Национальный исследовательский
Томский государственный университет

Кластеризация данных

Сергей Аксёнов, к.т.н., доцент кафедры
Теоретических основ информатики ИПМКН ТГУ

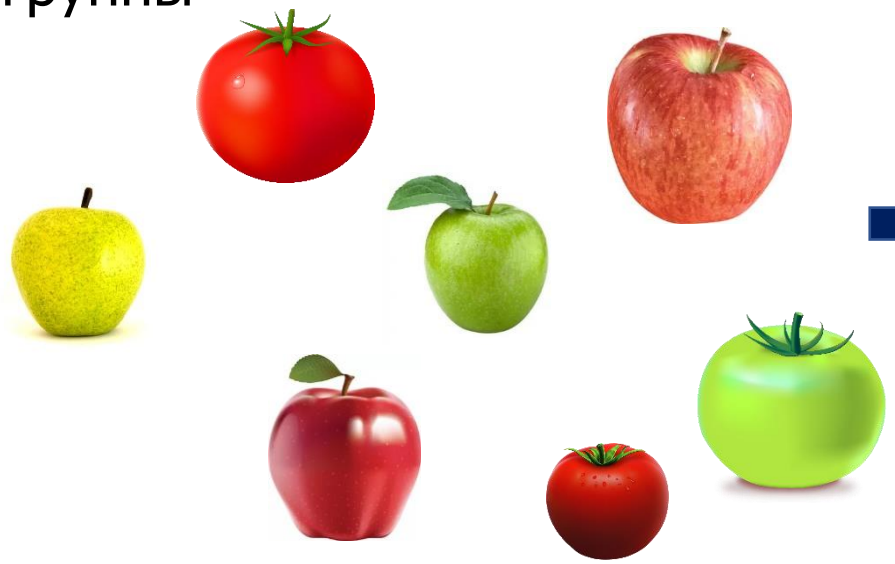
Томск-2023

Обучение без учителя

Меток класса нет. Метод используется для изучения данных.

Особенность: Субъективность кластеризации.

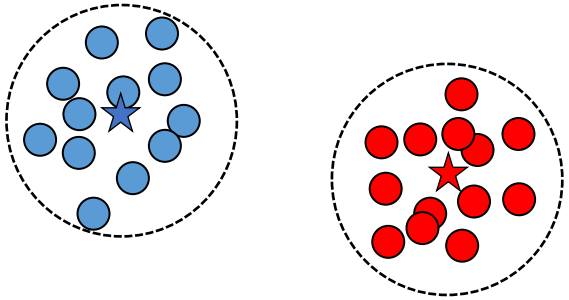
Задача: Разложить объекты на две группы



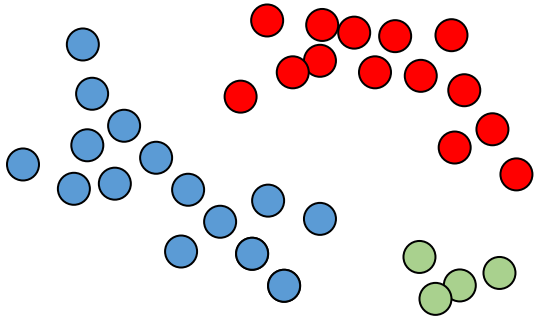
Цели кластеризации

1. Подготовка групп данных для последующей обработки
2. Нахождение аномалий
3. Снижение объема обрабатываемых данных
4. Построение иерархии объектов

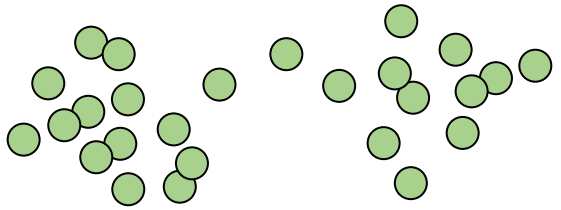
Примеры кластеров



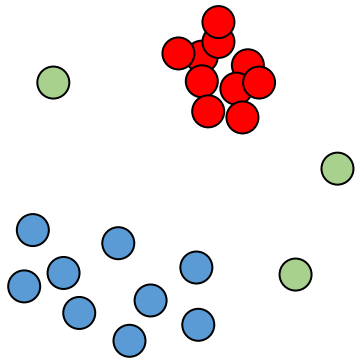
Кластеры с центроидами



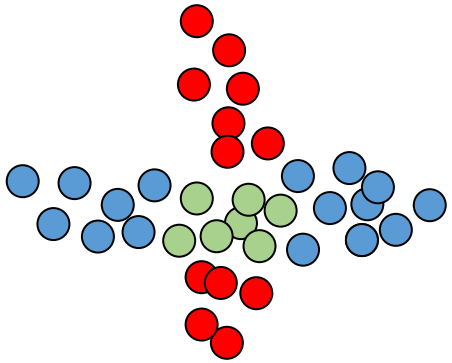
Кластеры произвольной формы



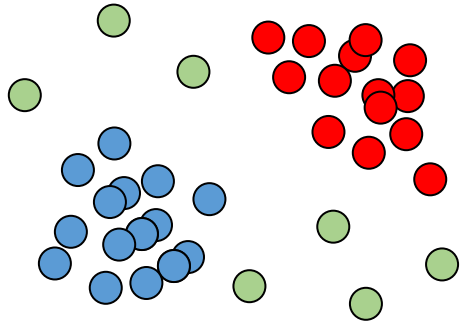
Кластеры с перемычками



Кластеры с разной плотностью объектов



Кластеры перекрываются



Кластеры с шумами

Проблемы кластеризации

1. Не существует однозначного критерия качества кластеризации.
    - Для хорошего решения нужно проводить оценку качества с помощью нескольких критериев

2. Число кластеров, как правило, заранее неизвестно и выбирается по субъективным критериям.
    - Число кластеров подбирается путём изменения параметров алгоритма(ов)

3. Результат кластеризации существенно зависит от метрики.

