

CPE 722 –

Redes Neurais Não Supervisionadas e Agrupamento

Luiz Calôba

caloba@ufrj.br , www.lps.ufrj.br/~caloba

2012 / 3º

1- Introdução

Redes Neurais Artificiais

Inteligência Artificial / Computacional

Cognitiva, Simbólica

IA “tradicional”

Evolucionista

AG, Vida Artificial, Enxame

Conexionista

Redes Neurais Artificiais

Redes Neurais Artificiais

Redes Feedforward

*Aproximadores e classificadores,
Redes “backpropagation”*

Redes Não Supervisionadas
Classificadores

Redes Realimentadas
Otimizadores

Redes Neurais Naturais

Redes Neurais não Supervisionadas

Redes Neurais com Treinamento não Supervisionado

uma família de redes:

Camada de Kohonen, SOM, ART,
Counterpropagation, LQV, etc.

Por que estudar isto ?

Propriedades importantes:

Aprendizado cego, não supervisionado (on line).
Plasticidade.

Estas redes realizam:

Classificadores por similaridade

Aplicações:

Reconhecimento de padrões;

Estatística de sinais;

Memórias associativas;

Filtragem não-linear;

Compressão de dados;

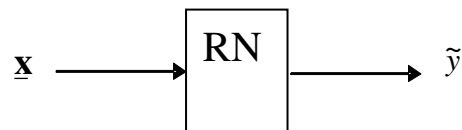
Mapas auto-organizáveis;

etc...

2 - Noções iniciais sobre Classificadores

Mapeadores

$$\underline{\mathbf{x}} \Rightarrow \underline{\mathbf{y}}$$



Mapeadores lineares / não lineares

$$\underline{\tilde{y}} = \underline{M} \underline{x}$$

$$\underline{\tilde{y}} = \varphi(\underline{x})$$

Tipos de mapeadores não lineares

Associadores

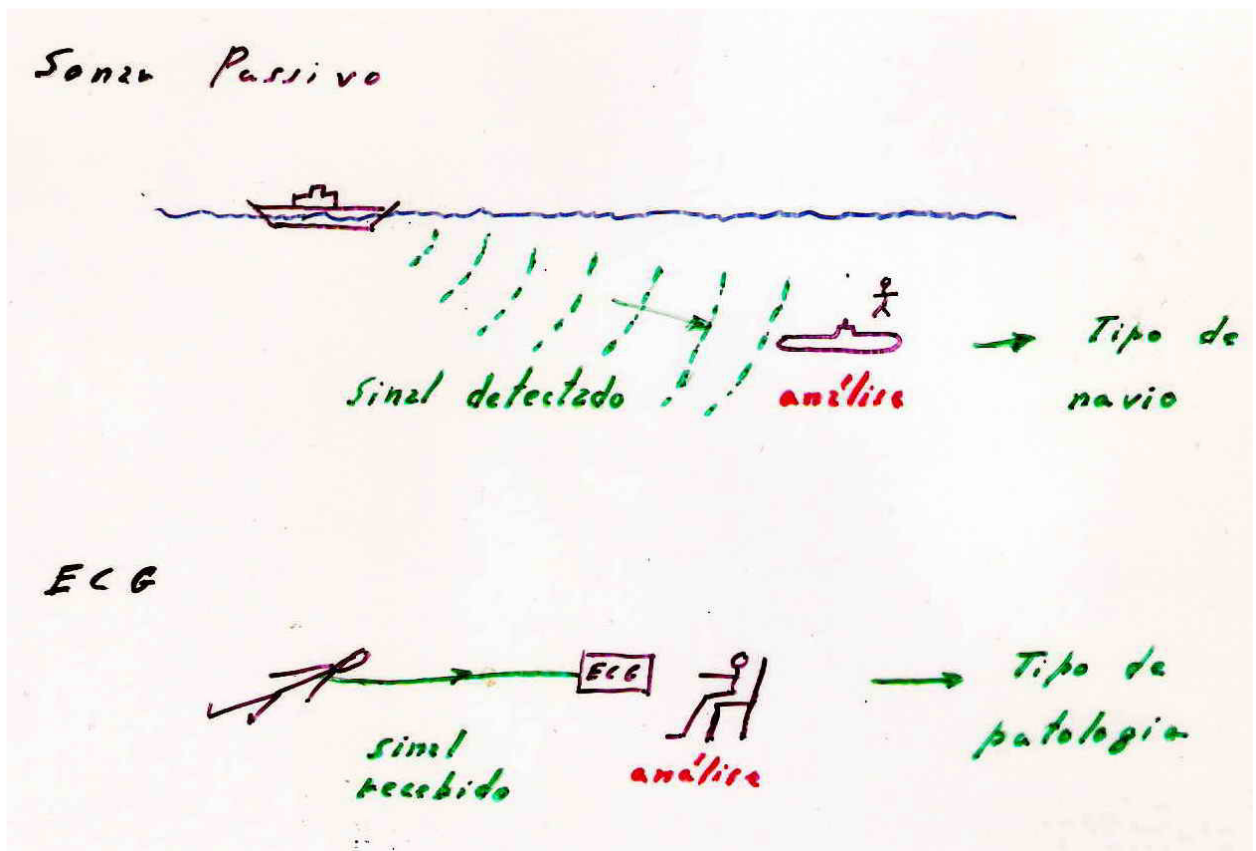
Associam saídas contínuas às entradas

Saídas contínuas $y_i \in (-1, +1)$

Classificadores

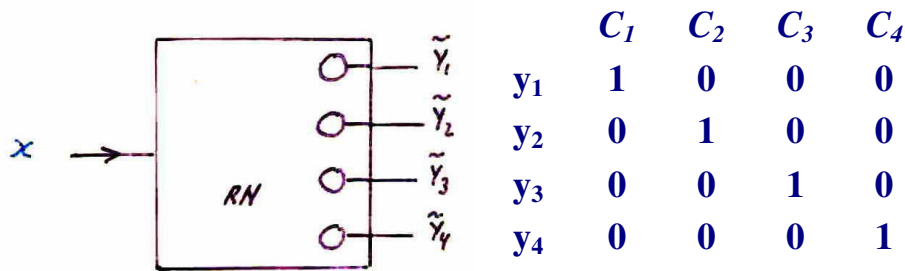
Associam saídas lógicas às entradas (classificam as entradas)

