Машинное обучение

Лекция 3. Методы кластеризации

Катя Тузова

Общие вопросы по домашке

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{TP + FN}$$

- TP — это количество элементов, которые классификатор верно отнёс к классу c,
- FP — количество элементов, которые классификатор неверно отнёс к классу c,
- FN — количество элементов, которые классификатор неверно отнёс к классу, отличному от c.

Общие вопросы по домашке

- Нужна ли нормировка признаков?
- В чем минус поиска ближайших соседей с помощью сортировки?

Общие вопросы по домашке

```
len(X_train) / (len(X_test) + len(X_train)) == ratio
len(y_train) / (len(y_test) + len(y_train)) == ratio
```

Как выбрать ratio?

Общие вопросы по домашке

- Надо ли было брать $k = 1$ в функции loocv?
- Можно ли не перебирать k от 1 до n?
- Можно ли остановиться при выборе k как только начнет увеличиваться LOO?
- Надо ли было использовать тестовые данные в функции loocv?

Пример зависимости k

картинка с machinelearning.ru

Быстрый поиск ближайшего соседа

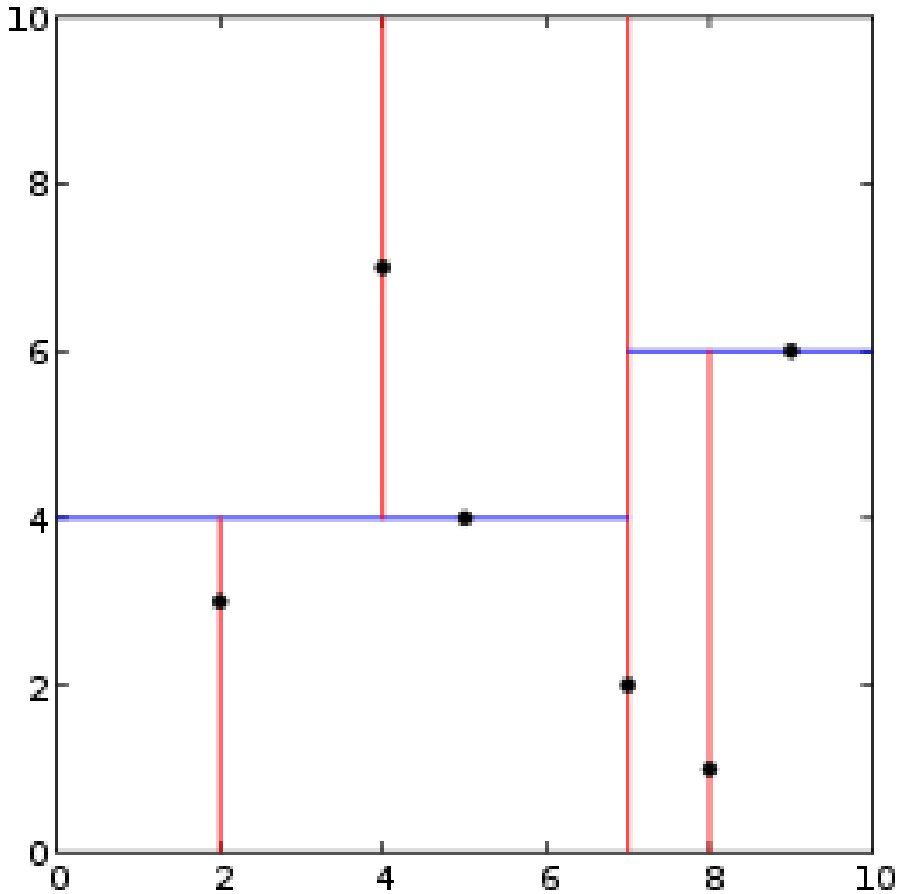
k-d дерево

Идея: разложим множество по поторому будем искать в бинарное дерево с простыми условиями и конкретными точками в узлах.

1. По циклу, или рандомно выбираем ось.
2. Ищем медиану (точку, разбивающую множество на как можно более равные части).
3. Повторяем 1-2 для каждого из получившихся подмножеств

Сложность построения: $O(n \log n)$
Сложность поиска: в лучшем случае $O(\log n)$, в худшем – $O(n)$

2-d дерево



k-d дерево. Особенности

+ Один из наиболее простых методов
– Работает только при малом количестве параметров
– Затратный алгоритм перестроения

Locality Sensitive Hash

Задача: Найти похожие документы в интернете

Locality Sensitive Hash

Проблема: Сколько сравнений нам понадобится для того, чтобы найти похожие среди N документов?

Locality Sensitive Hash

Проблема: Сколько сравнений нам понадобится для того, чтобы найти похожие среди N документов?

$$C = \frac{N(N-1)}{2}$$

$$N = 10^6 \Rightarrow C = 5 * 10^{11}$$

Locality Sensitive Hash

Идея:
Давайте от каждого документа (строки из нулей и единиц) возьмем хэш h:

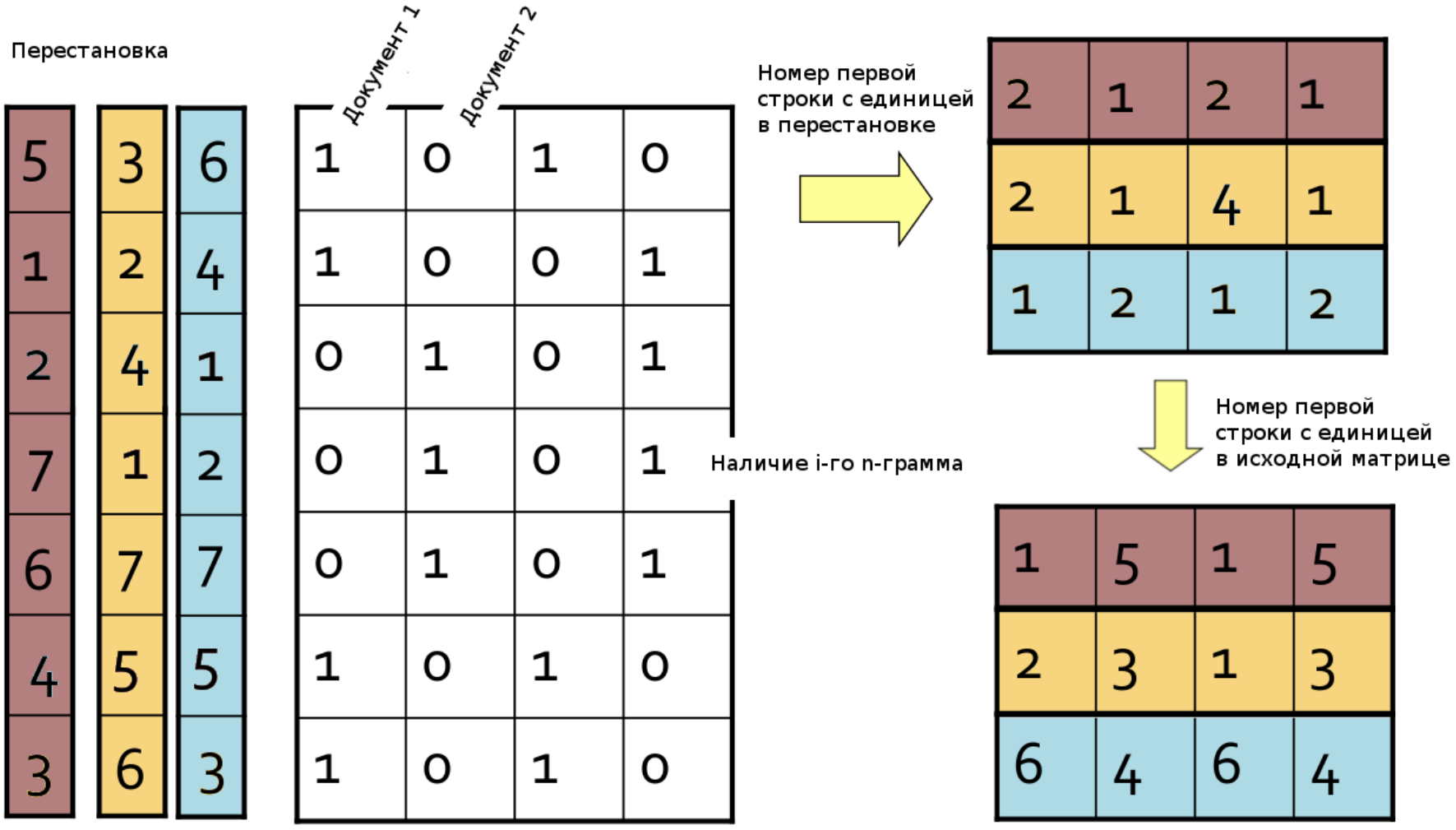
- Если документы C_1 и C_2 похожи, то с большой вероятностью $h(C1) == h(C2)$
- Иначе – с большой вероятностью $h(C1) \neq h(C2)$

Locality Sensitive Hash

Идея:

- Разбить документ на n-граммы
- Взять от каждого n-грамма хэш
- Получим представление документа в виде строки из нулей и единиц. Длина такого вектора = количество всевозможных n-грамм.
- Посчитаем документы похожими, если у них много совпадающих n-грамм

Locality Sensitive Hash



5	3	6
1	2	4
2	4	1
7	1	2
6	7	7
4	5	5
3	6	3

1	0	1	0
1	0	0	1
0	1	0	1
0	1	0	1
0	1	0	1
1	0	1	0
1	0	1	0

2	1	2	1
2	1	4	1
1	2	1	2

1	5	1	5
2	3	1	3
6	4	6	4

Задача кластеризации

Постановка задачи кластеризации

Кластеризация – задача разделения объектов одной природы на несколько групп так, чтобы объекты в одной группе обладали одним и тем же свойством.

Кластеризация – это обучение без учителя.

Постановка задачи кластеризации

X – пространство объектов
$\rho : X \times X \to [0, \infty)$ – функция расстояния между объектами

Найти:
Y – множество кластеров
$a : X \to Y$ – алгоритм кластеризации

Степени свободы в постановке задачи

Степени свободы в постановке задачи

- Критерий качества кластеризации
- Число кластеров неизвестно заранее
- Результат кластеризации существенно зависит от метрики

