



Definição dos clusters - Parte III

≡ Ciclo	Ciclo 08: Outros algoritmos Clusterização
# Aula	66
🕒 Created	@June 28, 2023 11:39 AM
☑ Done	<input type="checkbox"/>
☑ Ready	<input checked="" type="checkbox"/>

Objetivo da Aula:

- ☐ Recapitulando
- ☐ Exemplo prático
- ☐ Resumo
- ☐ Próxima aula

Conteúdo:

▼ 1. Recapitulando

Os passos para encontrar os grupos (clusters) formados pelos dados, usando o algoritmo de Affinity Propagation são os seguintes:

1. Definição da **métrica de similaridade**
2. Cálculo da similaridade entre todos os pontos do conjunto de dados, formando a **matriz de similaridade (S)**
3. **Até o número n de repetições** ser alcançada ou a variação dos valores das matrizes de responsabilidade e disponibilidade for menor do que um valor ϵ , faça:
 - a. Cálculo da **matriz de responsabilidade**
 - b. Cálculo da **matriz de disponibilidade**
4. Para cada ponto, some os valores da matriz de responsabilidade e disponibilidade, formando a **matriz de critério**
5. Atribua o mesmo **cluster** para os pontos que possuem **o mesmo valor de critério**.

▼ 2. Exemplo prático

A matriz abaixo mostra a avaliação individual de 5 pessoas para cada um dos seguintes filmes: Matrix Reloaded, Coringa, Interestelar, Vingadores: Ultimato e Gladiador. As notas variam de 1 a 5, sendo: 1 - muito ruim, 2 - ruim, 3 - razoável, 4 - bom e 5 - muito bom.

Participantes	Matrix Reloaded	Coringa	Interestelar	Vingadores: Ultimato	Gladiador
Alice	3	4	3	2	1
Bob	4	3	5	1	1
Cary	3	5	3	3	3

Doug	2	1	3	3	2
Edna	1	1	3	2	3

A sua tarefa é encontrar grupos de pessoas formadas a partir da similaridade entre as avaliações atribuídas a cada filme.

▼ Passo 1: Definição da métrica de similaridade

▼ Distância **Euclidiana**:

$$d_{euclidiana} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2}$$

▼ Distância **Manhattan**:

$$d_{manhattan} = \sum_{i=1}^n |x_i - y_i|$$

▼ Distância **Chebychev**:

$$d_{chebychev} = \max_{i=1}^n (|x_i - y_i|)$$

▼ Distância **Minkowski**:

$$d_{minkowski} = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{\frac{1}{p}}$$

▼ Distância **Cosine**:

$$d_{cosine} = 1 - \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}}$$

▼ Distância **Pearson**:

$$d_{pearson} = 1 - \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

▼ Distância **Mahalanobis**:

$$d_{mahalanobis} = \sqrt{(\vec{x} - \vec{y})^T S^{-1} (\vec{x} - \vec{y})}$$

▼ Distância **SED**:

$$d_{SED}(s, t) = \sum_{i=1}^n \begin{cases} 0 & \text{se } s_i = t_i \\ 1 & \text{se } s_i \neq t_i \end{cases}$$

▼ Distância **Jaccard**:

$$d_{jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

▼ Distância **Levenshtein**:

▼ Distância Sorensen-Dice:

▼ Distância Jensen-Shannon:

Onde:

▼ Distância **Canberra**:

▼ Distância Hamming:

▼ Distância Spearman:

onde: $s_d[i]$ é a diferença entre as posições dos elementos $s[i]$ em duas sequências ordenadas, e n é o tamanho das sequências.

▼ Distância Chi-Square:

▼ Passo 2: Matriz de similaridade (S)

Calculando a matriz de similaridade (S) com a distância Euclideana, temos:

▼ Exemplo:

▼ Matriz de avaliações



Participantes	Matrix Reloaded	Coringa	Interestelar	Vingadores: Ultinato	Gladiador
Alice	3	4	3	2	1
Bob	4	3	5	1	1
Cary	3	5	3	3	3
Doug	2	1	3	3	2
Edna	1	1	3	2	3

▼ Processo de construção da matriz similaridade (S)

Participantes	Matrix Reloaded	Coringa	Interestelar	Vingadores: Ultinato	Gladiador
Alice	3	4	3	2	1
Bob	4	3	5	1	1
Cary	3	5	3	3	3
Doug	2	1	3	3	2
Edna	1	1	3	2	3

▼ Matriz similaridade (S)

Participantes	Alice	Bob	Cary	Doug	Edna
Alice	-22	-7	-6	-12	-11
Bob	-7	-22	-17	-17	-15
Cary	-6	-17	-22	-18	-15
Doug	-12	-17	-18	-22	9
Edna	-17	-22	-21	-3	-19

▼ **Passo 3: Matriz de responsabilidade (R)**

$$R(i, k) = S(i, k) - \max_{k' \neq k} \{D(i, k') + S(i, k')\}$$

▼ **Exemplo:**

▼ Processo de construção da matriz responsabilidade (R)

▼ Matriz responsabilidade (R)

Participantes	Alice	Bob	Cary	Doug	Edna
Alice	-16	-1	1	-6	-11
Bob	10	-15	-10	-10	-15
Cary	11	-11	-16	-12	-15
Doug	-9	-14	-15	-19	9
Edna	-14	-19	-18	14	-19

▼ **Passo 4: Matriz de disponibilidade (D)**

Para calcular os valores da diagonal principal, use a seguinte fórmula:

$$D(k, k) = \sum_{i' \neq k} \max\{0, R(i', k)\}$$

Para calcular os valores fora da diagonal principal, use a fórmula:

$$D(i, k) = \min\{0, R(k, k) + \sum_{i' \neq k} \max\{0, R(i', k)\}\}$$

▼ Exemplo:

▼ Processo de construção da matriz disponibilidade (D)

▼ Matriz disponibilidade (D)

Participantes	Alice	Bob	Cary	Doug	Edna
Alice	21	-15	-16	-5	-10
Bob	-5	0	-15	-5	-10
Cary	-6	-15	1	-5	-10
Doug	0	-15	-15	14	-19
Edna	0	-15	-15	-19	9

▼ 3. Resumo

1. Definição da métrica de similaridade
2. Cálculo da similaridade entre todos os pontos do conjunto de dados, formando a matriz de similaridade (S)
3. Até o número n de repetições ser alcançada ou a variação dos valores das matrizes de responsabilidade de disponibilidade for menor do que um valor e, faça:
 - a. Cálculo da matriz de responsabilidade
 - b. Cálculo da matriz de disponibilidade

▼ 4. Próxima aula

Definição dos clusters - Parte IV