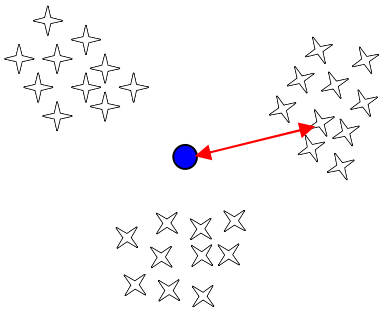
Características de um Agrupamento

1 - Características do conjunto de todos os elementos

- **Número total de elementos,** n_0
- **Média, baricentro global**

$$\vec{m}_0 = \underset{\forall i}{E}\vec{x}_i = \frac{1}{n_0}\sum_{i=1}^{N}\vec{x}_i$$



- **Dispersão, dissimilaridade total**

$$F_0 = \sum \left|\vec{x}_i - \vec{m}_0\right|^2$$

Desvio Padrão Total

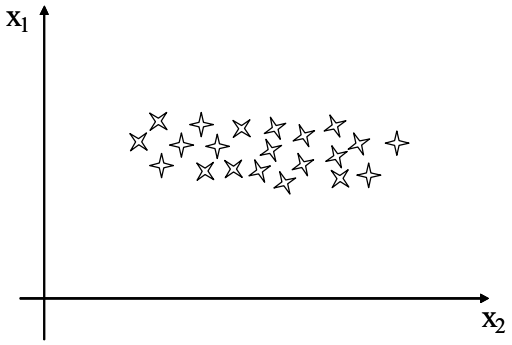
$$\sigma_0 = \sqrt{\frac{F_0}{n_0}} = \sqrt{\frac{1}{n_0}\sum \left|\vec{x}_i - \vec{m}_0\right|^2}$$

m_0, σ_0 e F_0 são independentes da clusterização usada

2 - Características de uma classe C_j

Características intra classe

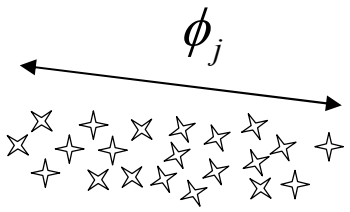
Classe C_j



- **Número de elementos da classe n_j**

- **Diâmetro da classe:**

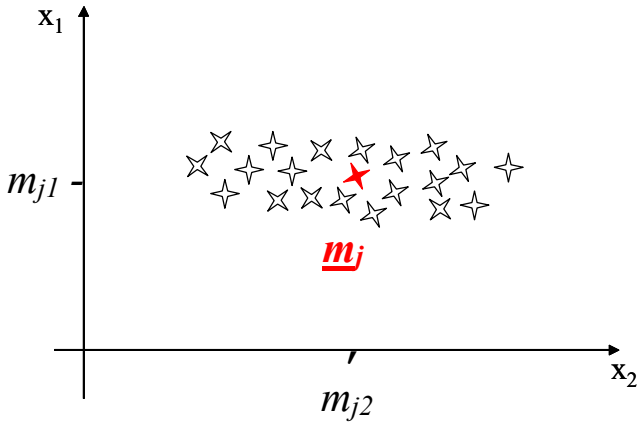
$$\phi_j = \underset{\forall \vec{x}_i, \vec{x}_k C_j}{Max} \left(\left| \vec{x}_i - \vec{x}_k \right| \right)$$



- **Baricentro (ou Média) da classe**

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

$$\vec{m}_j = \left[m_{j1} \ldots m_{jk} \ldots \right]^t$$



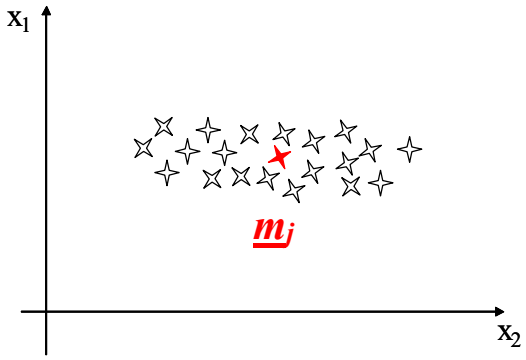
por componente k

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

- **Variância intra classe,**

Erro (médio quadrático) de representação da classe

$$\sigma_j^2 = \underset{\forall \vec{x} \in C_j}{E} \left| \vec{x} - \vec{m}_j \right|^2 = \frac{1}{n_j} \sum_{\forall \vec{x} \in C_j} \left| \vec{x} - \vec{m}_j \right|^2$$



por componente k

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

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

- **Desvio Padrão intra classe –**

Erro RMS (erro eficaz) de representação

$$\sigma_j = \sqrt{\sigma_j^2} = \left[\frac{1}{n_j} \sum_{\forall \vec{x} \in C_j} \left| \vec{x} - \vec{m}_j \right|^2 \right]^{1/2}$$

