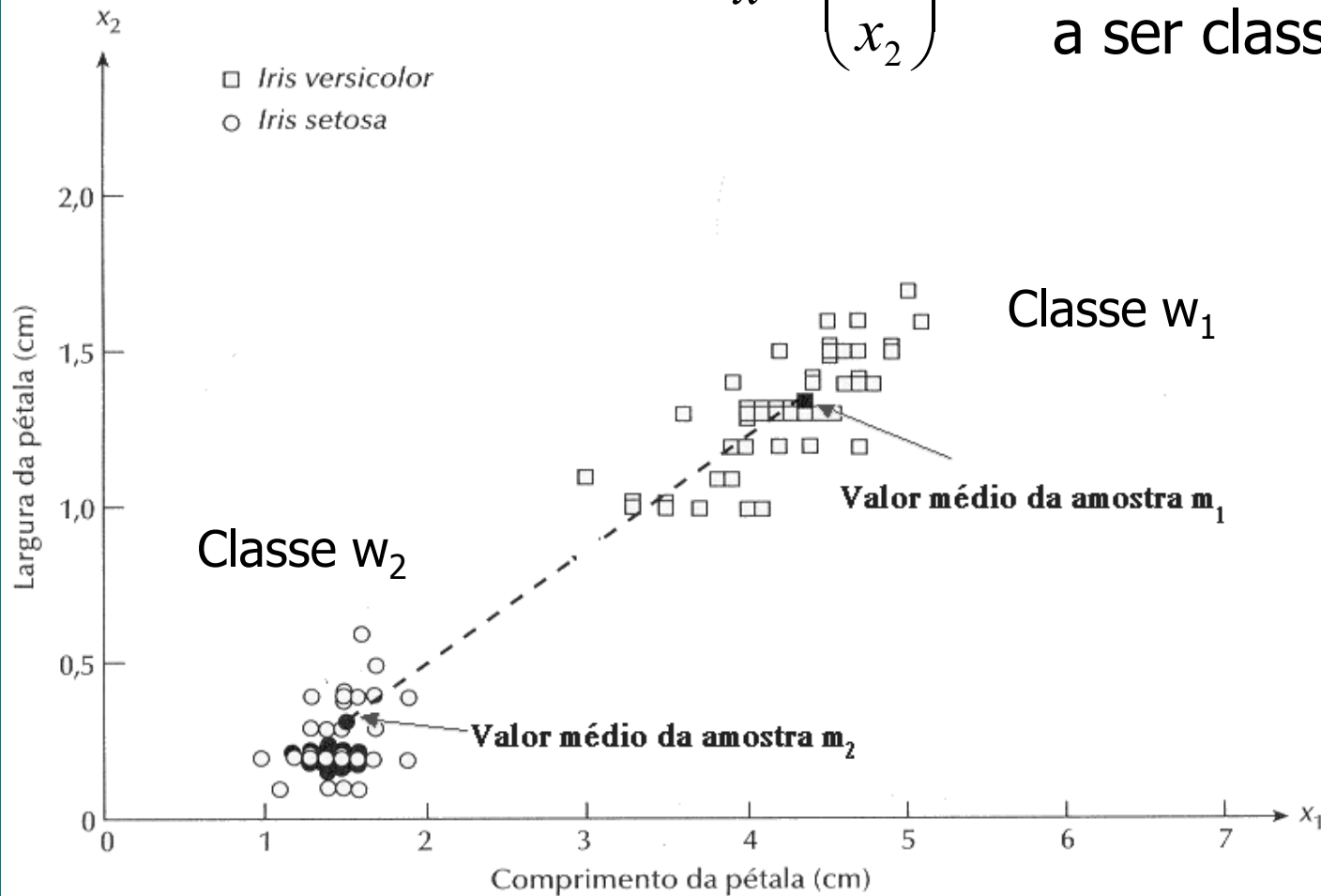
$n=3$  ----- um plano

$n>3$  ----- hiperplano

## Exemplo:

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

Vetor desconhecido a ser classificado



$$m_1 = (4.3, 1.3)^T$$
$$m_2 = (1.5, 0.3)^T$$



$$d_j(x) = x^T m_j - \frac{1}{2} m_j^T m_j \quad j = 1, 2, \dots, M$$

$$d_1(x) = x^T m_1 - \frac{1}{2} m_1^T m_1$$

$$(x_1 \quad x_2) \begin{pmatrix} 4.3 \\ 1.3 \end{pmatrix} - \frac{1}{2} (4.3 \quad 1.3) \begin{pmatrix} 4.3 \\ 1.3 \end{pmatrix} =$$

$$4.3x_1 + 1.3x_2 - \frac{1}{2} (4.3 \times 4.3 + 1.3 \times 1.3) =$$

$$4.3x_1 + 1.3x_2 - 10.09$$

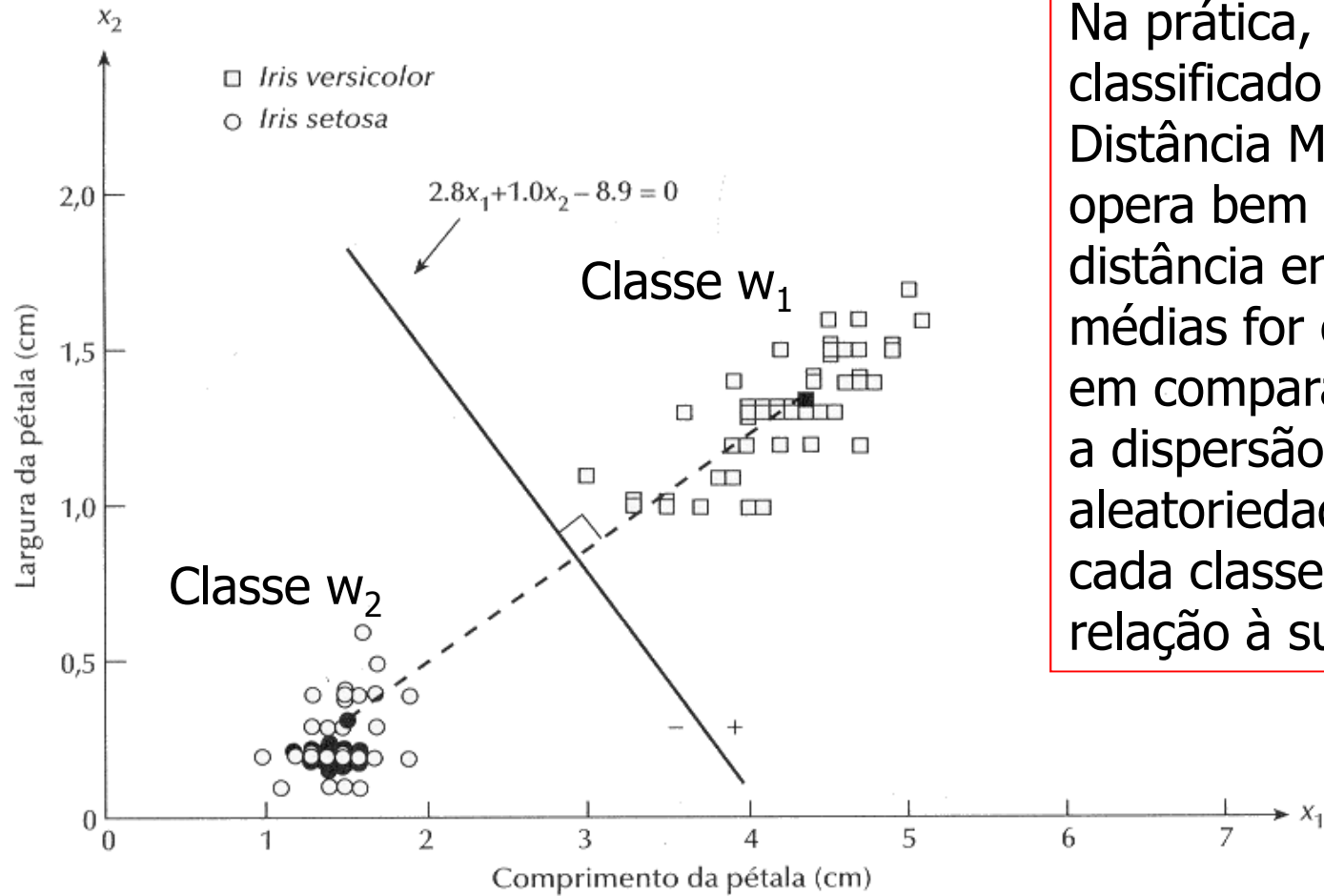
$$d_2(x) = x^T m_2 - \frac{1}{2} m_2^T m_2$$

$$d_2(x) = 1.5x_1 + 0.3x_2 - 1.17$$

Equação da Fronteira:

$$d_{12}(x) = d_1(x) - d_2(x) =$$

$$2.8x_1 + 1.0x_2 - 8.9 = 0$$



Na prática, o classificador de Distância Mínima opera bem quando a distância entre as médias for grande em comparação com a dispersão ou aleatoriedade de cada classe em relação à sua média.

