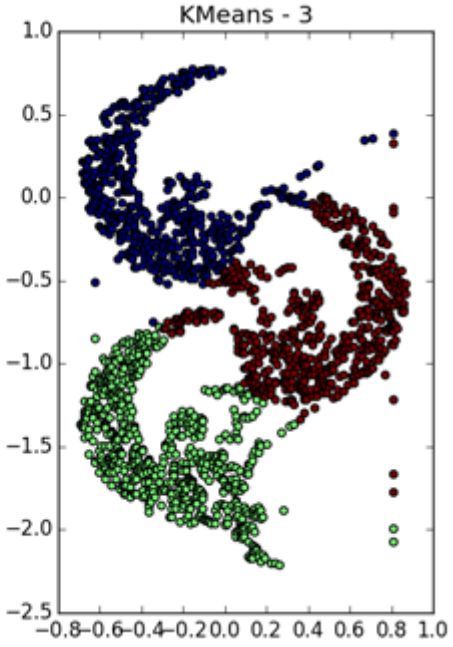
Задача кластеризации данных

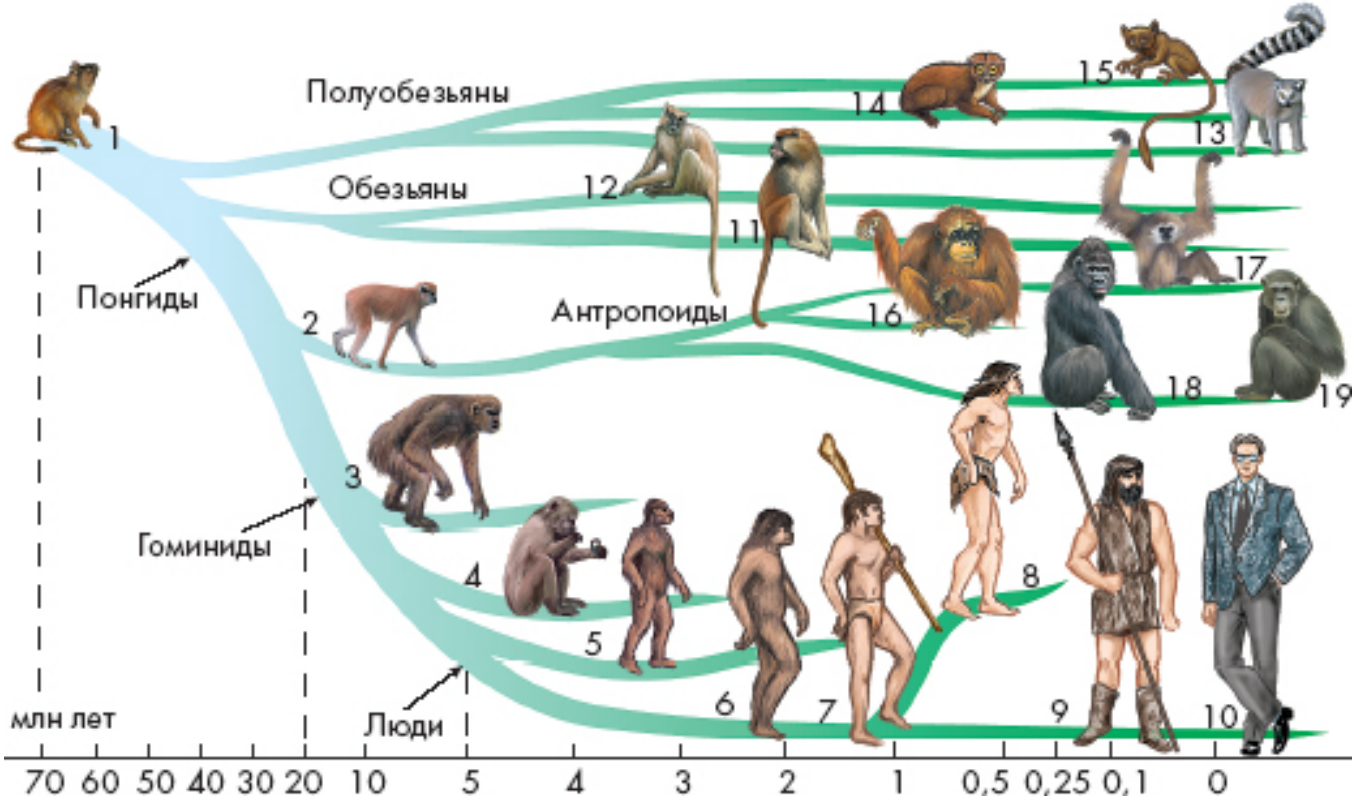


Группа людей, действуя совместно, может свершить такое, о чем поодиночке они не могли бы и мечтать.

Франклин Рузвельт

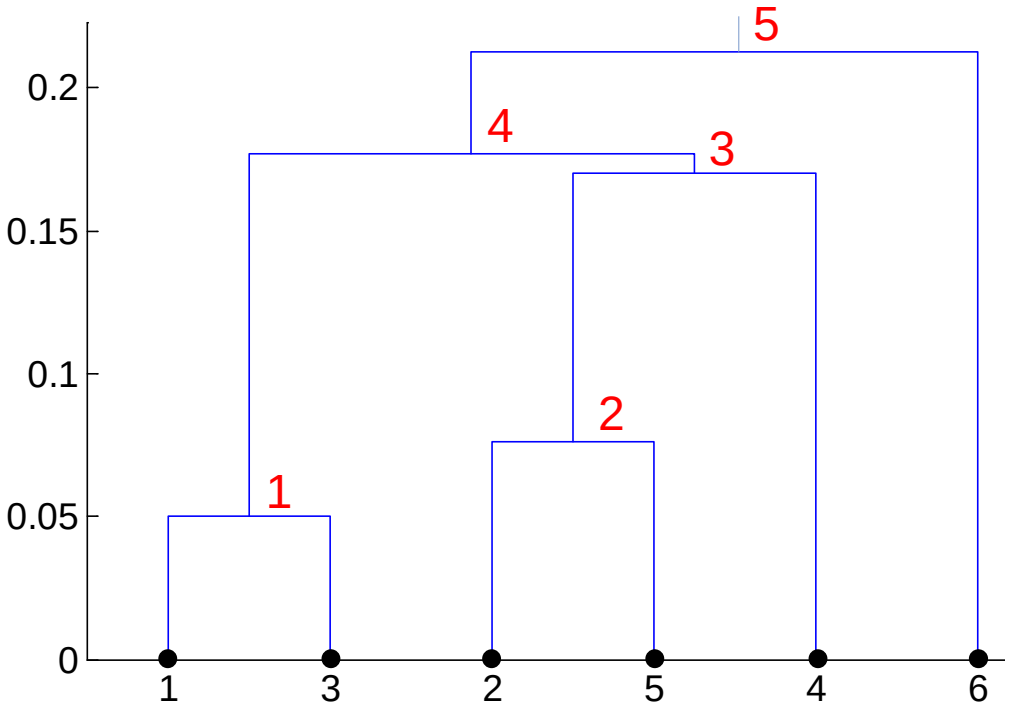
Содержание

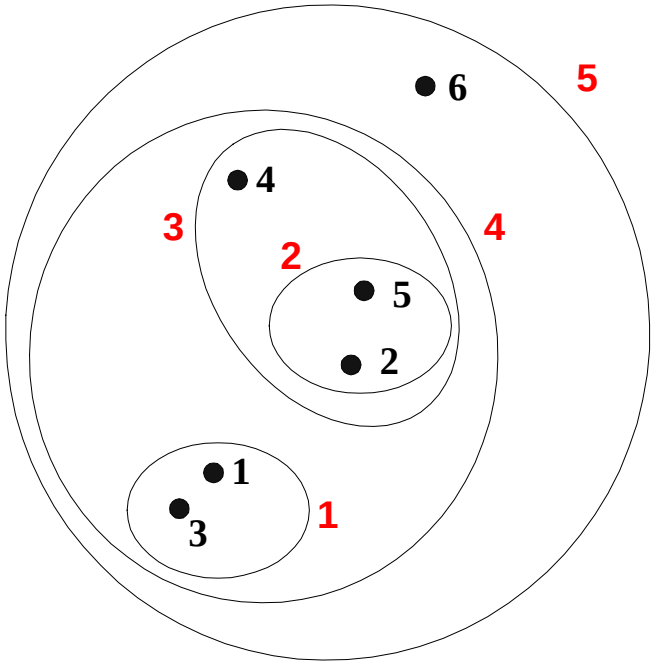
Иерархическая кластеризация



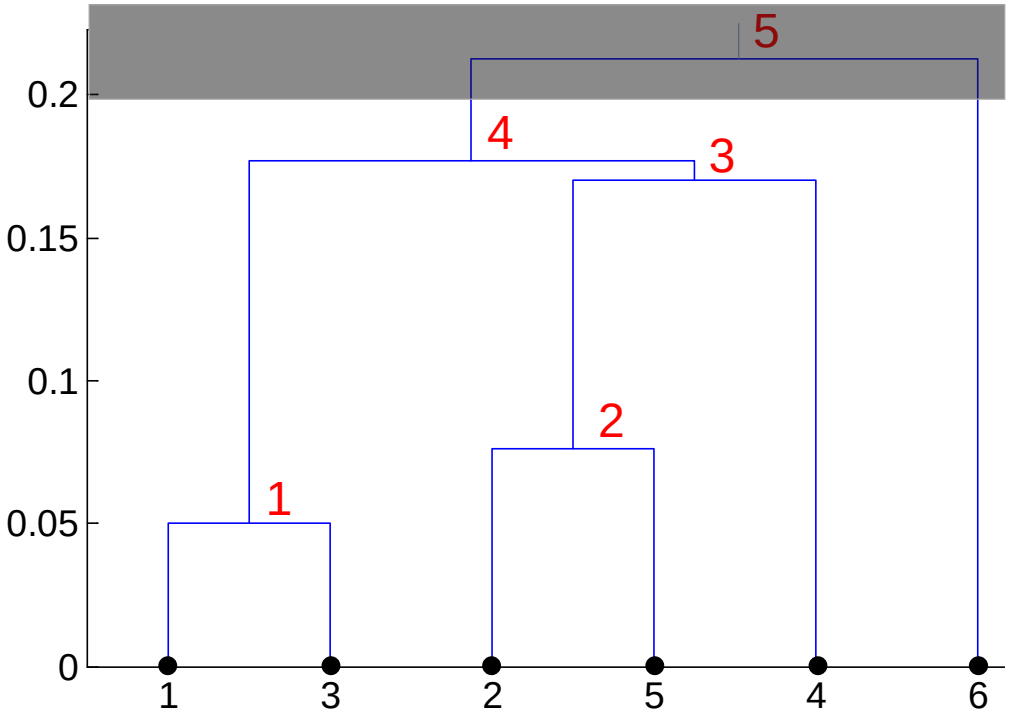
1 — плезиадацис (предок приматов); **2** — дрипитек африканский; **3** — рамапитек; **4** — австралопитек африканский; **5** — австралопитек; **6-7** — *homo erectus* (питекантроп, синантроп); **8** — неандерталец; **9** — *homo sapiens* (кроманьонец); **10** — современный человек; **11** — узконосые обезьяны; **12** — широконосые обезьяны; **13** — лемуры; **14** — лори; **15** — долгопяты; **16** — орангутаны; **17** — гиббоны; **18** — гориллы; **19** — шимпанзе

