Задание №1

Дан набор значений 2,4,10,12,3,20,30,11,25. Предположим количество кластеров k = 3, и выбраны начальные средние значения $m_1 = 2$, $m_2 = 4$, $m_3 = 6$. Покажите, какие кластеры будут после первой итерации алгоритма k-средних, и рассчитайте новые значения центров кластеров для следующей итерации.

											m1	m2	m3
D:	2	4	10	12	3	20	30	11	25		2	4	6
|D-m1|:	0	2	8	10	1	18	28	9	23				
|D-m2|:	2	0	6	8	1	16	26	7	21				
|D-m3|:	4	2	4	6	3	14	24	5	19		new		
C1:	2				3						m1:	2,5	
C2:		4									m2:	4	
C3:			10	12		20	30	11	25		m3:	18	

Задание №2

Дан набор точек x и вероятности из принадлежности к кластерам C_1 и C_2.

x	$P(C_1\|x)$	$P(C_2\|x)$
2	0.9	0.1
3	0.8	0.1
7	0.3	0.7
9	0.1	0.9
2	0.9	0.1
1	0.8	0.2

Выполните следующие задание:

A. Найдите оценку максимального правдоподобия для средних μ1 и μ2

$$\boldsymbol{\mu}_i = \frac{\sum_{j=1}^{n} w_{ij} \cdot \mathbf{x}_j}{\sum_{j=1}^{n} w_{ij}} \qquad w_{ij} = P(C_i|\mathbf{x}_j)$$

x	2	3	7	9	2	1	
P(C1|x)	0,9	0,8	0,3	0,1	0,9	0,8	3,8
P(C2|x)	0,1	0,1	0,7	0,9	0,1	0,2	2,1
P(C1|x) * x	1,8	2,4	2,1	0,9	1,8	0,8	9,8
P(C2|x) * x	0,2	0,3	4,9	8,1	0,2	0,2	13,9
m1	2,5789473						
m2	6,6190476						

B. Предположим, что μ1 = 2, μ2 = 7 и σ1 = σ2 = 1. Найдите вероятности принадлежности точки x = 5 к кластерам C_1 и C_2 . Априорные вероятности каждого кластера P(C_1) = P(C_2) = 0.5 и P(x = 5) = 0.029

$$\boldsymbol{\mu}_i = \frac{\sum_{j=1}^{n} w_{ij} \cdot \mathbf{x}_j}{\sum_{j=1}^{n} w_{ij}}$$

$$w_{ij} = P(C_i|\mathbf{x}_j) = \frac{f(\mathbf{x}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) P(C_i)}{f(\mathbf{x}_j)}$$

$$f_i(x) = f(x|\mu_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left\{-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right\}$$

i	mi	Si	P(Ci)	f(x|mi, Si)	f(x|mi, Si) * P(Ci)	P(Ci|x)
1	2	1	0,5	0,0044	0,0022	0,0759
2	7	1	0,5	0,0540	0,0270	0,9241
					0,0292	

Задание №3

Даны категориальные данные размерности 5

Point	X_1	X_2	X_3	X_4	X_5
$\mathbf{x}_1$	1	0	1	1	0
$\mathbf{x}_2$	1	1	0	1	0
$\mathbf{x}_3$	0	0	1	1	0
$\mathbf{x}_4$	0	1	0	1	0
$\mathbf{x}_5$	1	0	1	0	1
$\mathbf{x}_6$	0	1	1	0	0

Близость двух наблюдений определяется через количество совпадений и несовпадений значений признаков. Допустим, что n_{11} количество признаков одновременной равных 1 для наблюдений x_i и x_j, и n_{10} количество признаков равных 1 для наблюдения x_i и в то же время равных 0 для наблюдения x_j. По аналогии определяются значения n_{01} and n_{00}:

		$\mathbf{x}_j$	
		1	0
$\mathbf{x}_i$	1	n_{11}	n_{10}
	0	n_{01}	n_{00}

Определим следующие метрики:

Коэффициент простого совпадения

$$SMC(\mathbf{x}_i, \mathbf{x}_j) = \frac{n_{11}+n_{00}}{n_{11}+n_{10}+n_{01}+n_{00}}$$

Коэффициент Жаккара

$$JC(\mathbf{x}_i, \mathbf{x}_j) = \frac{n_{11}}{n_{11}+n_{10}+n_{01}}$$

Коэффициент Рассела и Рао

$$RC(\mathbf{x}_i, \mathbf{x}_j) = \frac{n_{11}}{n_{11}+n_{10}+n_{01}+n_{00}}$$

Постройте дендограммы полученные после иерархической кластеризации при следующих параметрах:

- Метод одиночной связи с метрикой RC

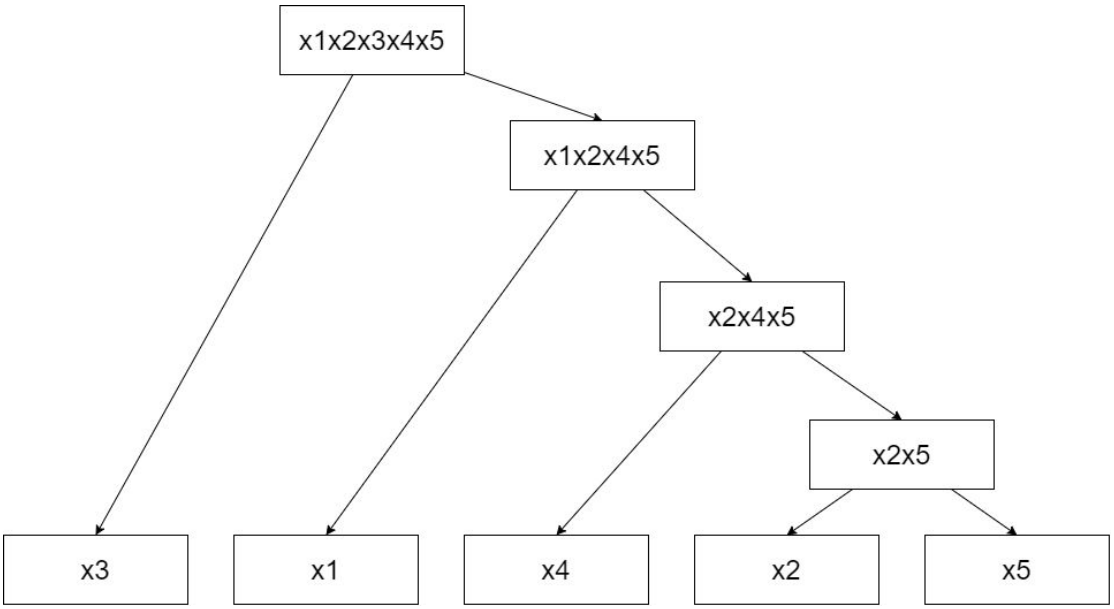
$$\delta(C_i, C_j) = \min\{\delta(\mathbf{x}, \mathbf{y}) \mid \mathbf{x} \in C_i, \mathbf{y} \in C_j\}$$

$$RC(\mathbf{x}_i, \mathbf{x}_j) = \frac{n_{11}}{n_{11}+n_{10}+n_{01}+n_{00}}$$

	x1	x2	x3	x4	x5
x1	-	1/6	1/3	1/3	1/6
x2		-	1/6	1/3	0
x3			-	1/3	1/6
x4				-	0
x5					-

