Практические задания №5. Григорьев И.С. 6304

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.

Кластер	Значения	Среднее
C_1	2, 3	2.5
C_2	4	4
C_3	10, 11, 12, 20, 25, 30	18

Задание №2

Дан набор точек x и вероятности из принадлежности к кластерам C_1и 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

1. Найдите оценку максимального правдоподобия для средних μ1 и μ2.

μ_i вычисляется как средневзвешенное из всех точек:

$$\mu_i = \frac{\sum_{j=1}^{n} w_{ij} x_j}{\sum_{j=1}^{n} w_{ij}}$$

где $w_{ij} = P(C_i|x_j)$ – вес или вклад точки x_j в кластер C_i.

```
w1 = np.array([0.9, 0.8, 0.3, 0.1, 0.9, 0.8])
w2 = np.array([0.1, 0.1, 0.7, 0.9, 0.1, 0.2])
X = np.array([2, 3, 7, 9, 2, 1])
```

```
m1 = (w1 * X).sum() / w1.sum()
m1
```

```
2.5789473684210535
```

```
m2 = (w2 * X).sum() / w2.sum()
m2
```

```
6.619047619047618
```

2. 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.

Апостериорные вероятности кластеров вычисляются с помощью ур-я:

$$P(C_i|\mathbf{x}_j) = \frac{f(\mathrm{x}_j|\mu_i, \sigma_i^2)P(C_i)}{\sum_{a=1}^{k} f(\mathrm{x}_j|\mu_a, \sigma_a^2)P(C_a)}$$

используем одномерные нормали для каждого кластера:

$$f_i(\mathbf{x}) = f(\mathrm{x}_j|\mu_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi}\sigma_j} exp\left\{-\frac{(\mathbf{x}_j - \boldsymbol{\mu}_i)^2}{2\sigma_i^2}\right\}$$

```
def f(x, mean, std):
    return np.exp(-(x - mean)**2 / (2 * std**2)) / (np.sqrt(2 * np.pi) * std)

pc1_x = f(5, 2, 1) * 0.5
pc2_x = f(5, 7, 1) * 0.5
x_in_c1 = pc1_x / (pc1_x + pc2_x)
x_in_c2 = pc2_x / (pc1_x + pc2_x)
print(f'P(C1|5) = {round(x_in_c1, 3)}')
print(f'P(C2|5) = {round(x_in_c2, 3)}')
```

```
P(C1|5) = 0.076
P(C2|5) = 0.924
```

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}:

	$\mathbf{x}_j$		
$\mathbf{x}_i$		1	0
	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

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

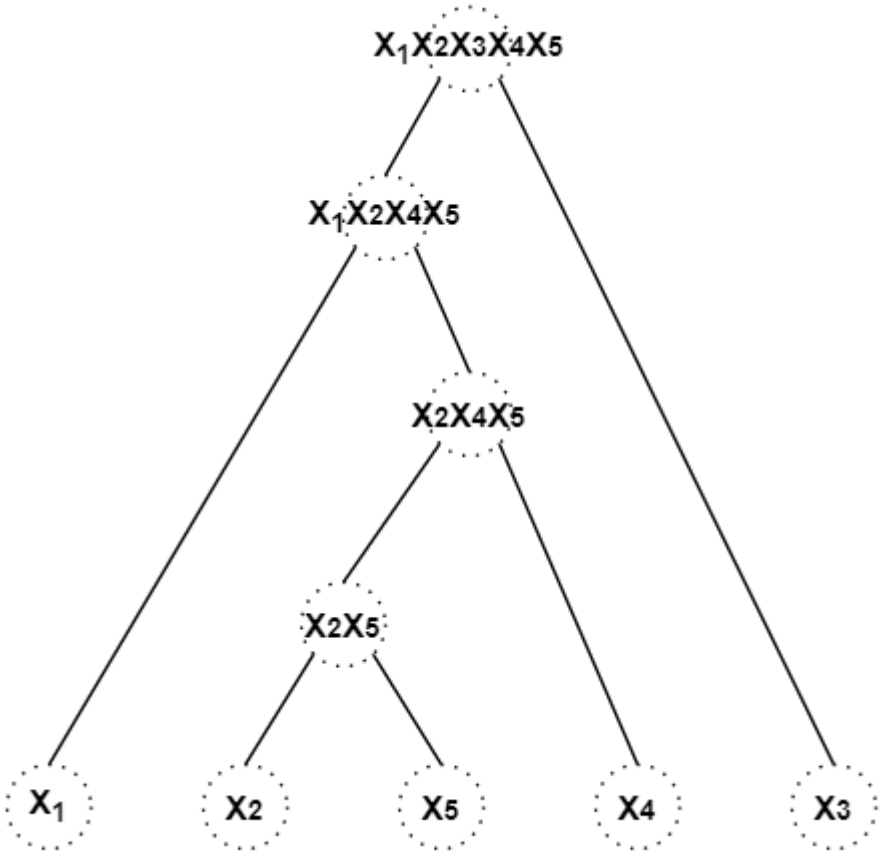
Постройте дендограммы полученные после иерархической кластеризации при следующих параметрах:

- Метод одиночной связи с метрикой RC

	X_3	$X_2X_4X_5$
X_1	1/3	1/6
X_3		1/6

	X_3	X_4	X_2X_5
X_1	1/3	1/3	1/6
X_3		1/3	1/6
X_4			0

	X_2	X_3	X_4	X_5
X_1	1/6	1/3	1/3	1/6
X_2		1/6	1/3	0
X_3			1/3	1/6
X_4				0

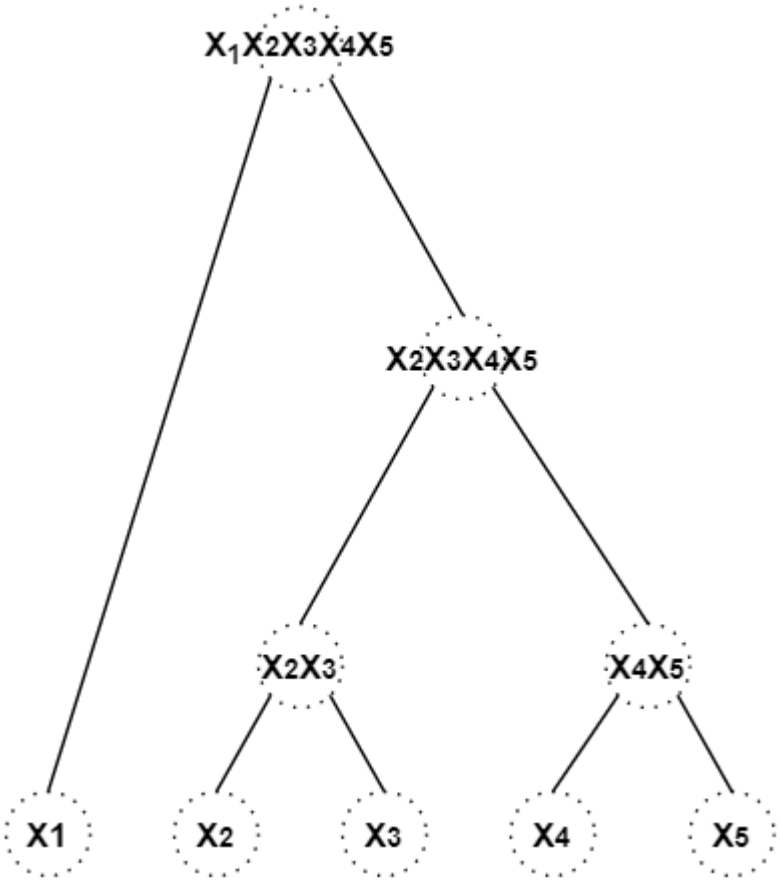


Метод полной связи с метрикой SMC

	X_2X_3	X_4X_5
X_1	1/2	2/3
X_2X_3		1/3

	X_2X_3	X_4	X_5
X_1	1/2	1/2	2/3
X_2X_3		1/2	1/2
X_4			1/6

	X_2	X_3	X_4	X_5
X_1	1/3	1/2	1/2	2/3
X_2		1/6	1/2	1/3
X_3			1/3	1/2
X_4				1/6



- Невзвешенный центроидный метод с метрикой JC

	$X_2X_4X_5$	X_3
X_1	59/180	2/5
$X_2X_4X_5$		3/4

	X_2X_5	X_3	X_4
X_1	7/24	2/5	2/5
X_2X_5		5/24	1/5
X_3			1/3

	X_2	X_3	X_4	X_5
X_1	1/4	2/5	2/5	1/3
X_2		1/6	2/5	0
X_3			1/3	1/4
X_4				0