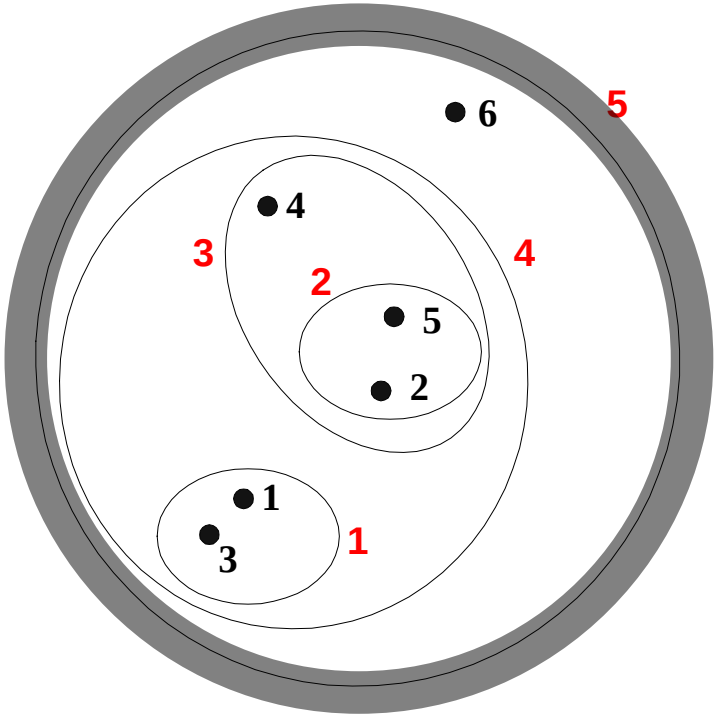
Дендрограммы



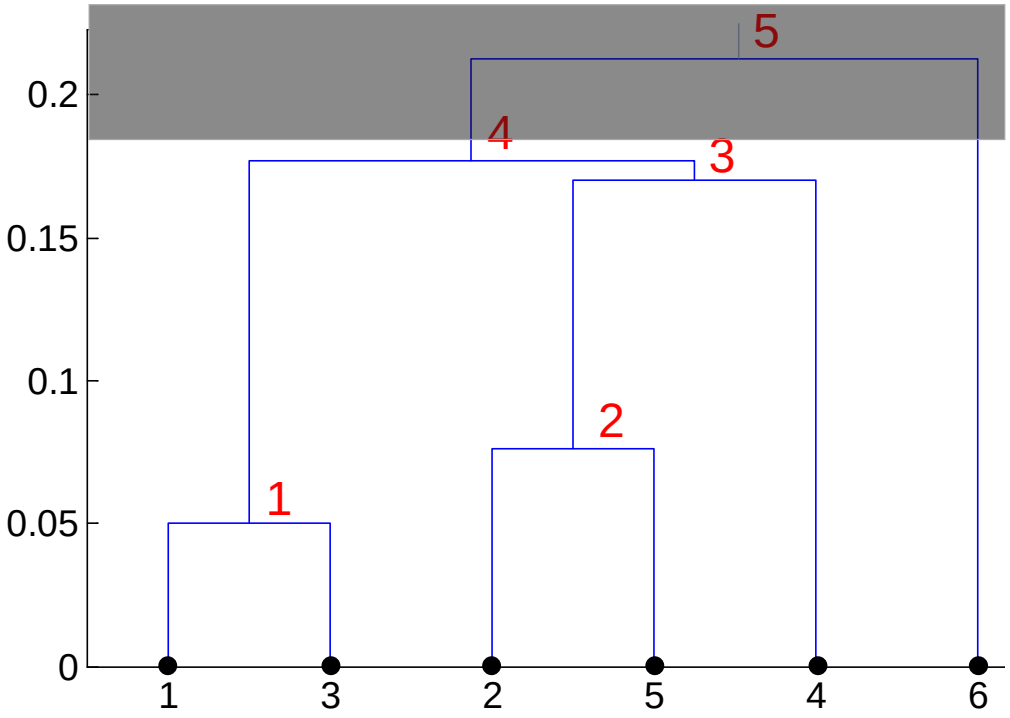


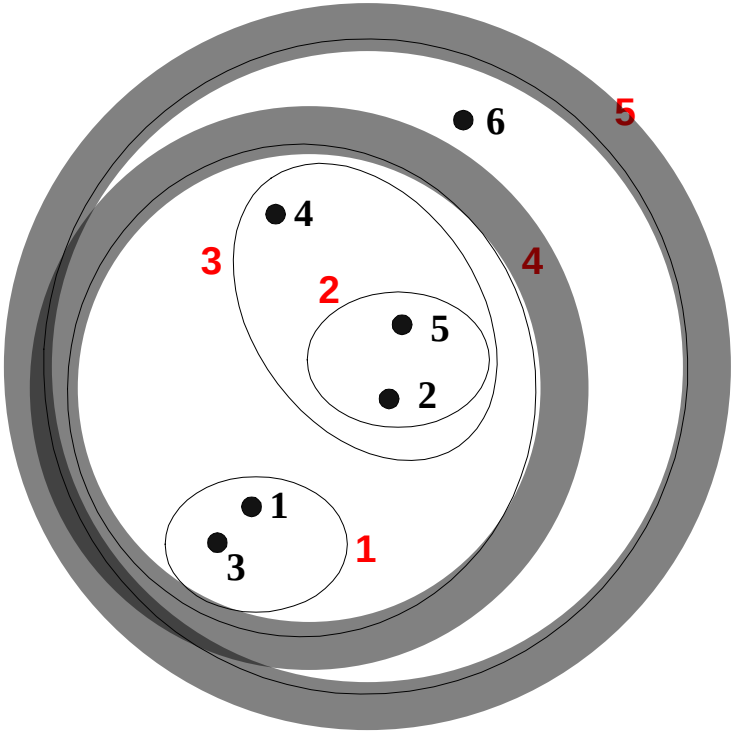
Дендрограммы



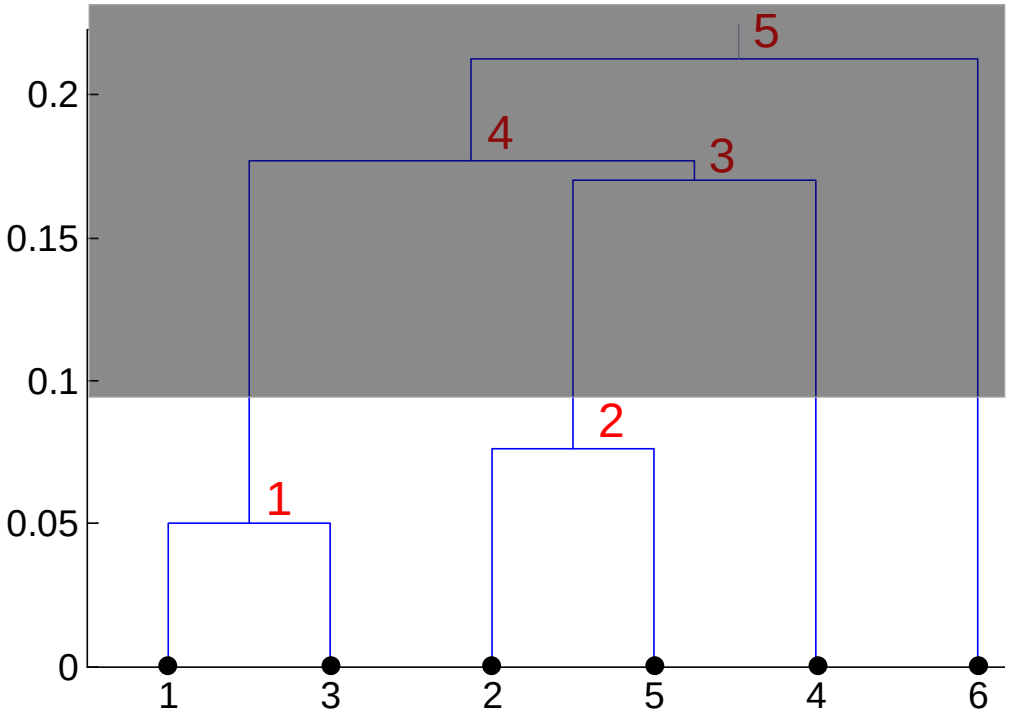


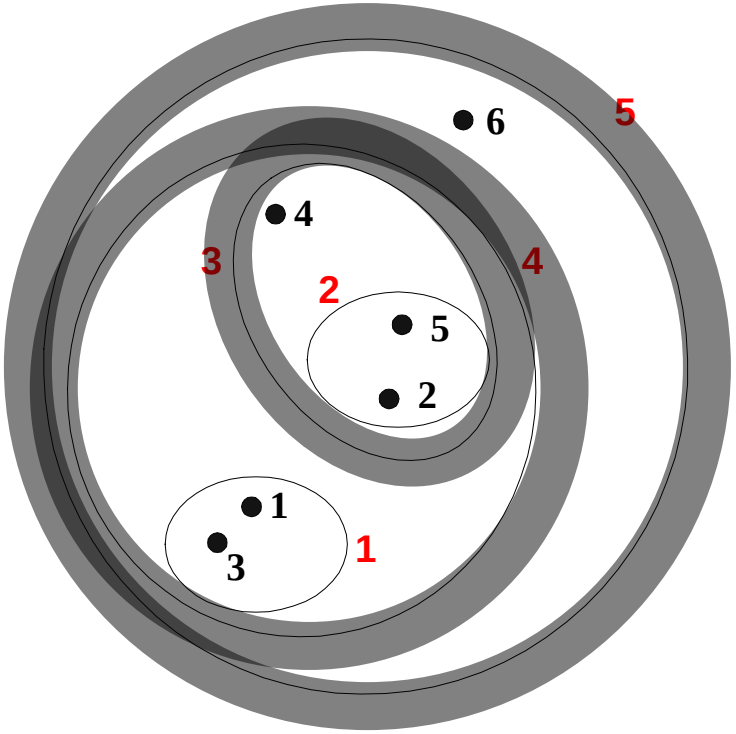
Дендрограммы



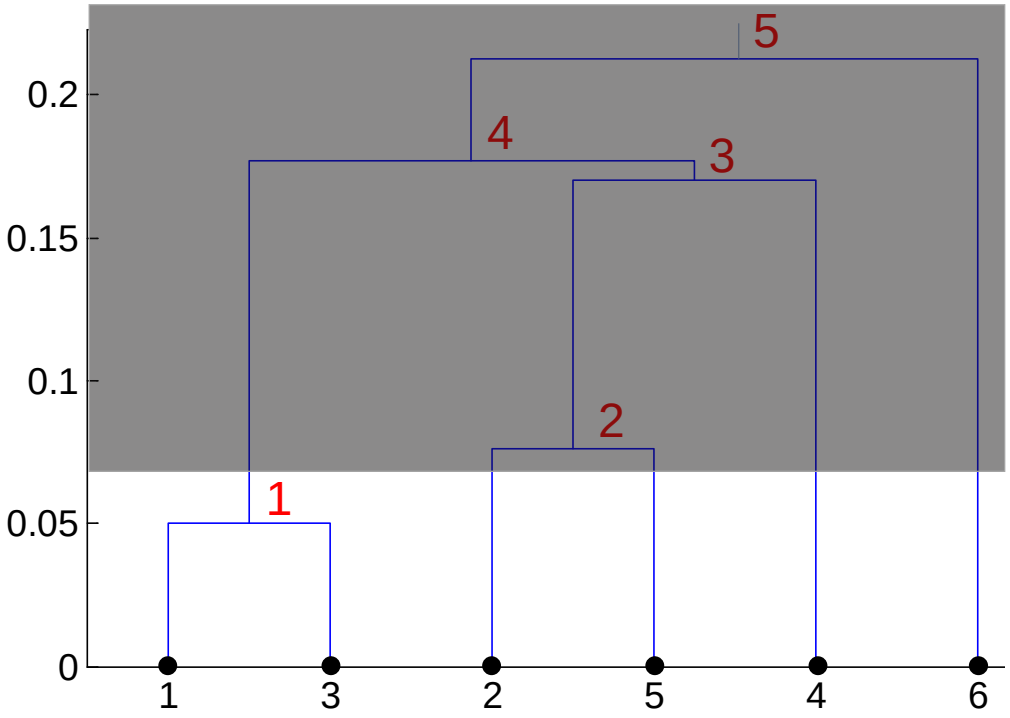


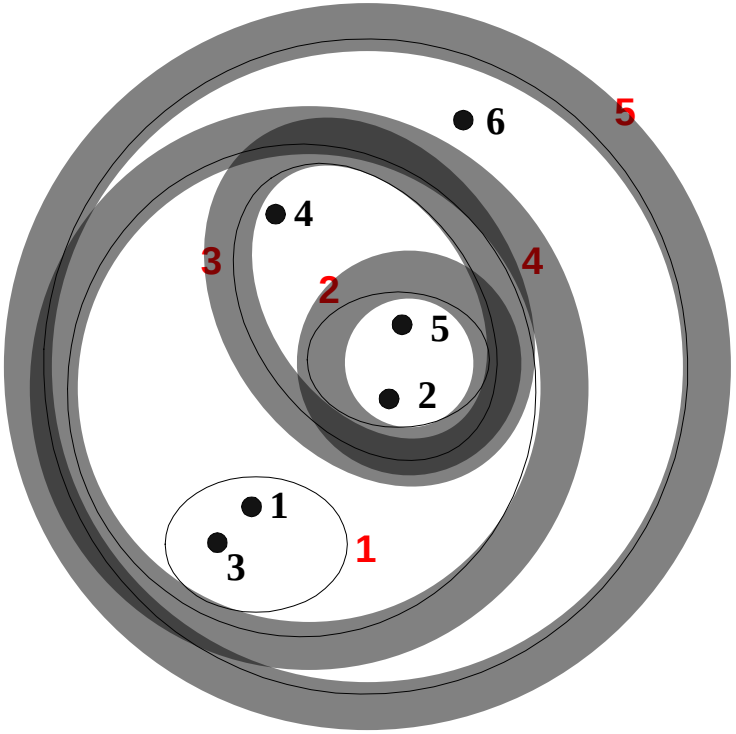
Дендрограммы



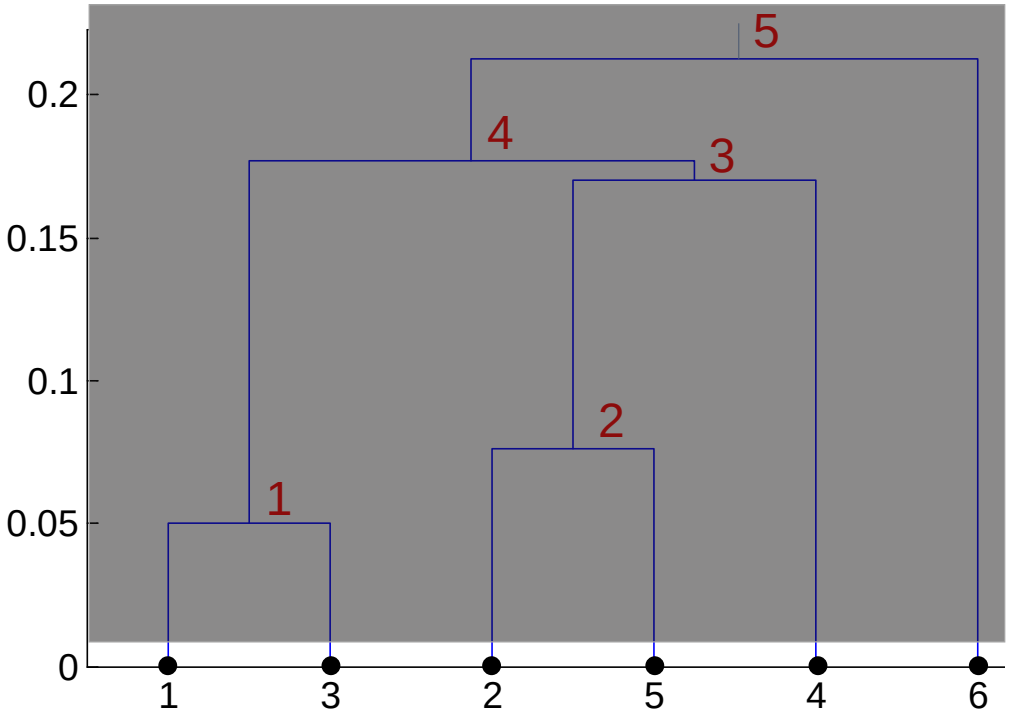


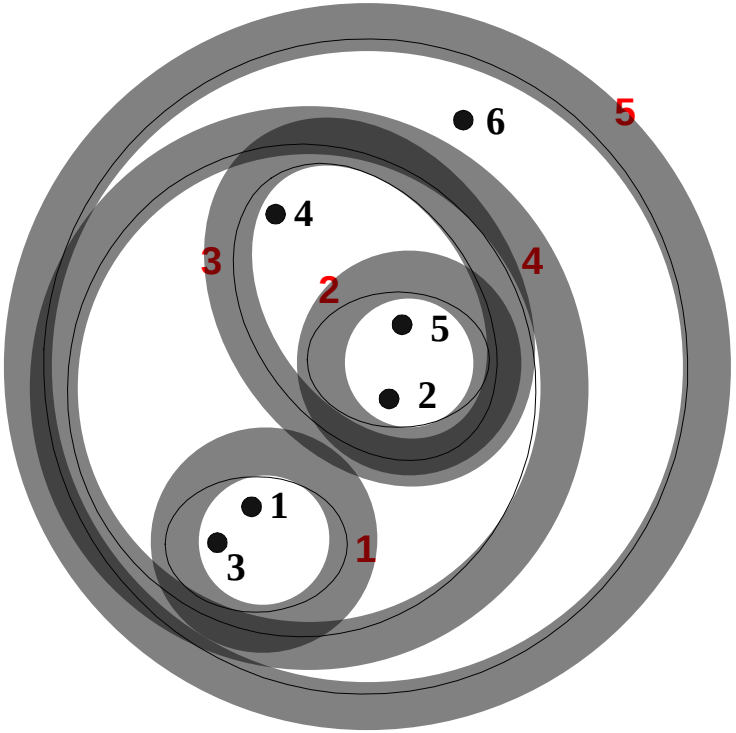
Дендрограммы



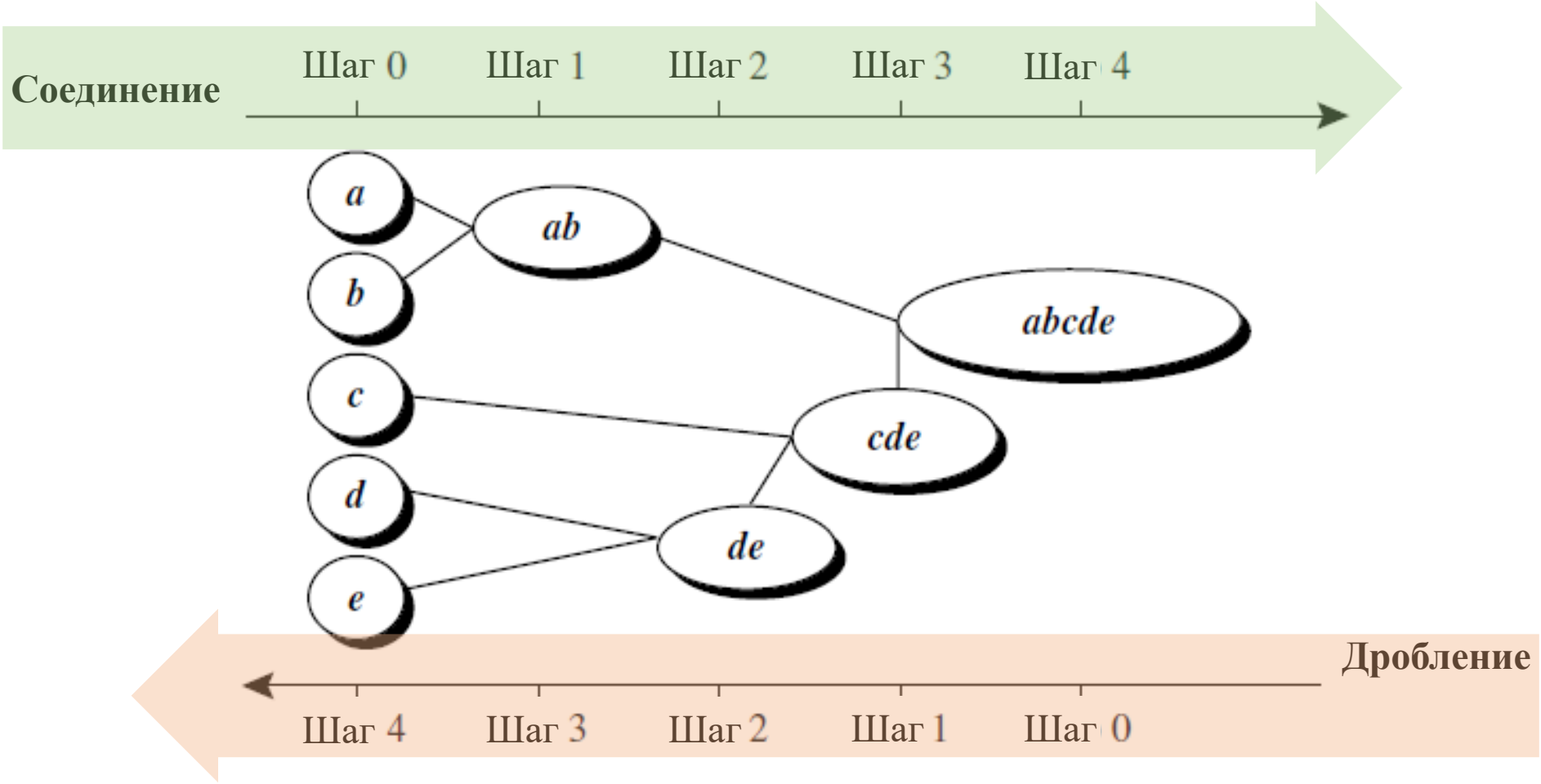


Дендрограммы





Соединение *vs.* дробление (agglomerative *vs.* divisive)

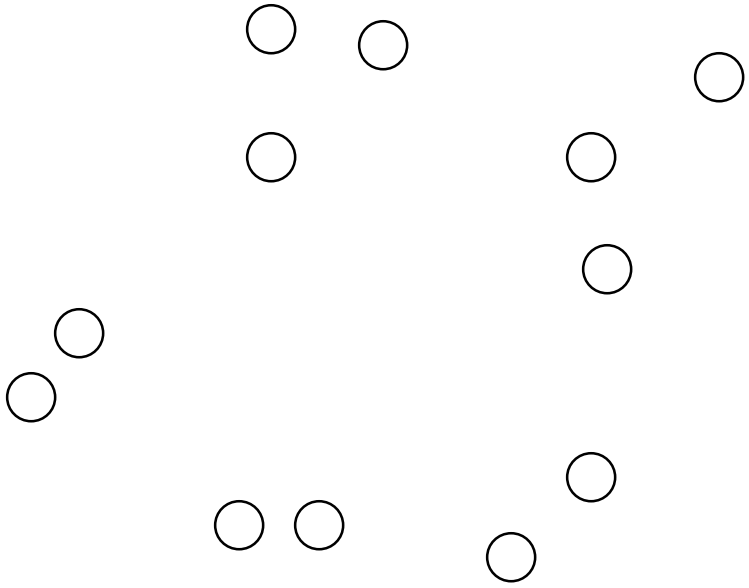


Агломеративная кластеризация

- Кластеризуемые объекты: $x_1, \ldots, x_n$
- Кластеры: $\boldsymbol{C} = \{C_1, \ldots, C_k\}$

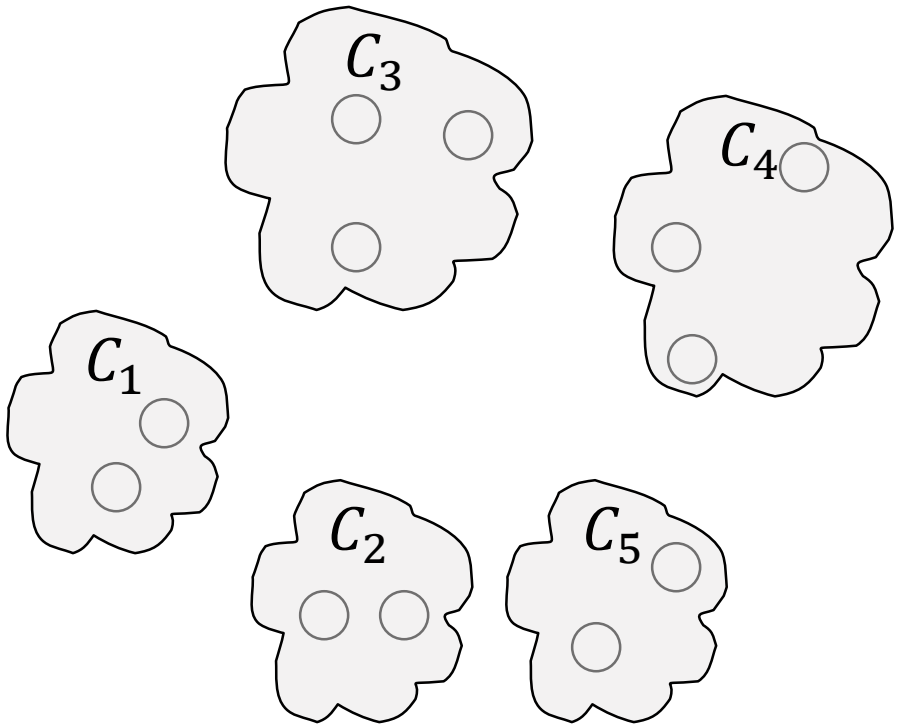