Qualquer padrão desconhecido  $x$  pode ser classificado observando-se o sinal de  $d_{12}$

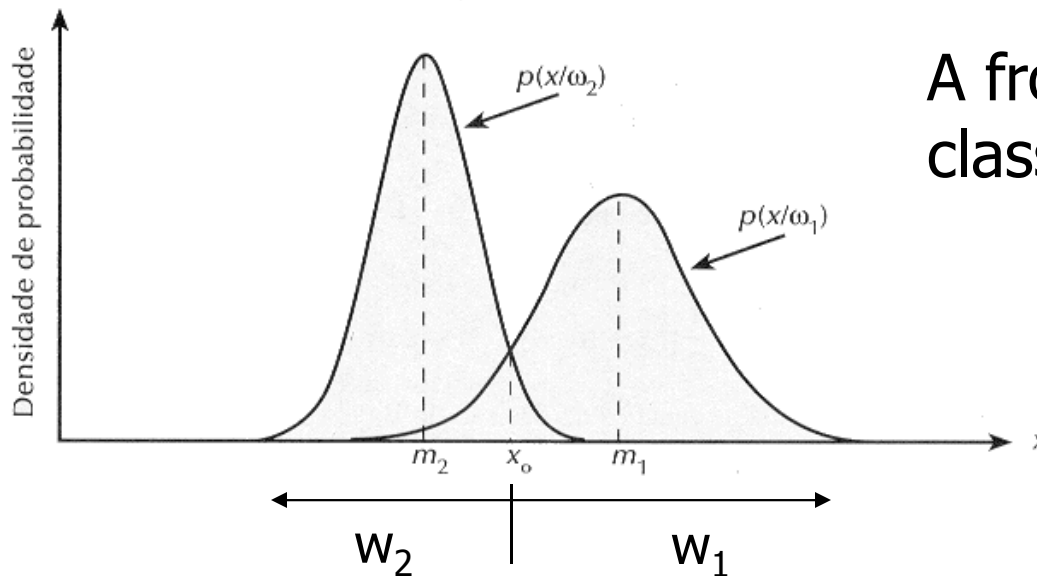
$$d_{12}(x) < 0 \quad \text{--- Classe } w_2$$

$$d_{12}(x) > 0 \quad \text{--- Classe } w_1$$

## Classificador Bayesiano:

Um problema envolvendo duas classes de padrões governadas por densidades gaussianas, com médias  $m_1$  e  $m_2$  e desvios padrão  $\sigma_1$  e  $\sigma_2$  respectivamente, pode ser resolvido usando-se as funções de decisão na forma:

$$d_j(x) = p(x/w_j)P(w_j) = \frac{1}{\sqrt{2\pi\sigma_j}} \exp\left[-\frac{(x-m_j)^2}{2\sigma_j^2}\right] P(w_j) \quad j = 1,2$$



A fronteira entre as classes é o ponto  $x_0$

## Outros Classificadores:

- Classificadores por Redes Neurais Artificiais
- Classificadores por Lógica Nebulosa ("Fuzzy Sets")

## Cluster Analysis: ( Análise de agrupamentos)

- É um método de Estatística Multivariada que identifica grupos em um grande número de objetos, baseado em suas características.
- Similarmente à Análise Discriminante, cada objeto tem múltiplas características que podem ser expressas como um vetor  $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p)$  com valores que variam de objeto para objeto.
- O Objetivo principal da Análise de Agrupamentos é identificar objetos similares baseada em suas características.

## Cluster Analysis: ( Análise de agrupamentos)

- “Cluster analysis” agrupa objetos similares em grupos tal que os objetos dentro de um grupo são similares e objetos entre os diferentes grupos são significativamente diferentes em suas características.
- Diferentemente da **Análise Discriminante** onde o número de grupos e seus nomes são conhecidos previamente, na **Análise de Agrupamentos** o número de grupos e suas características são desconhecidas antes da análise.

**Exemplo: Analisar o Agrupamento de cereais da Tabela de acordo com suas características nutricionais.**

Brand	Calories (Cal/oz)	Protein (g)	Fat (g)	Na (mg)	Fiber (g)	Carbs (g)	Sugar (g)	K (mg)
Cheerios	110	6	2	290	2.0	17.0	1	105
Cocoa Puffs	110	1	1	180	0.0	12.0	13	55
Honey Nut Cheerios	110	3	1	250	1.5	11.5	10	90
Kix	110	2	1	260	0.0	21.0	3	40
Lucky Charms	110	2	1	180	0.0	12.0	12	55
Oatmeal Raisin Crisp	130	3	2	170	1.5	13.5	10	120
Raisin Nut Bran	100	3	2	140	2.5	10.5	8	140
Total Corn Flakes	110	2	1	200	0.0	21.0	3	35
Total Raisin Bran	140	3	1	190	4.0	15.0	14	230
Trix	110	1	1	140	0.0	13.0	12	25
Wheaties Honey Gold	110	2	1	200	1.0	16.0	8	60
All-Bran	70	4	1	260	9.0	7.0	5	320
Apple Jacks	110	2	0	125	1.0	11.0	14	30
Corn Flakes	100	2	0	290	1.0	21.0	2	35
Corn Pops	110	1	0	90	1.0	13.0	12	20
Mueslix Crispy Blend	160	3	2	150	3.0	17.0	13	160
Nut & Honey Crunch	120	2	1	190	0.0	15.0	9	40
Nutri Grain Almond Raisin	140	3	2	220	3.0	21.0	7	130
Nutri Grain Wheat	90	3	0	170	3.0	18.0	2	90
Product 19	100	3	0	320	1.0	20.0	3	45
Raisin Bran	120	3	1	210	5.0	14.0	12	240
Rice Krispies	110	2	0	290	0.0	22.0	3	35
Special K	110	6	0	230	1.0	16.0	3	55
Life	100	4	2	150	2.0	12.0	6	95
Puffed Rice	50	1	0	0	0.0	13.0	0	15

O número de variáveis e quais serão selecionadas, afetarão o resultado final.

## Conjunto de Dados típicos em Cluster Analysis:

Objects	Variables			
	1	2	...	$p$
1	$x_{11}$	$x_{12}$	...	$x_{1p}$
2	$x_{21}$	$x_{22}$	...	$x_{2p}$
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$
$N$	$x_{N1}$	$x_{N2}$	...	$x_{Np}$

**Passo 1:** Selecionar as Variáveis de Agrupamento e a Medida de Distância.

**Passo 2:** Selecionar o algoritmo de Agrupamento (Hierárquico ou não-hierárquico)

**Passo 3:** Realizar a Análise de Agrupamento.

**Passo 4:** Interpretar os Agrupamentos.



## Medidas de Similaridade : Distâncias

Distância Euclidiana (DE):

$$d_{ik} = \sqrt{\sum_{j=1}^p (x_{ij} - x_{kj})^2}$$

Objects	Variables			
	1	2	...	$p$
1	$x_{11}$	$x_{12}$	...	$x_{1p}$
2	$x_{21}$	$x_{22}$	...	$x_{2p}$
⋮	⋮	⋮	...	⋮
$N$	$x_{N1}$	$x_{N2}$	...	$x_{Np}$

$d_{ik} \rightarrow$  DE entre o objeto  $i$  e o objeto  $k$  (*Vetor  $i$  e Vetor  $k$* )

- A escala numérica das variáveis pode variar significativamente.

No Exemplo:

Brand	Calories (Cal/oz)	Protein (g)	Fat (g)	Na (mg)	Fiber (g)	Carbs (g)	Sugar (g)	K (mg)
Cheerios	110	6	2	290	2.0	17.0	1	105
Cocoa Puffs	110	1	1	180	0.0	12.0	13	55
Honey Nut Cheerios	110	3	1	250	1.5	11.5	10	90

## Distância Euclidiana Normalizada: (Distância de Pearson)

- Se não se deseja que dados com maiores valores dominem o resultado, deve-se normalizar a escala.

