Постановка задачи кластеризации

Дано:

X — пространство объектов;

$X^\ell = \{x_i\}_{i=1}^\ell$ — обучающая выборка;

$\rho\colon X \times X \to [0, \infty)$ — функция расстояния между объектами.

Найти:

$y_i \in Y$ — метки кластеров объектов:

— каждый кластер состоит из близких объектов;

— объекты разных кластеров существенно различны.

Кластеризация — это *обучение без учителя*.

Как определить число кластеров?

Вместо этого можно строить иерархическую кластеризацию.

Агломеративная иерархическая кластеризация

Алгоритм Ланса-Уильямса [1967] основан на оценивании расстояний $R(U, V)$ между парами кластеров U, V.

1: сначала все кластеры одноэлементные:
$t := 1$; $C_t = \{\{x_1\}, \ldots, \{x_\ell\}\}$;
$R(\{x_i\}, \{x_j\}) := \rho(x_i, x_j)$;
2: **для всех** $t = 2, \ldots, \ell$ (t — номер итерации):
3: найти в C_{t-1} два ближайших кластера:
$(U, V) := \arg\min_{U \neq V} R(U, V)$;
$R_t := R(U, V)$;
4: слить их в один кластер:
$W := U \cup V$;
$C_t := C_{t-1} \cup \{W\} \setminus \{U, V\}$;
5: **для всех** $S \in C_t$
6: вычислить $R(W, S)$ по формуле Ланса-Уильямса;

Формула Ланса-Уильямса

Как определить расстояние $R(W, S)$
между кластерами $W = U \cup V$ и S,
зная расстояния $R(U, S)$, $R(V, S)$, $R(U, V)$?

Формула, обобщающая большинство разумных способов определить это расстояние [Ланс, Уильямс, 1967]:

$$\begin{aligned} R(U \cup V, S) = \alpha_U \cdot R(U, S) + \\ + \alpha_V \cdot R(V, S) + \\ + \beta \cdot R(U, V) + \\ + \gamma \cdot \big|R(U, S) - R(V, S)\big|, \end{aligned}$$

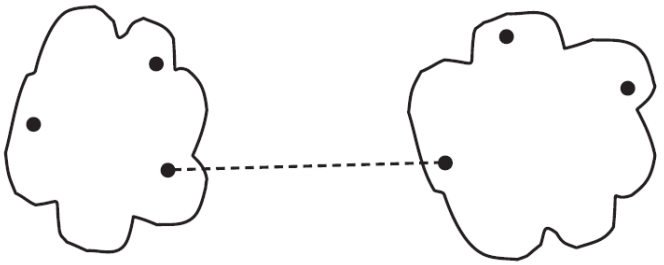
где α_U, α_V, β, γ — числовые параметры.

Частные случаи формулы Ланса-Уильямса

1. Расстояние ближнего соседа:

$$R^{б}(W,S) = \min_{w\in W, s\in S} \rho(w,s);$$

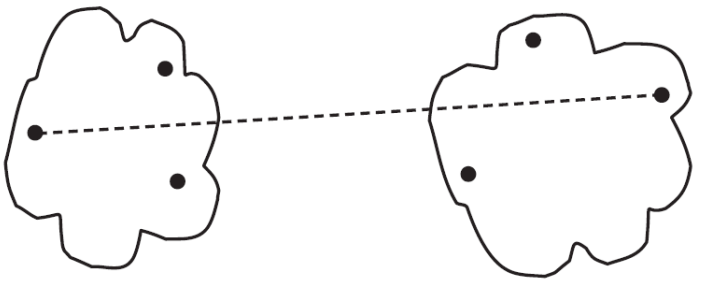
$$\alpha_U = \alpha_V = \tfrac{1}{2},\ \ \beta = 0,\ \ \gamma = -\tfrac{1}{2}.$$