1. Вычислить матрицу расстояний $\left(Dist_{ij}\right) := \text{dist}(x_i, x_j)$
2. $\forall k\ (1 \le k \le n)\ C_k := \{x_k\}$
3. **repeat**
4. $C_{\text{new}} := C_i \cup C_j$, где $\{C_i, C_j\} = \arg\min_{C_p \cap C_q = \emptyset} \text{dist}(C_p, C_q)$
5. Обновить $(Dist_{ij})$: $\forall k \ne i, j$ вычислить $\text{dist}(C_{\text{new}}, C_k)$
6. $\boldsymbol{C} := \boldsymbol{C} \cup C_{\text{new}}$; $\boldsymbol{C} := \boldsymbol{C} \setminus \{C_i, C_j\}$
7. **until** $|\boldsymbol{C}| = 1$

Агломеративная кластеризация (шаг 0)

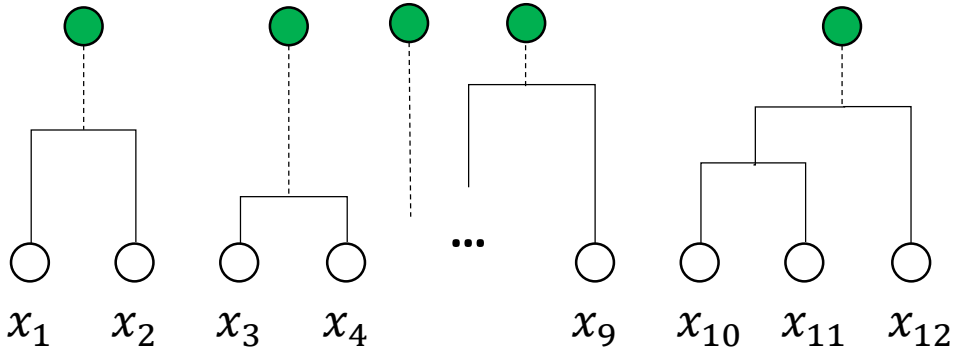


	x_1	x_2	x_3	x_4	x_5	...
x_1	0					
x_2		0				
x_3			0			
x_4				0		
x_5					0	
...						

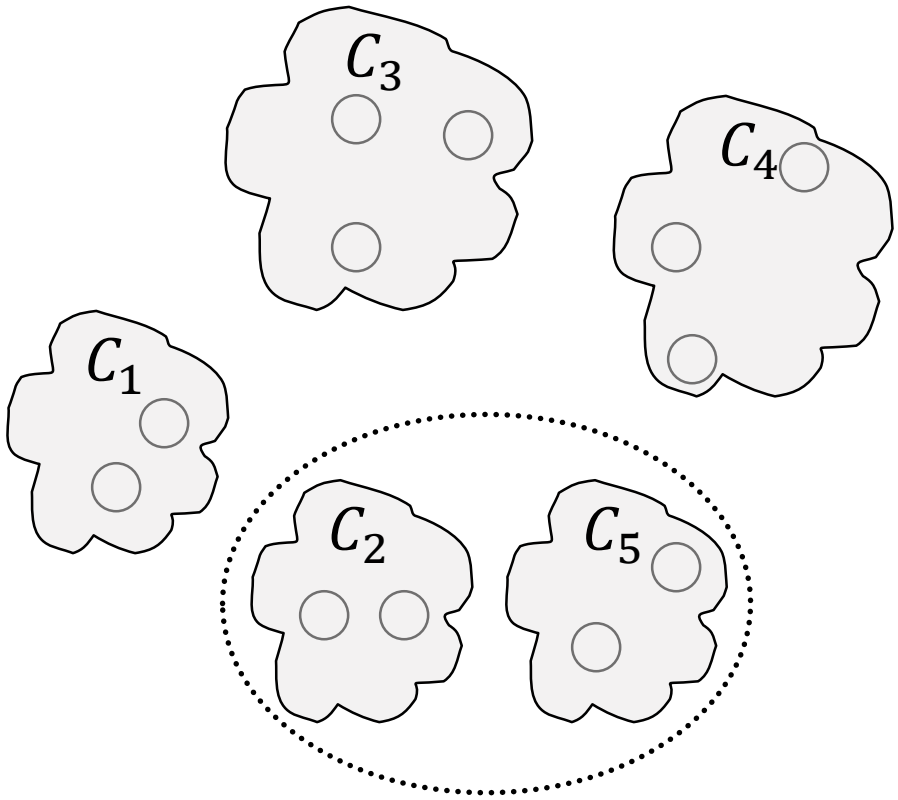
Агломеративная кластеризация (шаг k)