- Cada dado  $x_{ij}$  deve ser normalizado para  $z_{ij}$ : 
$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

Onde: 
$$\bar{x}_j = \frac{\sum_{k=1}^N x_{kj}}{N}$$
 É a Média de cada característica

$$s_j = \sqrt{\frac{\sum_{k=1}^N (x_{kj} - \bar{x}_j)^2}{N - 1}}$$
 É o Desvio Padrão das características

Logo, a **Distância Euclidiana Normalizada** (Distância de Pearson) entre cada Vetor ( $i$  e  $k$ ) será:

$$d_{ik} = \sqrt{\sum_{j=1}^p (z_{ij} - z_{kj})^2}$$

## Matriz de Distâncias:

As Distâncias entre cada objeto, ou melhor, entre todos os Vetores de toda a população de Vetores de Características, podem ser colocadas em uma Matriz de Distância para a análise.

$$\mathbf{D} = \begin{bmatrix} 0 & d_{12} & d_{13} & \cdots & d_{1N} \\ d_{21} & 0 & d_{23} & \cdots & d_{2N} \\ d_{31} & d_{32} & 0 & \cdots & d_{3N} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{N1} & d_{N2} & d_{N3} & \cdots & 0 \end{bmatrix}$$

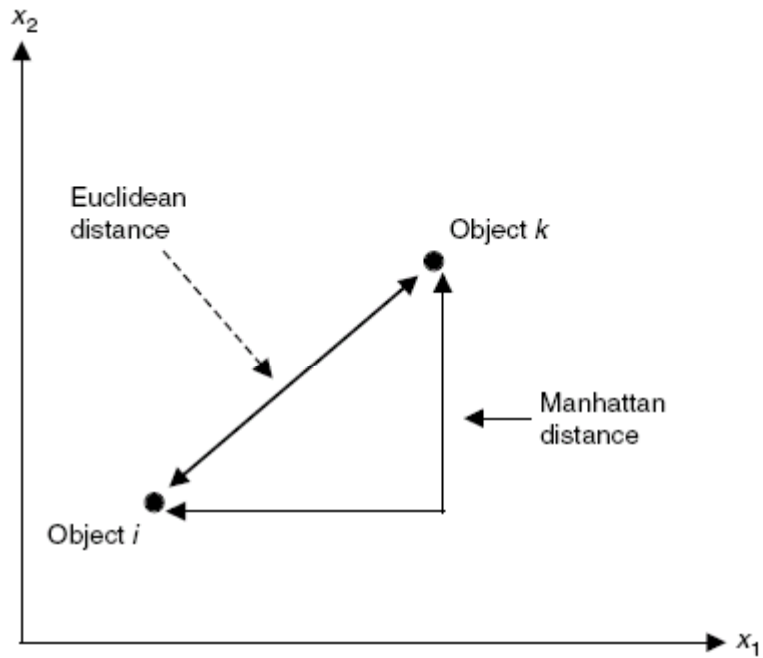
Distância Manhattan ou Distância City Block:

$$d_{ik} = \sum_{j=1}^p |x_{ij} - x_{kj}|$$

Distância Manhattan ou Distância City Block Normalizada:

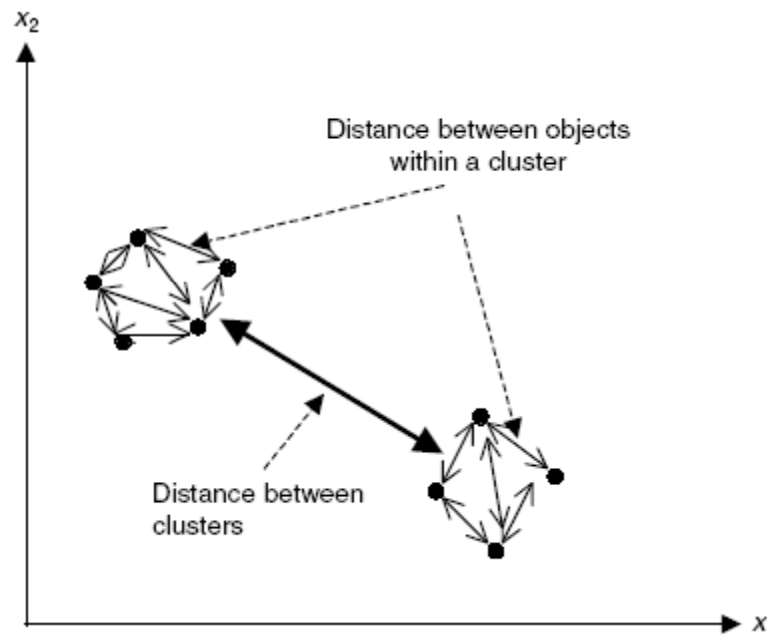
$$d_{ik} = \sum_{j=1}^p |z_{ij} - z_{kj}|$$

## Diferença entre a Distância Manhattan e a Distância Euclidiana:



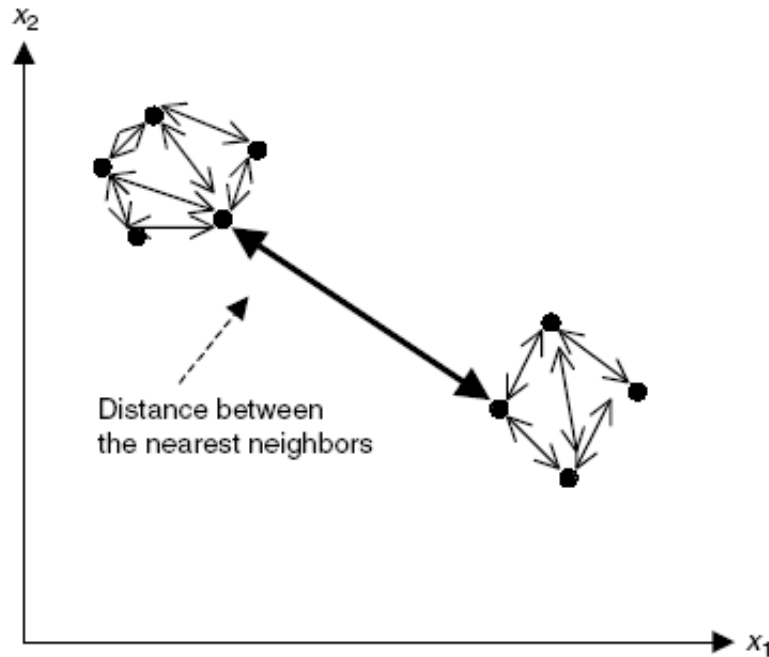
## Diferença entre Agrupamentos e Método de Ligação:

- Em "Cluster Analysis" é desejável que as distâncias entre os Vetores(objetos) dentro de um "cluster" (grupo) sejam pequenas e que as distâncias entre diferentes "clusters" sejam grandes.



- A Distância entre os grupos depende da relação definida entre eles.
- Esta relação é chamada de Método de Ligação ("Linkage Method")

## Método de Ligação Simples.



$$d_{(R)(S)} = \min\{d_{rs} | r \in R, s \in S\}$$

- A distância entre dois grupos ("clusters") é definida como a **Distância entre os Vizinhos mais Próximos**.

### Exemplo:

$$D = \begin{bmatrix} 1 & \cdot & \cdot & \cdot & 10 & 8 & 6 \\ 2 & \cdot & \cdot & \cdot & 6 & 9 & 5 \\ 3 & \cdot & \cdot & 0 & 13 & 11 & 8 \\ 4 & \cdot & \cdot & \cdot & 0 & \cdot & \cdot \\ 5 & \cdot & \cdot & \cdot & \cdot & 0 & \cdot \\ 6 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

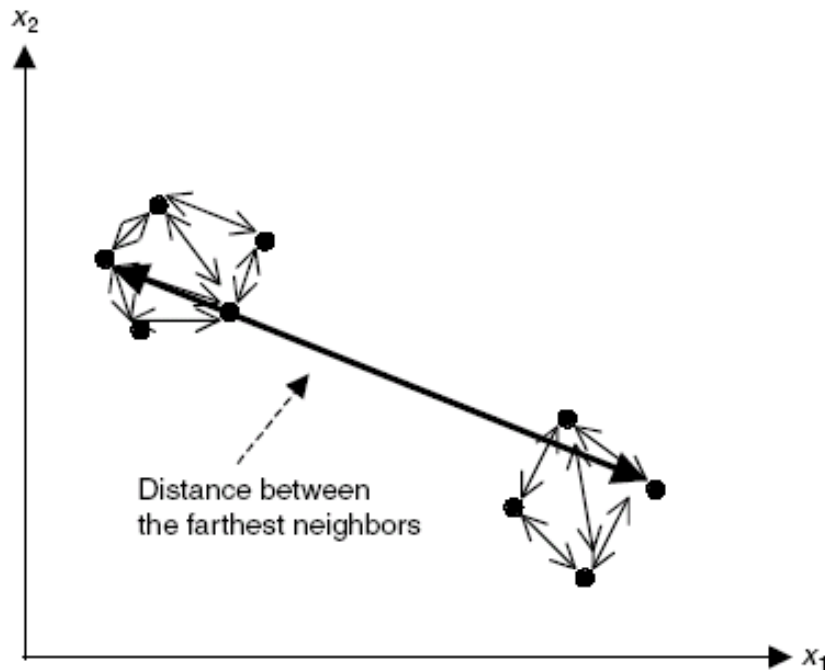
Cluster 1  
(Objetos  
1,2,3)

Cluster 2  
(Objetos  
4,5,6)

- A Distância entre os dois grupos (1 e 2) será:

$$\begin{aligned} d_{(1)(2)} &= \min\{d_{14}, d_{15}, d_{16}, d_{24}, d_{25}, d_{26}, d_{34}, d_{35}, d_{36}\} \\ &= \min\{10, 8, 6, 6, 9, 5, 13, 11, 8\} = 5 = d_{26} \end{aligned}$$

## Método de Ligação Completa:



$$d_{(R)(S)} = \max\{d_{rs} | r \in R, s \in S\}$$

- A distância entre dois grupos ("clusters") é definida como a **Distância entre os Vizinhos mais Distantes**.

