ЛЭТИ, гр. 8303, Гришин Константин
8 ноября 2021 г.

Машинное обучение
Практические задания №5

Задача 1. *Дан набор значений* $[2, 4, 10, 12, 3, 20, 30, 11, 25]$. *Предположим, что количество кластеров* $k = 3$, *и выбраны начальные средние значения* $m_1 = 2$, $m_2 = 4$, $m_3 = 6$. *Покажите, какие кластеры будут сформированы после первой итерации алгоритма k-средних, и рассчитайте новые значения центров кластеров для следующей итерации.*

Решение. Рассчитаем расстояние каждой точки данных до начальных значений.

	2	4	6	cluster
2	0	2	4	2
4	2	0	2	4
10	8	6	4	6
12	10	8	6	6
3	1	1	3	2
20	18	16	14	6
30	28	26	24	6
11	9	7	5	6
25	23	21	19	6

Получены кластеры данных, проведем расчет новых центров.

Cluster	Data	Mean
C_1	2, 3	2.5
C_2	4	4
C_3	10, 12, 20, 30, 11, 25	18

□

Задача 2. *Дан набор точек* x *и вероятности их 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*

A. Найдите оценку максимального правдоподобия для средних μ_1 μ_2.

B. Предположим, что $\mu_1 = 2$, $\mu_2 = 7$ *и* $\sigma_1 = \sigma_2 = 1$. *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* $P(x = 5) = 0.029$.

Решение.

A. μ_i - средневзвешенное всех точек:

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

```
>>> 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])
>>> (w1 * x).sum() / w1.sum()
2.5789473684210535
>>> (w2 * x).sum() / w2.sum()
6.619047619047618
```

$\mu_1 = 2.58$, $\mu_2 = 6.62$

B. Вероятность нахождения точки в кластере:

$$P(C_i|x_j) = \frac{f_i(x_j) \cdot P(C_i)}{\sum_{a=1}^{k} f_a(x_j) \cdot P(C_a)}$$

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

```
>>> f = lambda x, mean, std:
np.exp(-(x-mean)**2/(2*std**2))/(np.sqrt(2*np.pi)*std)
>>> pc1 = f(5, 2, 1) * 0.5
>>> pc2 = f(5, 7, 1) * 0.5
>>> pc1 / (pc1+pc2)
0.07585818002124355
>>> pc2 / (pc1+pc2)
0.9241418199787564
```

$P(C_1|5) = 0.076$, $P(C_2|5) = 0.924$

□

Задача 3. *Даны категориальные данные размерности:*

Point	X_1	X_2	X_3	X_4	X_5
x_1	*1*	*0*	*1*	*1*	*0*
x_2	*1*	*1*	*0*	*1*	*0*
x_3	*0*	*0*	*1*	*1*	*0*
x_4	*0*	*1*	*0*	*1*	*0*
x_5	*1*	*0*	*1*	*0*	*1*
x_6	*0*	*1*	*1*	*0*	*0*

Близость двух наблюдений определяется через количество совпадений и несовпадений значений признаков. Допустим, что n_{11} количество признаков одновременно 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 наблюдений x_j. По аналогии определяются значения для n_{01} и n_{00}:

		x_j	
		1	*0*
x_i	*1*	n_{11}	n_{10}
	0	n_{01}	n_{00}

Определим следующие метрики:

- *Коэффициент простого совпадения*

$$SMC(x_i, x_j) = \frac{n_{11} + n_{00}}{n_{11} + n_{10} + n_{01} + n_{00}}$$

- *Коэффициент Жаккара*

$$JC(x_i, x_j) = \frac{n_{11}}{n_{11} + n_{10} + n_{01}}$$

- *Коэффициент Рассела и Рао*

$$RC(x_i, x_j) = \frac{n_{11}}{n_{11} + n_{10} + n_{01} + n_{00}}$$

Постройте дендограммы полученные после иерархической кластеризации при следующих параметрах:

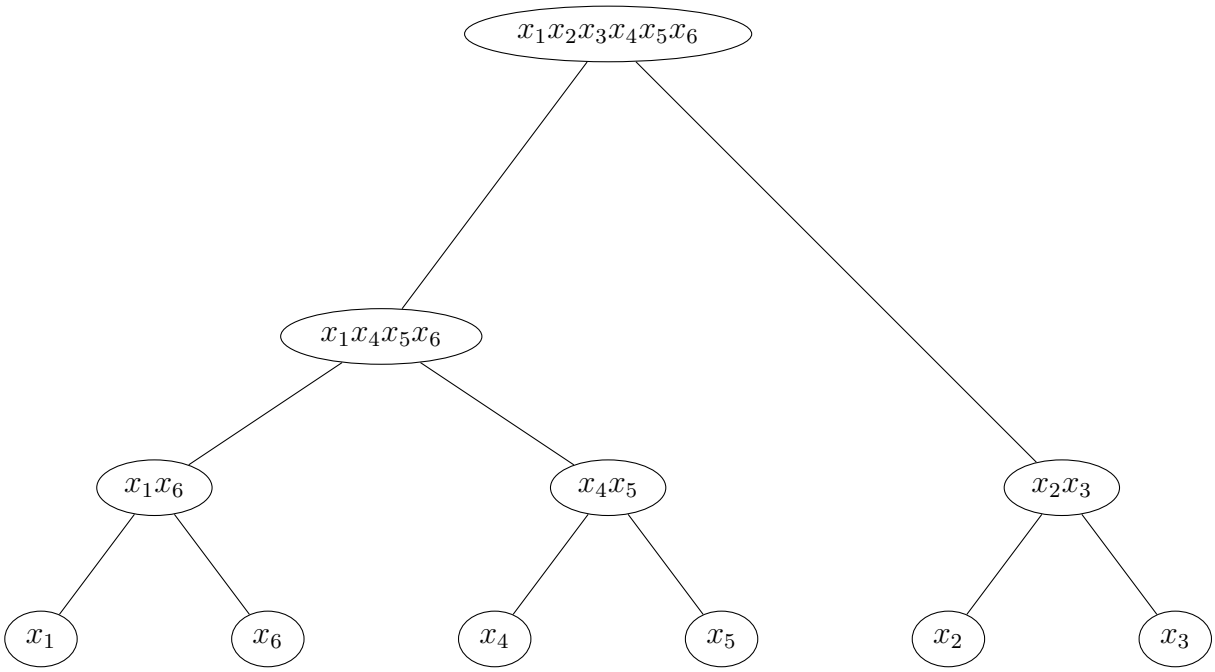
- *Метод одиночной связи с метрикой RC*
- *Метод полной связи с метрикой SMC*
- *Невзвешенный центройдный метод с метрикой JC*

Решение.

- Составим таблицу расстояний метрики RC

	x_1	x_2	x_3	x_4	x_5
x_2	0.4				
x_3	0.4	0.2			
x_4	0.2	0.4	0.2		
x_5	0.4	0.2	0.2	0.0	
x_6	0.2	0.2	0.2	0.2	0.2

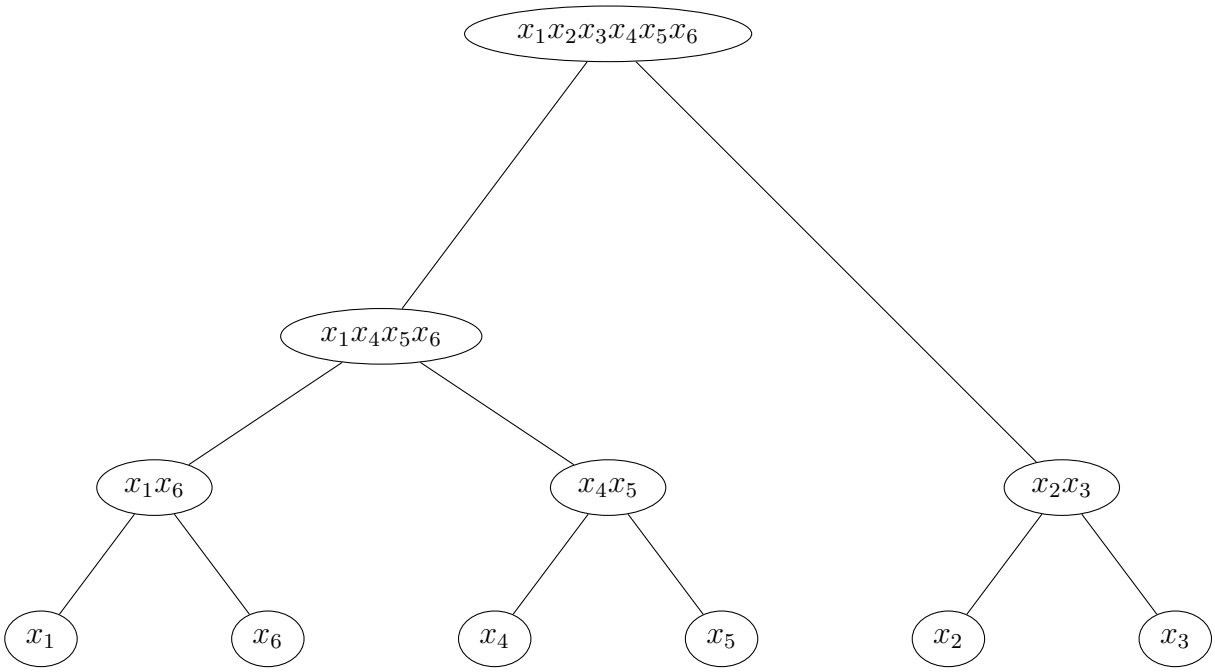
Построим дендограмму методом одиночной связи



