Definição dos clusters - Parte I

Ciclo	Ciclo 08: Outros algoritmos Clusterização
# Aula	64
Created	@June 27, 2023 8:27 AM
Done	☐
Ready	☑

Objetivo da Aula:

- ☐ Os 5 passos do treinamento
- ☐ Exemplo prático
- ☐ Resumo
- ☐ Próxima aula

Conteúdo:

▼ 1. Os 5 passos do treinamento

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. Calculo 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. 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 Negative Squared Euclidean:

$$d_{neg\_euclidiana} = -(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2$$

▼ Distância Euclidiana:

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

▼ 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:

$$d_{Levenshtein}(s,t) = \begin{cases} 0, & \text{if } s = t \\ d_{Levenshtein}(s[1..i], t[1..j-1]) + 1, & \text{if } s[i] \neq t[j] \\ d_{Levenshtein}(s[1..i-1], t[1..j-1]), & \text{if } s[i] = t[j] \\ d_{Levenshtein}(s[1..i-1], t[1..j]) + 1, & \text{if } s[i] = t[j] \\ d_{Levenshtein}(s[1..i], t[1..j-1]) + 1, & \text{if } s[i] = t[j] \end{cases}$$

▼ Distância Sorensen-Dice:

$$d_{sorensen}(A,B) = \frac{2|A \cap B|}{|A| + |B|}$$

▼ Distância Jensen-Shannon:

$$d_{JS}(P||Q) = \frac{1}{2}\left(D(P||M) + D(Q||M)\right)$$

Onde:

$$M = \frac{1}{2}(P+Q)$$
$$D(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

▼ Distância Canberra:

$$d_{canberra}(x,y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|}$$

▼ Distância Hamming:

$$d_{Hamming}(s,t) = \sum_{i=1}^{n} (s_i \neq t_i)$$

▼ Distância Spearman:

$$d_{spearman} = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

onde: d_i é a diferença entre as posições dos elementos i em duas sequências ordenadas, e n é o tamanho das sequências.

▼ Distância Chi-Square:

$$d_{\chi^2}(x,y) = \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{x_i + y_i}$$

▼ Passo 2: Matriz de similaridade (S)

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

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

▼ Exemplo:

▼ Matriz de avaliações

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

▼ Processo de construção da matriz similaridade (S)

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

▼ Matriz similaridade (S)

Participantes	Alice	Bob	Cary	Doug	Edna
Alice	-22	-7	-6	-12	-17
Bob	-7	-22	-17	-17	-22
Cary	-6	-17	-22	-18	-21
Doug	-12	-17	-18	-22	-3
Edna	-17	-22	-21	-3	-22

▼ 3. Resumo

1. Definição da métrica de similaridade
2. Calculo da similaridade entre todos os pontos do conjunto de dados, formando a matriz de similaridade (S)

▼ 4. Próxima aula

Definição dos clusters - Parte II