### Exemplo:

$$D = \begin{bmatrix} 1 & \cdot & \cdot & \cdot & 10 & 8 & 6 \\ 2 & \cdot & \cdot & \cdot & 6 & 9 & 5 \\ 3 & \cdot & \cdot & 0 & 13 & 11 & 8 \\ \hline 4 & \cdot & \cdot & \cdot & 0 & \cdot & \cdot \\ 5 & \cdot & \cdot & \cdot & \cdot & 0 & \cdot \\ 6 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

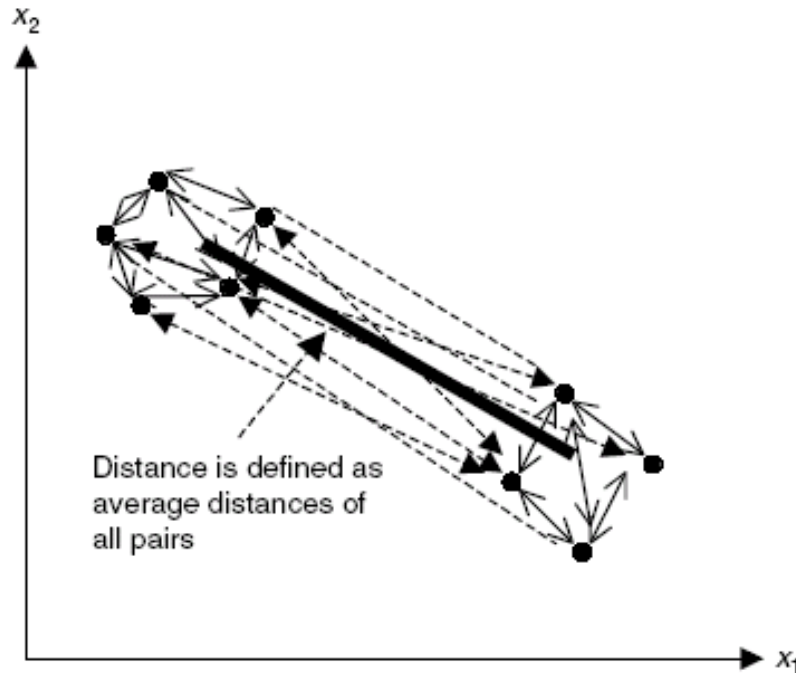
Cluster 1  
(Objetos  
1,2,3)

Cluster 2  
(Objetos  
4,5,6)

- A Distância entre os dois grupos (1 e 2) será:

$$\begin{aligned} d_{(1)(2)} &= \max\{d_{14}, d_{15}, d_{16}, d_{24}, d_{25}, d_{26}, d_{34}, d_{35}, d_{36}\} \\ &= \max\{10, 8, 6, 6, 9, 5, 13, 11, 8\} = 13 = d_{34} \end{aligned}$$

## Método de Ligação Média:



$$d_{(R)(S)} = \frac{\sum_r \sum_s d_{rs}}{n_R n_S}$$

- A distância entre dois grupos ("clusters") é definida como a **A Média de todas as distâncias entre os dois grupos.**

### Exemplo:

$$D = \begin{bmatrix} 1 & \cdot & \cdot & \cdot & 10 & 8 & 6 \\ 2 & \cdot & \cdot & \cdot & 6 & 9 & 5 \\ 3 & \cdot & \cdot & 0 & 13 & 11 & 8 \\ \hline 4 & \cdot & \cdot & \cdot & 0 & \cdot & \cdot \\ 5 & \cdot & \cdot & \cdot & \cdot & 0 & \cdot \\ 6 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

Cluster 1  
(Objetos  
1,2,3)

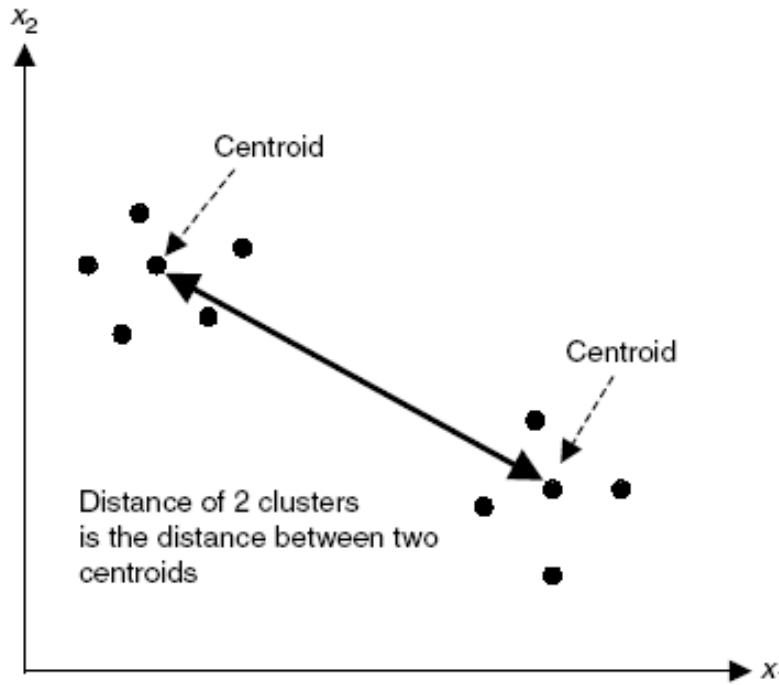
Cluster 2  
(Objetos  
4,5,6)

- A Distância entre os dois grupos (1 e 2) será:

$$\begin{aligned} d_{(1)(2)} &= \frac{d_{14} + d_{15} + d_{16} + d_{24} + d_{25} + d_{26} + d_{34} + d_{35} + d_{36}}{3 \times 3} \\ &= \frac{10 + 8 + 6 + 6 + 9 + 5 + 13 + 11 + 8}{9} = 8.44 \end{aligned}$$



## Método de Ligação Centróide:



- A distância entre dois grupos (“clusters”) é definida como a distância entre os Centros Gravitacionais de cada grupo.

- Sejam os grupos: R com  $n_R$  objetos e S com  $n_S$  objetos:

- A coordenada do Centro de Gravidade de cada Grupo será:

$$\bar{\mathbf{x}}_R = \frac{\sum_r \mathbf{x}_r}{n_R} = \begin{bmatrix} \bar{x}_{r1} \\ \bar{x}_{r2} \\ \vdots \\ \bar{x}_{rp} \end{bmatrix} \quad \bar{\mathbf{x}}_S = \frac{\sum_s \mathbf{x}_s}{n_S} = \begin{bmatrix} \bar{x}_{s1} \\ \bar{x}_{s2} \\ \vdots \\ \bar{x}_{sp} \end{bmatrix}$$

- Logo, a Distância Euclidiana entre os dois grupos será:

$$d_{(R)(S)} = \sqrt{(\bar{x}_{r1} - \bar{x}_{s1})^2 + \dots + (\bar{x}_{rp} - \bar{x}_{sp})^2}$$

## Similaridade:

- **Similaridade** é a diferença entre dois objetos (vetores) ou entre dois Grupos de Objetos ("clusters").
- Quanto maior é a **Similaridade** menor é a distância entre eles.
- Dados dois objetos  $\mathbf{x}_r$  e  $\mathbf{x}_s$ , a Similaridade é dada por  $s_{rs}$  e obedece às seguintes condições:

1.  $0 \leq s_{rs} \leq 1$

2.  $s_{rs} = 1$  if and only if  $\mathbf{x}_r = \mathbf{x}_s$

3.  $s_{rs} = s_{sr}$

## Similaridade:

- A medida de Similaridade pode ser dada por:

$$s_{rs} = 1 - \frac{d_{rs}}{d_{\max}}$$

Onde,  $d_{\max}$  é a Máxima Distância na Matriz de Distâncias D.