Saídas lógicas $y_i \in \{0, 1\}$



2.1 Convenção / Estratégia:

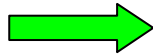
Codificação da Saída (maximamente esparsa)



Cada saída separa a sua classe das demais, i.e.

classe versus não classe (um contra todos)

Separação de N classes

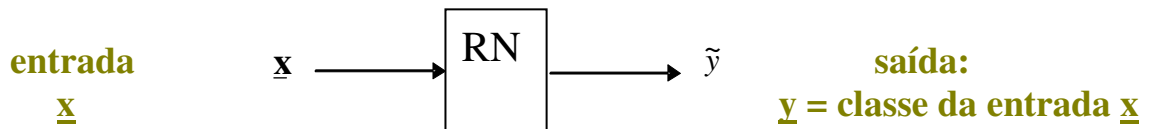


Separação de duas classes:

classe vs. não classe

3 - Tipos de classificação / Tipos de aprendizado

Classificadores:



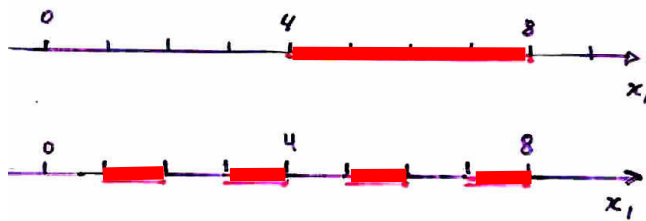
Critérios de classificação

Critérios “não conhecidos”, arbitrários

Critérios conhecidos, determinados

Classificadores por similaridade

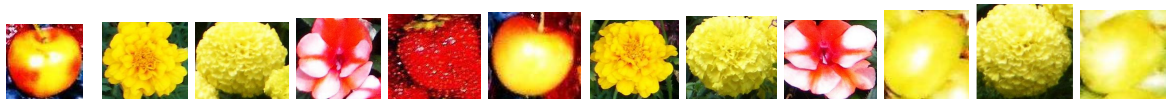
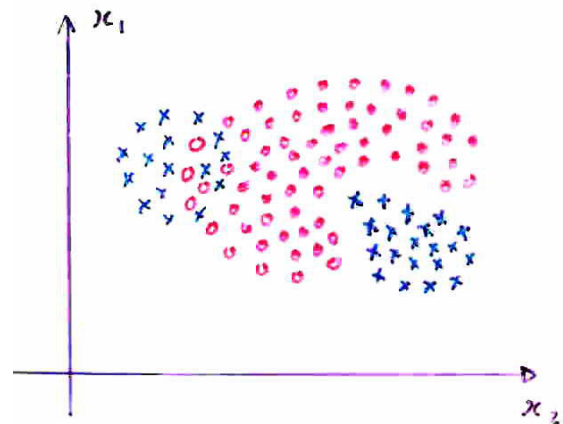
Classificações Arbitrárias



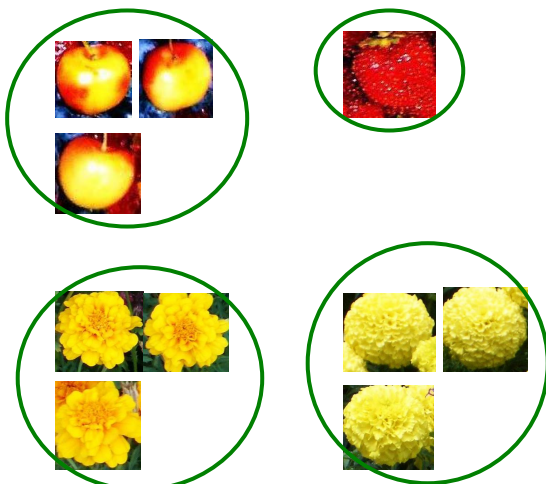
Regiões de fusão

(confusão)

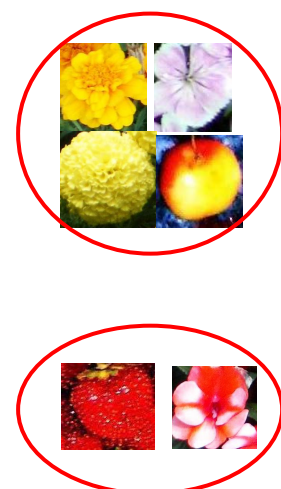
entre classes



Agrupamentos por similaridade

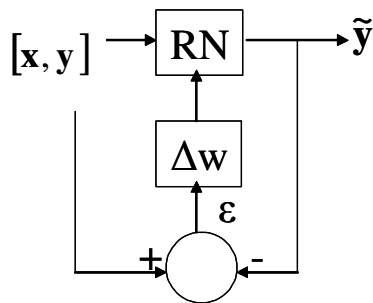


Agrupamentos arbitrários

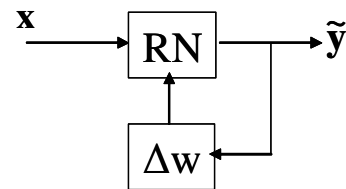


3.1 - Tipo de aprendizado

Treinamento supervisionado / não-supervisionado



Pode realizar
classificações arbitárias



Necessita algum critério:
Similaridade

4 - Classificação, Agrupamento ou Clusterização por Similaridade

Agrupar, “de forma natural”,
conjuntos de dados com “similaridade interna”

Classes agrupam elementos `similares` entre si ‘de forma natural’ mas

O que significa “de forma natural ?”