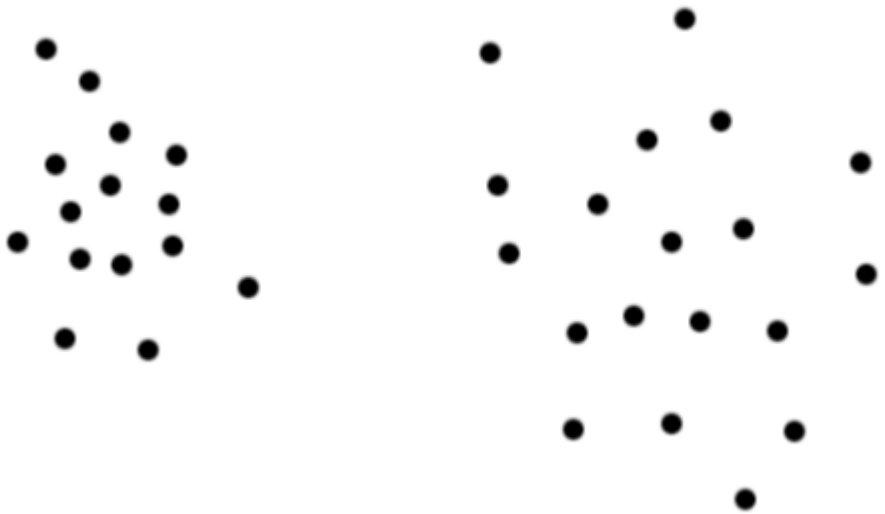
Цели кластеризации

Цели кластеризации

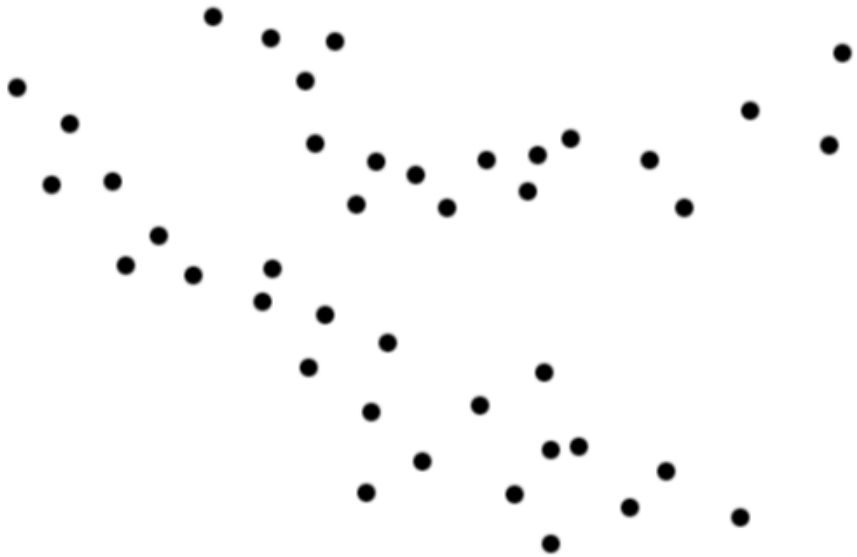
- Сократить объём хранимых данных
- Выделить нетипичные объекты
- Упростить дальнейшую обработку данных
- Построить иерархию множества объектов

Какие бывают кластеры?