2. Расстояние дальнего соседа:

$$R^{д}(W,S) = \max_{w\in W, s\in S} \rho(w,s);$$

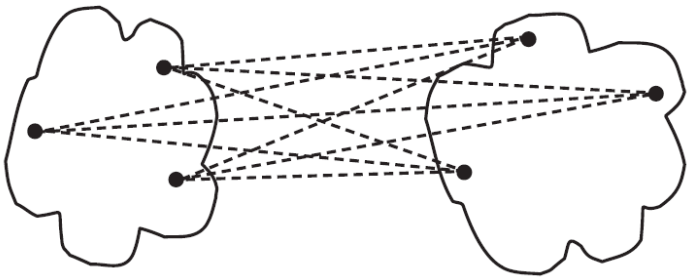
$$\alpha_U = \alpha_V = \tfrac{1}{2},\ \ \beta = 0,\ \ \gamma = \tfrac{1}{2}.$$



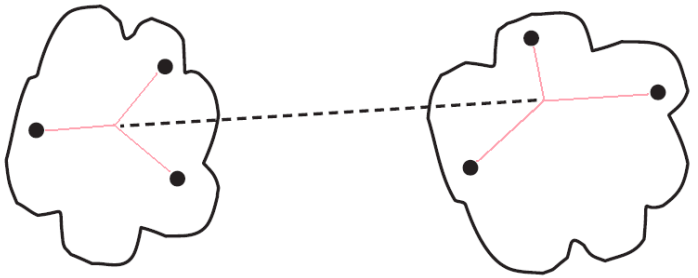
3. Групповое среднее расстояние:

$$R^{г}(W,S) = \frac{1}{|W||S|} \sum_{w\in W} \sum_{s\in S} \rho(w,s);$$

$$\alpha_U = \frac{|U|}{|W|},\ \ \alpha_V = \frac{|V|}{|W|},\ \ \beta = \gamma = 0.$$



4. Расстояние между центрами:



$$R^{\text{ч}}(W,S) = \rho^2\Big(\sum_{w\in W}\frac{w}{|W|}, \sum_{s\in S}\frac{s}{|S|}\Big);$$

$$\alpha_U = \frac{|U|}{|W|}, \quad \alpha_V = \frac{|V|}{|W|},$$

$$\beta = -\alpha_U\alpha_V, \quad \gamma = 0.$$

5. Расстояние Уорда:

$$R^{\text{у}}(W,S) = \frac{|S||W|}{|S|+|W|}\,\rho^2\Big(\sum_{w\in W}\frac{w}{|W|}, \sum_{s\in S}\frac{s}{|S|}\Big);$$

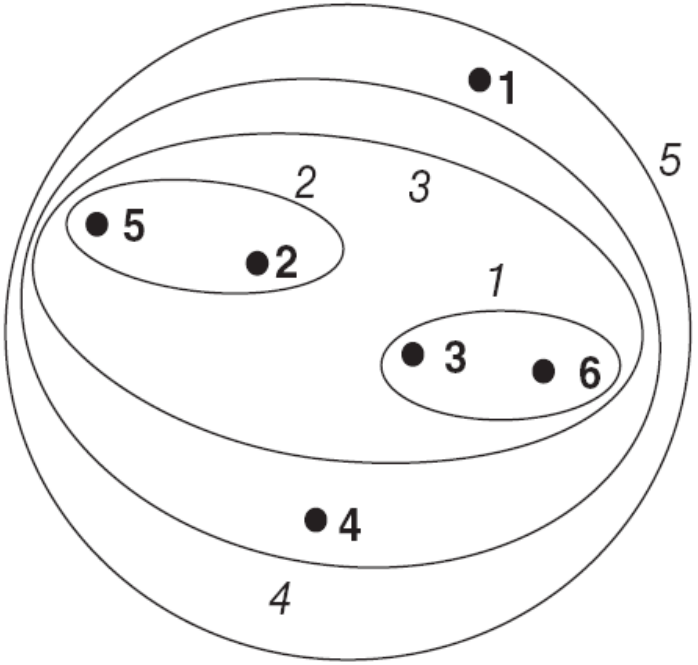
$$\alpha_U = \frac{|S|+|U|}{|S|+|W|}, \quad \alpha_V = \frac{|S|+|V|}{|S|+|W|}, \quad \beta = \frac{-|S|}{|S|+|W|}, \quad \gamma = 0.$$