Обозначения

X - Кластеризуемое множество

N - Количество элементов в X

c

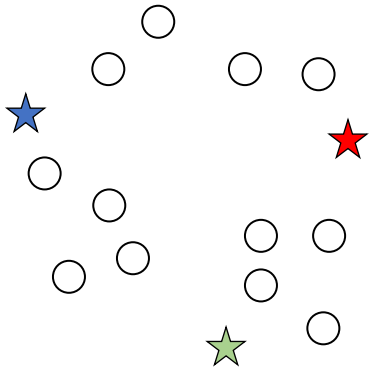
n_{c_j} - Число элементов в кластере c_j

v_j - Центр кластера c_j: $v_j = \sum_{x_i \in c_j} x_i / n_{c_j}$

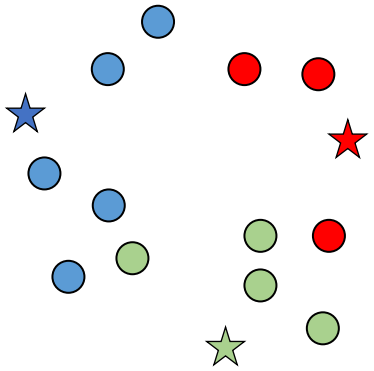
dim - Размерность множества X

u_{ij} - Степень принадлежности x_i кластеру c_j

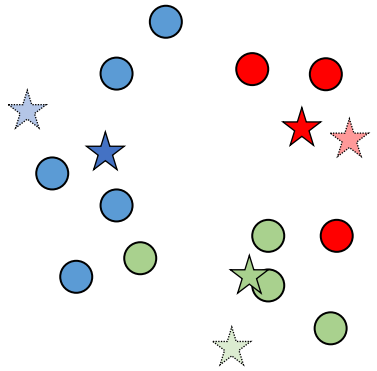
K–Средних: Пример



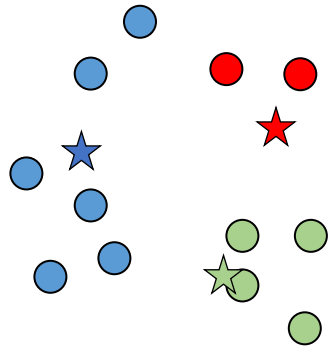
Инициализация



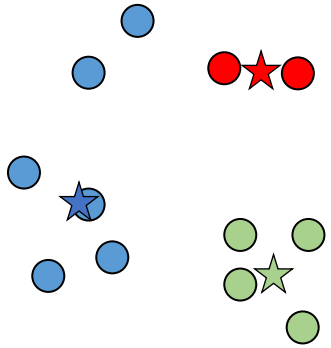
Связывание наблюдений



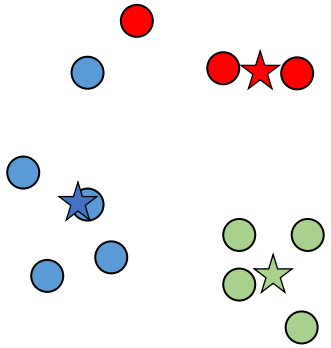
Сдвиг центроидов