O que significa “similaridade interna ?”

Quais são as classes ?

Classificadores supervisionados

As classes são informadas

Ação: apenas classifica as entradas

Classificadores **não** supervisionados

As classes **não** são informadas

Ações: **determina as classes e**
classifica as entradas

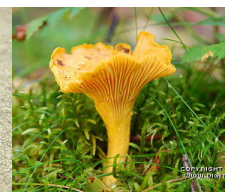
Sobrevivência vs.

Classificação por similaridade

Aprendizado não supervisionado

Plasticidade

Animais e vegetais venenosos ou comestíveis ?



4.1 – Dados reais vs. Representação matemática na

Classificação por Similaridade

objetos físicos O_1 O_2  objetos matemáticos
vetores $\underline{x}_1$ $\underline{x}_2$

similaridade física  similaridade matemática

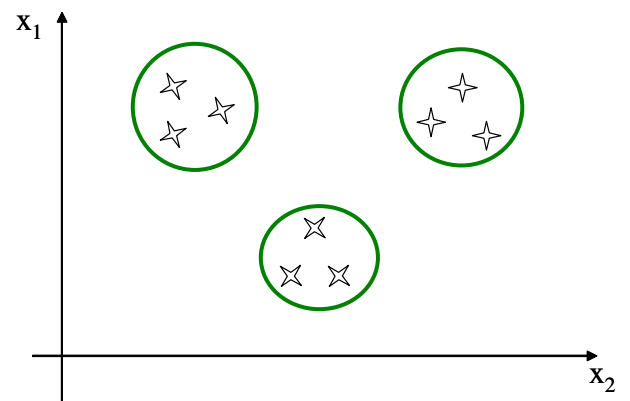
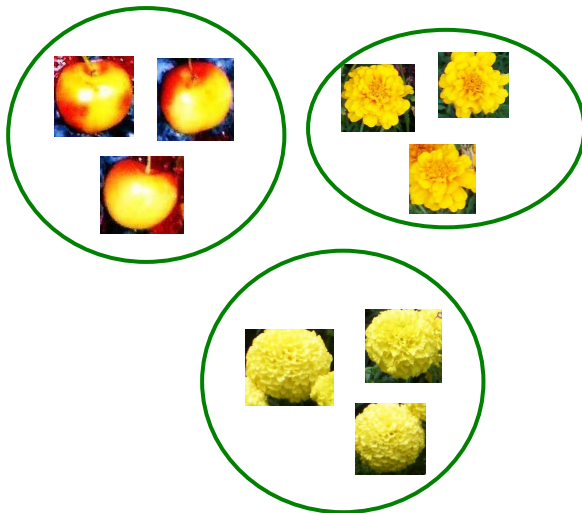
$$O_1 \approx O_2$$

$$\underline{x}_1 \cong \underline{x}_2 \quad \text{ou} \quad |\underline{x}_1 - \underline{x}_2| \ll \ll$$

$$O_1 \approx O_2$$



$$\underline{x}_1 \cong \underline{x}_2$$



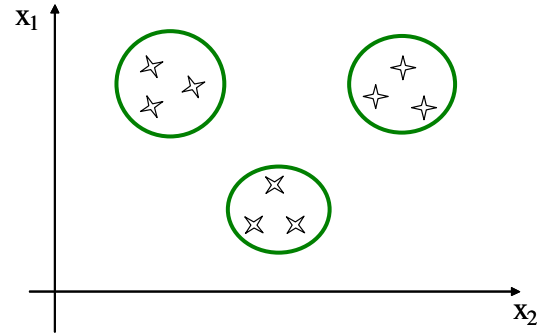
**Classificar por similaridade : alocar na mesma classe
vetores que estão relativamente próximos entre si.**

Como traduzir matematicamente

“relativamente próximos entre si ?”

Como determinar as classes ?

Como realizar a alocação ?



Noções iniciais sobre

Clusterização

Agrupamento ou

Classificação

por similaridade

5 - Características de um Agrupamento

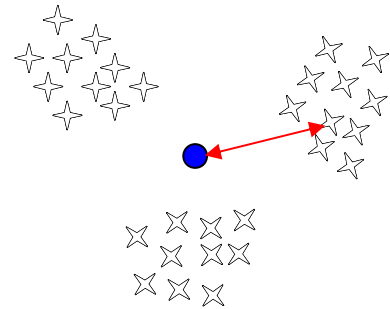
5.1 - Características do conjunto de todos os elementos

Média

$$\vec{m} = E_{\forall i} \vec{x}_i = \frac{1}{N} \sum_{i=1}^N \vec{x}_i$$

Dispersão, dissimilaridade total

$$F_0 = \sum |\vec{x}_i - \vec{m}|^2$$



m e F_0 são independentes da clusterização

5.2 Características de uma classe C_j

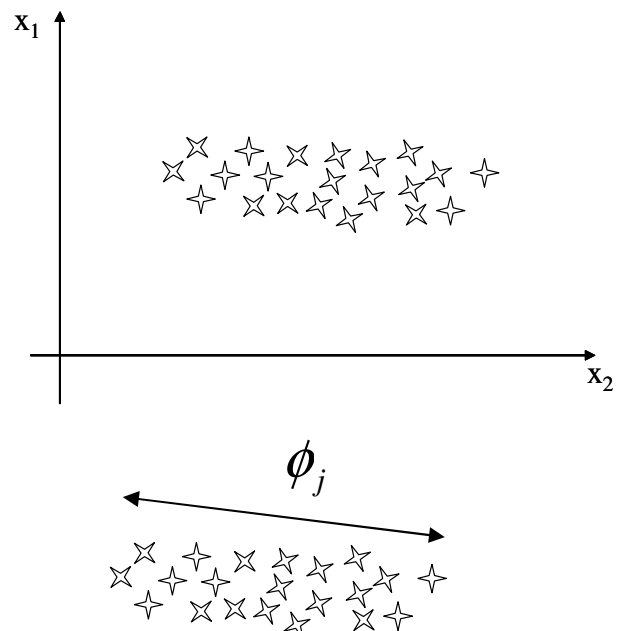
Características intra classe

Classe C_j

número de elementos n_j

Diâmetro da classe:

$$\phi_j = \underset{\forall \vec{x}_i, \vec{x}_k \in C_j}{\text{Max}} (|\vec{x}_i - \vec{x}_k|)$$



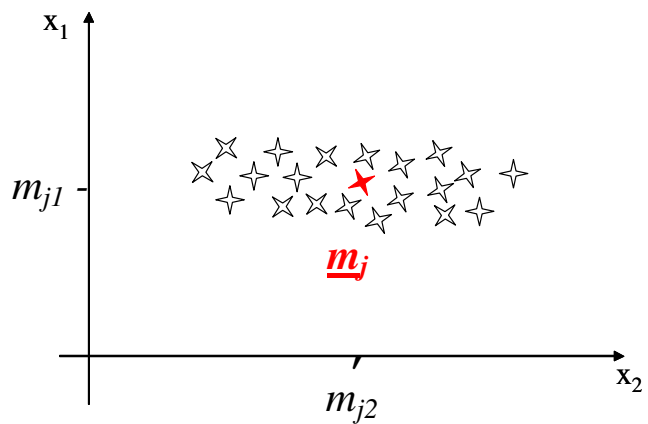
Centro, Baricentro ou Padrão da classe

$$\vec{m}_j = E_{\forall \vec{x}_i \in C_j} \vec{x}_i = \frac{1}{n_j} \sum_{\forall \vec{x}_i \in C_j} \vec{x}_i$$

$$\vec{m}_j = [m_{j1} \dots m_{jk} \dots]^t$$

por componente k

$$m_{jk} = E_{\forall \vec{x}_i \in C_j} x_{ik} = \frac{1}{n_j} \sum_{\forall \vec{x}_i \in C_j} x_{ik}$$

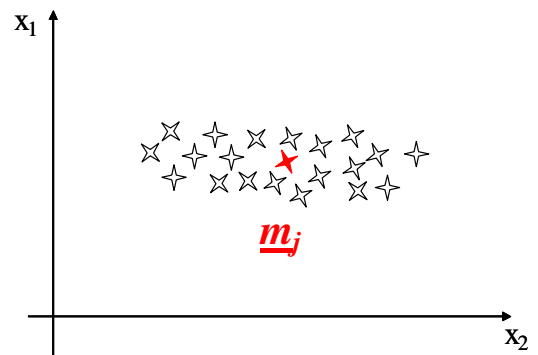


Dispersão média intra classe, Erro (médio quadrático) de representação

$$\sigma_j^2 = E_{\forall \vec{x}_i \in C_j} \|\vec{x}_i - \vec{m}_j\|^2 = \frac{1}{n_j} \sum_{\forall \vec{x}_i \in C_j} \|\vec{x}_i - \vec{m}_j\|^2$$

por componente k

$$\sigma_{jk}^2 = E_{\forall \vec{x}_i \in C_j} (x_{ik} - m_{jk})^2 = \frac{1}{n_j} \sum_{\forall \vec{x}_i \in C_j} (x_{ik} - m_{jk})^2$$



$$\text{e } \sigma_j^2 = \sum_{\forall k} \sigma_{jk}^2$$

Obs: Valor do padrão p_j que minimiza a dispersão intra classe da classe j :

$$\sigma_j^2 = \sum_{\forall k} \sigma_{jk}^2 \quad \sigma_{jk}^2 = \frac{1}{n_j} \sum_{\forall \vec{x} \in C_j} (x_{jk} - p_{jk})^2$$

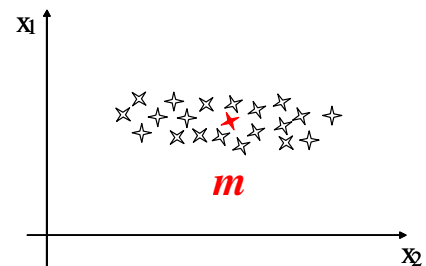
após pequena álgebra

$$\frac{\partial \sigma_{jk}^2}{\partial p_{jk}} = 0 \quad \Rightarrow \quad p_{jk} = m_{jk} \quad \Rightarrow \quad \vec{p}_j = \vec{m}_j$$

o padrão que minimiza a dispersão intra classe (ou erro de representação) de uma classe é o seu baricentro

Dispersão total intra classe da classe C_j :

$$F_j = n_j \sigma_j^2 = \sum_{\forall \vec{x}_i \in C_j} \|\vec{x}_i - \vec{m}_j\|^2$$