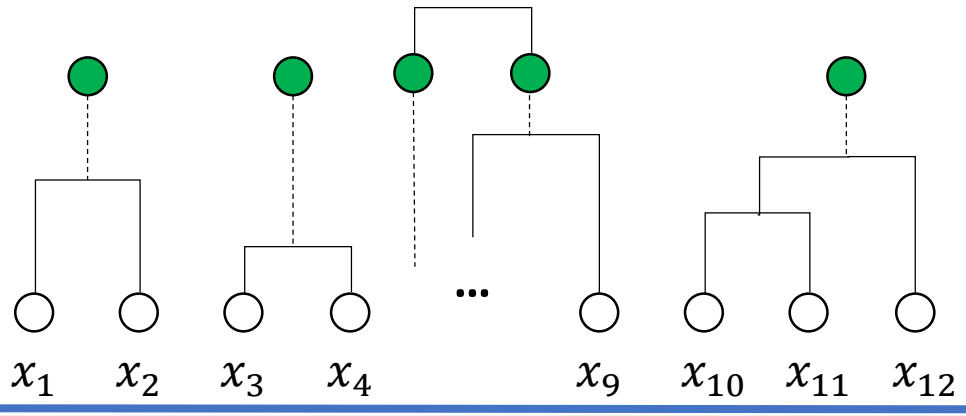
	c_1	c_2	c_3	c_4	c_5	…
c_1	0					
c_2		0				
c_3			0			
c_4				0		
c_5					0	
…						



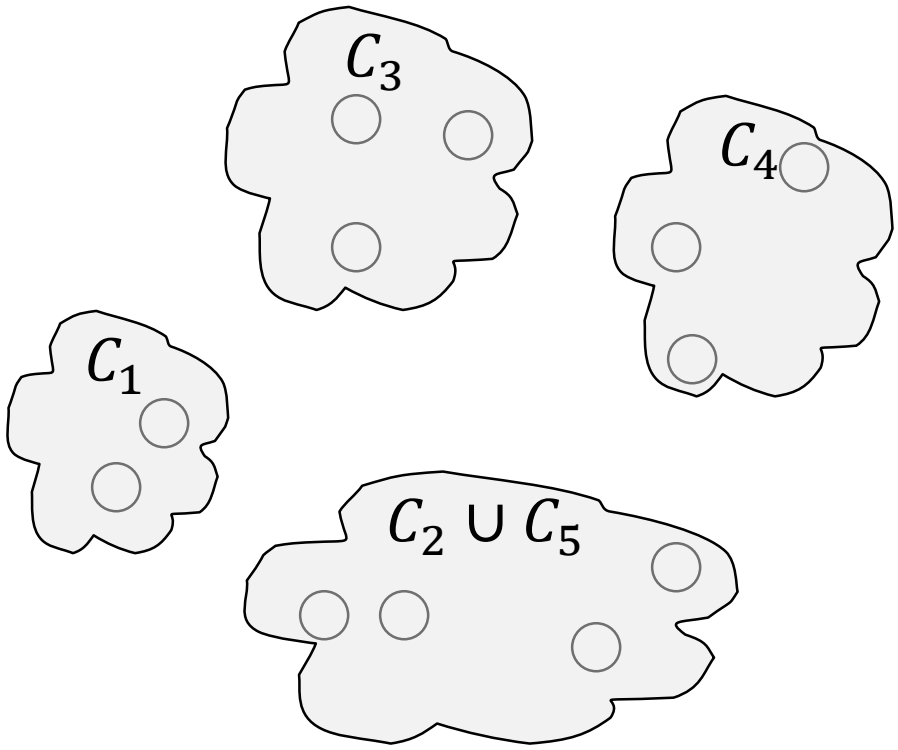
Агломеративная кластеризация (шаг $k+1$)



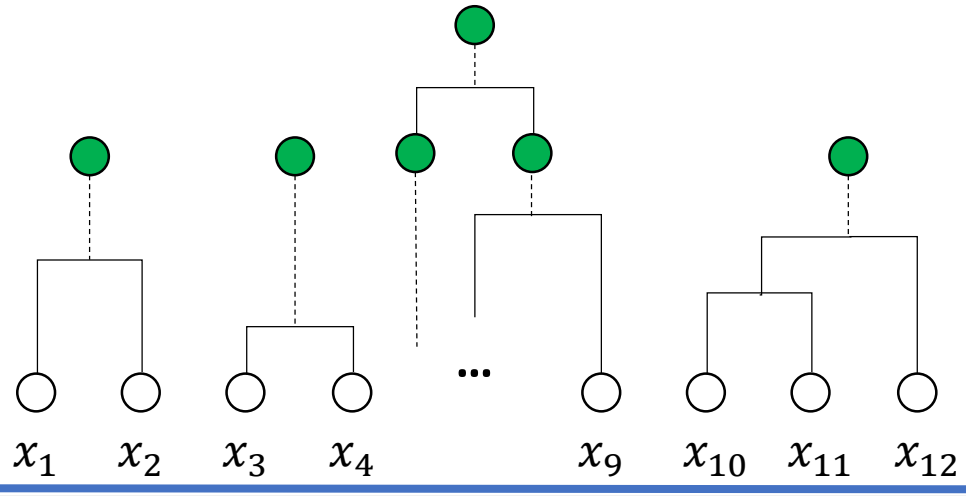
	c_1	c_2	c_3	c_4	c_5	…
c_1	0					
c_2		0				
c_3			0			
c_4				0		
c_5					0	
…						



Агломеративная кластеризация (шаг $k+1$)

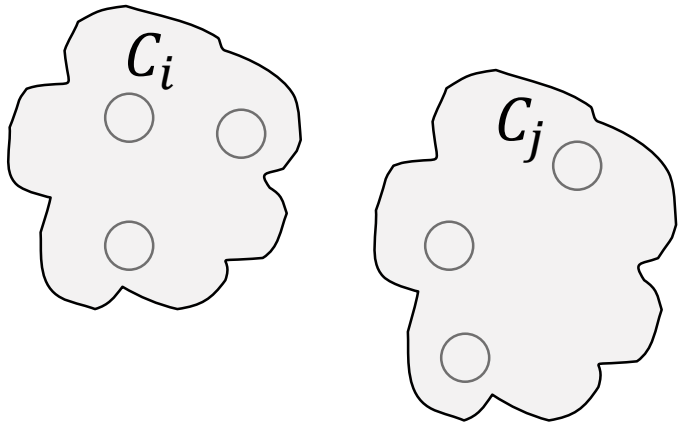


	c_1	$c_2 \cup c_5$	c_3	c_4	…
c_1	0				
$c_2 \cup c_5$		0			
c_3			0		
c_4				0	
…					



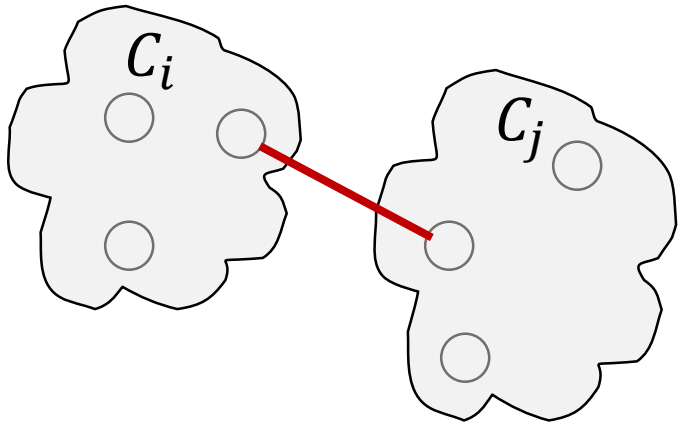
Меры схожести кластеров

- Min (Single linkage)
- Max (Complete linkage)
- Avg (Group average)
- CAvg (Centroid Average)
- Расстояние Уорда (Ward)



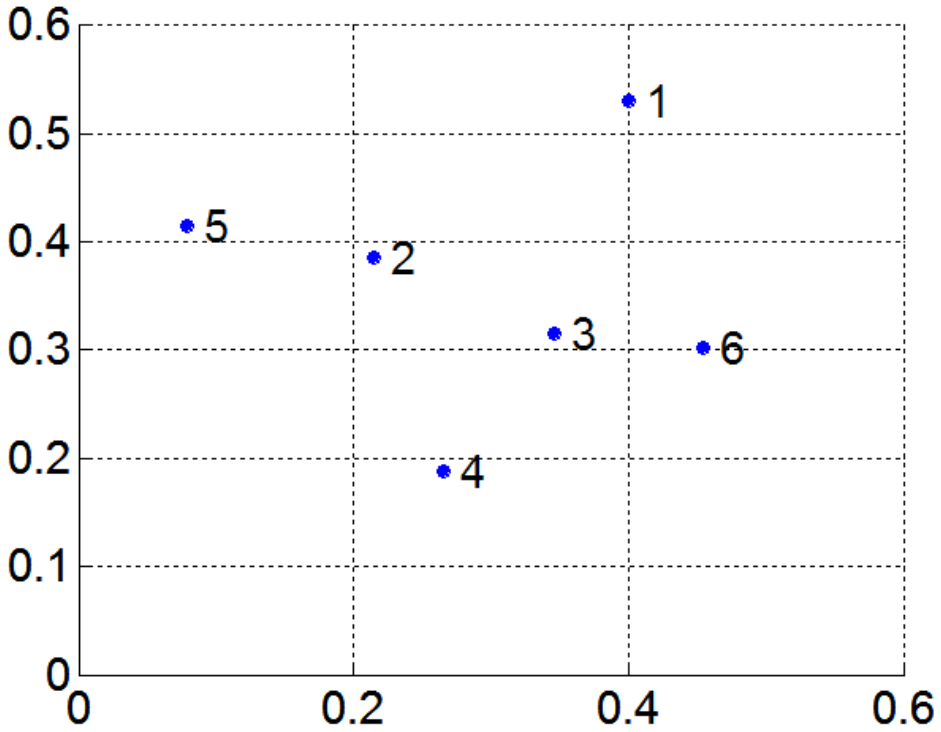
Меры схожести кластеров

- **Min (Single linkage)**
- Max (Complete linkage)
- Avg (Group average)
- CAvg (Centroid Average)
- Расстояние Уорда (Ward)



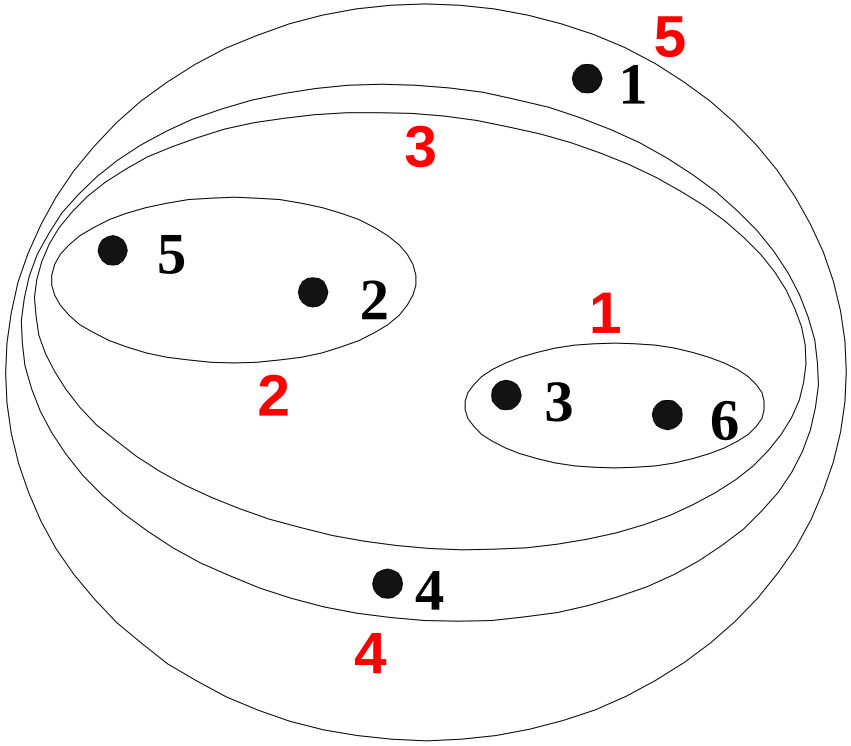
$$Dist\left(C_i, C_j\right) = \min_{x \in C_i,\, y \in C_j} \operatorname{dist}(x, y)$$

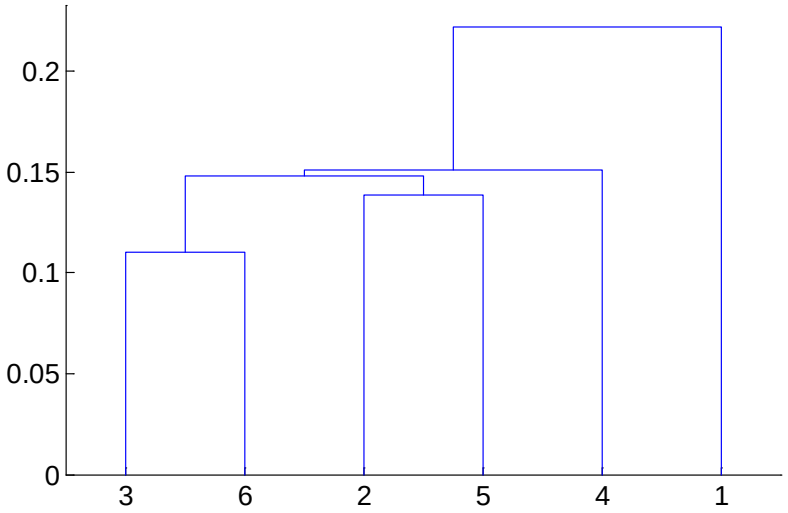
Min (Single linkage): пример



	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Min (Single linkage): пример