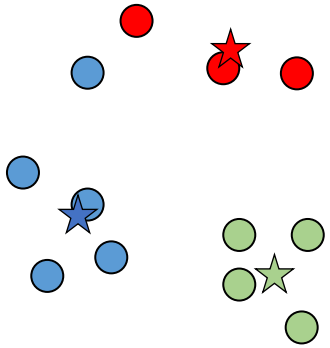
Связывание наблюдений



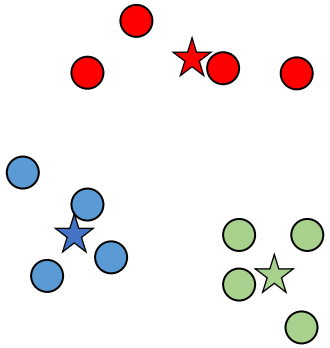
Сдвиг центроидов



Связывание наблюдений

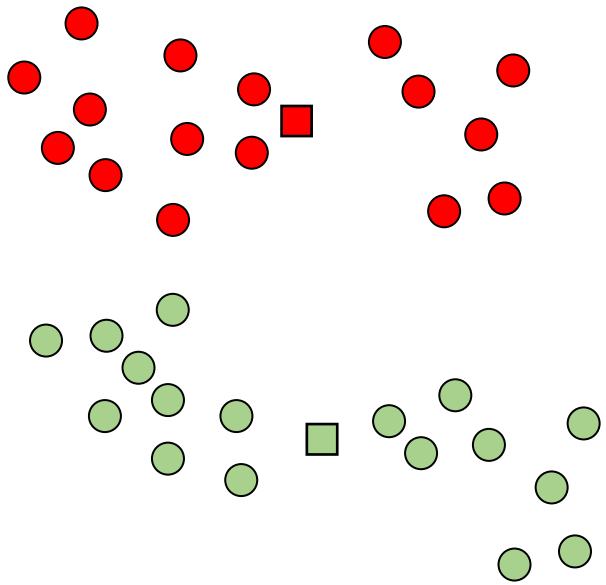


Сдвиг центроидов

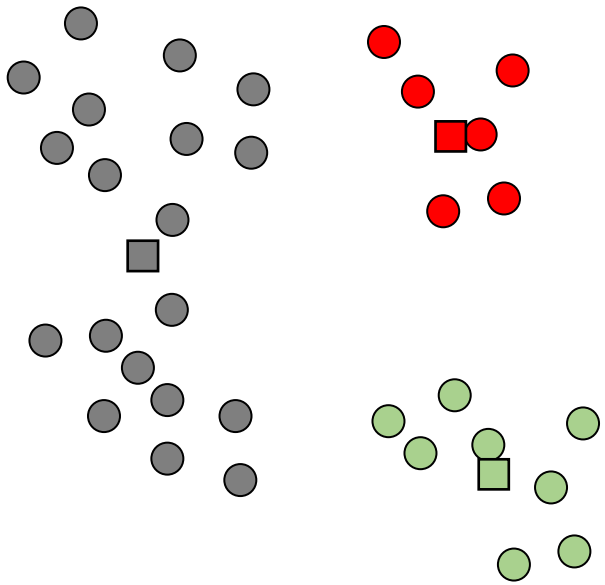


Сдвиг центроидов

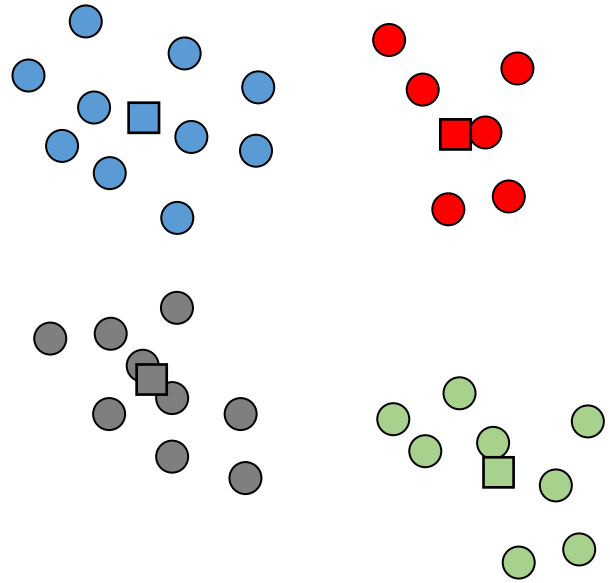
Число кластеров



2 Кластера



3 Кластера



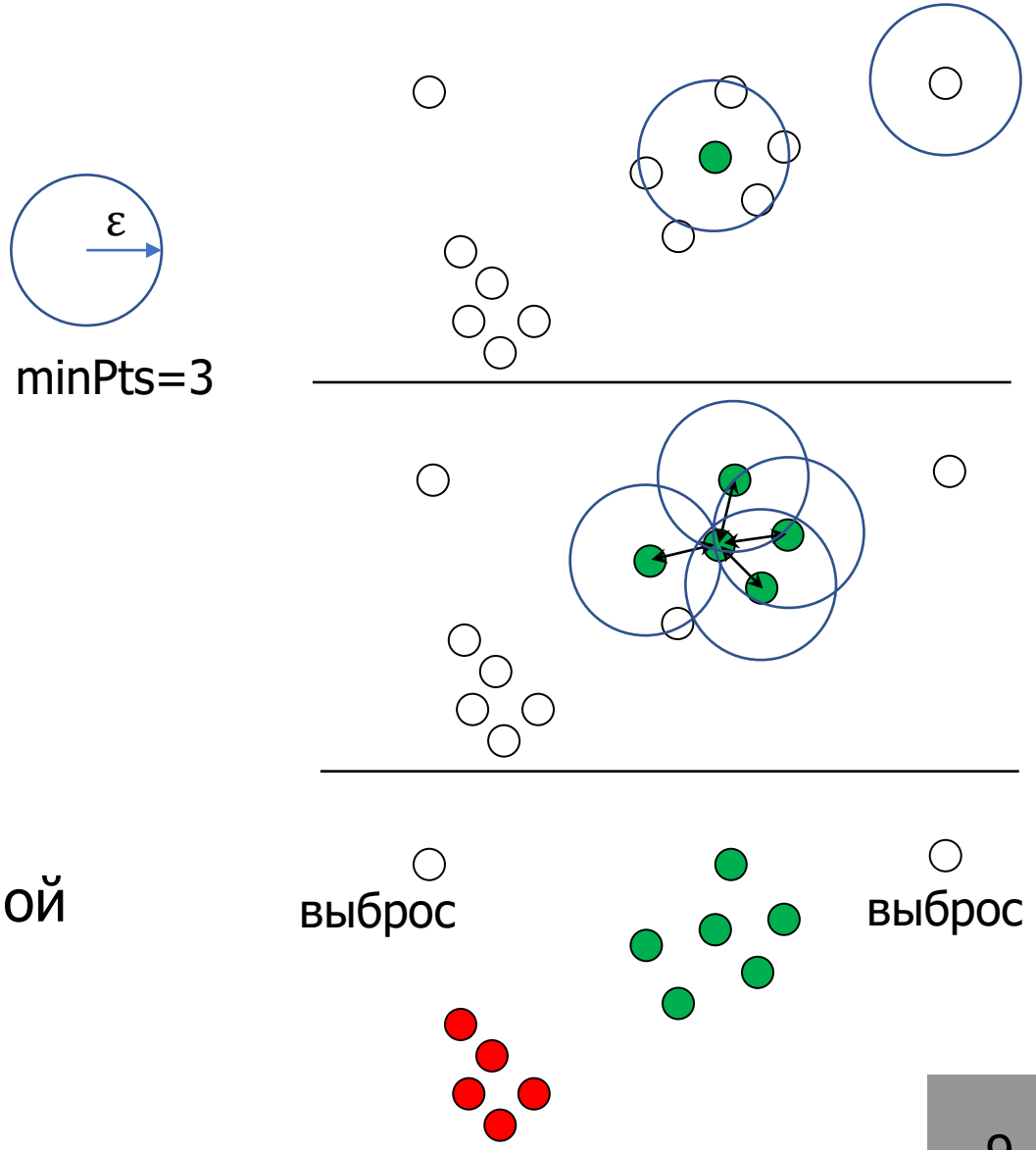
4 Кластера

Кластеризация, основанная на плотности DBSCAN

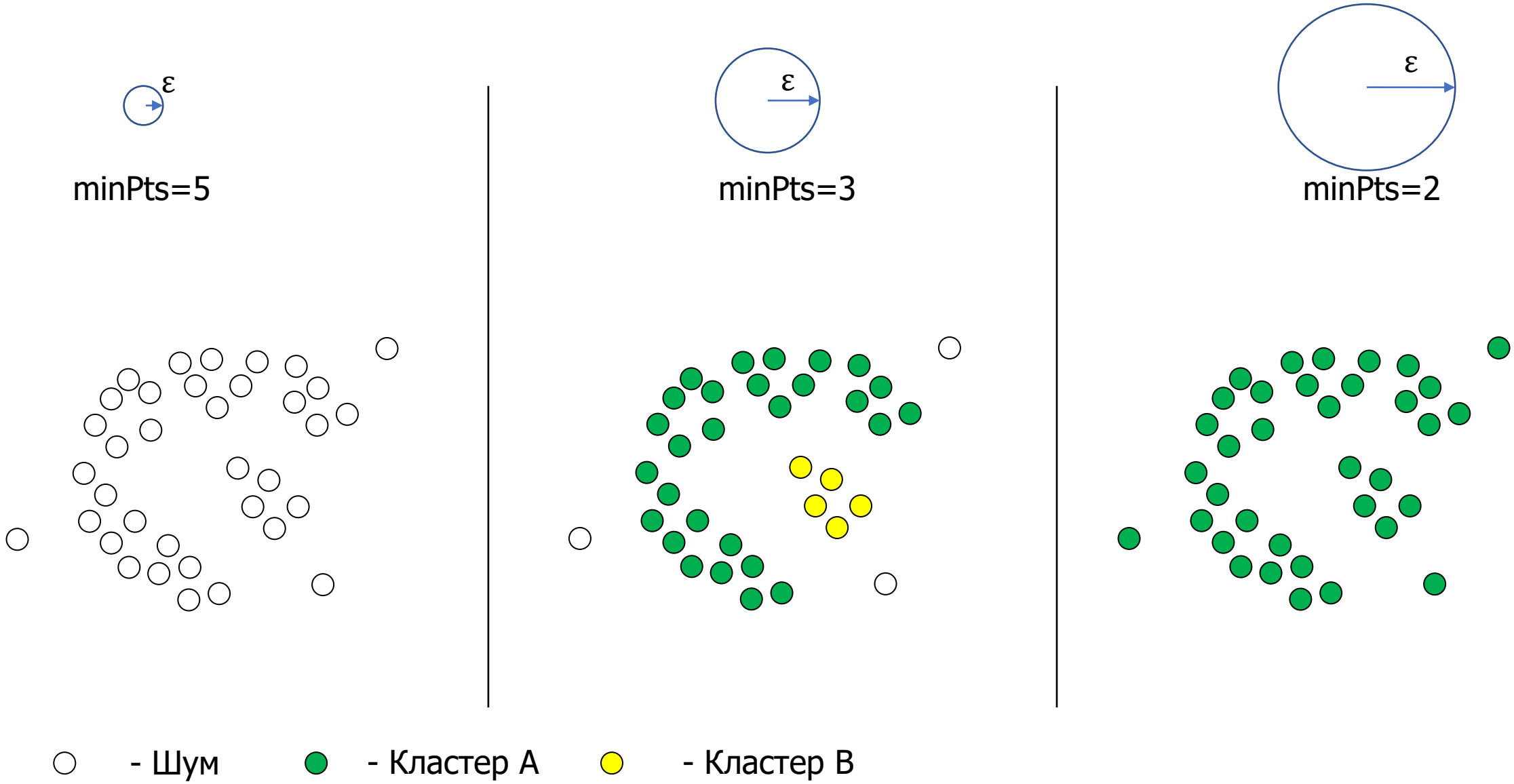
Параметры: minPts – число объектов, для формирования точки кластера, ε – радиус соседства