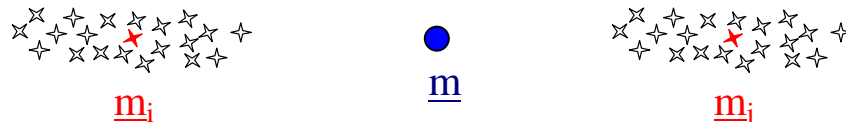


Dispersão total intra classe para todas as classes:

$$F_{in} = \sum_{\forall C_j} F_j \quad \geq 0$$

F_{in} - parâmetro à minimizar

5.3 – Características (medidas) Inter Classes



M – número de classes

Dissimilaridade total inter classes

$$F_{out} = \sum_{\forall j} n_j \left\| \vec{m}_j - \vec{m} \right\|^2 \geq 0$$

F_{out} - parâmetro à maximizar

Para um bom agrupamento

escolher as classes de forma a

Minimizar a dispersão intra classe total F_{in}

Maximizar a dissimilaridade inter classes total F_{out}

Com alguma álgebra é possível mostrar que

$$F_{in} + F_{out} = F_0 = \text{constante, independente da clusterização}$$

Logo a clusterização que minimiza F_{in}

ao mesmo tempo maximiza F_{out} , i.e.,

basta que o processo de clusterização minimize F_{in}

Obs 1:

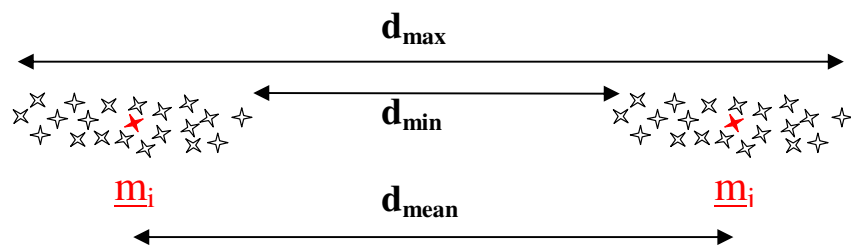
Usamos como medida inter classes a maximizar a

Dissimilaridade total inter classes

$$F_{out} = \sum_{\forall j} n_j \left\| \vec{m}_j - \vec{m} \right\|^2 \geq 0$$

mas existem outras medidas de dissimilaridade (distância, separação) inter classes:

Outras medidas de dissimilaridade (distância, separação) inter classes:



vizinho + próximo

$$d_{\min}(\mathcal{X}_i, \mathcal{X}_j) = \min_{\mathbf{x} \in \mathcal{X}_i, \mathbf{x}' \in \mathcal{X}_j} \|\mathbf{x} - \mathbf{x}'\|$$

vizinho + distante

$$d_{\max}(\mathcal{X}_i, \mathcal{X}_j) = \max_{\mathbf{x} \in \mathcal{X}_i, \mathbf{x}' \in \mathcal{X}_j} \|\mathbf{x} - \mathbf{x}'\|$$

distância média

$$d_{\text{avg}}(\mathcal{X}_i, \mathcal{X}_j) = \frac{1}{n_i n_j} \sum_{\mathbf{x} \in \mathcal{X}_i} \sum_{\mathbf{x}' \in \mathcal{X}_j} \|\mathbf{x} - \mathbf{x}'\|$$

distância entre médias

$$d_{\text{mean}}(\mathcal{X}_i, \mathcal{X}_j) = \|\mathbf{m}_i - \mathbf{m}_j\|.$$

(assim como existem também outras medidas de dispersão intra classe)

6 – Processos de Agrupamento

Como realizar a

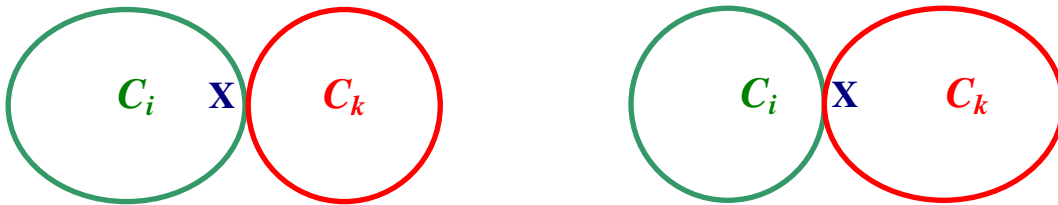
Minimização interativa de F_{in}

(e simultaneamente a maximização de F_{out}) ?

$$F_{in} = \sum_j F_j$$

Rearranjando os elementos entre as classes para reduzir F_{in}

Deslocando um elemento $\hat{x}$ da classe C_i para a classe C_k



Após alguma álgebra:



$$\bar{m}_k = \frac{1}{n_k} \sum_{\forall \vec{x} \in C_k} \vec{x} \quad \Rightarrow \quad \bar{m}'_k = \frac{1}{n'_k} \sum_{\forall \vec{x} \in C'_k} \vec{x} = \frac{1}{n_k + 1} (\hat{x} + n_k \bar{m}_k) = \bar{m}_k + \frac{\hat{x} - \bar{m}_k}{n_k + 1}$$

$$F_k = \sum_{\forall \vec{x} \in C_k} |\vec{x} - \bar{m}_k|^2 \quad \Rightarrow \quad F'_k = \sum_{\forall \vec{x} \in C'_k} |\vec{x} - \bar{m}'_k|^2 = \dots = F_k + \frac{n_k}{n_k + 1} |\hat{x} - \bar{m}_k|^2$$

$$\bar{m}_i = \frac{1}{n_i} \sum_{\forall \vec{x} \in C_i} \vec{x} \quad \Rightarrow \quad \bar{m}'_i = \frac{1}{n'_i} \sum_{\forall \vec{x} \in C'_i} \vec{x} = \frac{1}{n_i - 1} (-\hat{x} + n_i \bar{m}_i) = \bar{m}_i - \frac{\hat{x} - \bar{m}_i}{n_i - 1}$$

$$F_i = \sum_{\forall \vec{x} \in C_i} |\vec{x} - \bar{m}_i|^2 \quad \Rightarrow \quad F'_i = \sum_{\forall \vec{x} \in C'_i} |\vec{x} - \bar{m}'_i|^2 = \dots = F_i - \frac{n_i}{n_i - 1} |\hat{x} - \bar{m}_i|^2$$

$$\Delta F_{in} = (F'_k - F_k) + (F'_i - F_i) = \frac{n_k}{n_k + 1} |\hat{x} - \bar{m}_k|^2 - \frac{n_i}{n_i - 1} |\hat{x} - \bar{m}_i|^2$$