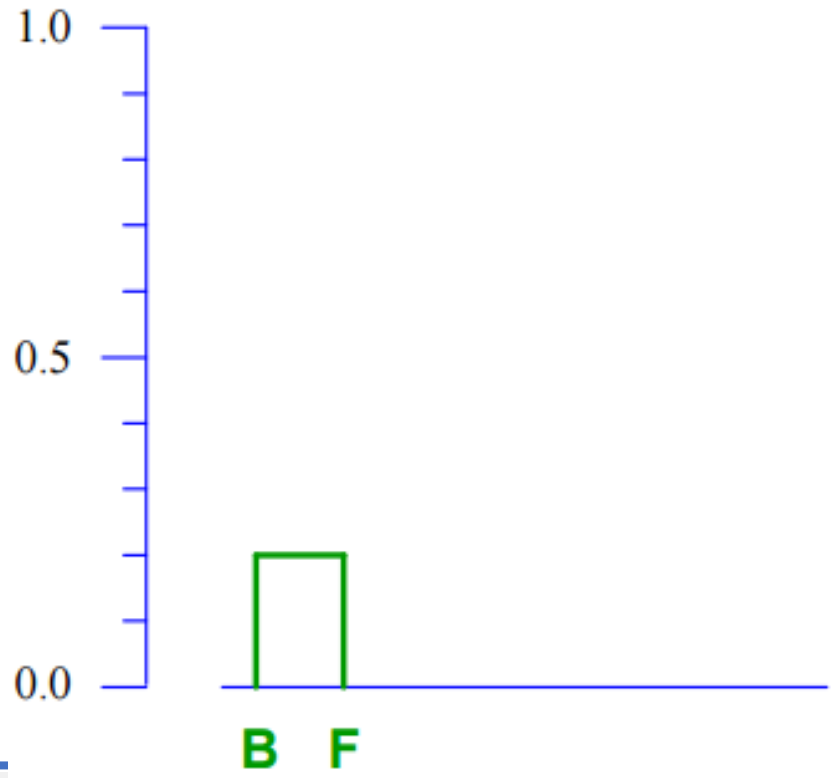
Иерархическая кластеризация: пример

samples	A	B	C	D	E	F	G
A	0	0.5000	0.4286	1.0000	0.2500	0.6250	0.3750
B	0.5000	0	0.7143	0.8333	0.6667	0.2000	0.7778
C	0.4286	0.7143	0	1.0000	0.4286	0.6667	0.3333
D	1.0000	0.8333	1.0000	0	1.0000	0.8000	0.8571
E	0.2500	0.6667	0.4286	1.0000	0	0.7778	0.3750
F	0.6250	0.2000	0.6667	0.8000	0.7778	0	0.7500
G	0.3750	0.7778	0.3333	0.8571	0.3750	0.7500	0

Матрица расстояний

Иерархическая кластеризация: пример (1)

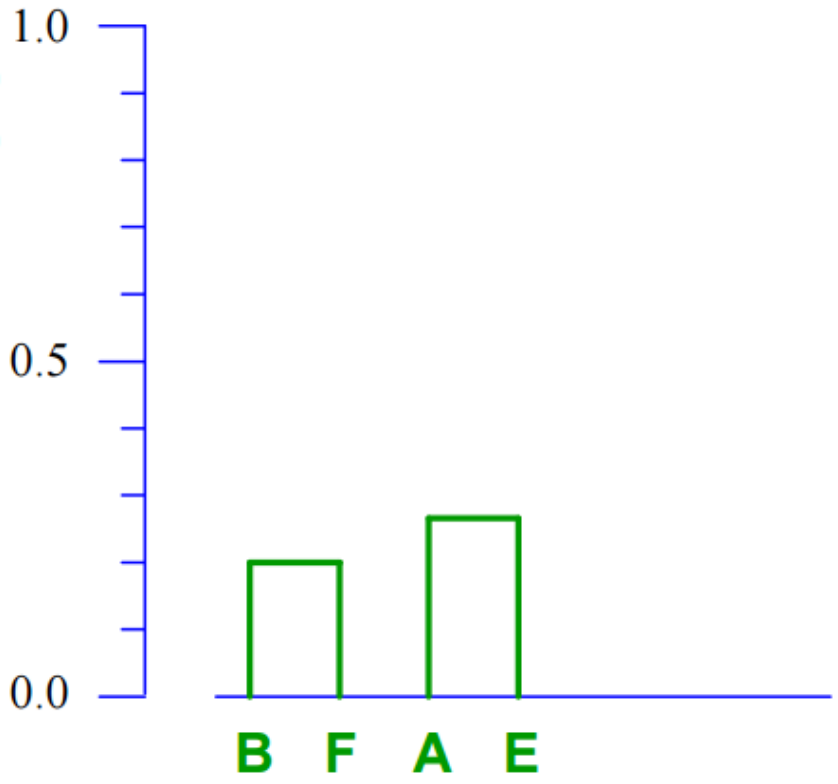
samples	A	B	C	D	E	F	G
A	0	0.5000	0.4286	1.0000	0.2500	0.6250	0.3750
B	0.5000	0	0.7143	0.8333	0.6667	0.2000	0.7778
C	0.4286	0.7143	0	1.0000	0.4286	0.6667	0.3333
D	1.0000	0.8333	1.0000	0	1.0000	0.8000	0.8571
E	0.2500	0.6667	0.4286	1.0000	0	0.7778	0.3750
F	0.6250	0.2000	0.6667	0.8000	0.7778	0	0.7500
G	0.3750	0.7778	0.3333	0.8571	0.3750	0.7500	0



- Min (Single linkage)

Иерархическая кластеризация: пример (2)

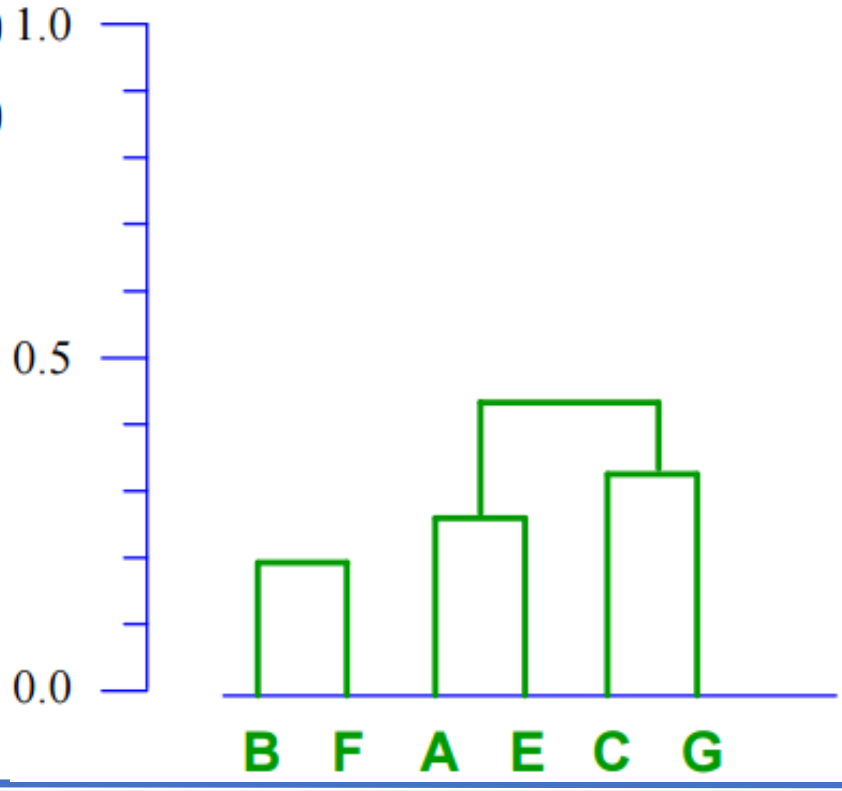
samples	A	(B,F)	C	D	E	G
A	0	0.6250	0.4286	1.0000	0.2500	0.3750
(B,F)	0.6250	0	0.7143	0.8333	0.7778	0.7778
C	0.4286	0.7143	0	1.0000	0.4286	0.3333
D	1.0000	0.8333	1.0000	0	1.0000	0.8571
E	0.2500	0.7778	0.4286	1.0000	0	0.3750
G	0.3750	0.7778	0.3333	0.8571	0.3750	0



- Min (Single linkage)

Иерархическая кластеризация: пример (3)

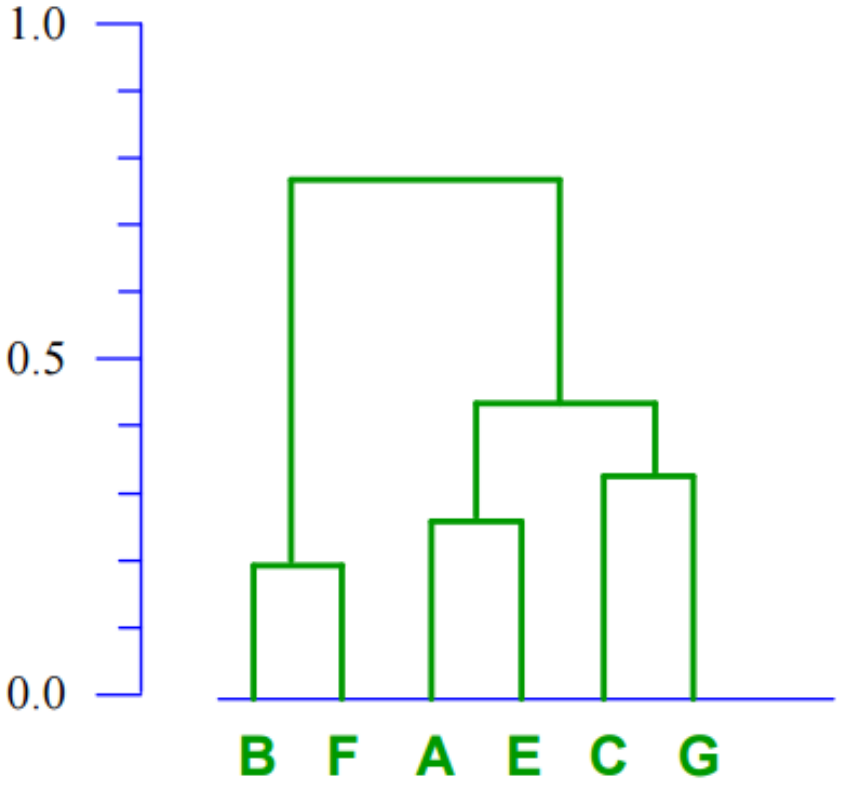
samples	(A,E)	(B,F)	(C,G)	D
(A,E)	0	0.7778	0.4286	1.0000
(B,F)	0.7778	0	0.7778	0.8333
(C,G)	0.4286	0.7778	0	1.0000
D	1.0000	0.8333	1.0000	0



- Min (Single linkage)

Иерархическая кластеризация: пример (4)

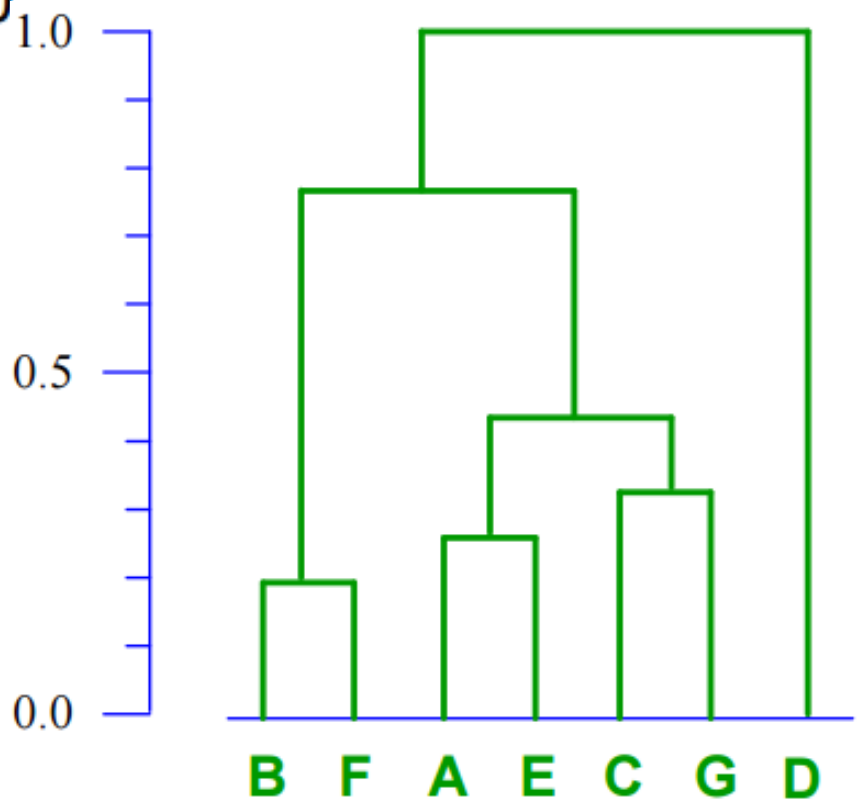
samples	(A,E,C,G)	(B,F)	D
(A,E,C,G)	0	0.7778	1.0000
(B,F)	0.7778	0	0.8333
D	1.0000	0.8333	0



- Min (Single linkage)