Проблема выбора

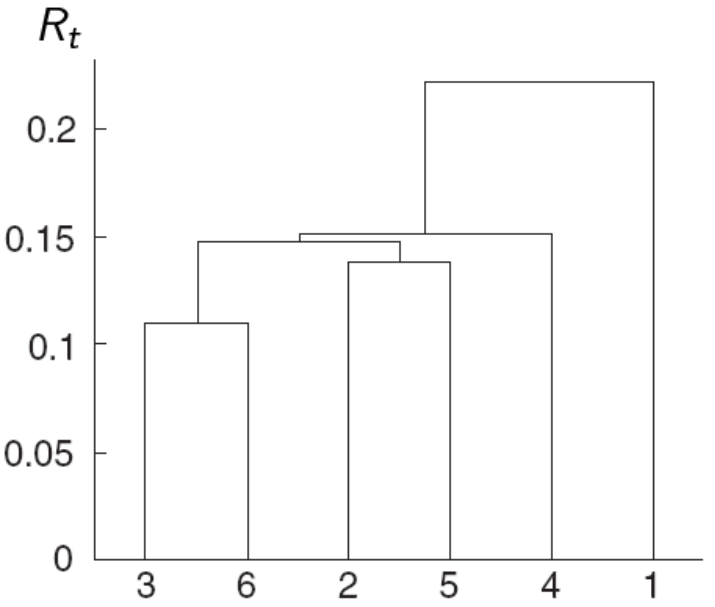
Какая функция расстояния лучше?

1. Расстояние ближнего соседа:

Диаграмма вложения

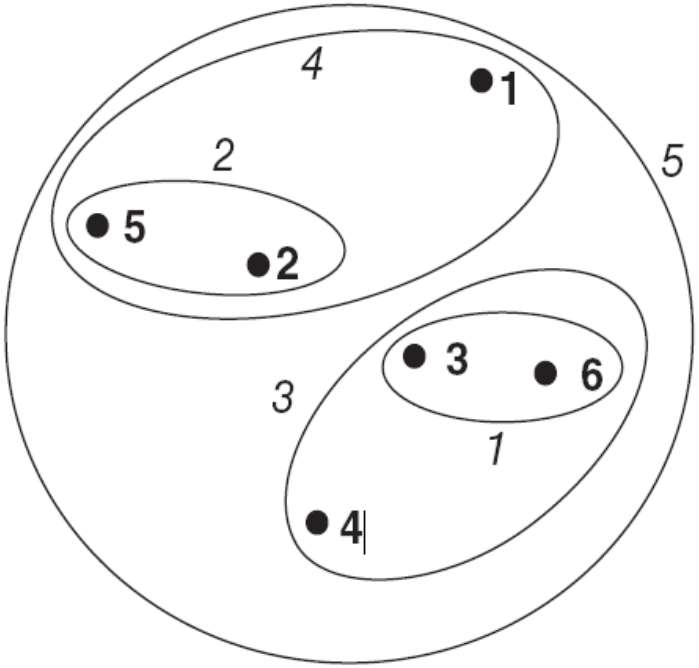


Дендрограмма

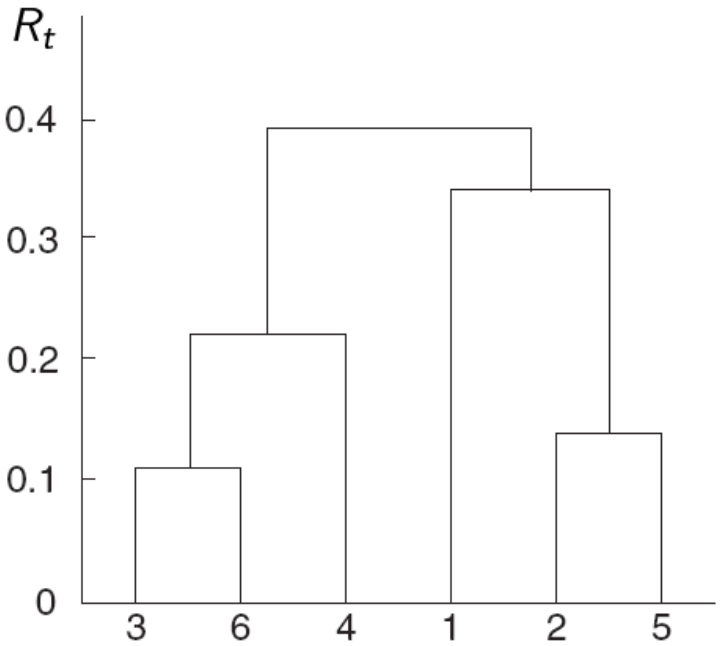


2. Расстояние дальнего соседа:

Диаграмма вложения

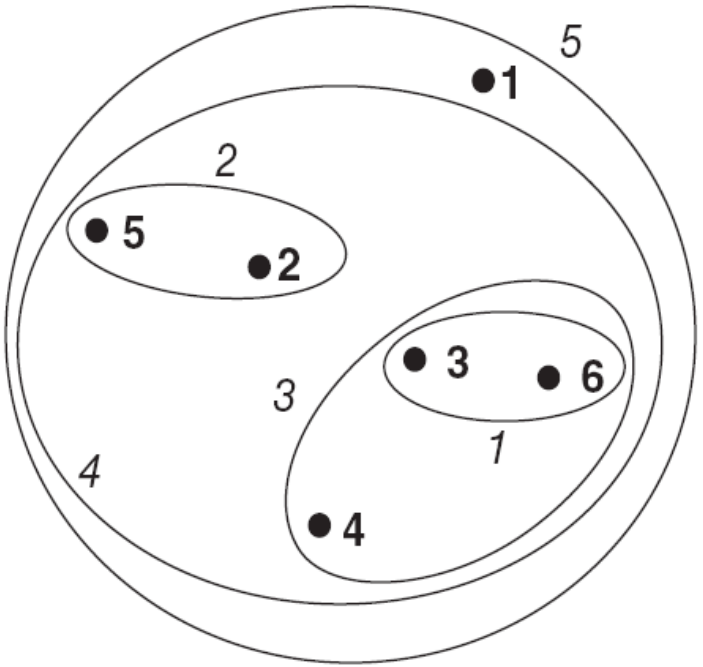


Дендрограмма

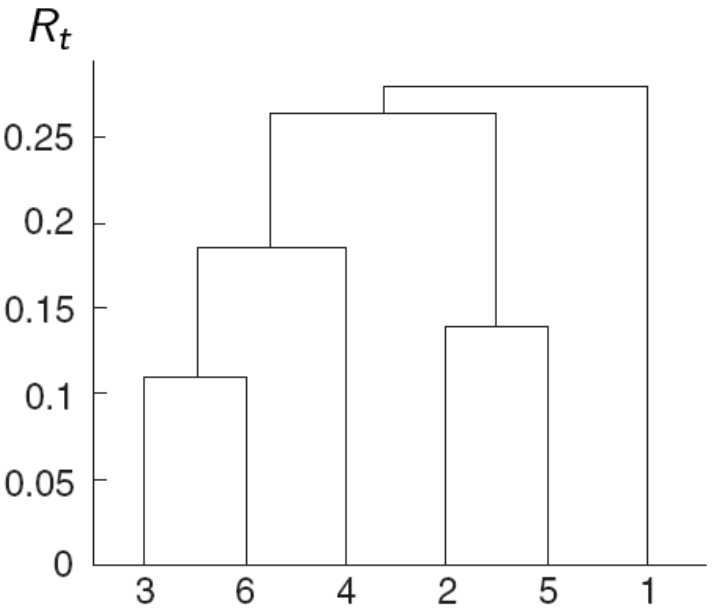


3. Групповое среднее расстояние:

Диаграмма вложения

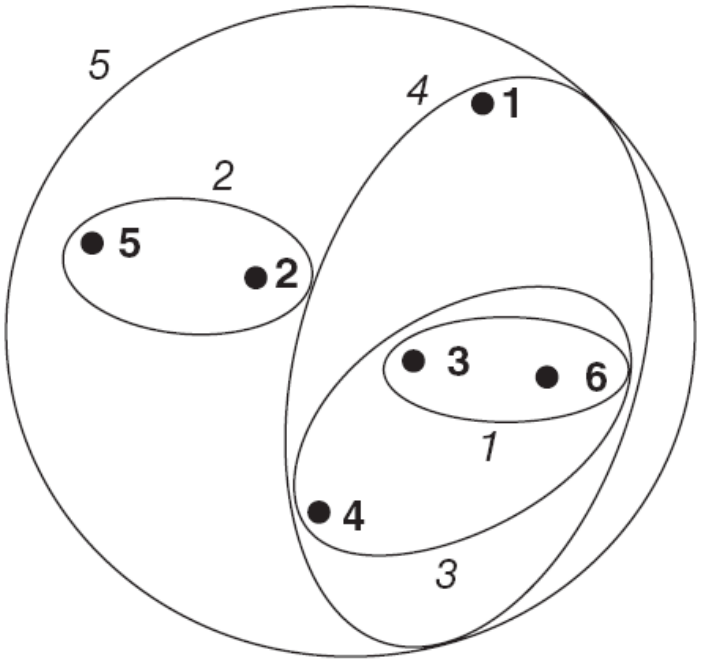


Дендрограмма

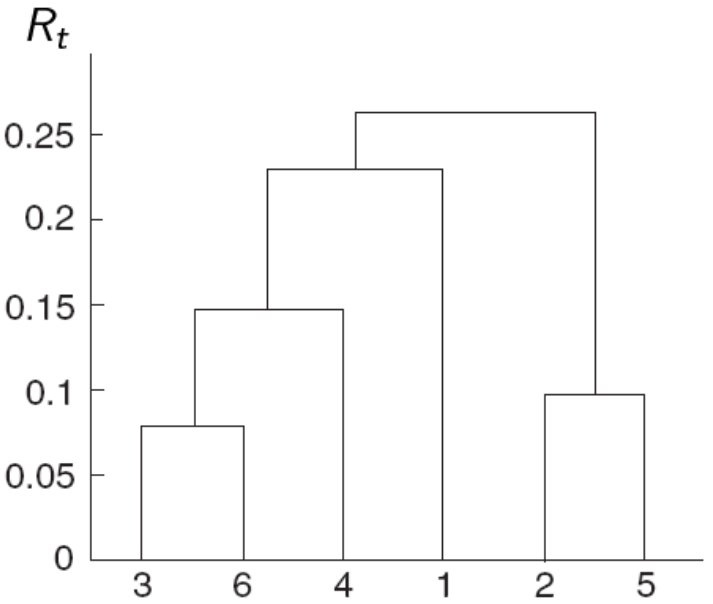


5. Расстояние Уорда:

Диаграмма вложения



Дендрограмма



Основные свойства иерархической кластеризации

- *Монотонность*: дендрограмма не имеет самопересечений, при каждом слиянии расстояние между объединяемыми кластерами только увеличивается: $R_2 \leqslant R_3 \leqslant \ldots \leqslant R_\ell$.
  Достаточное условие монотонности:

  $$\alpha_U \geqslant 0, \quad \alpha_V \geqslant 0, \quad \alpha_U + \alpha_V + \beta \geqslant 1, \quad \min\{\alpha_U, \alpha_V\} + \gamma \geqslant 0.$$

- *Сжимающее расстояние*: $R_t \leqslant \rho(\mu_U, \mu_V), \ \forall t$.
- *Растягивающее расстояние*: $R_t \geqslant \rho(\mu_U, \mu_V), \ \forall t$

$R^{ч}$ не монотонно; $R^{б}$, $R^{д}$, $R^{г}$, $R^{у}$ — монотонны.

$R^{б}$ — сжимающее; $R^{д}$, $R^{у}$ — растягивающие;

Рекомендации и выводы

- рекомендуется пользоваться расстоянием Уорда $R^{\text{У}}$;
- обычно строят несколько вариантов и выбирают лучший визуально по дендрограмме;
- определение числа кластеров — по максимуму $|R_{t+1} - R_t|$, тогда результирующее множество кластеров $:= C_t$.