F_{in} é reduzido quando

deslocamos um elemento $\hat{x}$ da classe C_i para a classe C_k

se

$$\frac{n_i}{n_i - 1} \left| \hat{x} - \vec{m}_i \right| > \frac{n_k}{n_k + 1} \left| \hat{x} - \vec{m}_k \right|^2$$

como $\frac{n_i}{n_i - 1} > 1 > \frac{n_k}{n_k + 1}$ podemos simplificar a condição, basta que

$$\left| \hat{x} - \vec{m}_i \right| > \left| \hat{x} - \vec{m}_k \right|^2$$

i.e. que a distância do elemento ao baricentro da nova classe seja menor que ao da antiga.

6.1 Um processo de clusterização iterativo simples e eficaz:

K-means, C-means, Isodata ou Lloyd

entradas: $\vec{x}_i \quad i = 1, 2, \dots, N$

centros de classe: $\vec{m}_j \quad j = 1, 2, \dots, K$

Inicialização: Arbitre K centros $\vec{m}_j \quad j = 1, 2, \dots, K$

sugestão: use as K primeiras entradas $\vec{x}_i \quad i = 1, 2, \dots, K$
 não muito próximas entre si (à discutir mais tarde)

Loop**Reclassificação das entradas**

$$\forall \vec{x}_i \quad i = 1, 2, \dots, N$$

Se $\vec{m}_j$ é a melhor representação de $\vec{x}_i$, i.e. se

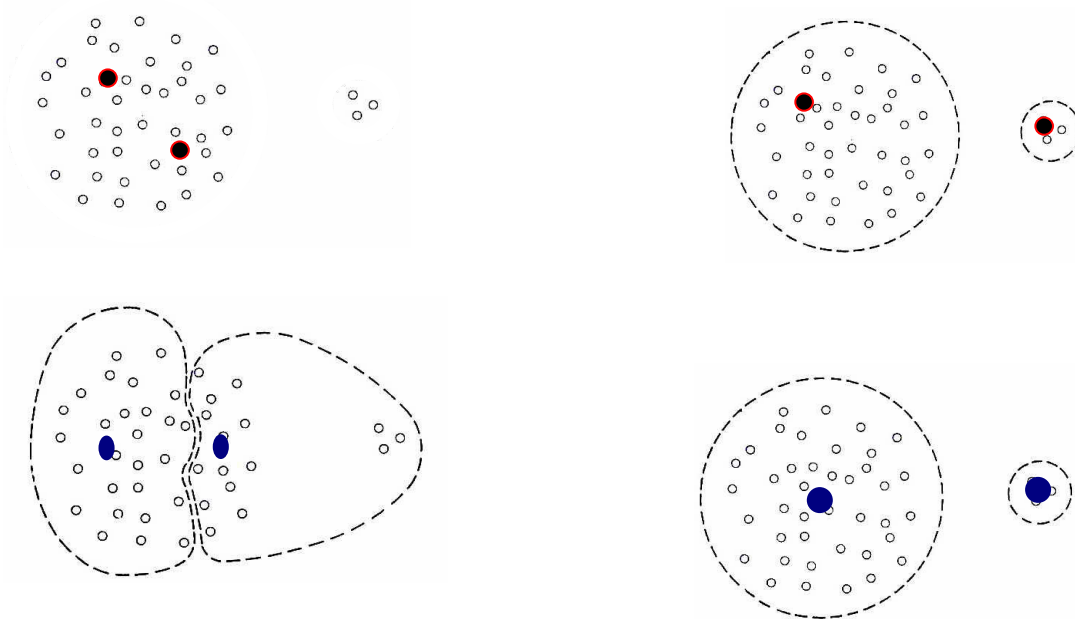
$$|\vec{x}_i - \vec{m}_j| < |\vec{x}_i - \vec{m}_k| \quad \forall k \neq j \quad \textbf{então} \quad \vec{x}_i \in C_j$$

Atualização dos centros de classe

$$\forall \quad j = 1, 2, \dots, K$$

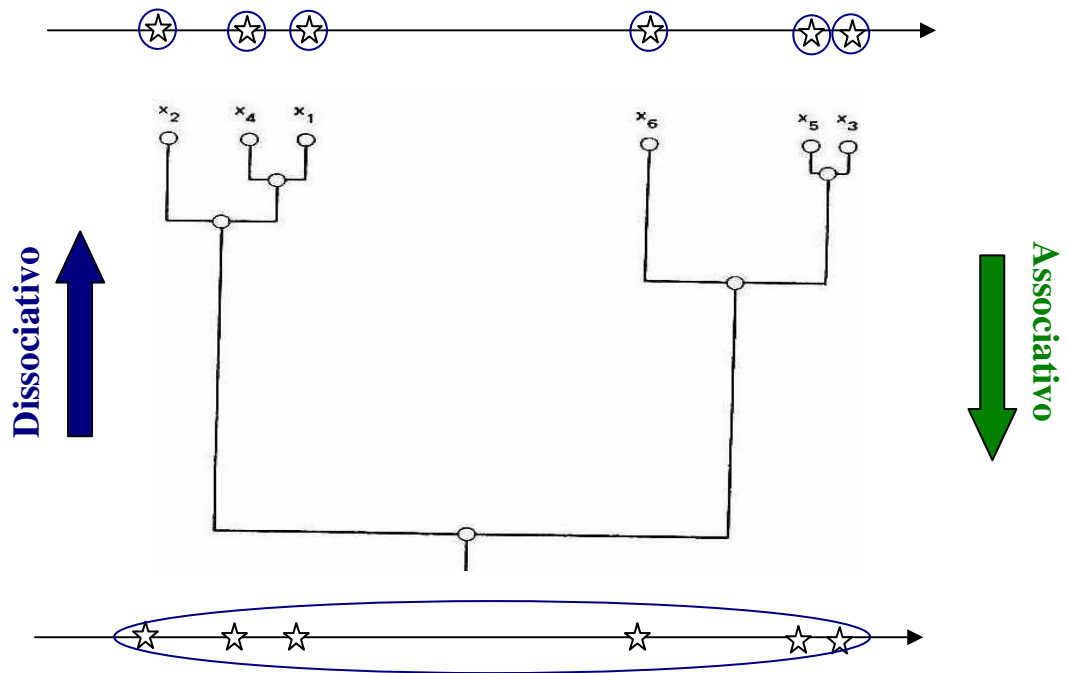
$$\vec{m}_j = \frac{1}{n_j} \sum_{\vec{x} \in C_j} \vec{x}$$