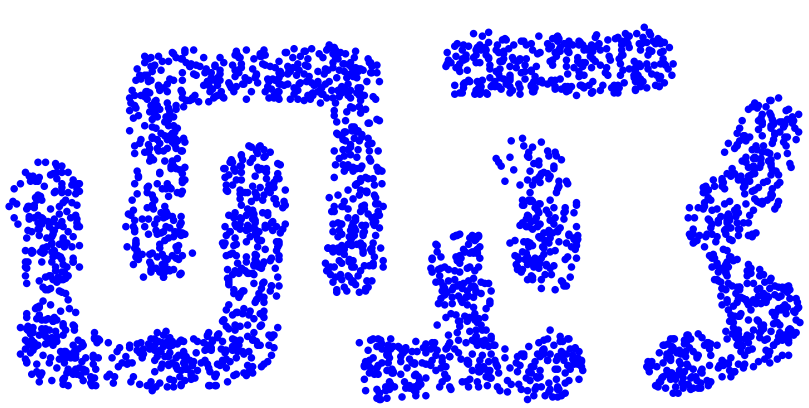
Иерархическая кластеризация: пример (5)

samples	(A,E,C,G,B,F)	D
(A,E,C,G,B,F)	0	1.0000
D	1.0000	0

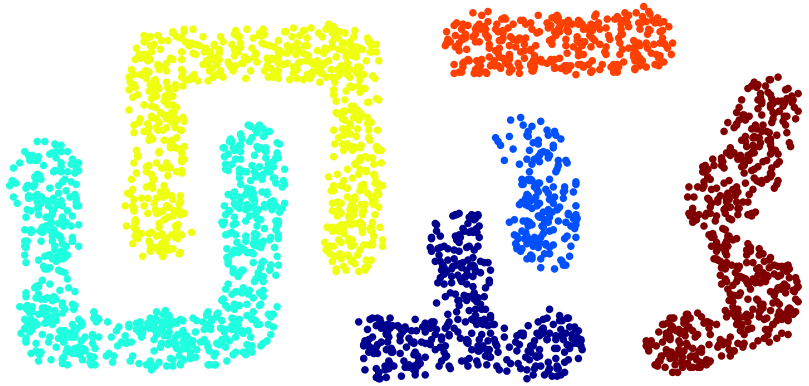


- Min (Single linkage)

Min (Single linkage): преимущества

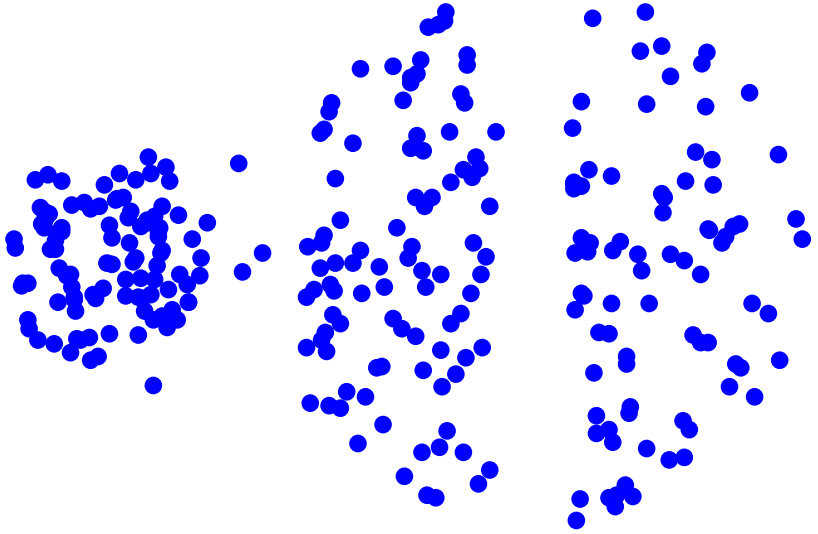


Исходное множество

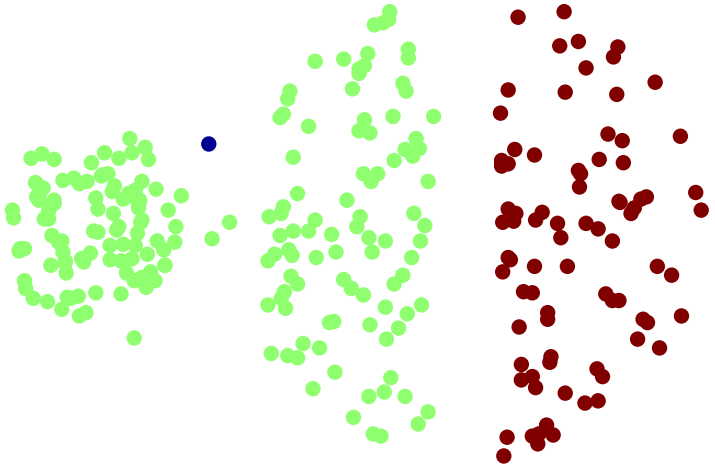


Результат кластеризации с применением меры MIN (6 кластеров)

Min (Single linkage): недостатки



Исходное множество

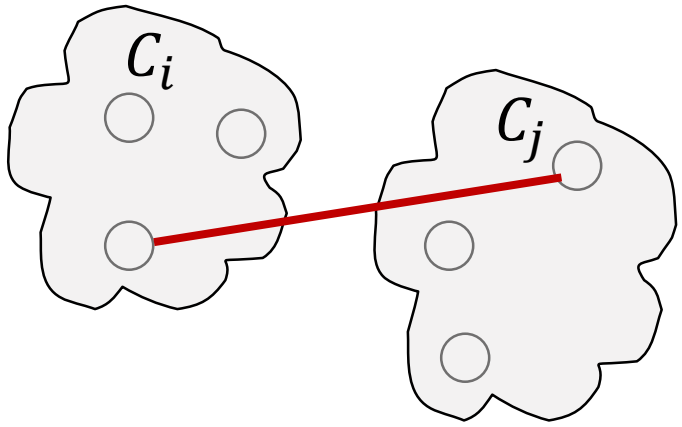


3 кластера

Чувствительность к шуму и выбросам

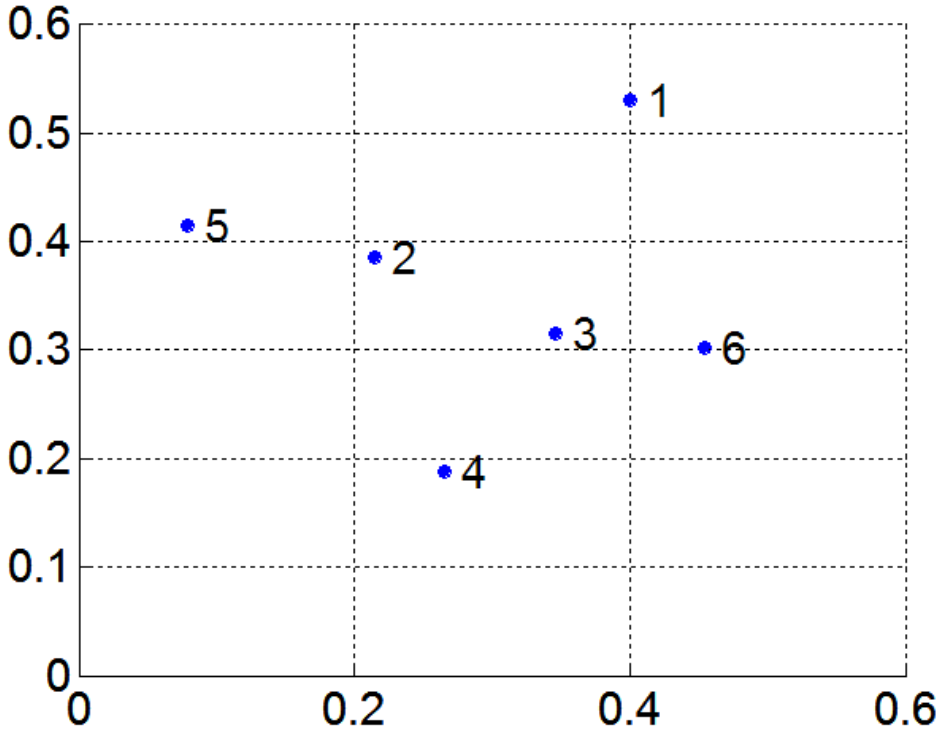
Меры схожести кластеров

- Min (Single linkage)
- **Max (Complete linkage)**
- Avg (Group average)
- CAvg (Centroid average)
- Расстояние Уорда (Ward)



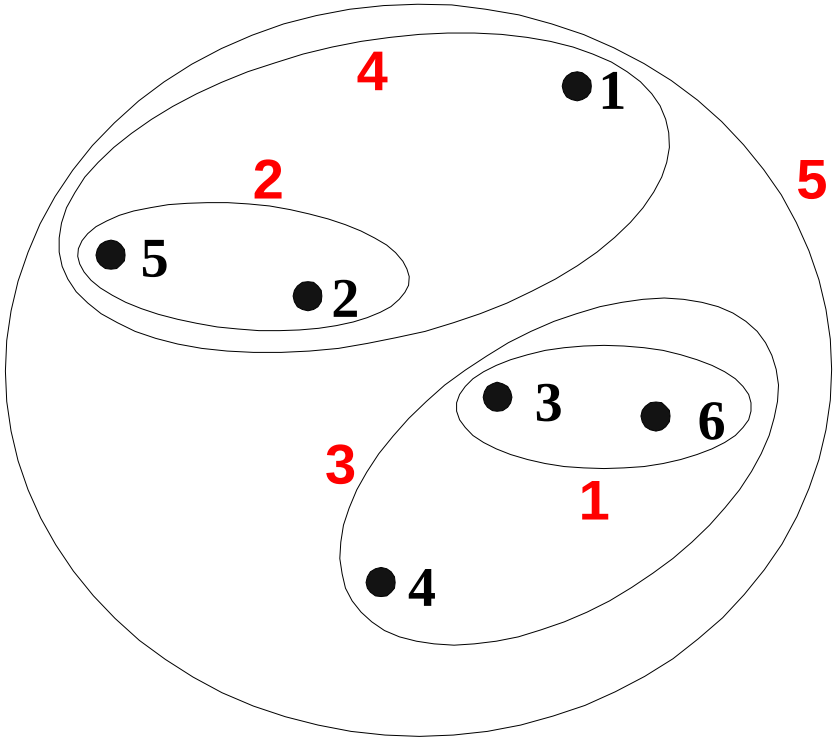
$$Dist\left(C_i, C_j\right) = \max_{x \in C_i,\, y \in C_j} \operatorname{dist}(x, y)$$

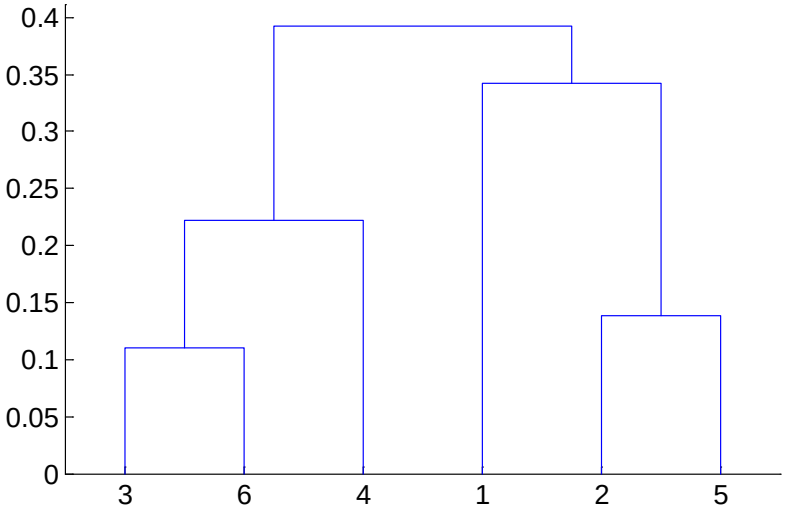
Max (Complete linkage): пример



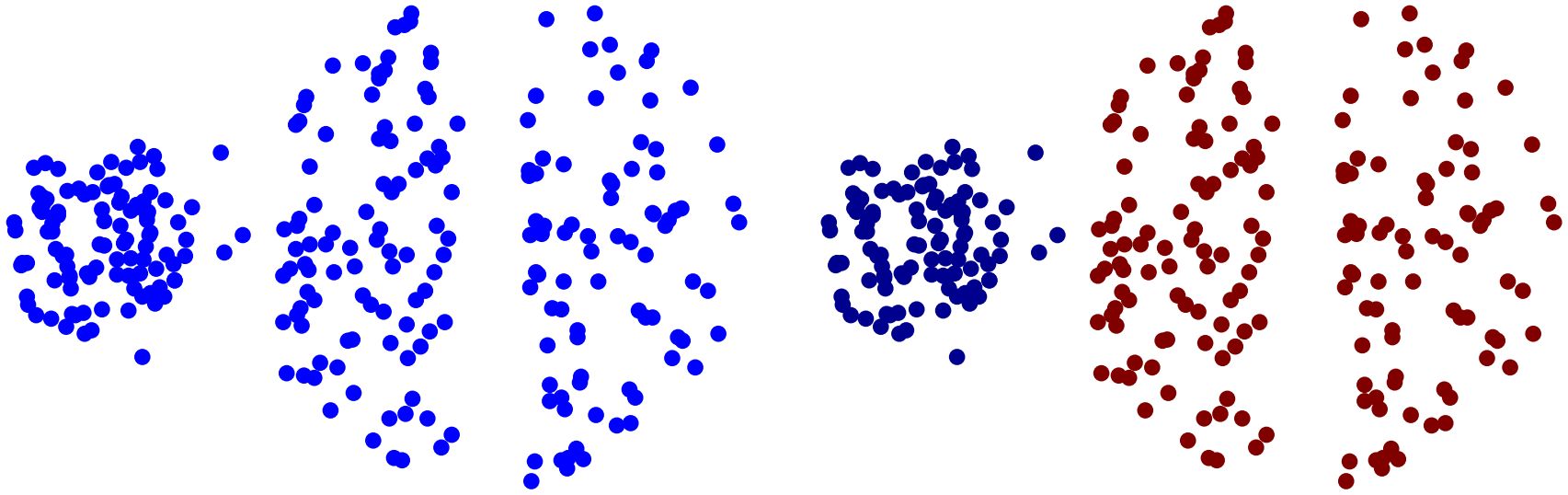
	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Max (Complete linkage): пример