Преимущества:
- Кластеры произвольной формы
- Не требуется задавать число кластеров
- Имеется понятие шума

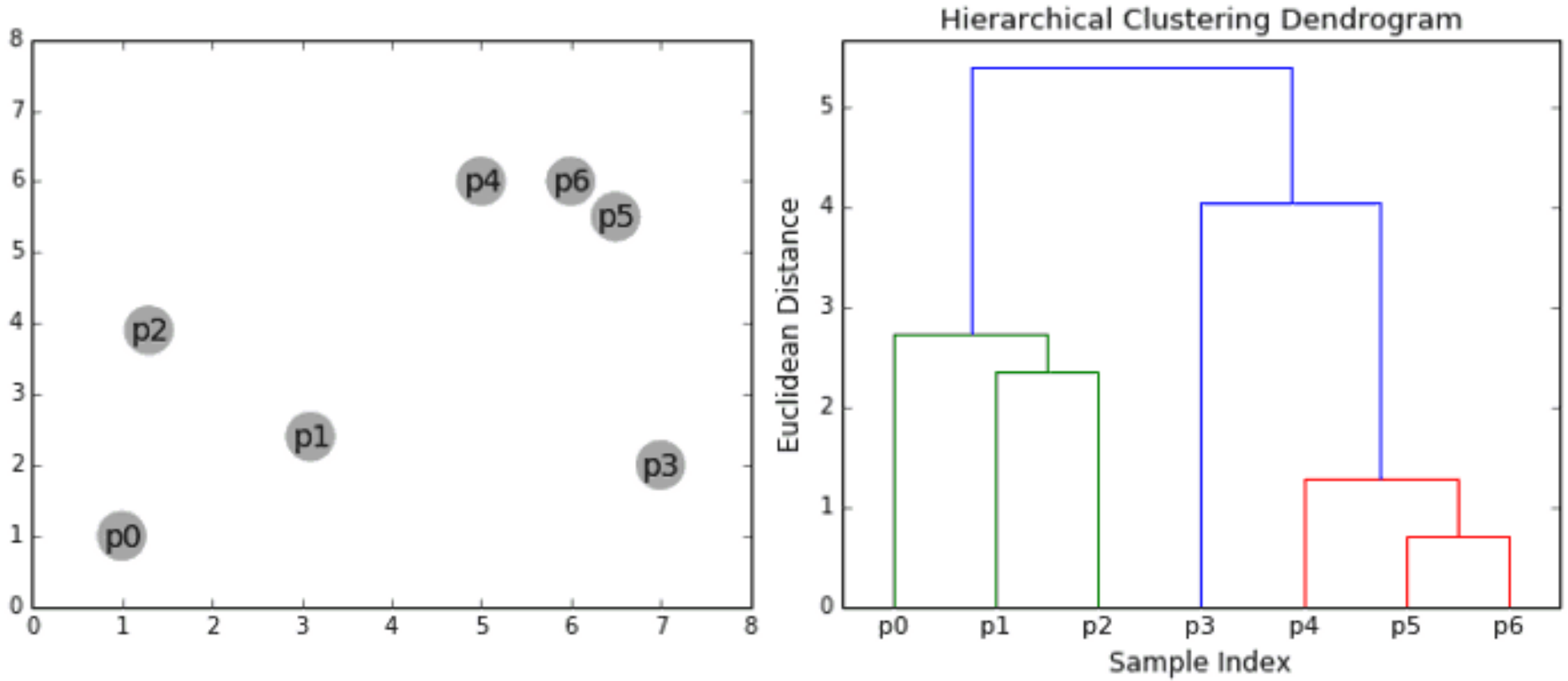
Недостатки:
- Проблема краевых точек
- Трудность подбора параметров
- Плохо кластеризует наборы данных с большой разницей в плотности



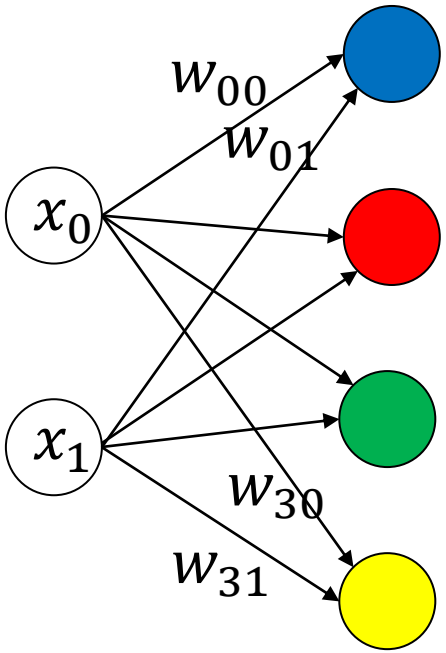
Пример кластеризации DBSCAN



Иерархическая кластеризация



Сеть Кохонена (Kohonen network)