- Uma outra maneira de medir a **Similaridade** é através da **Correlação de Momentos do Produto de Pearson**

$$q_{rs} = \frac{\sum_{j=1}^p (x_{rj} - \bar{x}_r)(x_{sj} - \bar{x}_s)}{\left[ \sum_{j=1}^p (x_{rj} - \bar{x}_r)^2 \sum_{j=1}^p (x_{sj} - \bar{x}_s)^2 \right]}$$

# Agrupamento Hierárquico

- **Agrupamento Hierárquico** (Hierarchical clustering) é uma maneira de investigar o agrupamento dos dados, simultaneamente em várias escalas, através da geração de uma **Árvore de Grupos** (Cluster Tree).
- A **Árvore de Grupos** não é apenas um simples conjunto de grupos, mas uma Hierarquia em multi-nível onde grupos em um nível são unidos a grupos em um próximo nível mais alto.
- Isto permite decidir qual nível ou escala de agrupamento é mais apropriada para cada aplicação.

# Agrupamento Hierárquico

- O número de Agrupamentos (“clusters”) e quais são eles é desconhecido.
- Usa a Matriz de Distâncias para construir um gráfico de Árvore de Grupo chamado de **Dendrograma**.
- Algoritmo:
  1. Considerar inicialmente todos os elementos (Vetores/Objetos) individuais como um cluster formado por ele mesmo.
  2. Combinar em um cluster dois objetos com a menor distância.
  3. Computar as distâncias entre os objetos e o novo cluster formado.
  4. Repetir o processo até que o número de clusters seja reduzido a 1.
  5. Decidir o número de grupos para solucionar o problema.

## Exemplo de Análise através de Agrupamento Hierárquico:

- Comportamento do tempo em uma cidade Norte-americana no mês de Fevereiro entre os anos de 1982 e 1990

Year	$x_1$ , Mean temp.	$x_2$ , Max. temp.	$x_3$ , Min. temp.	$x_4$ , Soil temp. (@ 10 cm)	$x_5$ , Monthly rainfall (mm)	$x_6$ , Max. rain in a day	$x_7$ , Days with snow
1982	4.2	13.3	-5.3	4.0	23	6	0
1983	1.0	7.8	-5.3	3.0	34	11	8
1984	2.9	11.4	-5.1	3.2	65	17	0
1985	1.6	10.2	-6.0	2.9	7	2	5
1986	-1.1	2.7	-9.0	1.5	22	5	24
1987	3.3	13.4	-7.3	2.7	46	15	2
1988	4.5	13.0	-2.9	3.7	89	22	4
1989	5.7	13.5	-2.7	5.2	92	16	0
1990	6.6	14.9	-0.6	5.5	131	29	0

7 Características



9 Vetores de Características ou Objetos

### Passo 0 : Cálculo da Matriz de Distâncias

$$D = \begin{bmatrix} 0.00 & 2.44 & 1.90 & 1.86 & 5.32 & 1.82 & 2.70 & 2.56 & 4.48 \\ 2.44 & 0.00 & 1.92 & 1.49 & 3.31 & 2.16 & 2.98 & 3.69 & 5.23 \\ 1.90 & 1.92 & 0.00 & 2.45 & 4.98 & 1.25 & 1.57 & 2.36 & 3.78 \\ 1.86 & 1.48 & 2.45 & 0.00 & 3.75 & 2.19 & 3.64 & 4.01 & 5.84 \\ 5.32 & 3.31 & 4.98 & 3.75 & 0.00 & 4.73 & 5.90 & 6.71 & 8.13 \\ 1.82 & 2.16 & 1.25 & 2.19 & 4.73 & 0.00 & 2.40 & 3.10 & 4.59 \\ 2.70 & 2.98 & 1.57 & 3.64 & 5.90 & 2.40 & 0.00 & 1.57 & 2.42 \\ 2.56 & 3.69 & 2.36 & 4.01 & 6.71 & 3.10 & 1.57 & 0.00 & 2.05 \\ 4.48 & 5.23 & 3.78 & 5.83 & 8.13 & 4.59 & 2.42 & 2.05 & 0.00 \end{bmatrix}$$

Distância Euclidiana Normalizada

## Exemplo de Análise através de Agrupamento Hierárquico:

Passo 1 : Encontrar a menor distância e criar um novo cluster.

$$D = \begin{bmatrix} 0.00 & 2.44 & 1.90 & 1.86 & 5.32 & 1.82 & 2.70 & 2.56 & 4.48 \\ 2.44 & 0.00 & 1.92 & 1.49 & 3.31 & 2.16 & 2.98 & 3.69 & 5.23 \\ 1.90 & 1.92 & 0.00 & 2.45 & 4.98 & 1.25 & 1.57 & 2.36 & 3.78 \\ 1.86 & 1.48 & 2.45 & 0.00 & 3.75 & 2.19 & 3.64 & 4.01 & 5.84 \\ 5.32 & 3.31 & 4.98 & 3.75 & 0.00 & 4.73 & 5.90 & 6.71 & 8.13 \\ 1.82 & 2.16 & 1.25 & 2.19 & 4.73 & 0.00 & 2.40 & 3.10 & 4.59 \\ 2.70 & 2.98 & 1.57 & 3.64 & 5.90 & 2.40 & 0.00 & 1.57 & 2.42 \\ 2.56 & 3.69 & 2.36 & 4.01 & 6.71 & 3.10 & 1.57 & 0.00 & 2.05 \\ 4.48 & 5.23 & 3.78 & 5.83 & 8.13 & 4.59 & 2.42 & 2.05 & 0.00 \end{bmatrix}$$

- Menor distância  $\rightarrow$  entre o objeto(Vetor) 3 e o 6

- Combinar o objeto 3 e o 6 em um único cluster.

Passo 2 : Atualizar as distâncias.

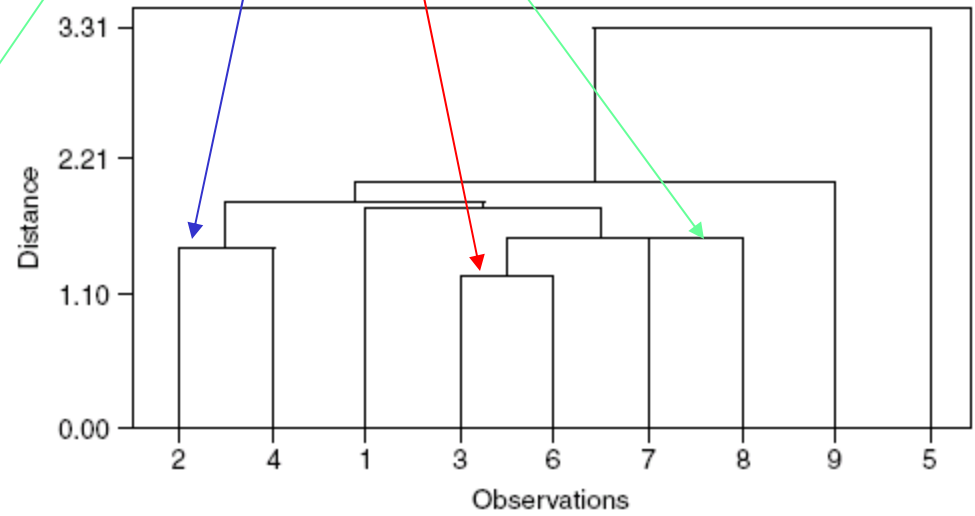
$$\begin{aligned} d_{1,(3,6)} &= \min(d_{13}, d_{16}) = \min(1.90, 1.82) = 1.82 \\ d_{2,(3,6)} &= \min(d_{23}, d_{26}) = \min(1.92, 2.16) = 1.92 \\ d_{4,(3,6)} &= \min(d_{43}, d_{46}) = \min(2.45, 2.19) = 2.19 \\ d_{5,(3,6)} &= \min(d_{53}, d_{56}) = \min(4.98, 4.73) = 4.73 \\ d_{7,(3,6)} &= \min(d_{73}, d_{76}) = \min(1.57, 2.40) = 1.57 \\ d_{8,(3,6)} &= \min(d_{83}, d_{86}) = \min(2.36, 3.10) = 2.36 \\ d_{9,(3,6)} &= \min(d_{93}, d_{96}) = \min(3.78, 4.59) = 3.78 \end{aligned}$$

- Foi utilizado o Método de Ligação Simples.

# Exemplo de Análise através de Agrupamento Hierárquico:

Passo 3 : Repetir os Passos 1 e 2 estabelecendo um **Dendrograma**.