Max (Complete linkage): преимущества

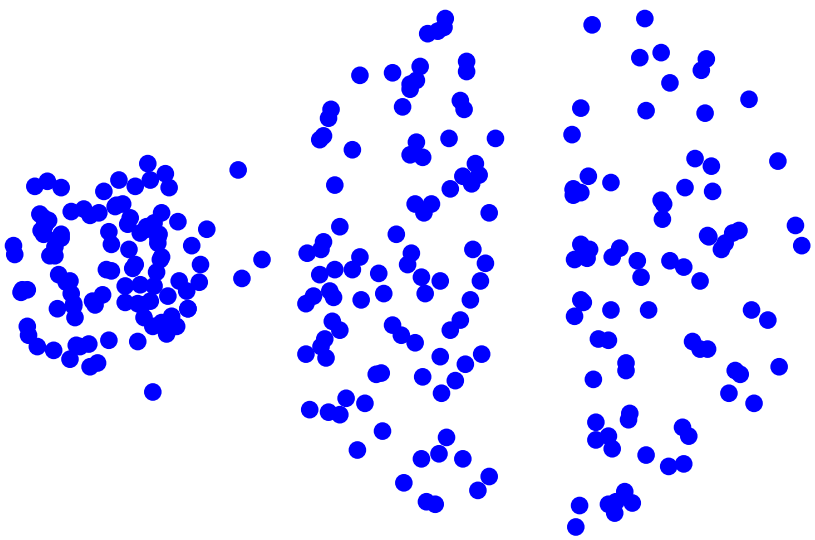


Исходное множество

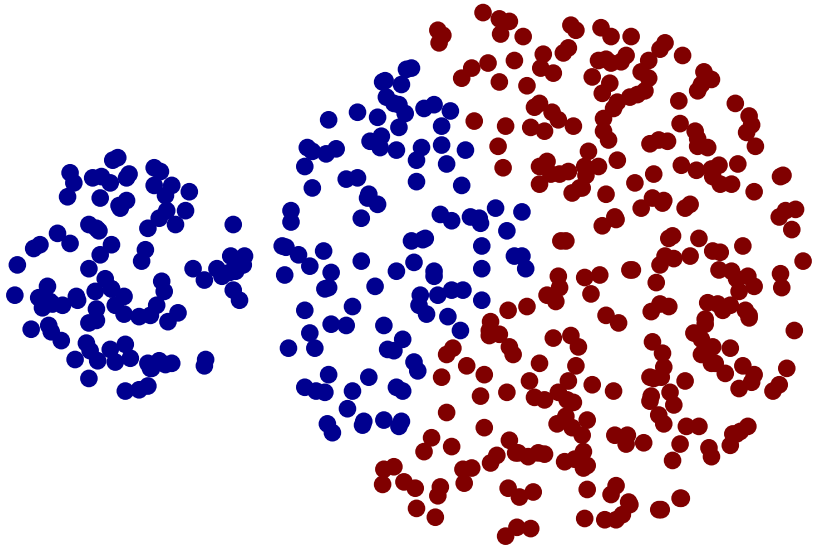
2 кластера

Низкая чувствительность к выбросам

Max (Complete linkage): недостатки



Исходное множество



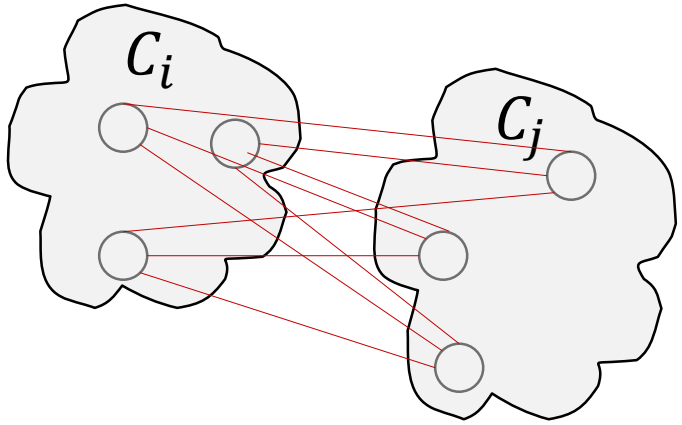
2 кластера

Тенденция к разбиению кластеров большой мощности

Тяготеет к кластерам выпуклой формы

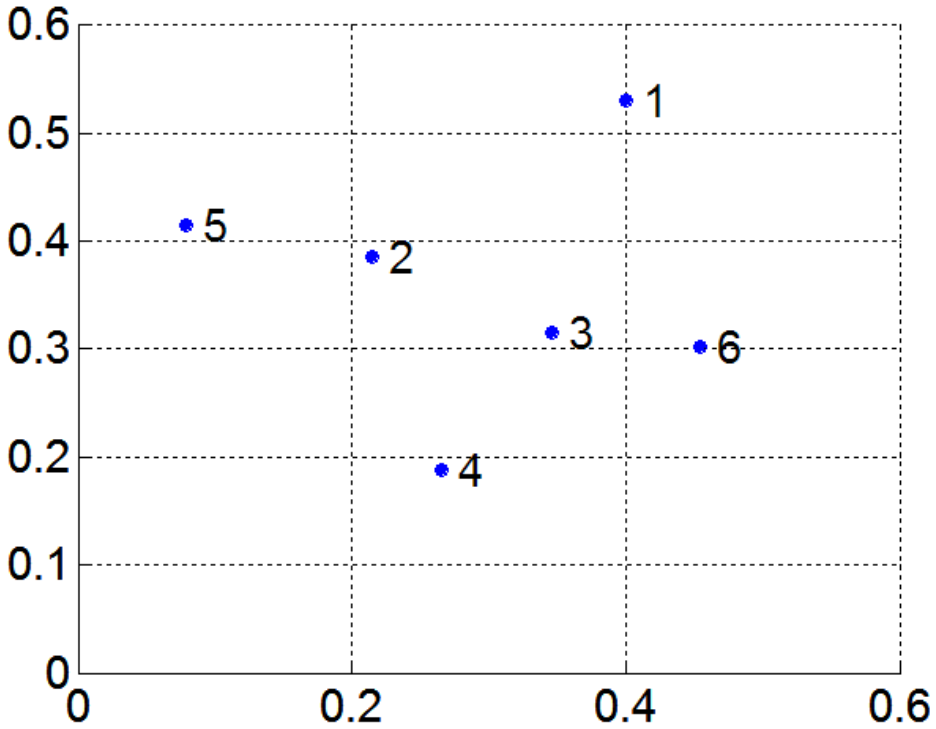
Меры схожести кластеров

- Min (Single linkage)
- Max (Complete linkage)
- **Avg (Group average)**
- Центроидная
- Расстояние Уорда (Ward)



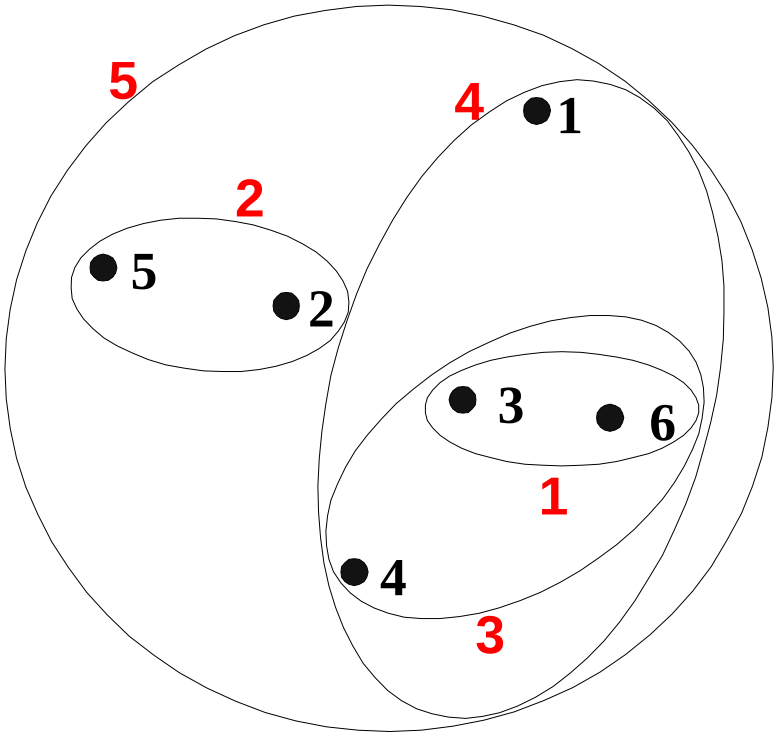
$$Dist\left(C_i, C_j\right) = \frac{\sum_{x \in C_i, y \in C_j} dist(x, y)}{|C_i| \cdot |C_j|}$$

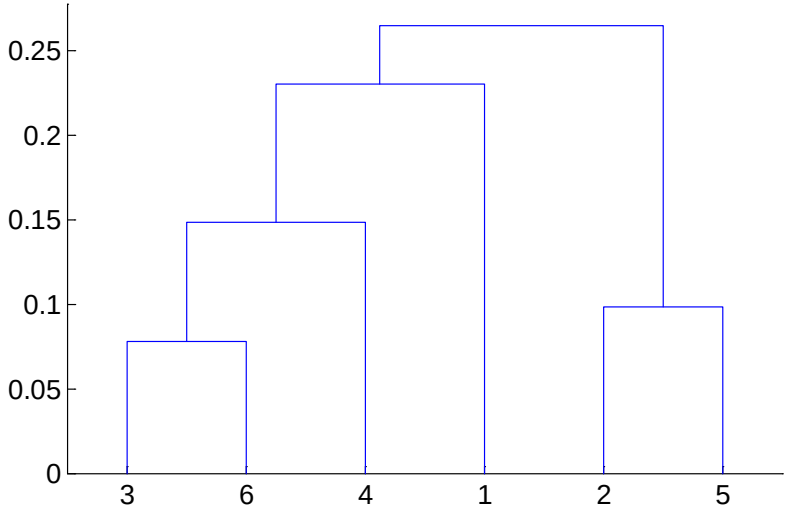
Avg (Group average): пример



	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Avg (Group average): пример



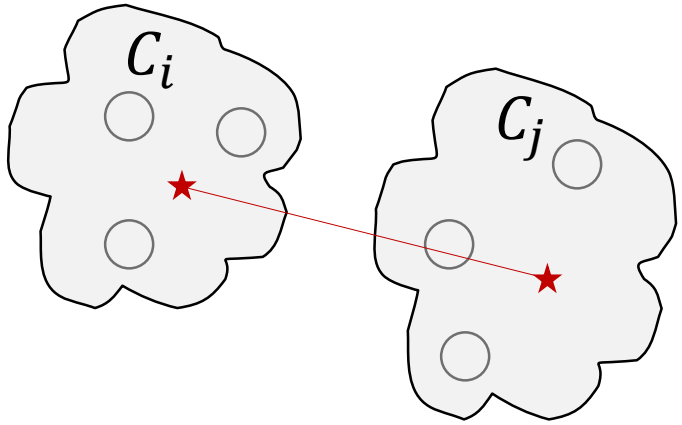


Avg (Group average): преимущества и недостатки

- Компромисс между Min (Single linkage) и Max (Complete linkage)
- Низкая чувствительность к шумам и выбросам
- Предпочтение выпуклым кластерам

Меры схожести кластеров

- Min (Single linkage)
- Max (Complete linkage)
- Avg (Group average)
- **CAvg (Centroid average)**
- Расстояние Уорда (Ward)



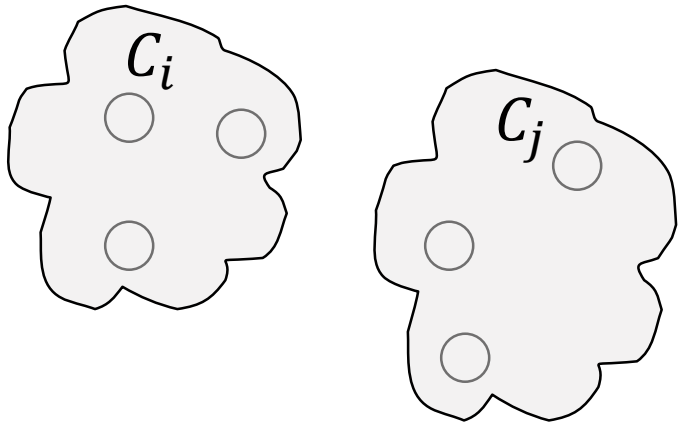
$$Dist\left(C_i, C_j\right) = dist^2\left(\sum_{x \in C_i} \frac{x}{|C_i|}, \sum_{y \in C_j} \frac{y}{|C_j|}\right)$$

Метод CAvg (Centroid average): недостатки

- Похож на k-means, но хуже, чем расстояние Уорда, поскольку допускает инверсии: два объединяемых кластера могут быть более похожими (менее удаленными), чем пара кластеров, которые были объединены на предыдущем шаге
- Для других методов приращение расстояния между объединенными кластерами неотрицательное по мере перехода от кластеров-синглтонов к одному всеобъемлющему кластеру

Меры схожести кластеров

- Min (Single linkage)
- Max (Complete linkage)
- Avg (Group average)
- CAvg (Centroid average)
- **Расстояние Уорда (Ward)**