σ_j - parâmetro à minimizar

- **Padrão da Classe**

Critério: representa os pontos da classe com o menor erro, minimiza a dispersão intra classe (variância) da classe j:

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

$$\frac{\partial \sigma_{jk}^2}{\partial p_{jk}} = 0 \qquad p_{jk} = ?$$

$$\frac{\partial \sigma_{jk}^2}{\partial p_{jk}} = \frac{\partial}{\partial p_{jk}} \frac{1}{n_j} \sum_{\forall \vec{x} \in C_j} \left(x_{jk} - p_{jk} \right)^2 = \frac{\partial}{\partial p_{jk}} \frac{1}{n_j} \sum_{\forall \vec{x} \in C_j} \left(x_{jk}^2 - 2 p_{jk} x_{jk} + p_{jk}^2 \right) =$$

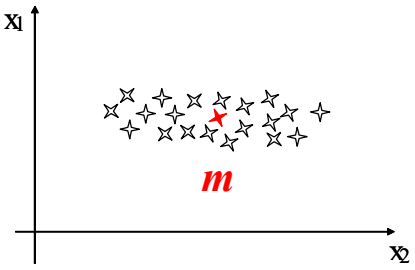
$$= \frac{\partial}{\partial p_{jk}} \left[\underset{\forall \vec{x} \in C_j}{E} x_{jk}^2 - 2 p_{jk} \underset{\forall \vec{x} \in C_j}{E} x_{jk} + p_{jk}^2 \right] = \frac{\partial}{\partial p_{jk}} \left[\underset{\forall \vec{x} \in C_j}{E} x_{jk}^2 - 2 p_{jk} m_{jk} + p_{jk}^2 \right] =$$

$$= -2 m_{jk} + 2 p_{jk} = 0 \quad \Rightarrow \quad p_{jk} = m_{jk} \quad \Rightarrow \quad \boxed{\vec{p}_j = \vec{m}_j}$$

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

- **Dispersão total intra classe da classe *Cj:***

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

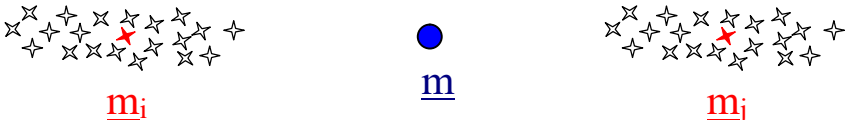


Dispersão total intra classe para todas as classes:

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

F_{in} - parâmetro à minimizar

3 – Características (medidas) Inter Classes



- **Número de classes M**

- **Dissimilaridade total inter classes**

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

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

$F_{in} + F_{out} = F_0$ = 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:

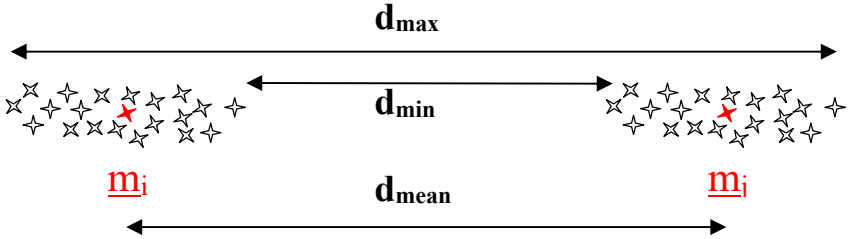
Critério: usamos como medida inter classes a maximizar a

Dissimilaridade total inter classes

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

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

Outras medidas de disssimilaridade (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)

Em resumo:

Para um bom agrupamento

escolher as classes de forma a

Minimizar a dispersão intra classe total F_{in}