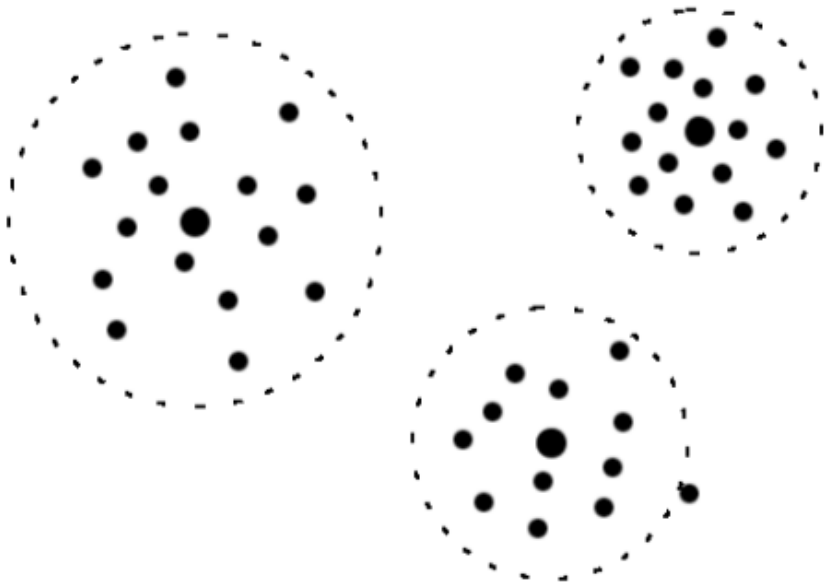
Типы кластерных структур. Сгущения



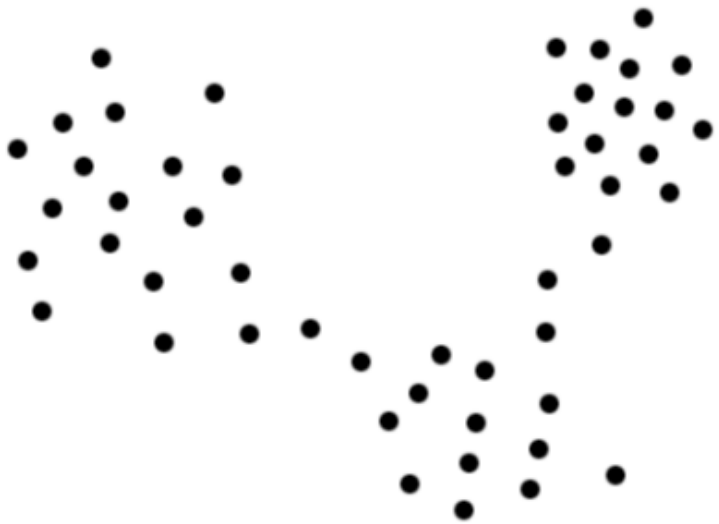
Типы кластерных структур. Ленты



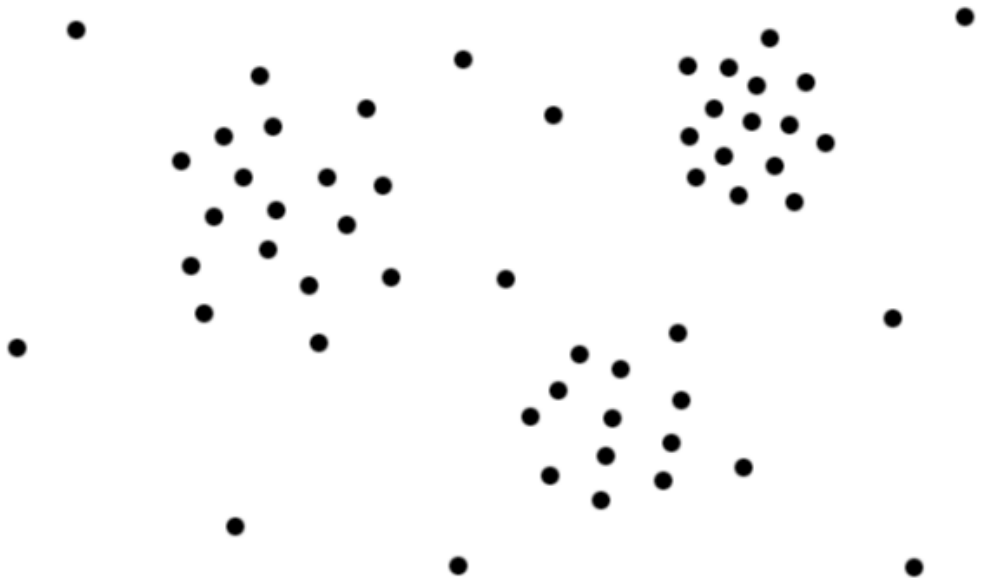
Типы кластерных структур. С центром



Типы кластерных структур. С перемычками

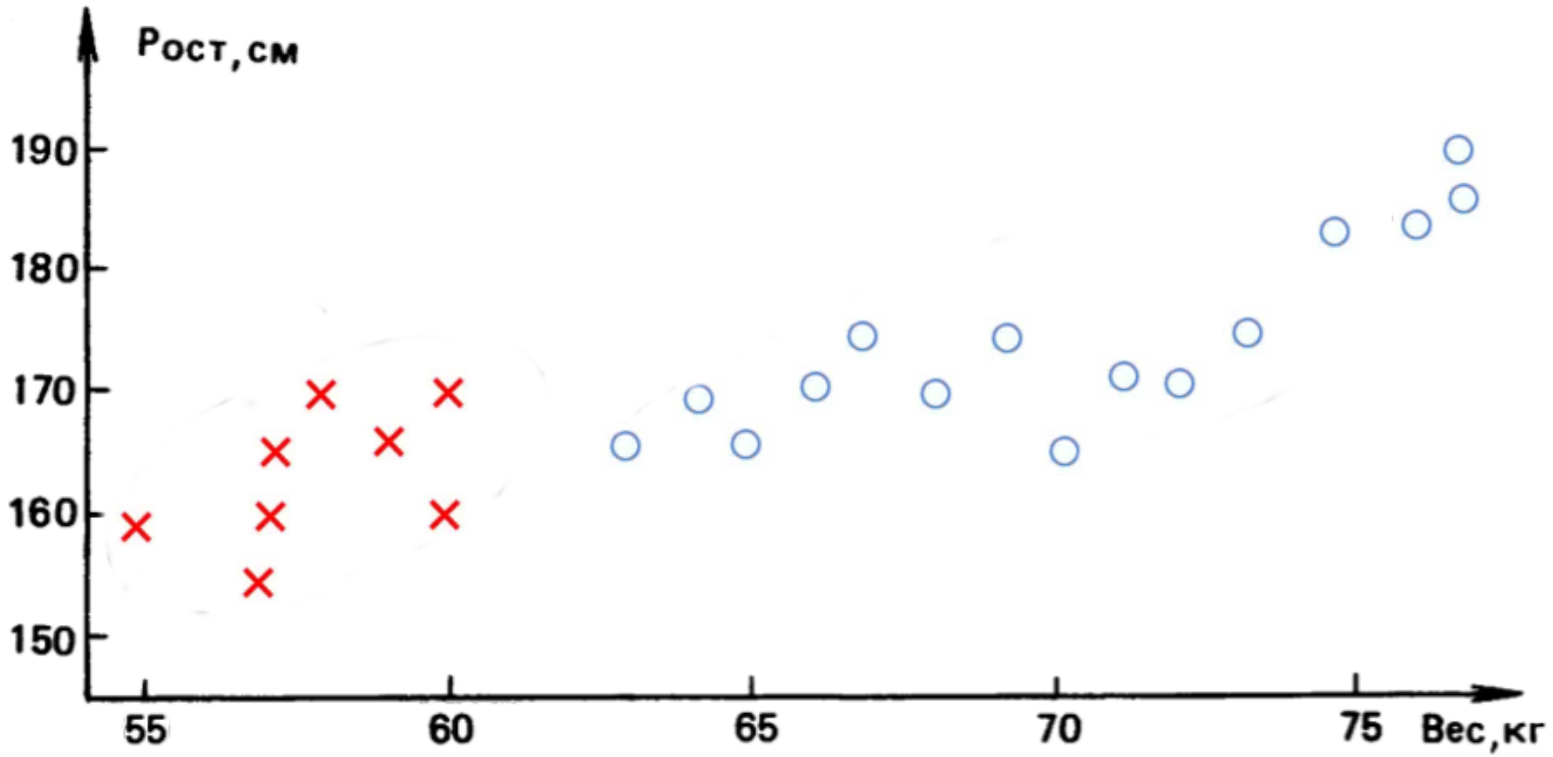


Типы кластерных структур. На фоне

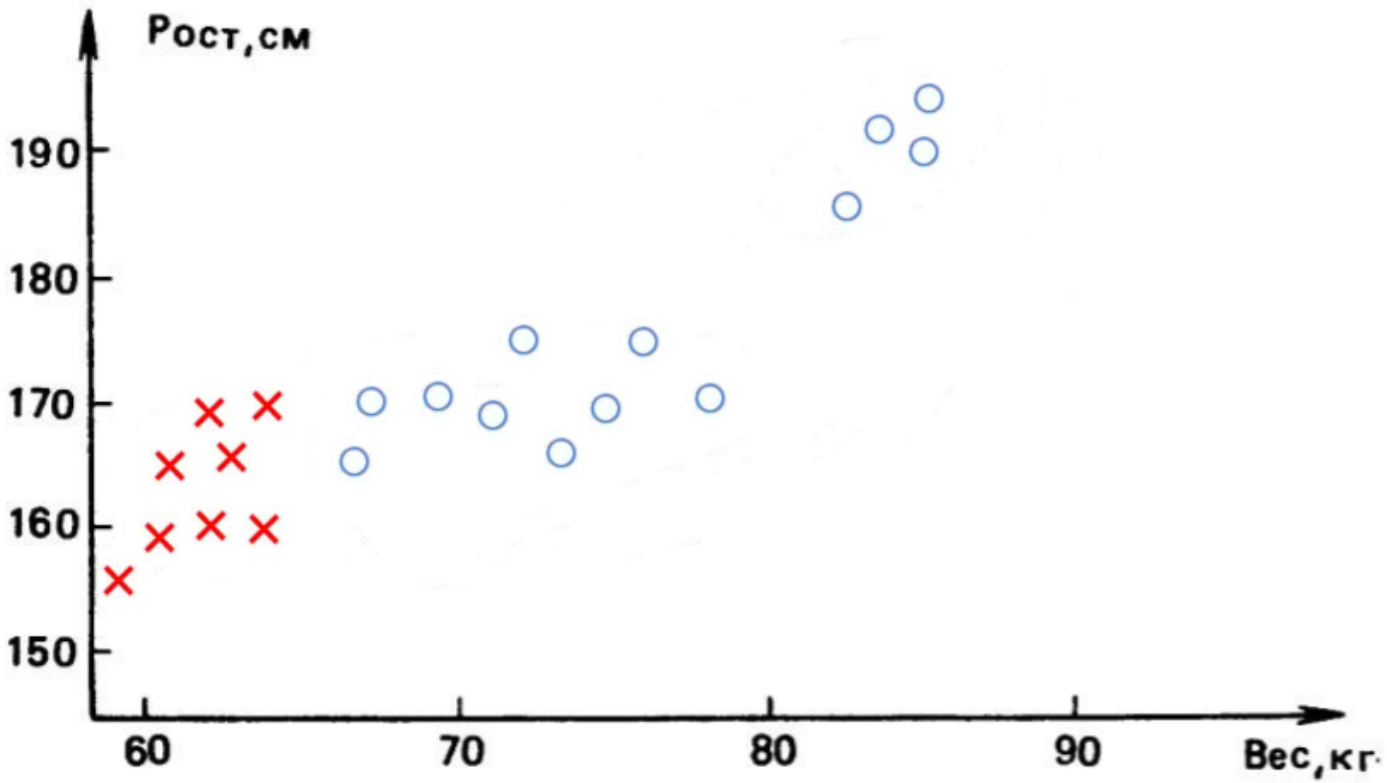


Типы кластерных структур. Перекрывающиеся

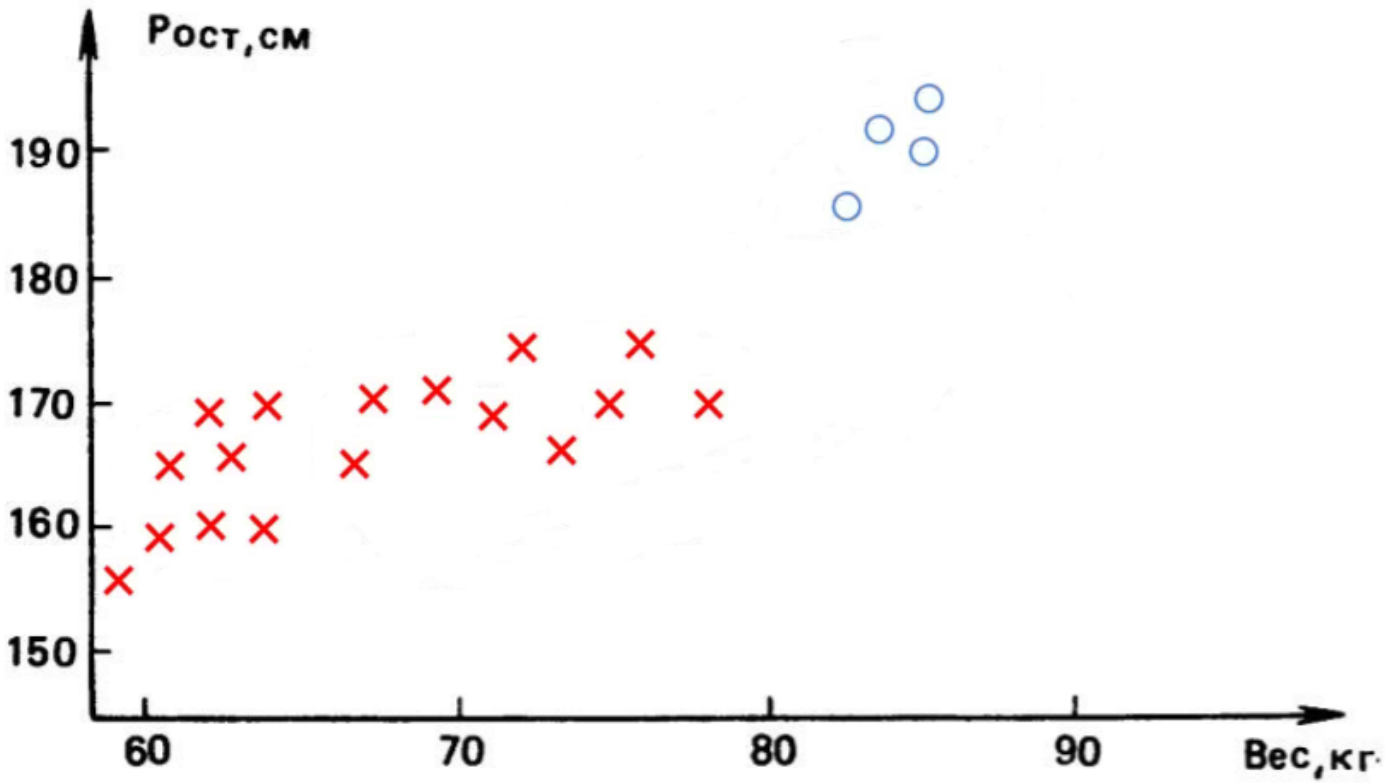
Чувствительность к выбору метрики



Чувствительность к выбору метрики



Чувствительность к выбору метрики



Оценка качества кластеризации

Есть несколько разбиений на кластеры. Как их сравнить?

Оценка качества кластеризации

- Минимизировать среднее внутрикластерное расстояние

$$\frac{\sum\limits_{a(x_i)=a(x_j)} \rho(x_i,x_j)}{\sum\limits_{a(x_i)=a(x_j)} 1} \to \min$$

- Максимизировать среднее межкластерное расстояние

$$\frac{\sum\limits_{a(x_i)\neq a(x_j)} \rho(x_i,x_j)}{\sum\limits_{a(x_i)\neq a(x_j)} 1} \to \max$$

Методы кластеризации

- Иерархические
- Графовые
- Статистические

Иерархическая кластеризация

Агломеративный алгоритм Ланса-Уильямса

Идея:

- Считаем каждую точку кластером.
- Затем объединяем ближайшие точки в новый кластер.
- Повторяем.