- Составим таблицу расстояний метрики SMC:

	x_1	x_2	x_3	x_4	x_5
x_2	0.6				
x_3	0.8	0.4			
x_4	0.4	0.8	0.6		
x_5	0.6	0.2	0.4	0.0	
x_6	0.4	0.4	0.6	0.6	0.4

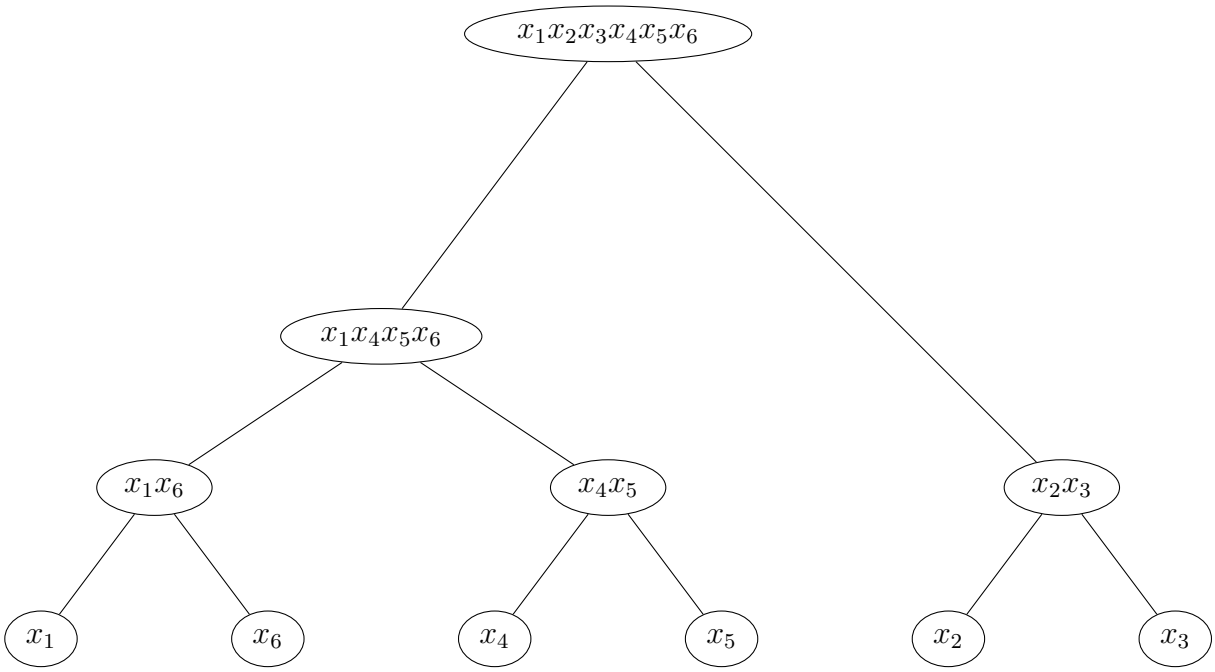
Построим дендограмму методом полной связи:



- Составим таблицу расстояний метрики JC

	x_1	x_2	x_3	x_4	x_5
x_2	0.50				
x_3	0.67	0.25			
x_4	0.25	0.67	0.33		
x_5	0.50	0.20	0.25	0.0	
x_6	0.25	0.25	0.33	0.33	0.25

Построим дендограмму методом полной связи:



Для указанных комбинаций метрики и метода определения расстояния между кластерами, разбиение на классы одинаково. Различаются только итоговые расстояния между точками и кластерами. □