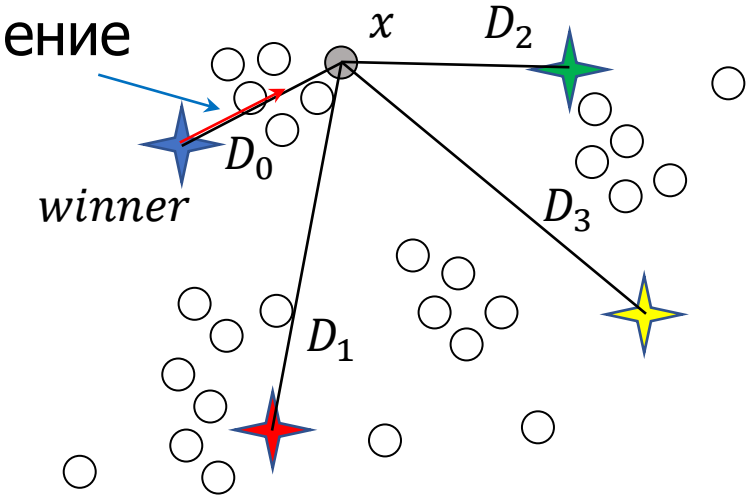


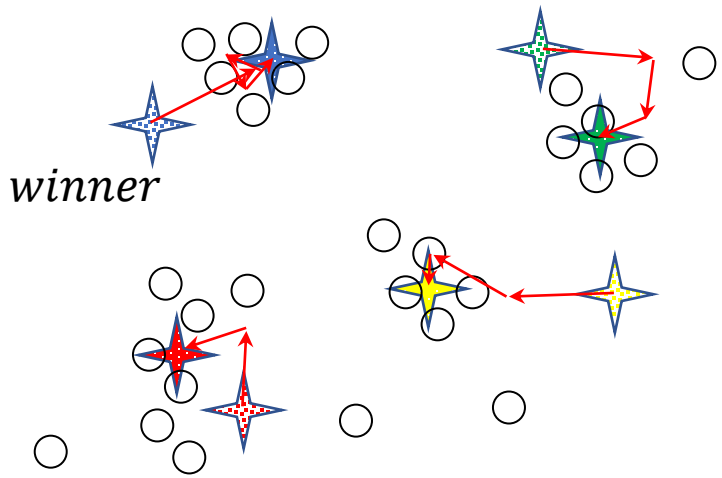
$$D_j = \sum_{i=1}^{dim} \left(x_i - w_{ij}\right)^2$$

$winner = k$, где $D_k = \min_j(D_j)$

$0 < \eta < 1$ - Скорость настройки

Обучение





Алгоритм обучения

Шаг 1. Установка случайных весов w_{ij}

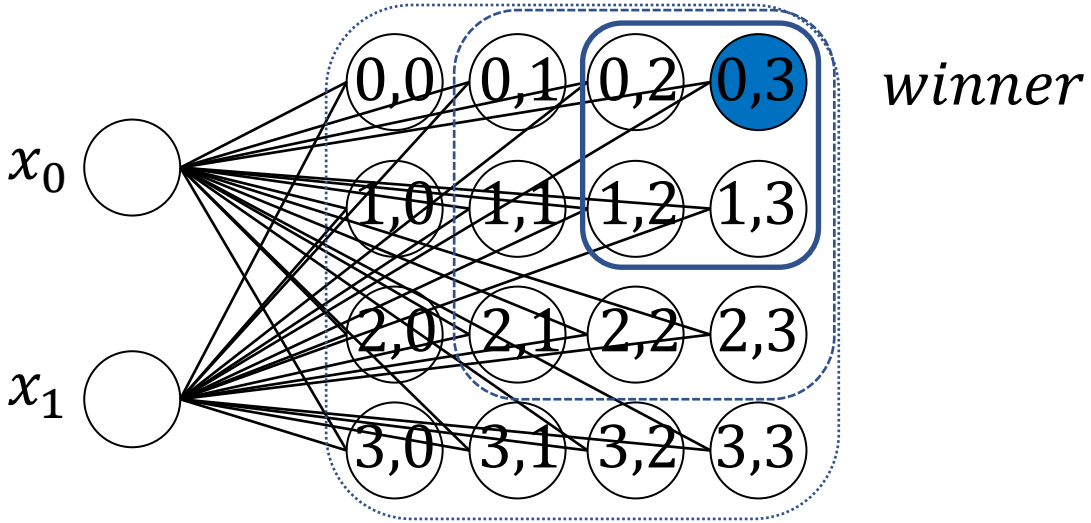
Шаг 2. Активация сети примером из выборки (расчет D_j)

Шаг 3. Корректировка весов победителя по правилу:

$$w_{ki}(t+1) = w_{ki}(t) + \eta \cdot [x_i - w_{ki}(t)]$$

Шаг 4. Оценка критерия останова. В противном случае на шаг 2.

Самоорганизующаяся карта признаков (SOM)



Корректировка всех весов сети на этапе настройки.

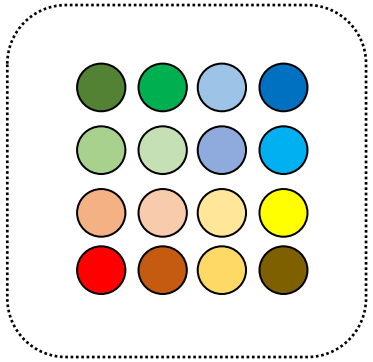
$$w_{ji}(t+1) = w_{ji}(t) + \theta(j, winner, t) \cdot \eta \cdot \left[x_i - w_{ji}(t)\right]$$

$\theta(j, winner, t)$ - Функция расстояния, при $j = winner$: $\theta(j, j, t) = 1$

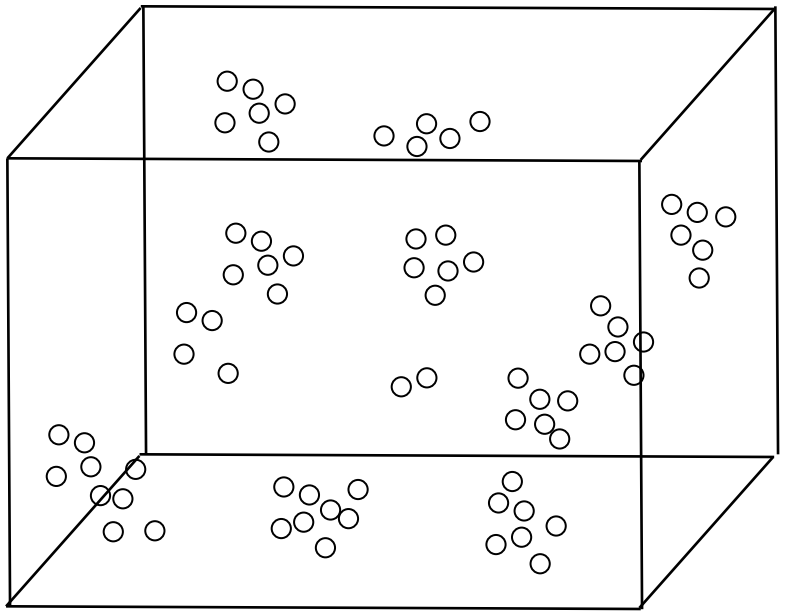
при $j \neq winner$: $0 < \theta(j, j, t) < 1$

Чем дальше настраиваемый узел от победителя, тем ближе $\theta(j, winner, t)$ к 0.