Step	Number of clusters	Similarity level	Distance level	Clusters joined	New cluster	Number of obs. in new cluster
1	8	84.67	1.246	3 6	3	2
2	7	81.71	1.486	2 4	2	2
3	6	80.70	1.569	7 8	7	2
4	5	80.64	1.573	3 7	3	4
5	4	77.63	1.818	1 3	1	5
6	3	77.12	1.860	1 2	1	7
7	2	74.83	2.046	1 9	1	8
8	1	59.28	3.309	1 5	1	9

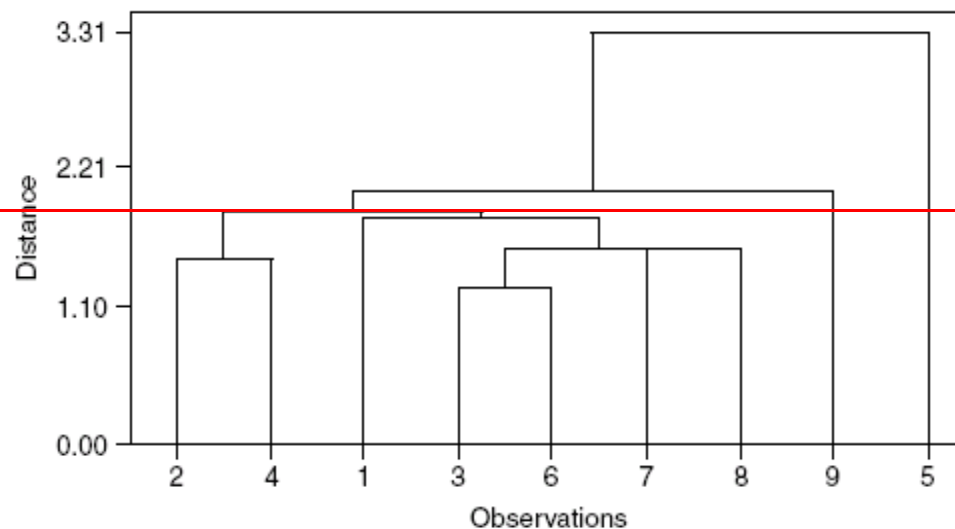
$$D = \begin{bmatrix} 0.00 & 2.44 & 1.90 & 1.86 & 5.32 & 1.82 & 2.70 & 2.56 & 4.48 \\ 2.44 & 0.00 & 1.92 & 1.49 & 3.31 & 2.16 & 2.98 & 3.69 & 5.23 \\ 1.90 & 1.92 & 0.00 & 2.45 & 4.98 & 1.25 & 1.57 & 2.36 & 3.78 \\ 1.86 & 1.48 & 2.45 & 0.00 & 3.75 & 2.19 & 3.64 & 4.01 & 5.84 \\ 5.32 & 3.31 & 4.98 & 3.75 & 0.00 & 4.73 & 5.90 & 6.71 & 8.13 \\ 1.82 & 2.16 & 1.25 & 2.19 & 4.73 & 0.00 & 2.40 & 3.10 & 4.59 \\ 2.70 & 2.98 & 1.57 & 3.64 & 5.90 & 2.40 & 0.00 & 1.57 & 2.42 \\ 2.56 & 3.69 & 2.36 & 4.01 & 6.71 & 3.10 & 1.57 & 0.00 & 2.05 \\ 4.48 & 5.23 & 3.78 & 5.83 & 8.13 & 4.59 & 2.42 & 2.05 & 0.00 \end{bmatrix}$$




## Exemplo de Análise através de Agrupamento Hierárquico:

- Observando-se o Dendrograma pode-se decidir que 4 clusters solucionam o problema, ou seja, cortando-se o gráfico na distância 1.818 tem-se os clusters: (2 e 4) (1,3, 6, 7 e 8) (9) (5)

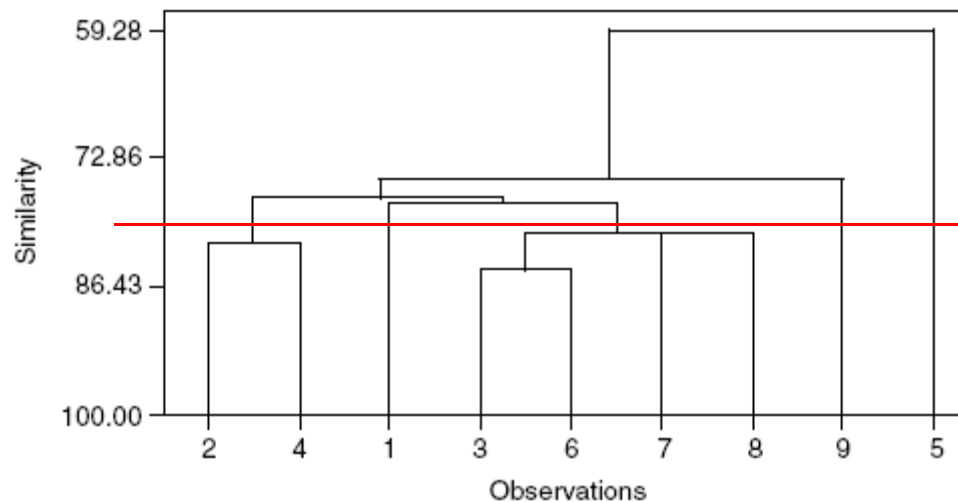
Step	Number of clusters	Similarity level	Distance level	Clusters joined	New cluster	Number of obs. in new cluster
1	8	84.67	1.246	3 6	3	2
2	7	81.71	1.486	2 4	2	2
3	6	80.70	1.569	7 8	7	2
4	5	80.64	1.573	3 7	3	4
5	4	77.63	1.818	1 3	1	5
6	3	77.12	1.860	1 2	1	7
7	2	74.83	2.046	1 9	1	8
8	1	59.28	3.309	1 5	1	9



## Exemplo de Análise através de Agrupamento Hierárquico:

- Utilizando a Similaridade ao invés da Distância, pode-se agrupar os vetores através do índice de similaridade entre eles.

Step	Number of clusters	Similarity level	Distance level	Clusters joined	New cluster	Number of obs. in new cluster
1	8	84.67	1.246	3 6	3	2
2	7	81.71	1.486	2 4	2	2
3	6	80.70	1.569	7 8	7	2
4	5	80.64	1.573	3 7	3	4
5	4	77.63	1.818	1 3	1	5
6	3	77.12	1.860	1 2	1	7
7	2	74.83	2.046	1 9	1	8
8	1	59.28	3.309	1 5	1	9



$$S_{rs} = 1 - \frac{d_{rs}}{d_{\max}}$$

- Similaridade de 80%  
→ 5 clusters

## Exemplo de Análise através de Agrupamento Hierárquico:

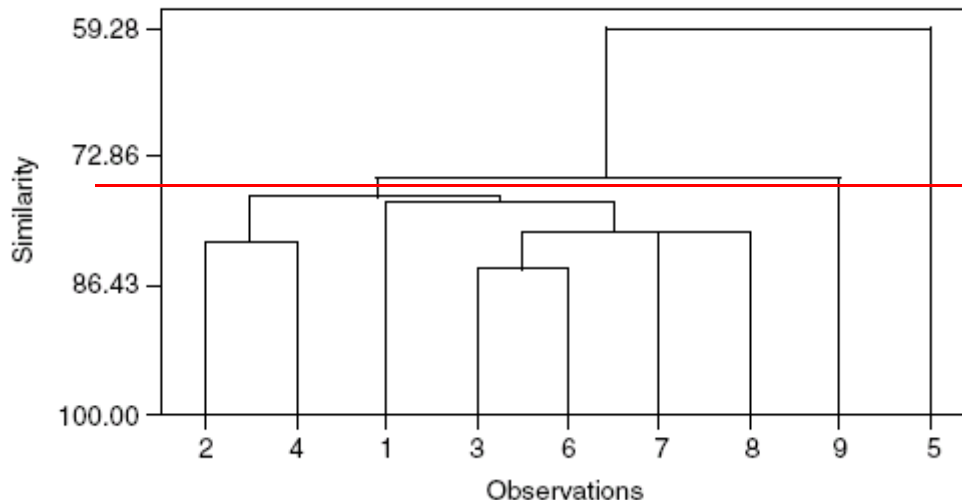
- Interpretação dos resultados: 75% de similaridade (3 grupos)

Year	$x_1$ , Mean temp.	$x_2$ , Max. temp.	$x_3$ , Min. temp.	$x_4$ , Soil temp. (@ 10 cm)	$x_5$ , Monthly rainfall (mm)	$x_6$ , Max. rain in a day	$x_7$ , Days with snow
1982	4.2	13.3	-5.3	4.0	23	6	0
1983	1.0	7.8	-5.3	3.0	34	11	8
1984	2.9	11.4	-5.1	3.2	65	17	0
1985	1.6	10.2	-6.0	2.9	7	2	5
1986	-1.1	2.7	-9.0	1.5	22	5	24
1987	3.3	13.4	-7.3	2.7	46	15	2
1988	4.5	13.0	-2.9	3.7	89	22	4
1989	5.7	13.5	-2.7	5.2	92	16	0
1990	6.6	14.9	-0.6	5.5	131	29	0

- **Cluster 1:** (2,4,1,3,6,7,8) (1982,1983,1984,1985,1987,1988,1989) → Fevereiro típico (não muito frio, não muito quente, neve e chuva na média)

- **Cluster 2:** (5) (1986) → Fevereiro frio e com neve

- **Cluster 3:** (9) (1990) → Fevereiro quente e chuvoso



## Agrupamento Não-Hierárquico.

- No método de **Agrupamento Não-hierárquico** o analista deve primeiramente especificar o número de “clusters” desejados.