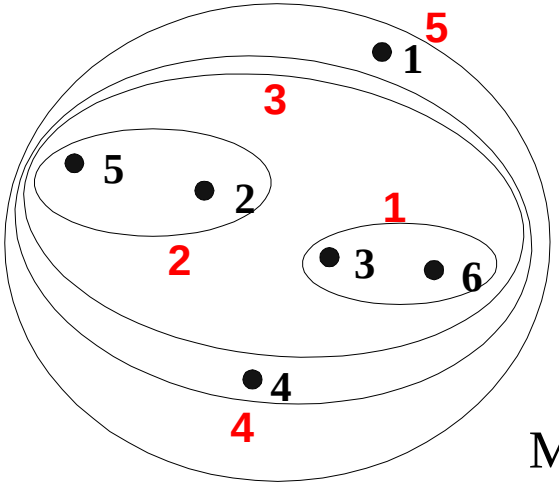


$$Dist\left(C_i, C_j\right) = \frac{|C_i| \cdot |C_j|}{|C_i| + |C_j|} \cdot dist^2 \left(\sum_{x \in C_i} \frac{x}{|C_i|}, \sum_{y \in C_j} \frac{y}{|C_j|}\right)$$

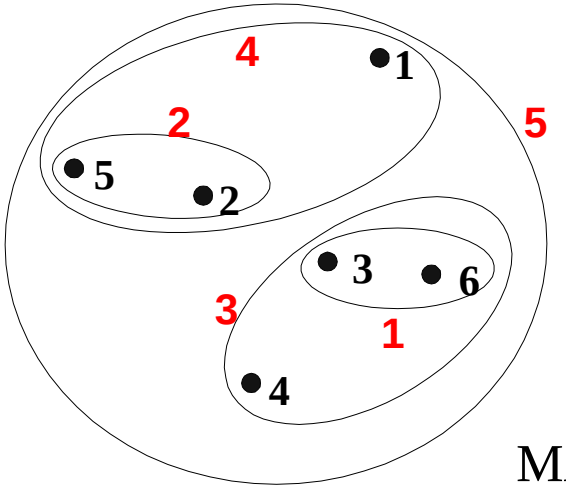
Метод Уорда

- Сходство двух кластеров вычисляется как прирост суммы квадратов расстояний объектов до центра кластера, получаемого в результате их объединения. Объединяются два кластера, приводящие к минимальному увеличению дисперсии
- Особенности
  - Применяется в задачах с близко расположенными кластерами
  - Низкая чувствительность к шумам и выбросам
  - Предпочтение выпуклым кластерам
  - Аналогично AVG (Group average), если расстояние между точками равно квадрату расстояния
  - Является иерархическим аналогом k-Means и может быть использован для инициализации центроидов в k-Means

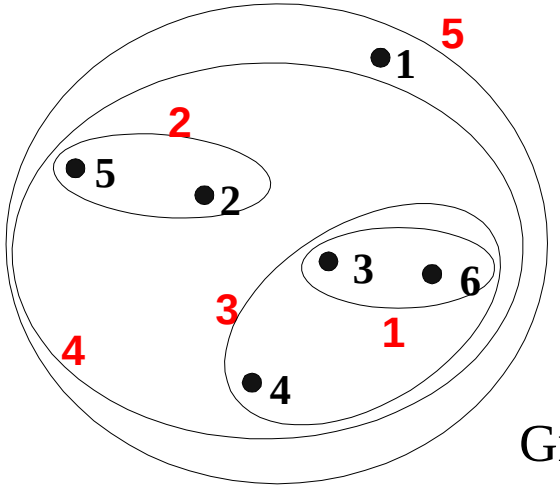
Сравнение мер



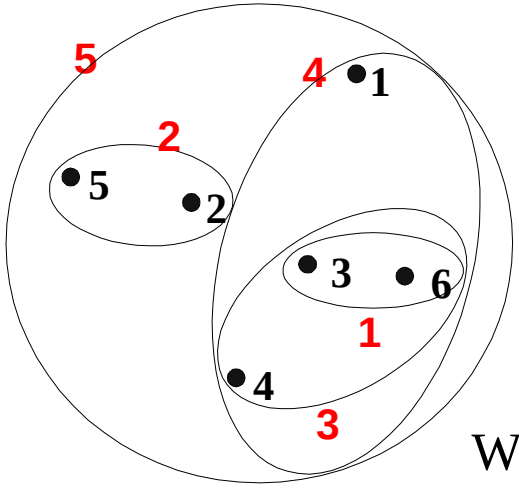
MIN



MAX

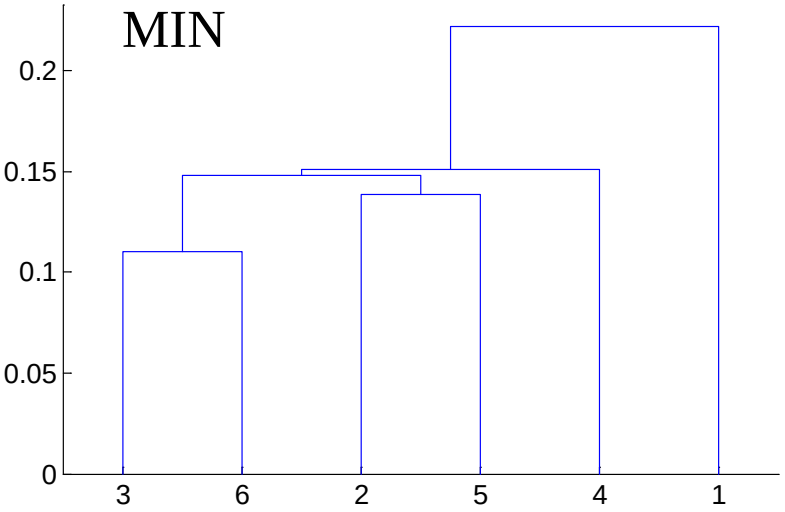


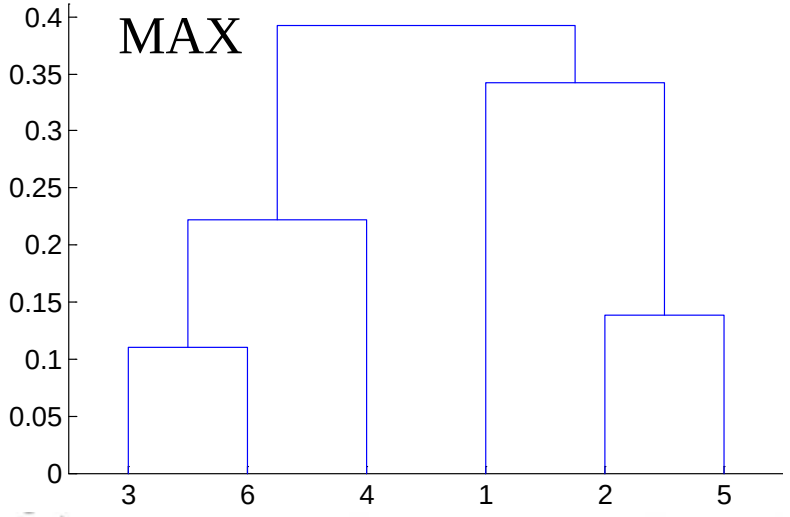
Group Average

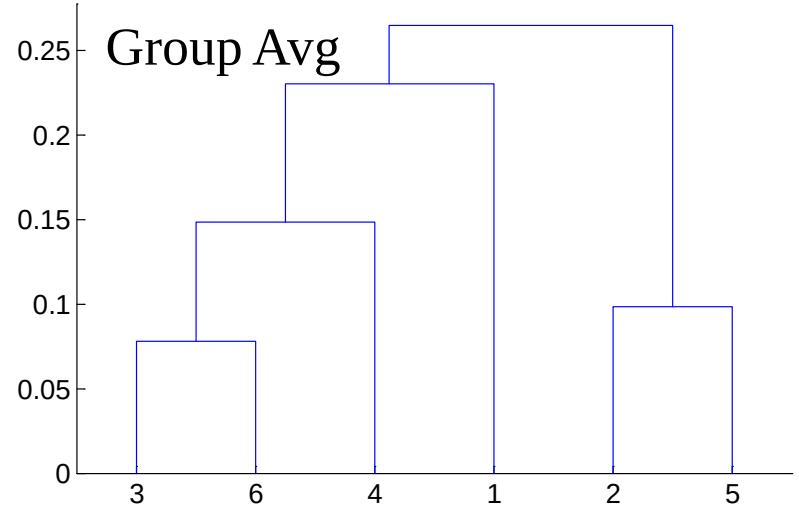


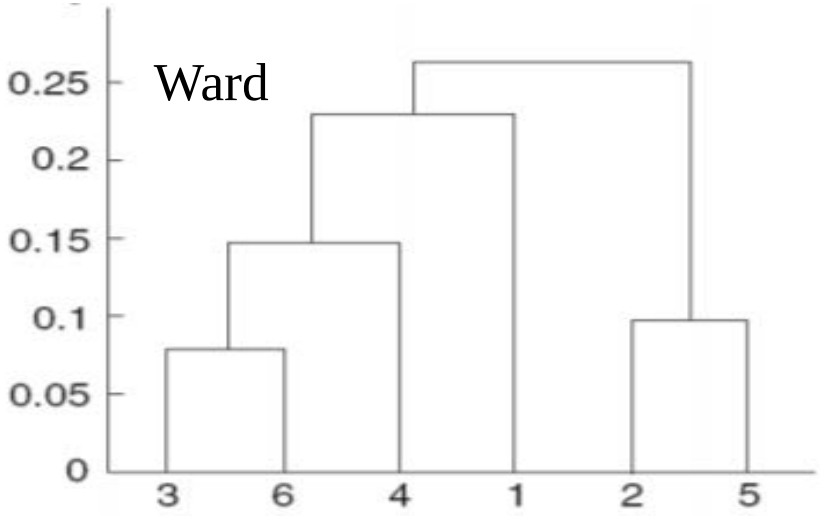
Ward

Сравнение мер









Сравнение мер: MIN (набор Iris)


setosa


versicolor


virginica

Длина чашелистика	Ширина чашелистика	Длина лепестка	Ширина лепестка	**Вид ириса**
…	…	…	…	setosa
…	…	…	…	versicolor
…	…	…	…	virginica

Сравнение мер: MAX (набор Iris)


setosa


versicolor


virginica

Длина чашелистика	Ширина чашелистика	Длина лепестка	Ширина лепестка	**Вид ириса**
…	…	…	…	setosa
…	…	…	…	versicolor
…	…	…	…	virginica

Сравнение мер: Avg (набор Iris)


setosa


versicolor


virginica

Длина чашелистика	Ширина чашелистика	Длина лепестка	Ширина лепестка	**Вид ириса**
…	…	…	…	setosa
…	…	…	…	versicolor
…	…	…	…	virginica

Сравнение мер: Уорд (набор Iris)


setosa


versicolor


virginica

Длина чашелистика	Ширина чашелистика	Длина лепестка	Ширина лепестка	**Вид ириса**
…	…	…	…	setosa
…	…	…	…	versicolor
…	…	…	…	virginica

Лучшие результаты среди всех мер.
Точно выделен вид *Iris setosa*
и заметно отделены виды
Iris virginica и *Iris versicolor*

Формула Ланса-Уильямса

$$p(R,Q) = \alpha_A\, p(A,Q) + \alpha_B\, p(B,Q) + \beta\, p(A,B) + \gamma\, |p(A,Q) - p(B,Q)|$$

Clustering Method	α_A	α_B	β	γ
Single Link	$1/2$	$1/2$	0	$-1/2$
Complete Link	$1/2$	$1/2$	0	$1/2$
Group Average	$\frac{m_A}{m_A+m_B}$	$\frac{m_B}{m_A+m_B}$	0	0
Centroid	$\frac{m_A}{m_A+m_B}$	$\frac{m_B}{m_A+m_B}$	$\frac{-m_A m_B}{(m_A+m_B)^2}$	0
Ward's	$\frac{m_A+m_Q}{m_A+m_B+m_Q}$	$\frac{m_B+m_Q}{m_A+m_B+m_Q}$	$\frac{-m_Q}{m_A+m_B+m_Q}$	0

- Обозначения
  $p(\cdot,\cdot)$ – расстояние между кластерами
  $R = A \cup B$: кластер R – результат слияния A и B
  m_A, m_B, m_Q – мощность кластеров
- Формула позволяет не хранить исходные точки данных, вместо этого обновляется матрица расстояний по мере выполнения кластеризации

Монотонность иерархической кластеризации

- R_t – расстояние между кластерами, выбранными на шаге t для объединения. Дендрограмма показывает иерархию на декартовой плоскости Объекты $\times R_t$
- Дендрограмма не имеет самопересечений (наглядна), если на оси объектов любой кластер представлен в виде непрерывного отрезка
- Условием наглядности дендрограммы является *монотонность расстояния* R_t: $R_2 \leq R_3 \leq \cdots \leq R_m$
- Теорема Миллигана: Расстояние R_t монотонно, если
  1. $\alpha_A \geq 0, \alpha_B \geq 0$
  2. $\alpha_A + \alpha_B + \beta \geq 1$
  3. $\min(\alpha_A, \alpha_B) + \gamma \geq 0$

Из рассмотренных расстояний теореме Миллигана удовлетворят все методы, *кроме центроидного*

Иерархическая кластеризация: резюме

- Пространственная сложность: $O(n^2)$ (хранение матрицы расстояний)
- Временная сложность: как правило, $O(n^3)$
  - n шагов, на каждом из которых n^2 операций с матрицей расстояний
  - существуют приемы для сокращения сложности до $O(n^2 \log n)$
- Жадный алгоритм
- Отсутствует глобальная целевая функция, которая явно минимизируется
- Прочие сложности
  - Чувствительность к шумам и выбросам
  - Обработка кластеров различных мощностей и невыпуклых форм

Литература

- Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. 3rd Edition. Morgan Kaufmann, 2011. 740 p. ISBN 978-0123814791
  - 10.3. Hierarchical Methods, pp. 457-470
- Tan P.-N., Steinbach M., Karpatne A., Kumar V. Introduction to Data Mining. 2nd Edition. Pearson, 2019. 839 p. ISBN: 978-0-13-312890-1
  - 7.3 Agglomerative Hierarchical Clustering, pp. 554-565