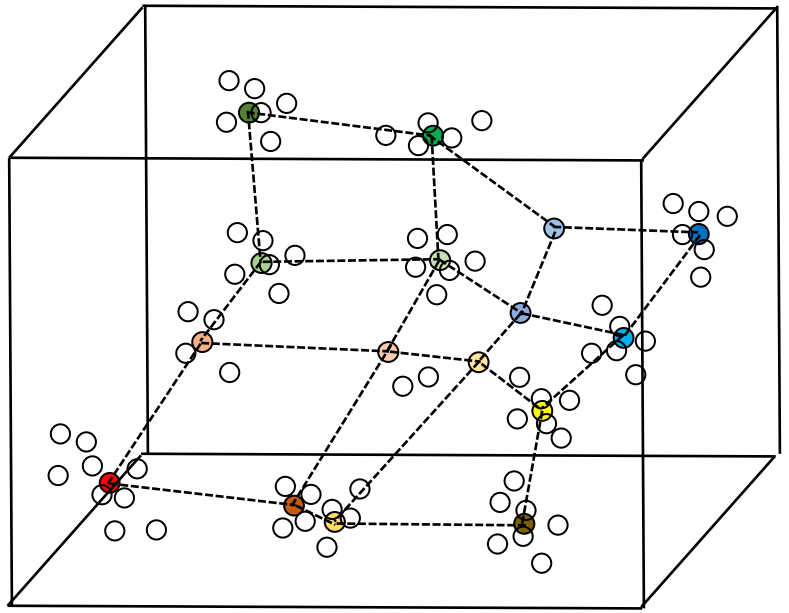
Пример SOM



$4 \times 4\ SOM$



$3D \rightarrow 2D$



Пространство признаков

«Мертвые» нейроны:

Относительные методы оценки качества четкой кластеризации

Оценка решения производится путём сравнения нескольких структур:

- Повторный запуск одного и того же алгоритма
- Запуск одного и того же алгоритма с разными значениями параметров
- Запуск разных алгоритмов

Критерии оценки качества:

Компактность – элементы, находящиеся в одном кластере, должны быть как можно ближе друг к другу

Отделимость – элементы, располагающиеся в разных кластерах, должны быть как можно дальше друг от друга

Индекс Данна (Dunn Index)

$$Dunn\_index = min_{i,j\in\{1,\dots,c\},i\neq j}\left\{\frac{d(c_i,c_j)}{max_{k\in\{1,\dots,c\}}diam(c_k)}\right\}$$

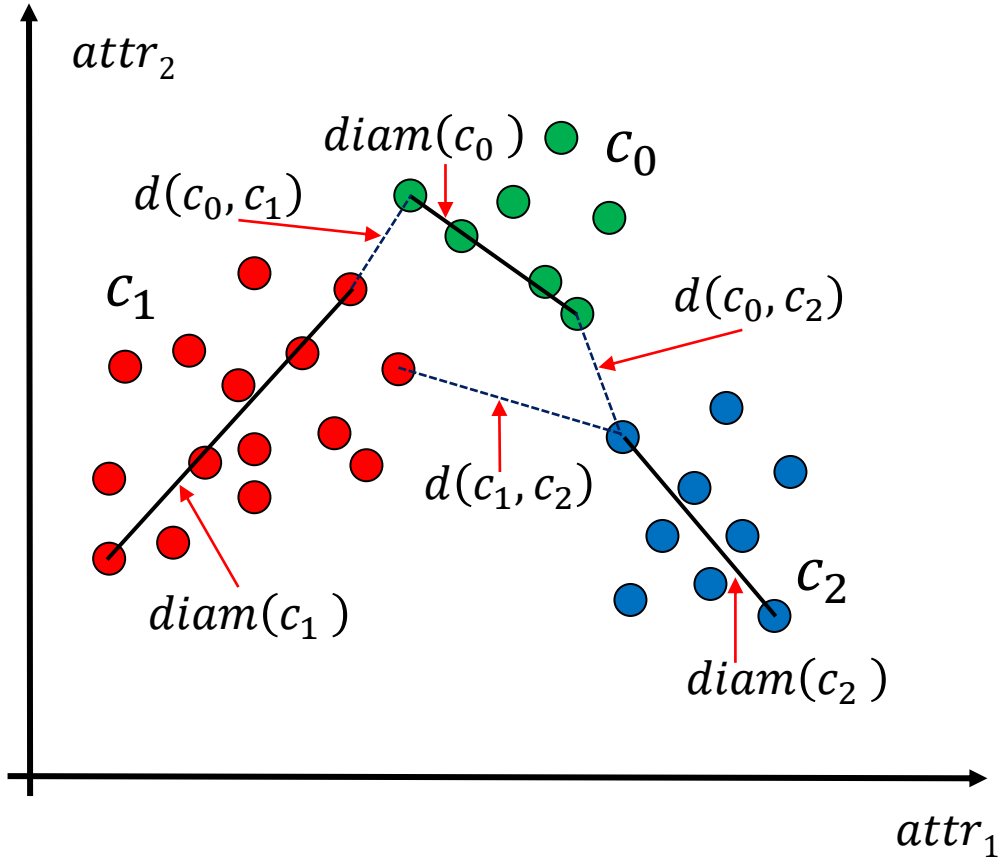
Диаметр кластера:

$$diam(c_i) = max_{x,y\in c_i}\|x-y\|$$

Расстояние между кластерами:

$$d(c_i,c_j) = min_{x\in c_i,y\in c_j}\|x-y\|$$

$$Dunn\_index \rightarrow Max$$



Индекс Дэвиса-Болдуина (Davis-Baldwin Index)