	x1	x2x5	x3	x4
x1	-	1/6	1/3	1/3
x2x5		-	1/6	0
x3			-	1/3
x4				-

	x1	x2x5x4	x3
x1	-	1/6	1/3
x2x5x4		-	1/6
x3			-



- Метод полной связи с метрикой SMC

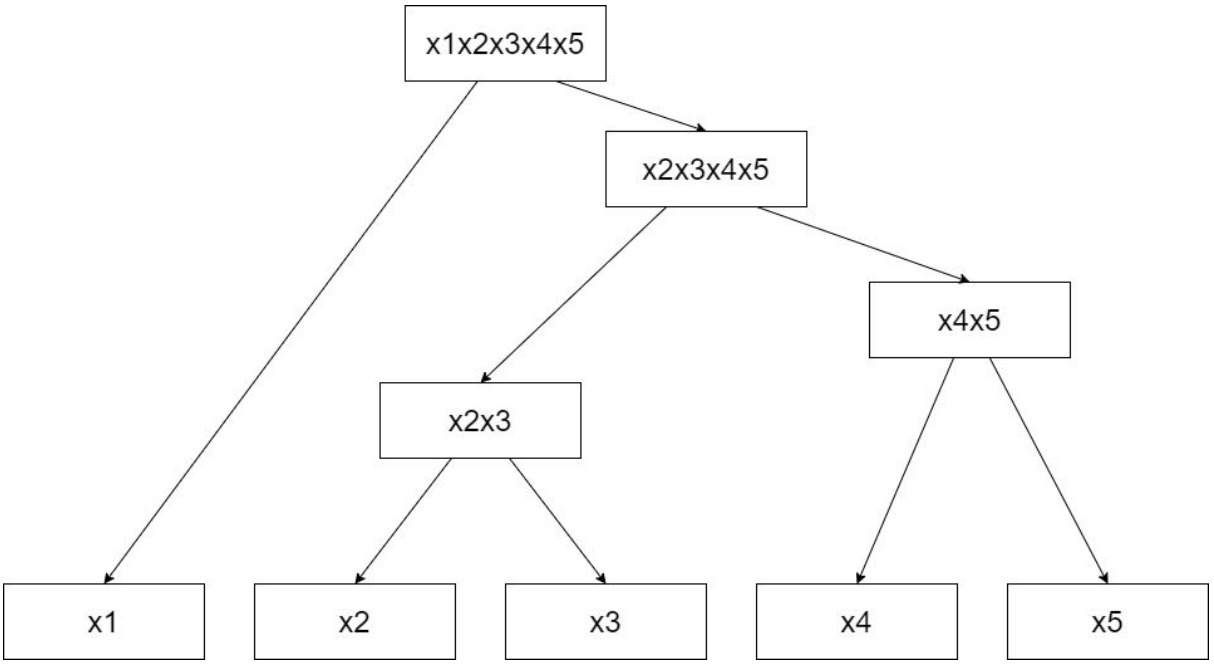
$$\delta(C_i, C_j) = \max\{\delta(\mathbf{x}, \mathbf{y}) \mid \mathbf{x} \in C_i, \mathbf{y} \in C_j\}$$

$$SMC(\mathbf{x}_i, \mathbf{x}_j) = \frac{n_{11}+n_{00}}{n_{11}+n_{10}+n_{01}+n_{00}}$$

	x1	x2	x3	x4	x5
x1	-	1/3	1/2	1/2	2/3
x2		-	1/6	1/2	1/3
x3			-	1/3	1/2
x4				-	1/6
x5					-

	x1	x2x3	x4	x5
x1	-	1/2	1/2	2/3
x2x3		-	1/2	1/2
x4			-	1/6
x5				-

	x1	x2x3	x4x5
x1	-	1/2	2/3
x2x3		-	1/3
x4x5			-



- Невзвешенный центроидный метод с метрикой JC

$$\delta(C_i, C_j) = \frac{\sum_{\mathbf{x} \in C_i} \sum_{\mathbf{y} \in C_j} \delta(\mathbf{x}, \mathbf{y})}{n_i \cdot n_j}$$

$$JC(\mathbf{x}_i, \mathbf{x}_j) = \frac{n_{11}}{n_{11}+n_{10}+n_{01}}$$

	x1	x2	x3	x4	x5
x1	-	1/4	2/5	2/5	1/3
x2		-	1/6	2/5	0
x3			-	1/3	1/4
x4				-	0
x5					-

	x1	x2x5	x3	x4
x1	-	7/24	2/5	2/5
x2x5		-	5/24	1/5
x3			-	1/3
x4				-

	x1	x2x5x4	x3
x1	-	1/3	2/5
x2x5x4		-	3/4
x3			-