- **Método K-Means (K-Médias)**

- **Passo 1:** Especificar inicialmente k sementes cada uma delas como um cluster. Calcular seus centróides.
- **Passo 2:** Calcular a distância de cada objeto (Vetor) para o centróide de cada cluster. Atribuir o objeto ao cluster mais próximo. Re-atribuir se necessário.
- **Passo 3:** Recalcular o centróide baseado nas re-atribuições e repetir o Passo 2. Parar se nenhum objeto puder ser re-atribuído a um cluster.

# Agrupamento Não-Hierárquico.

- Algumas dificuldades com o [K-Means](#).
  1. A composição dos grupos é muito sensível às sementes iniciais. Para diferentes sementes pode-se ter diferentes tipos de clusters. Não há garantia que convirja para uma solução ótima.
  2. Algumas vezes é difícil escolher um bom número de grupos antes de analisar os dados.
    - Pode-se combinar os métodos hierárquicos e não-hierárquicos para identificar as sementes e o número de grupos. Os resultados podem então ser usados no agrupamento não-hierárquico para refinar a solução.

# Exemplo: K-Means (K-Médias)

- Valores Unidimensionais

- $V = \{3, 1, 2, 0, 2, 10, 12, 9, 8, 11\}$

- Início:

- $M1 = 1$

- $M2 = 3$

- Iteração

- 1ª Iteração

- $G1 = \{1, 2, 0, 2\}$

- $G2 = \{3, 10, 12, 9, 8, 11\}$

- $M1 = 1.25$

- $M2 = 8.8$

- 2ª Iteração

- $G1 = \{3, 1, 2, 0, 2\}$

- $G2 = \{10, 12, 9, 8, 11\}$

- $M1 = 1.6$

- $M2 = 10$

- $K = 2$

- Distâncias entre cada objeto e as Médias (M1 e M2)

- Matriz de Distâncias

1a. iteração

	3	1	2	0	2	10	12	9	8	11
M1=1	2	0	1	1	1	9	11	8	7	10
M2=3	0	2	1	3	1	7	9	6	5	8

$$M1 = (1 + 2 + 0 + 2) / 4 = \frac{5}{4} = 1.25$$

$$M2 = (3 + 10 + 12 + 9 + 8 + 11) / 6 = \frac{53}{6} = 8.8$$

2a. iteração

	3	1	2	0	2	10	12	9	8	11
M1=1.25	1.75	0.25	0.75	1.25	0.75	8.75	10.75	7.75	6.75	9.75
M2=8.8	5.8	7.8	6.8	8.8	6.8	1.2	3.2	0.2	0.8	2.2

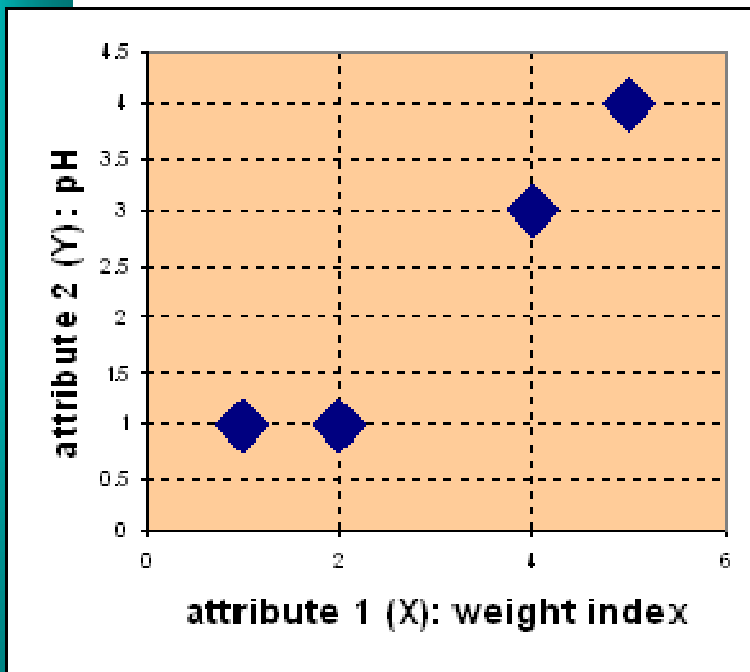
3a. iteração

	3	1	2	0	2	10	12	9	8	11
M1=1.6	1.4	0.6	0.4	1.6	0.4	8.4	10.4	7.4	6.4	9.4
M2=10	7	9	8	10	8	0	2	1	2	1

# Exemplo: K-Means (K-Médias)

- Valores Bi-dimensionais

Objeto	Atributo_1(X): Índice de Peso	Atributo_2(Y): pH
Produto_A	1	1
Produto_B	2	1
Produto_C	4	3
Produto_D	5	4



- Vetores de Características (X Y)

$$A = (1 \ 1)$$

$$B = (2 \ 1)$$

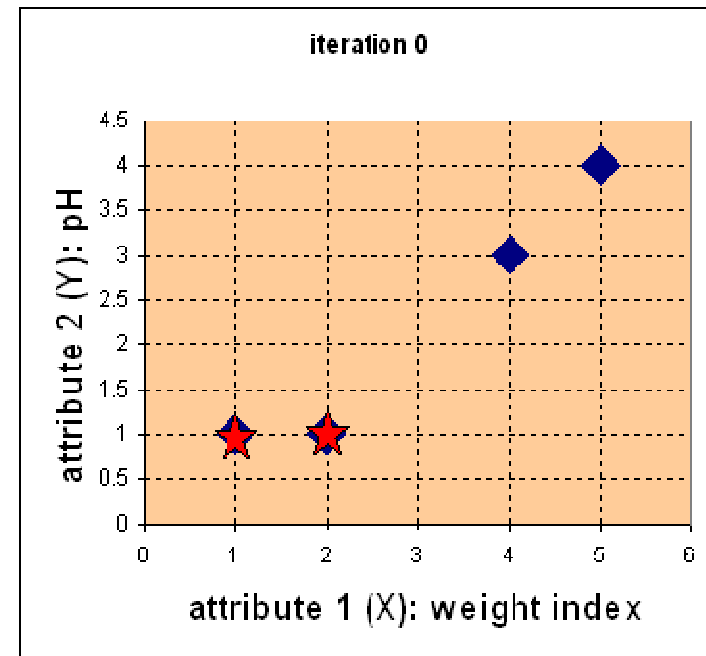
$$C = (4 \ 3)$$

$$D = (5 \ 4)$$

- Centróides Iniciais:

$$C_1 = (1 \ 1)$$

$$C_2 = (2 \ 1)$$



# Exemplo: K-Means (K-Médias)

- $D^0$  = Matriz de Distâncias na iteração 0.

Coluna → Objeto      Linha → Distância ao centróide

$$D^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix}$$

$$C_1 = (1,1) \quad \textit{grupo\_1}$$

$$C_2 = (2,1) \quad \textit{grupo\_2}$$

- Iteração 0.