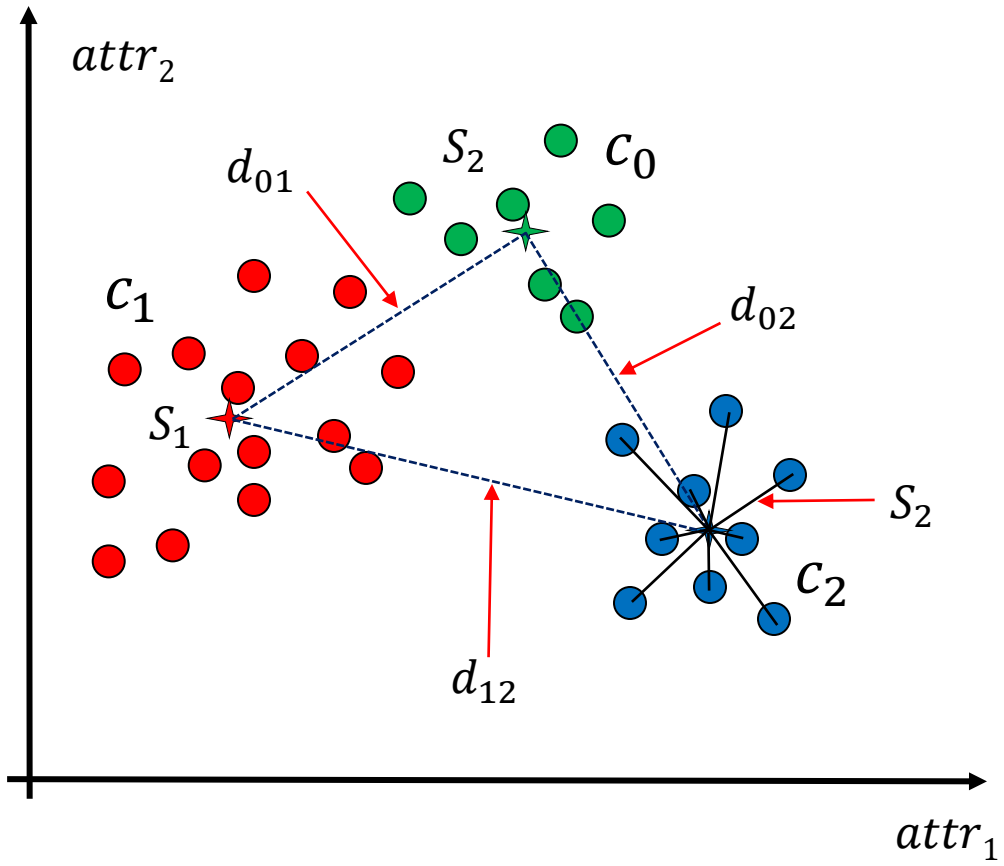
$$S_i = \left\{ \frac{1}{n_{c_i}} \sum_{x \in c_i} \|x - v_i\|^q \right\}^{\frac{1}{q}}$$

$$d_{ij} = \left\{ \sum_{k=1}^{dim} \left| v_i^k - v_j^k \right|^p \right\}^{\frac{1}{p}}$$

$$R_{ij} = \frac{S_i + S_j}{d_{ij}}$$

$$R_i = max_{i,j \in \{1,\ldots,c\}}(R_{ij})$$

$$DB\_Index = \frac{1}{c} \sum_{i=1}^{c} R_i$$



$$DB\_Index \rightarrow Min$$

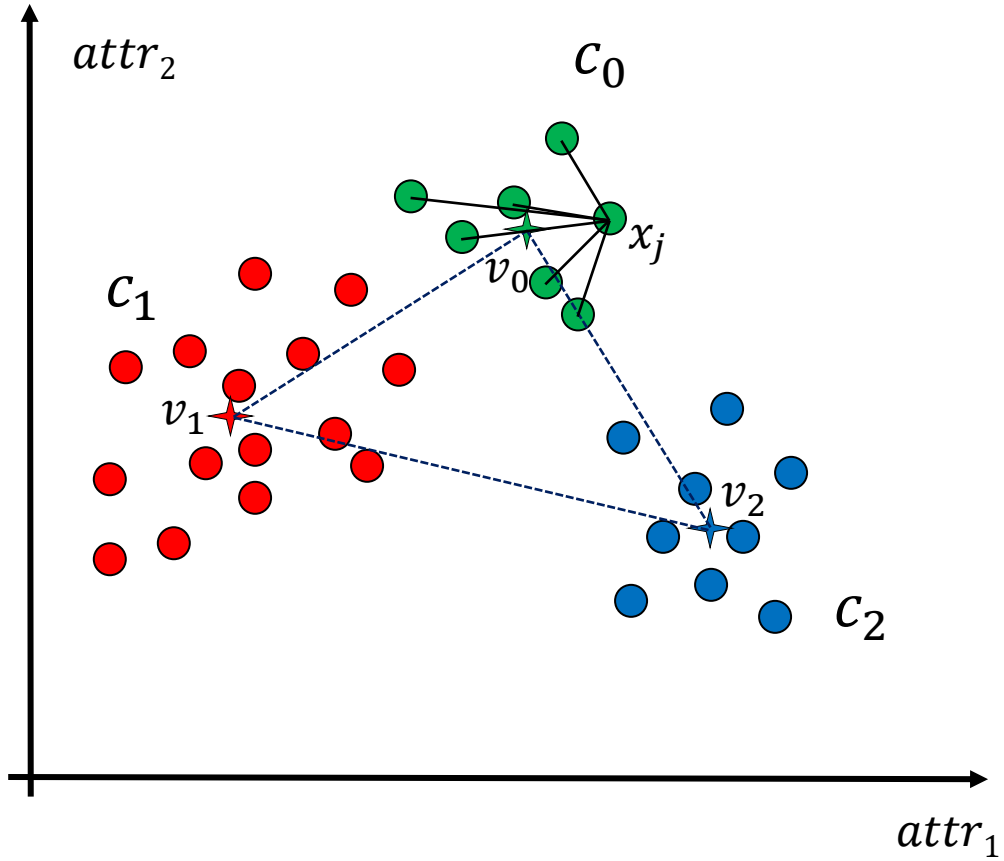
CS Индекс (CS Index)

$$CS\_Index = \frac{\sum_{i=1}^{c}\left\{\sum_{x_j, x_k \in c_i} \max(\|x_j - x_k\|)\right\}}{\sum_{i=1}^{c} min_{i \neq j}(\|v_i - v_j\|)}$$

Компактность – сумма расстояний между всеми объектами в группе.

