K-means ou clustering

Aprendizado não supervisionado

Márcio Koch - lobokoch@gmail.com

K-means (k-médias)

Método de classificação que distribui os objetos em um número k preestabelecido de classes

Método de agregação em torno dos centróides móveis

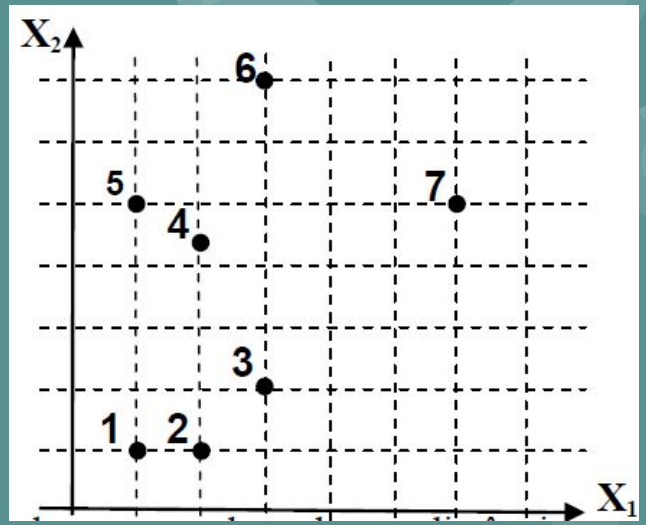
Centroides

Centroide: Ponto médio em relação a um determinado grupo de pontos.

Casos para agrupamento

Caso	Variáveis		Ponto
	X_1	X_2	$x_i = (X_1; X_{i2})$
1	1	1	(1 ; 1)
2	2	1	(2 ; 1)
3	3	2	(3 ; 2)
4	2	4,5	(2; 4,5)
5	1	5	(1 ; 5)
6	3	7	(3 ;7)
7	6	5	(6 ; 5)

Objetos a serem agrupados no plano cartesiano

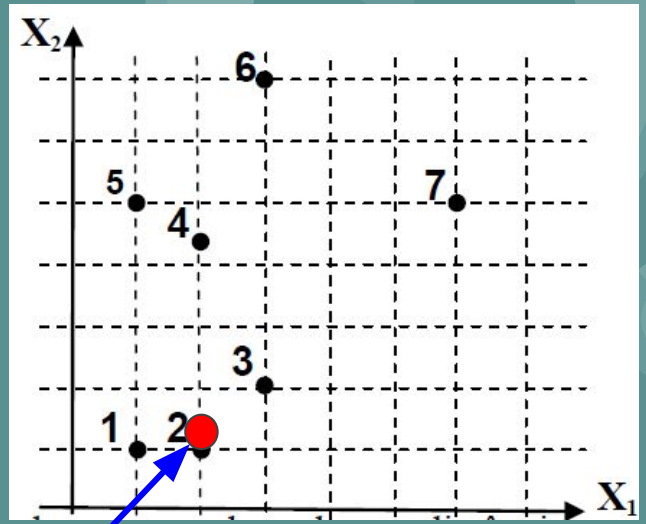


Centroides

Casos para agrupamento

Caso	Variáveis		Ponto
	X_1	X_2	$x_i = (X_1; X_{i2})$
1	1	1	(1 ; 1)
2	2	1	(2 ; 1)
3	3	2	(3 ; 2)
4	2	4,5	(2; 4,5)
5	1	5	(1 ; 5)
6	3	7	(3 ;7)
7	6	5	(6 ; 5)

Objetos a serem agrupados no plano cartesiano



$C_1 = (x_1 = 2, x_2 = 1{,}33)$

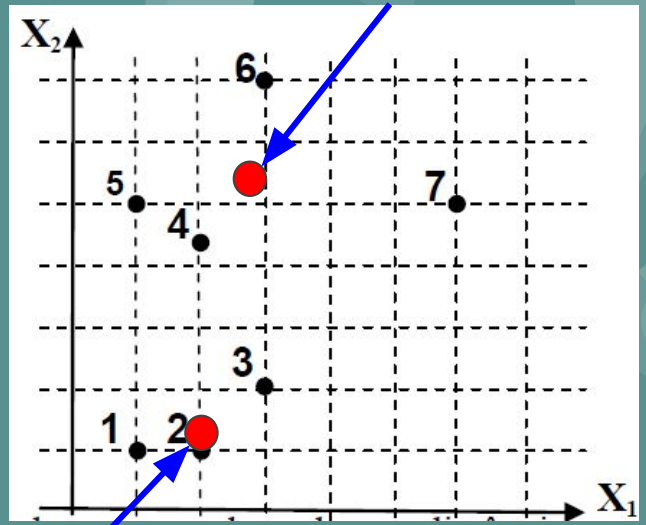
Centroides

Casos para agrupamento

Caso	Variáveis		Ponto
	X_1	X_2	$x_i = (X_1; X_{i2})$
1	1	1	(1 ; 1)
2	2	1	(2 ; 1)
3	3	2	(3 ; 2)
4	2	4,5	(2; 4,5)
5	1	5	(1 ; 5)
6	3	7	(3 ;7)
7	6	5	(6 ; 5)

Objetos a serem agrupados no plano cartesiano

$C_2 = (x_1 = 2{,}75, x_2 = 5{,}37)$



$C_1 = (x_1 = 2, x_2 = 1{,}33)$

Distância Euclidiana

Permite calcular a distância entre dois pontos em um espaço multi-dimensional.