Алгоритм Ланса-Уильямса

$C_1 = \{\{x_1\}, \{x_2\}, \ldots, \{x_l\}\}$
for $t = 2, \ldots, l$:
    $(U, V) = \arg\min_{U \neq V} \rho(U, V)$
    $W = U \cup V$
    $C_t = C_{t-1} \cup \{W\} \setminus \{U, V\}$
    foreach $S \in C_t$
        вычислить $\rho(W, S)$

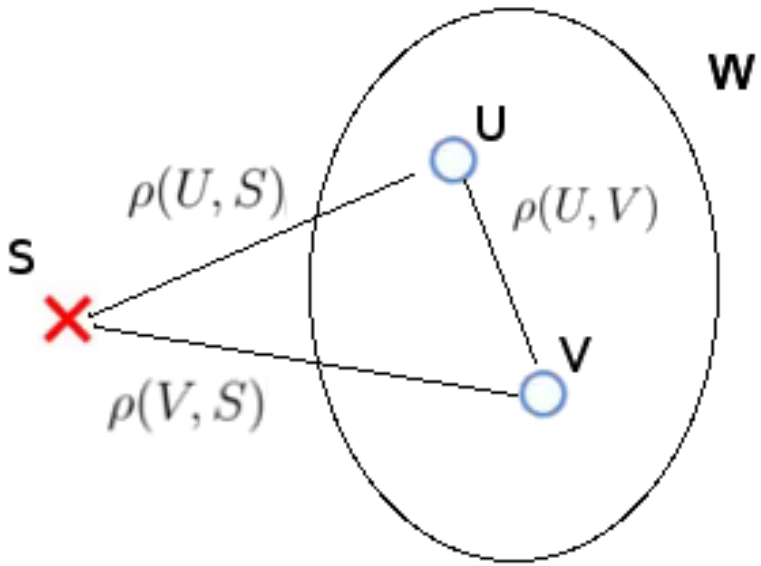
Алгоритм Ланса-Уильямса

Чего не хватает?

Формула Ланса-Уильямса

Расстояние $\rho(W, S)$?
$W = \{U \cup V\}$

Знаем:
$\rho(U, S), \rho(V, S), \rho(U, V)$



Формула Ланса-Уильямса

Расстояние $\rho(W, S)$?
$W = \{U \cup V\}$

Знаем:
$\rho(U, S), \rho(V, S), \rho(U, V)$

$$\rho(U \cup V, S) = \alpha_U \rho(U, S) + \alpha_V \rho(V, S) + \\ + \beta \rho(U, V) + \gamma |\rho(U, S) - \rho(V, S)|$$

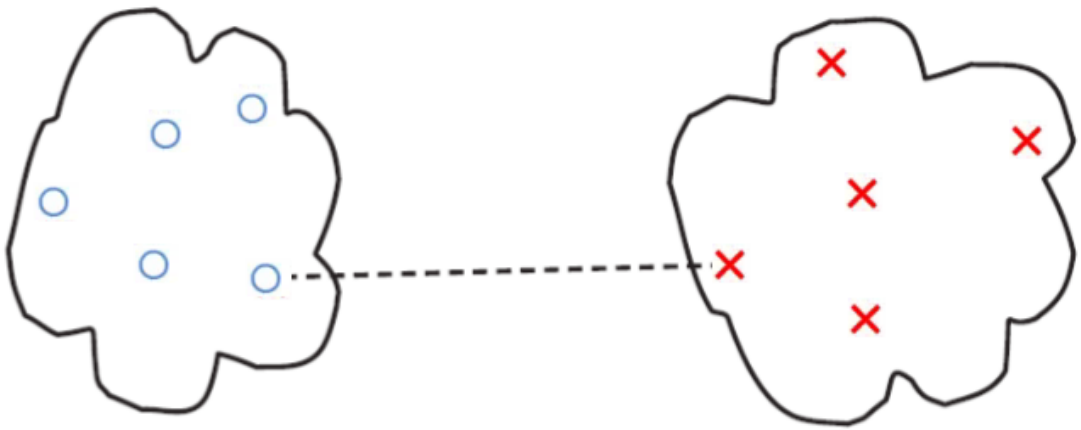
$\alpha_U, \alpha_V, \beta, \gamma$ – числовые параметры

Формула Ланса-Уильямса

Значения параметров $\alpha_U, \alpha_V, \beta, \gamma$?

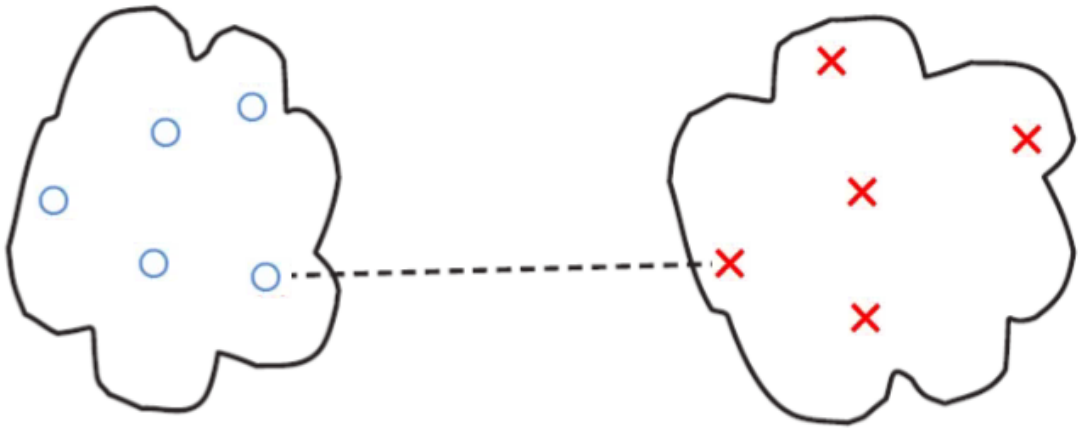
Формула Ланса-Уильямса

Расстояние ближнего соседа:



Формула Ланса-Уильямса

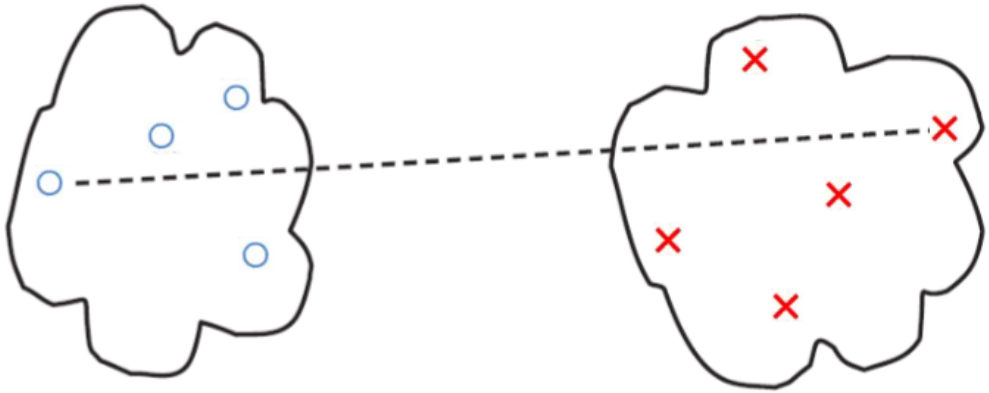
Расстояние ближнего соседа:



$\alpha_U = \alpha_V = \frac{1}{2}$
$\beta = 0$
$\gamma = -\frac{1}{2}$

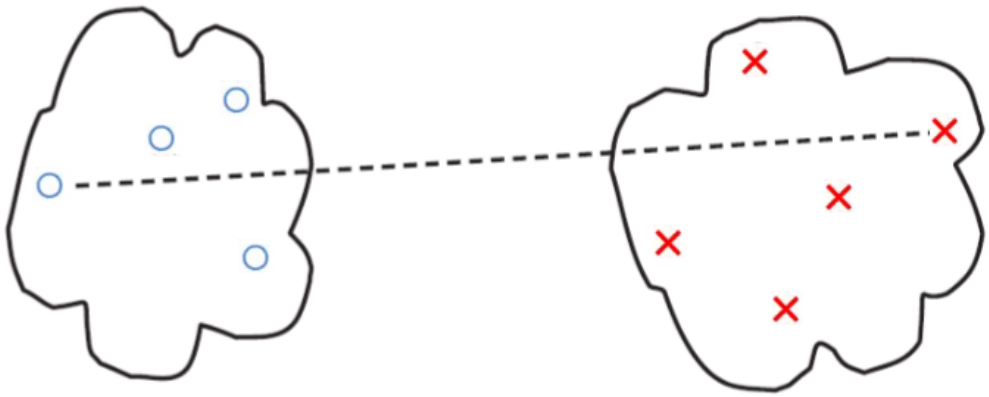
Формула Ланса-Уильямса

Расстояние дальнего соседа:



Формула Ланса-Уильямса

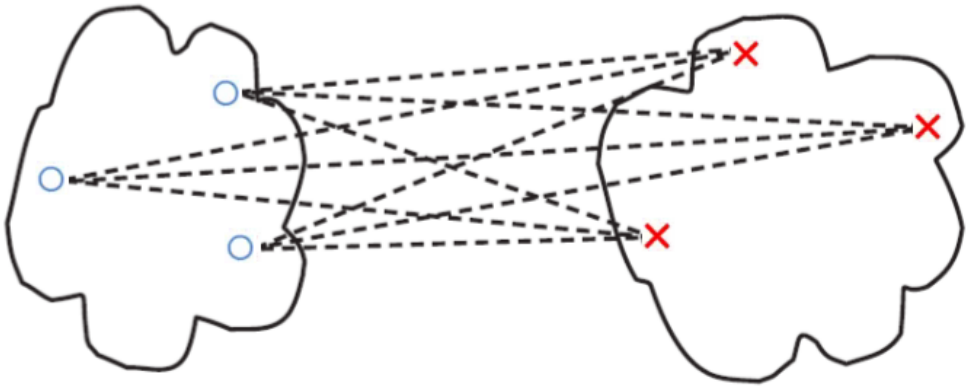
Расстояние дальнего соседа:



$\alpha_U = \alpha_V = \frac{1}{2}$
$\beta = 0$
$\gamma = \frac{1}{2}$

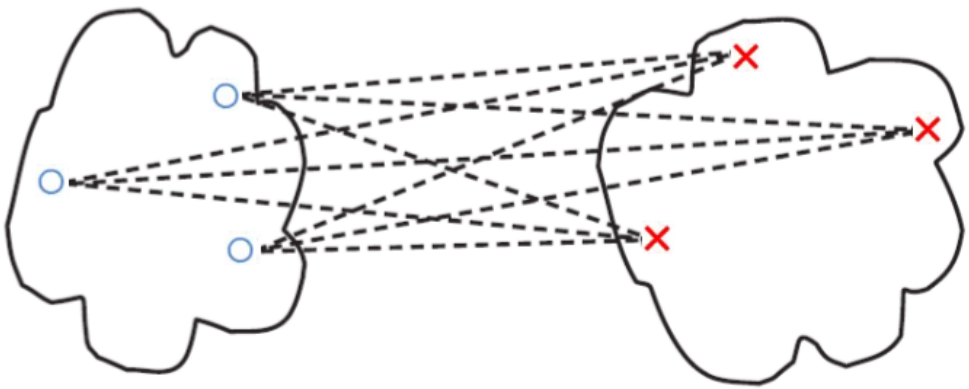
Формула Ланса-Уильямса

Групповое среднее:



Формула Ланса-Уильямса

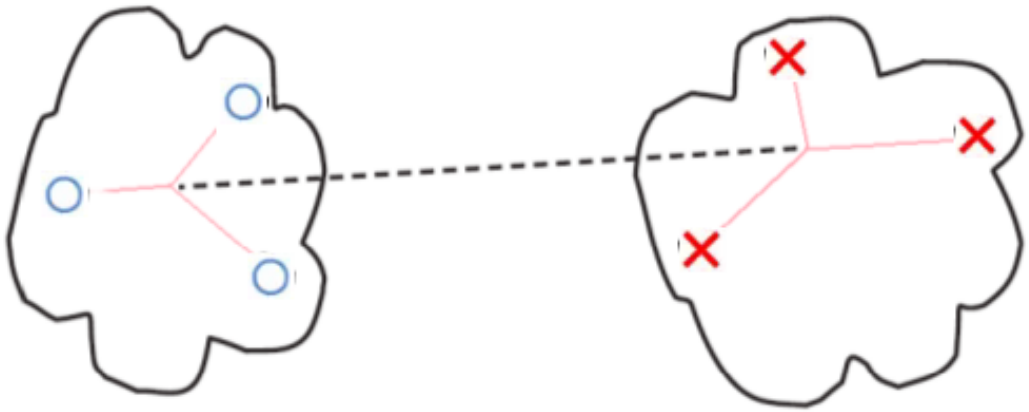
Групповое среднее:



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

Формула Ланса-Уильямса

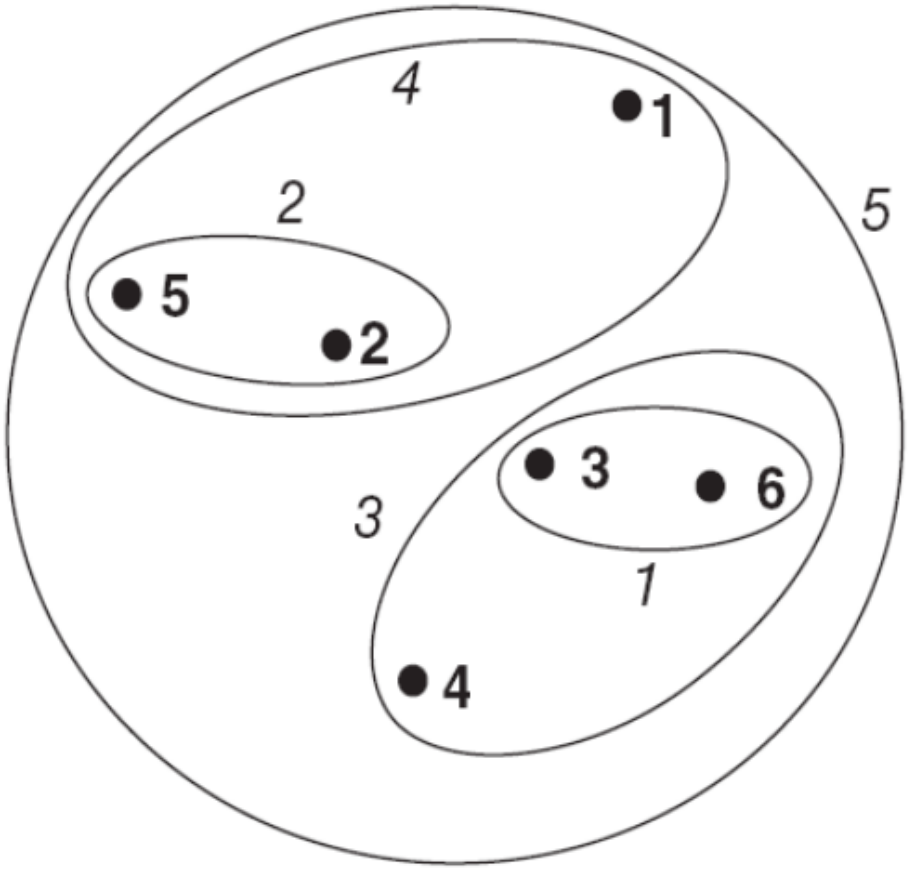
Расстояние Уорда:



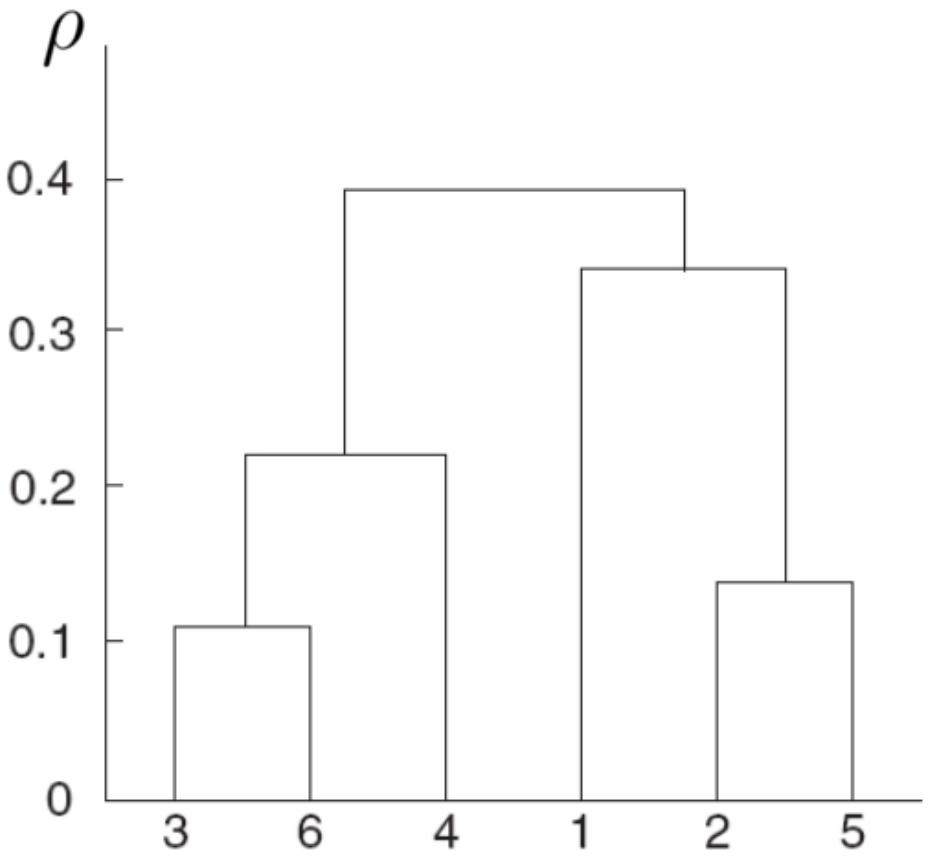
$$\alpha_U = \frac{|S|+|U|}{|S|+|W|}$$
$$\alpha_V = \frac{|S|+|V|}{|S|+|W|}$$
$$\beta = \frac{-|S|}{|S|+|W|}$$
$$\gamma = 0$$

Визуализация кластеров

Диаграмма вложения



Дендрограмма



Дендрограмма

Может ли так случиться, что дендрограмма имеет самопересечения?

Дендрограмма

Может ли так случиться, что дендрограмма имеет самопересечения?

Как избежать?

Свойство монотонности

Кластеризация монотонна, если на каждом шаге расстояние ρ между объединяемыми кластерами не уменьшается.

$\rho_2 \leq \rho_3 \leq \cdots \leq \rho_l$

На следующей лекции

- Задача кластеризации
- Графовые алгоритмы кластеризации
- Статистические алгоритмы кластеризации