Для отделимости используется сумма наименьших расстояний между центрами кластеров.

$$CS\_Index \rightarrow Min$$



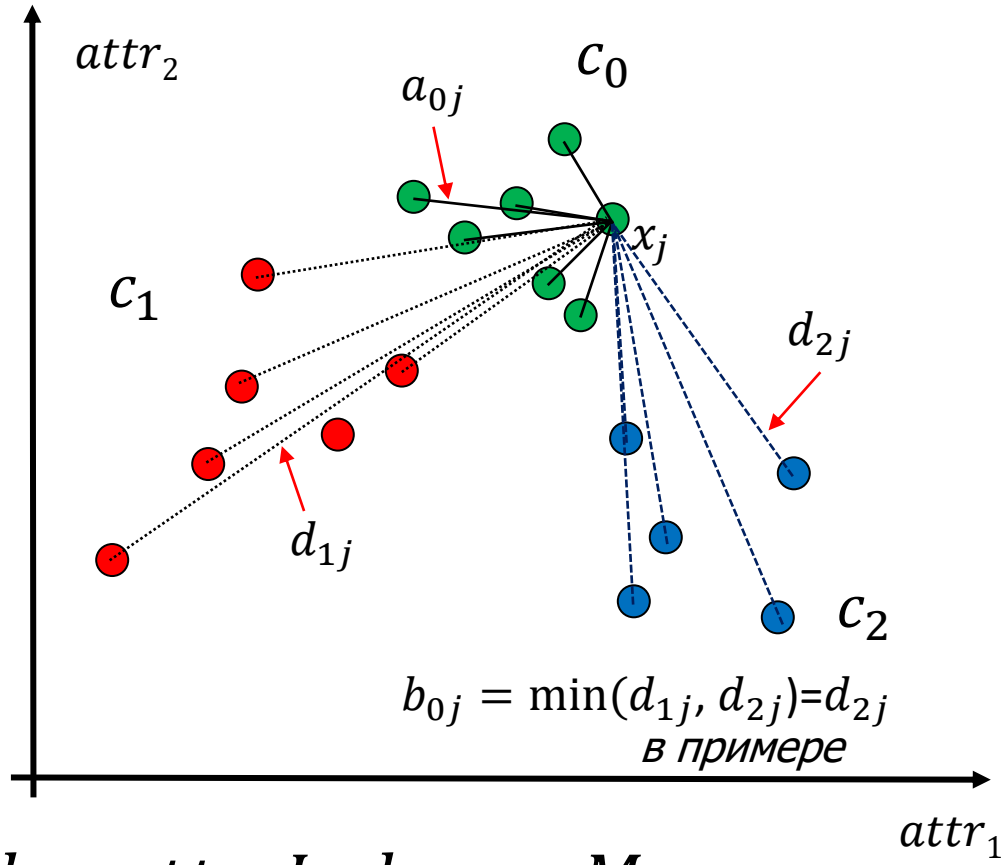
Индекс оценки Силуэта (Silhouette Index)

$$a_{pj} = \frac{1}{n_{c_p} - 1} \sum_{x_k \in c_p} \|x_j - x_k\|$$

$$d_{qj} = \frac{1}{n_{c_q}} \sum_{x_k \in c_q} \|x_j - x_k\|$$

$$b_{pj} = min_{q \neq p} d_{qj}$$

$$S_{x_j} = \frac{b_{pj} - a_{pj}}{max(a_{pj}, b_{pj})}$$



$$Silhouette\_Index = \frac{1}{N} \sum_{j=1}^{N} S_{x_j}$$

$$Silhouette\_Index \rightarrow Max$$

Нечеткая кластеризация: GK и Fuzzy K-means

$$v_j = \frac{\sum_{j=1}^{N}(u_{ij})^m \cdot x_i}{\sum_{i=1}^{N}(u_{ij})^m}$$

$$D_{ij} = \sqrt{\left\|x_i - v_j\right\|^2}$$

$$A_j = \frac{\sum_{i=1}^{N}(u_{ij})^m \cdot \left(x_i - v_j\right)^T \cdot \left(x_i - v_j\right)}{\sum_{i=1}^{N}(u_{ij})^m}$$

$$D_{A_j} = \left(x_i - v_j\right) \cdot \left[\left(det\left(A_j\right)\right)^{\frac{1}{N}} \cdot A_j^{-1}\right] \cdot \left(x_i - v_j\right)^T$$

$$u_{ij} = \left(\left(D_{ij}\right)^2 \sum_{k=1}^{c} \frac{1}{(D_{ik})^2}\right)^{-\frac{1}{m-1}}$$

Индекс Си-Бени (XB - Xie-Beni Index)

$$Xie - Beni\_Index = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}(u_{ij})^{m}\left\|x_j - v_i\right\|^2}{N \cdot min_{l\neq s}\left\|v_l - v_s\right\|^2}$$

$$Xie - Beni\_Index \rightarrow Min$$

Нечеткий упрощенный Силуэт (Fuzzy Simplified Silhouette)

$$S_j = \frac{b_{pj} - a_{pj}}{max\{a_{pj}, b_{pj}\}}$$

$$F\_S\_Silhoutte\_Index = \frac{\sum_{j=1}^{N}(u_{pj} - u_{qj})^{\alpha} S_j}{\sum_{j=1}^{N}(u_{pj} - u_{qj})^{\alpha}}$$

$$F\_S\_Silhoutte\_Index \rightarrow Max$$

Индекс Квона (Kwon Index)

$$Kwon\_Index = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}(u_{ij})^{m}\left\|x_j - v_i\right\|^2 + \frac{1}{c}\sum_{i=1}^{c}\|v_i - \bar{x}\|^2}{min_{l\neq s}\|v_l - v_s\|^2}$$

$$\bar{x} = \frac{1}{N}\sum_{j=1}^{N} x_j$$

$$Kwon\_Index \rightarrow Min$$

TSS индекс (Tang-Sun-Sun Index)

$$TSS\_Index = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}(u_{ij})^{m}\left\|x_j - v_i\right\|^2 + \frac{1}{c(c-1)}\sum_{i=1}^{c}\sum_{l=1,l\neq i}^{c}\|v_i - \bar{x}\|^2}{min_{l\neq s}\|v_l - v_s\|^2 + 1/k}$$

$$TSS\_Index \rightarrow Min$$

Индекс Фукуямы-Сугено (Fukuyama-Sugeno Index)

$$FS\_Index = \sum_{i=1}^{c}\sum_{j=1}^{N}(u_{ij})^{m}\left\|x_j - v_i\right\|^2 + \sum_{i=1}^{c}\sum_{j=1}^{N}(u_{ij})^{m}\|v_i - \bar{x}\|^2$$

$$FS\_Index \rightarrow Min$$