Quando nenhum elemento trocar de classe: Fim.



inicializar com entradas não muito próximas

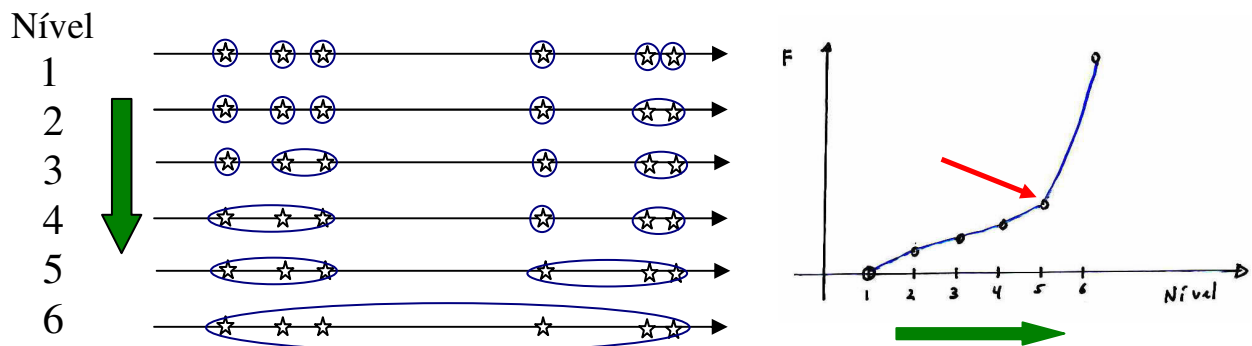
7 - Algoritmos Hierárquicos: dendogramas, árvores



7.1 Algoritmo Aglomerativo, Associativo:

n elementos c classes

$$\begin{array}{ccccccc}
 c = & n & & n-1 & & n-2 & \dots & 2 & & 1 \\
 F_{in} = & F_n=0 & < & F_{n-1} & < & F_{n-2} & < & F_2 & < & F_1
 \end{array}$$



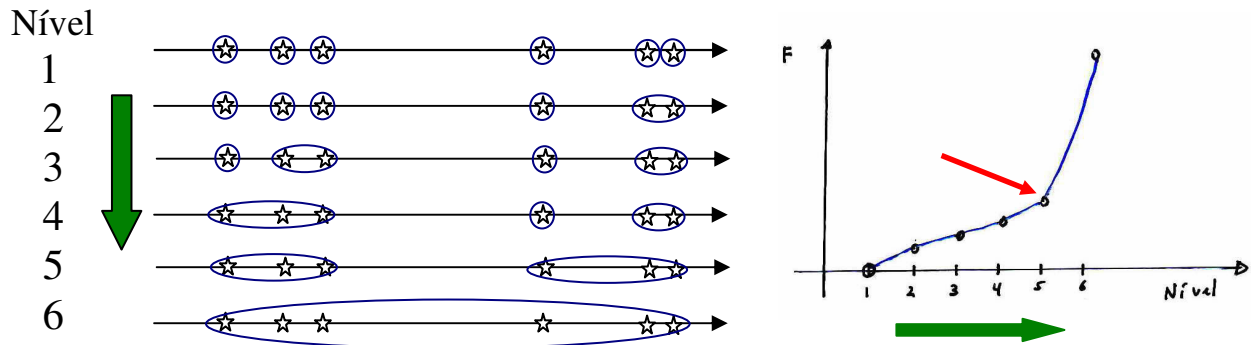
Algoritmo Hierárquico Associativo:

Inicialização: n elementos $\implies n$ clusters

Loop:

Junte os dois clusters que aumentam F_{in} o mínimo possível

Repetir até chegar ao número de clusters desejado ou F_{in} crescer muito



Obs: se $\underline{x}_i$ e $\underline{x}_j$ estão em um mesmo cluster, permanecerão juntos. É dito um **algoritmo hierárquico**.

Que clusters juntar ?

$$\begin{aligned}
 C_j \quad \vec{m}_j &= \frac{1}{n_j} \sum_{\forall \vec{x} \in C_j} \vec{x} & F_j &= \sum_{\forall \vec{x} \in C_j} |\vec{x}|^2 - n_j |\vec{m}_j|^2 \\
 C_k \quad \vec{m}_k &= \frac{1}{n_k} \sum_{\forall \vec{x} \in C_k} \vec{x} & F_k &= \sum_{\forall \vec{x} \in C_k} |\vec{x}|^2 - n_k |\vec{m}_k|^2 \\
 C_{jk} = C_j \cup C_k \quad \vec{m}_{jk} &= \frac{n_j \vec{m}_j + n_k \vec{m}_k}{n_j + n_k} & F_{jk} &= \sum_{\forall \vec{x} \in C_j, C_k} |\vec{x}|^2 - (n_j + n_k) |\vec{m}_{jk}|^2 \\
 \Delta F &= F_{jk} - (F_j + F_k) = \frac{n_j n_k}{n_j + n_k} |\vec{m}_j - \vec{m}_k|^2
 \end{aligned}$$