- Distâncias Euclidianas

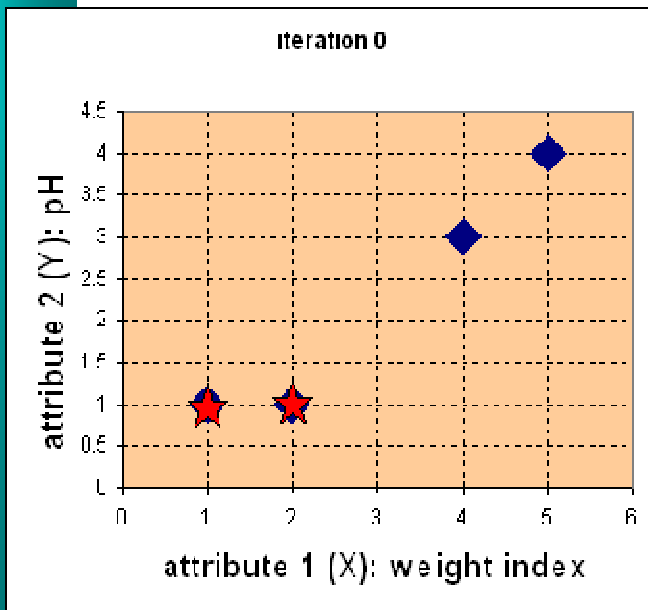
$$D(A, C_1) = \sqrt{(1-1)^2 + (1-1)^2} = 0$$

$$D(B, C_1) = \sqrt{(2-1)^2 + (1-1)^2} = 1$$

$$D(C, C_1) = \sqrt{(4-1)^2 + (3-1)^2} = \sqrt{9+4} = \sqrt{13} = 3.61$$

$$D(D, C_1) = \sqrt{(5-1)^2 + (4-1)^2} = \sqrt{16+9} = \sqrt{25} = 5$$

	A	B	C	D	
	1	2	4	5	X
	1	1	3	4	Y



$$D(A, C_2) = \sqrt{(1-2)^2 + (1-1)^2} = 1$$

$$D(B, C_2) = \sqrt{(2-2)^2 + (1-1)^2} = 0$$

$$D(C, C_2) = \sqrt{(4-2)^2 + (3-1)^2} = \sqrt{4+4} = \sqrt{8} = 2.83$$

$$D(D, C_2) = \sqrt{(5-2)^2 + (4-1)^2} = \sqrt{9+9} = \sqrt{18} = 4.24$$



## Exemplo: K-Means (K-Médias)

$$D^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix}$$

$$G^0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix}$$

A B C D

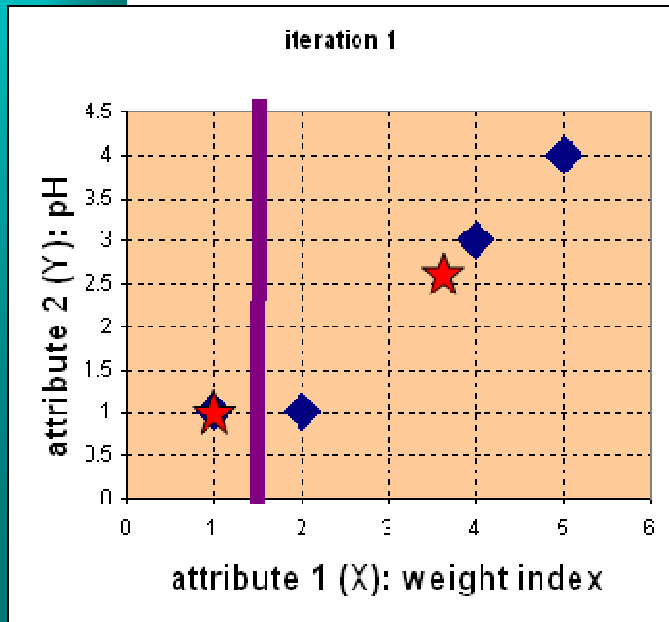
$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \begin{matrix} X \\ Y \end{matrix}$$

Grupo\_1 = (A)

Grupo\_2 = (B C D)

- $G^0$  = Matriz de Grupos na iteração 0.

• Observando-se a Matriz de Distâncias  $D^0$ , atribui-se o valor 1 na Matriz de Grupos  $G^0$  à posição de menor distância de cada objeto.



- Iteração 1.

- Novos Centróides

$$C_1 = (1, 1)$$

$$C_2 = \left( \frac{2+4+5}{3}, \frac{1+3+4}{3} \right) = \left( \frac{11}{3}, \frac{8}{3} \right) = (3.67 \quad 2.67)$$

$$C_1 = (1, 1) \quad \textit{grupo\_1}$$

$$C_2 = (3.67, 2.67) \quad \textit{grupo\_2}$$

## Exemplo: K-Means (K-Médias)

- Iteração 1.

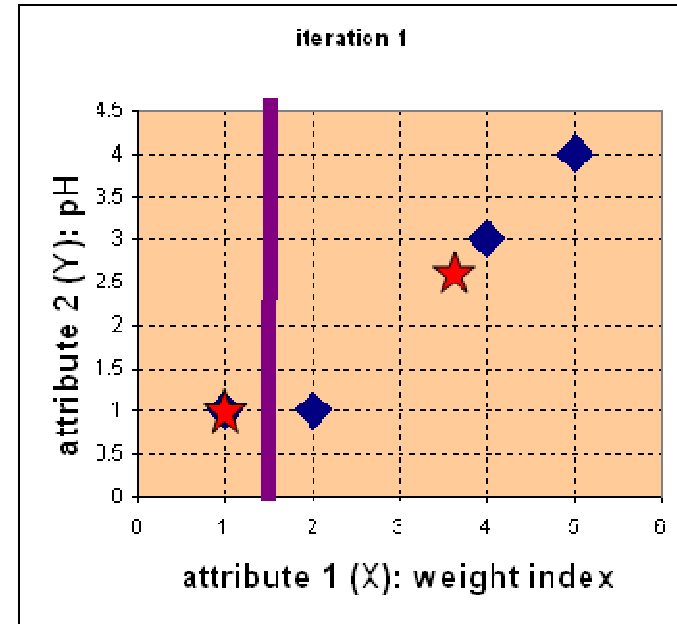
$$D^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix}$$

	A	B	C	D	
	1	2	4	5	$X$
	1	1	3	4	$Y$

$$G^1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

Grupo\_1 = (A B)

Grupo\_2 = (C D)



$$C_1 = (1, 1)$$

*grupo\_1*

$$C_2 = (3.67, 2.67)$$

*grupo\_2*

- Novos Centróides

$$C_1 = \left( \frac{1+2}{2}, \frac{1+1}{2} \right) = (1.5, 1)$$

$$C_2 = \left( \frac{4+5}{2}, \frac{3+4}{2} \right) = (4.5, 3.5)$$

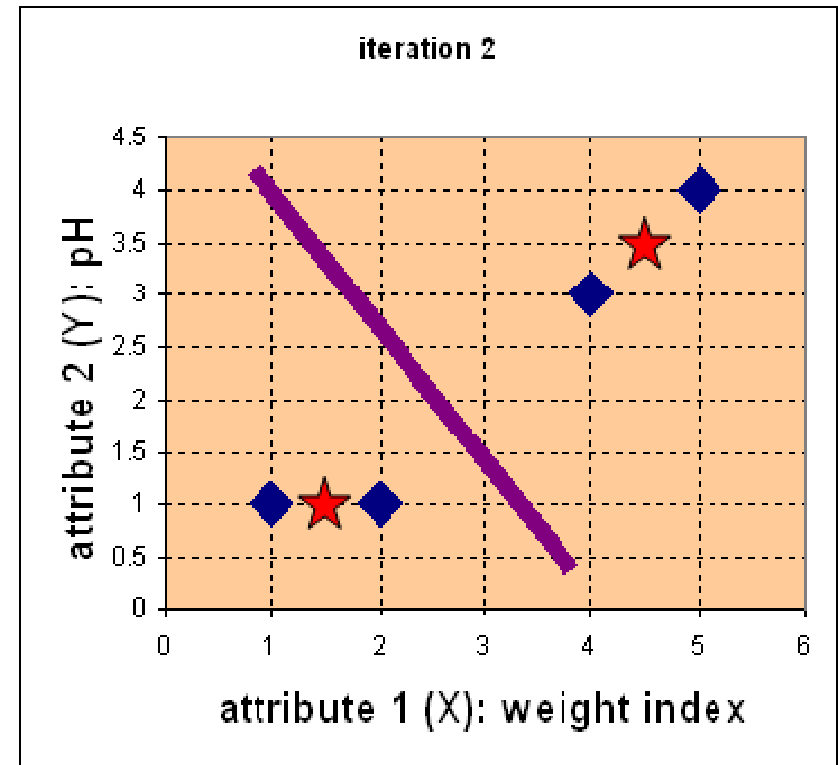
## Exemplo: K-Means (K-Médias)

- Iteração 2.

$$D^2 = \begin{bmatrix} 0.5 & 0.5 & 3.2 & 4.61 \\ 4.3 & 3.54 & 0.71 & 0.71 \end{bmatrix}$$

A	B	C	D	
1	2	4	5	$X$
1	1	3	4	$Y$

$$G^2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$



$$C_1 = (1.5 \ 1)$$

$$C_2 = (4.5 \ 3.5)$$

- Como  $G^2 = G^1$  os objetos não mais se moverão entre os grupos, logo a partição que agrupa os Produtos Similares é:

$$\text{Grupo}_1 = (A \ B)$$

$$\text{Grupo}_2 = (C \ D)$$