$P = (p_1, p_2, \ldots, p_n)$ e $Q = (q_1, q_2, \ldots, q_n)$

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

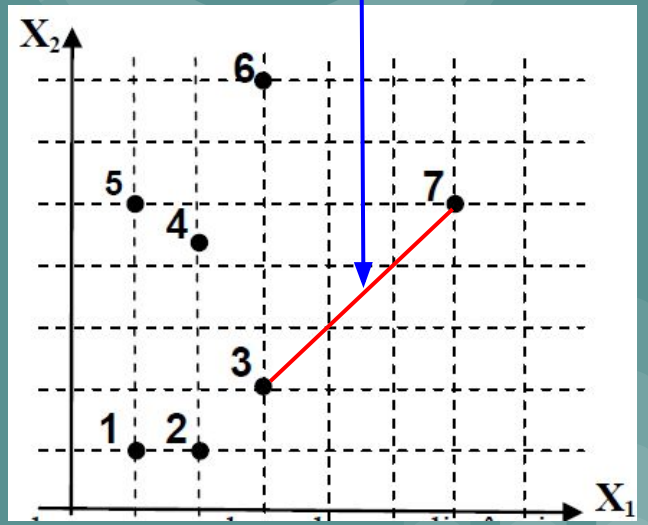
Distância euclidiana

$$P = (p_1, p_2, \ldots, p_n) \text{ e } Q = (q_1, q_2, \ldots, q_n)$$

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

Caso	Variáveis		Ponto
	X_1	X_2	$x_i = (X_1; X_{i2})$
1	1	1	(1 ; 1)
2	2	1	(2 ; 1)
3	3	2	(3 ; 2)
4	2	4,5	(2; 4,5)
5	1	5	(1 ; 5)
6	3	7	(3 ;7)
7	6	5	(6 ; 5)

	X1	X2
P3	3	2
P7	6	5
	(3 - 6)^2	(2 - 5)^2
	9	9
	sum(9 + 9) = 18	
	sqrt(18)	
	4,242640687	



K-means - Algoritmo

1. Padronize os dados (opcional);
2. Selecione aleatoriamente **k** objetos de observação como centroides iniciais (ou escolha os centroides iniciais de alguma forma);
3. Forme **k** classes colocando cada objeto a seu centroide mais próximo, de acordo com a medida de distância adotada;
4. Calcule o centroide de cada classe;
5. Repita os passos 3 e 4 até que os centroides não apresentem mais mudanças.

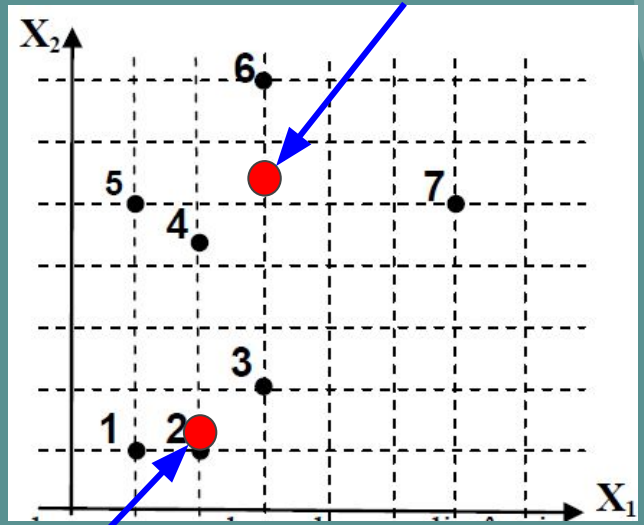
K-means - Algoritmo

Coordenadas dos casos			Iteração 1			Iteração 2			Iteração 3			Iteração 4		
			$\overline{C}_1 = (1 ; 1)$	$\overline{C}_2 = (3 ; 2)$		$\overline{C}_1 = (1{,}5 ; 1)$	$\overline{C}_2 = (3 ; 4{,}7)$		$\overline{C}_1 = (2; 1{,}3)$	$\overline{C}_2 = (3 ; 5{,}375)$		$\overline{C}_1 = (2; 1{,}3)$	$\overline{C}_2 = (3 ; 5{,}375)$	
X_i	x_1	x_2	$d(x_i, \overline{C}_1)$	$d(x_i, \overline{C}_2)$	classe	$d(x_i, \overline{C}_1)$	$d(x_i, \overline{C}_2)$	classe	$d(x_i, \overline{C}_1)$	$d(x_i, \overline{C}_2)$	classe	$d(x_i, \overline{C}_1)$	$d(x_i, \overline{C}_2)$	classe
1	**1**	**1**	0,000	2,236	1	0,500	4,206	1	1,044	4,810	1	Não há variação nos centroides		
2	2	1	1,000	1,414	1	0,500	3,833	1	0,300	4,488	1			
3	**3**	**2**	2,236	0,000	2	1,803	2,700	1	1,221	3,375	1			
4	2	4,5	3,640	2,693	2	3,536	1,020	2	3,200	1,329	2			
5	1	5	4,000	3,606	2	4,031	2,022	2	3,833	2,035	2			
6	3	7	6,325	5,000	2	6,185	2,300	2	5,787	1,625	2			
7	6	5	6,403	4,243	2	6,021	3,015	2	5,449	3,023	2			

K-means - Algoritmo

Cálculo dos centroides

$C_2 = (x_1 = 3, x_2 = 5{,}375)$



$C_1 = (x_1 = 2, x_2 = 1{,}3)$

K-means - Algoritmo

Implementação em Java.