$$\Delta F = \frac{n_j n_k}{n_j + n_k} \left| \vec{m}_j - \vec{m}_k \right|^2 \quad \text{deve ser pequeno}$$

escolher centros próximos:

$$\vec{m}_j \approx \vec{m}_k$$

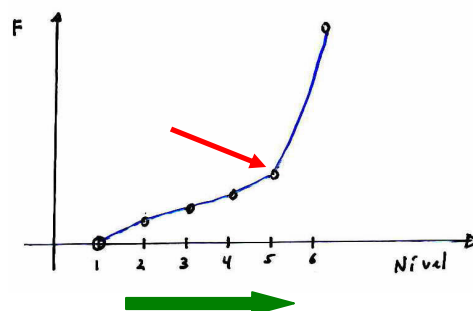
de preferência com populações desiguais:

$$\frac{n_j n_k}{n_j + n_k} \approx \begin{cases} n_j & \text{se } n_j \ll n_k \\ \frac{n_j}{2} & \text{se } n_j \approx n_k \end{cases}$$

Quando parar o processo ?

Um agrupamento “natural” (um número “bom” de classes) é alcançado no algoritmo associativo imediatamente antes de F_{in} apresentar um grande acréscimo em um único passo.

Por este critério o ponto de parada está indicado pelas seta vermelha no exemplo anterior.

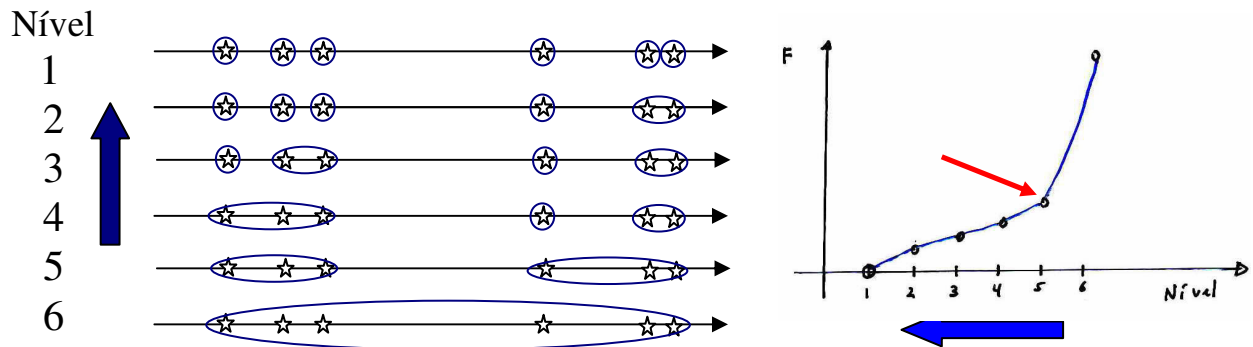


7.2 Algoritmo Divisivo, Dissociativo:

n elementos **c** classes

$$c = \quad 1 \quad 2 \quad 3 \quad \dots \quad n-1 \quad n$$

$$F_{in} = \quad F_1 < F_2 < F_3 \quad < \quad F_{n-1} < F_n=0$$



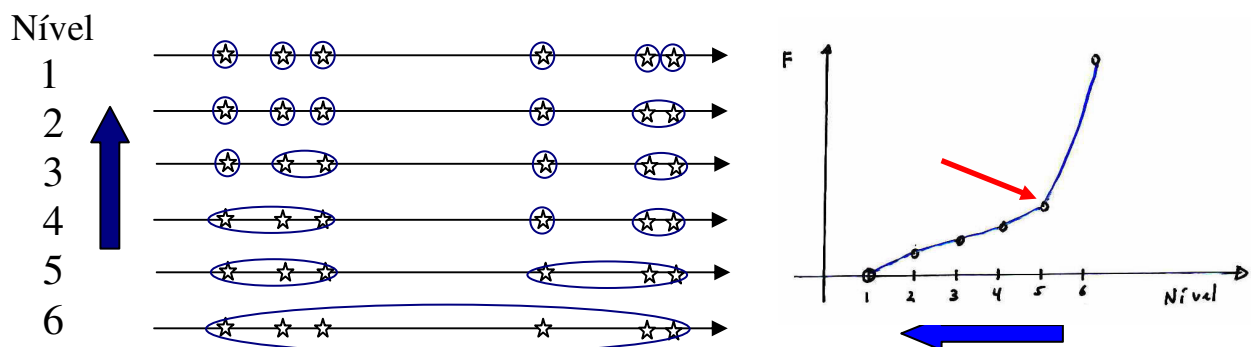
Algoritmo Hierárquico Dissociativo:

Inicialização: **n** elementos $\Rightarrow$ 1 cluster

Loop:

Divida o cluster que reduz F_{in} o máximo possível

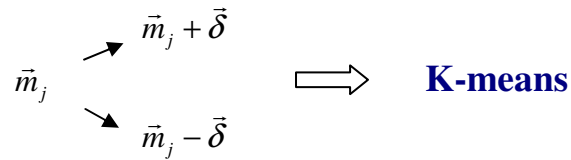
Repetir até chegar ao número de clusters desejado ou F_{in} quase não decair



Que cluster dividir ?

Critério: maior F_j , por exemplo

Método:



Quando parar o processo ?

Um agrupamento “natural” (um número “bom” de classes) é obtido no algoritmo dissociativo imediatamente após F_{in} apresentar um grande decréscimo em um único passo.

Por este critério o ponto de parada está indicado pela seta vermelha no exemplo anterior.

