

Основы машинного обучения

Поляк Марк Дмитриевич

2025

Обучение без учителя и частичное обучение

Лекция 10

Постановка задачи

Кластеризация, частичное обучение

Постановка задачи кластеризации

Дано:

- X – пространство объектов;
- $X^\ell = \{x_1, \dots, x_\ell\}$ – обучающая выборка;
- $\rho: X \times X \rightarrow [0, \infty)$ – функция расстояния между объектами

Найти:

- Y – множество кластеров,
- $g: X \rightarrow Y$ – алгоритм кластеризации, такой что:
 - каждый кластер состоит из близких объектов;
 - объекты разных кластеров существенно различны.

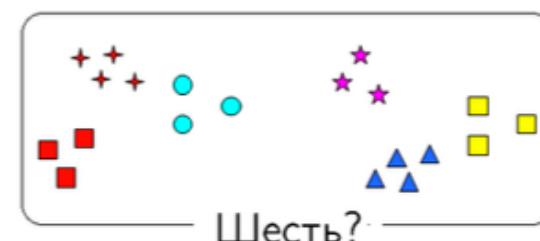
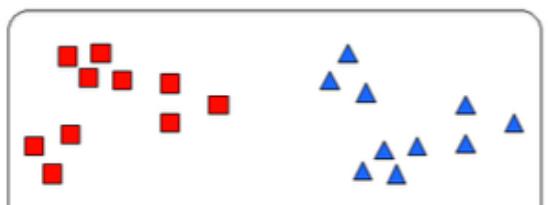
Это задача *обучения без учителя* (unsupervised learning).

Некорректность задачи кластеризации

Решение задачи кластеризации принципиально неоднозначно:

- точной постановки задачи кластеризации нет;
- существует много критериев качества кластеризации;
- существует много эвристических методов кластеризации;
- число кластеров $|Y|$, как правило, неизвестно заранее;
- результат кластеризации сильно зависит от метрики ρ , выбор которой также является эвристикой.

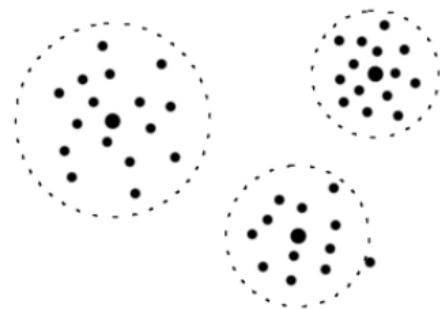
Пример: сколько здесь кластеров?



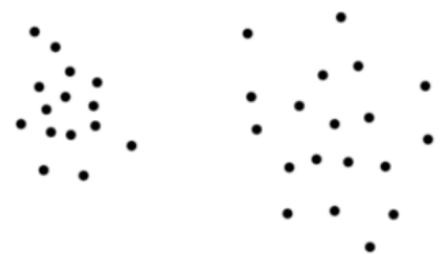
Цели кластеризации

- **Упростить дальнейшую обработку данных,**
разбить множество X^ℓ на группы схожих объектов
чтобы работать с каждой группой в отдельности
(задачи классификации, регрессии, прогнозирования).
- **Сократить объём хранимых данных,**
оставив по одному представителю от каждого кластера
(задачи сжатия данных).
- **Выделить нетипичные объекты,**
которые не подходят ни к одному из кластеров
(задачи одноклассовой классификации).
- **Построить иерархию множества объектов,**
пример — классификация животных и растений К.Линнея
(задачи таксономии).

Типы кластерных структур



кластеры с центрами



внутрикластерные расстояния
меньше межкластерных



ленточные кластеры

Типы кластерных структур



перемычки между кластерами



разреженный фон
из нетипичных объектов



перекрывающиеся кластеры

Типы кластерных структур



кластеры могут вообще отсутствовать

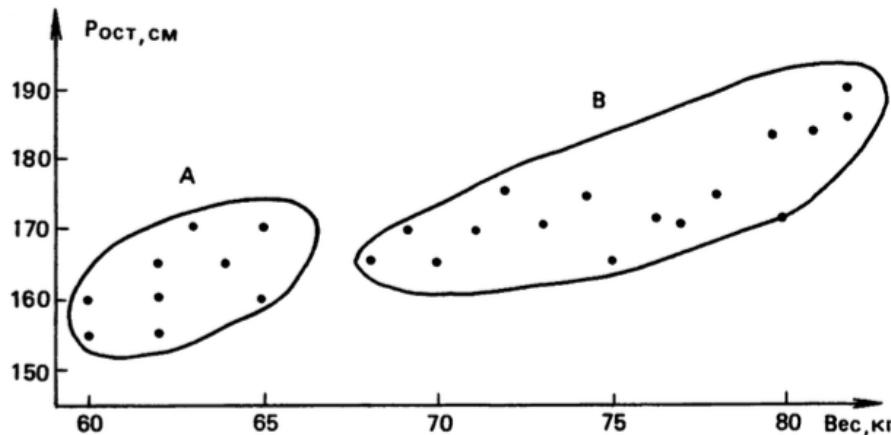


а это вообще не кластеры

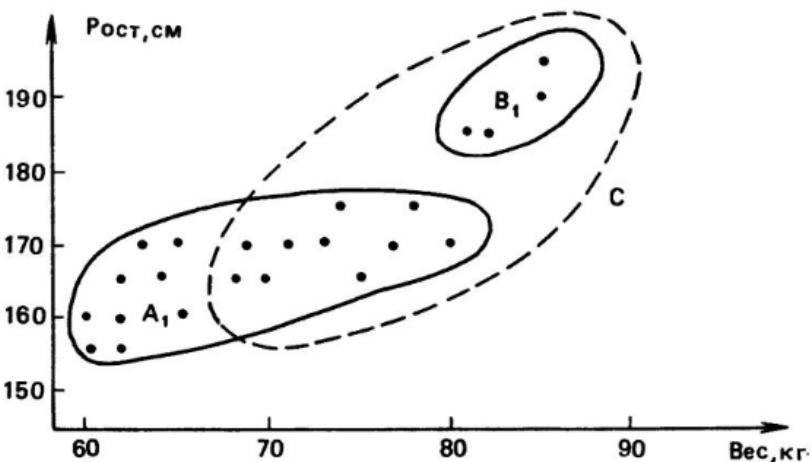
- Каждый метод кластеризации имеет свои ограничения и выделяет кластеры лишь некоторых типов.
- Понятие «тип кластерной структуры» зависит от метода и также не имеет формального определения.

Проблема чувствительности к выбору метрики

Результат зависит от нормировки признаков:



A — студентки,
B — студенты



после перенормировки
(сжали ось «вес» вдвое)

Постановка задачи частичного обучения

Дано:

множество объектов X , множество классов Y ;

$X^k = \{x_1, \dots, x_k\}$ – размеченные объекты (labeled data);
 $\{y_1, \dots, y_k\}$

$U = \{x_{k+1}, \dots, x_\ell\}$ – неразмеченные объекты (unlabeled data);

Два варианта постановки задачи:

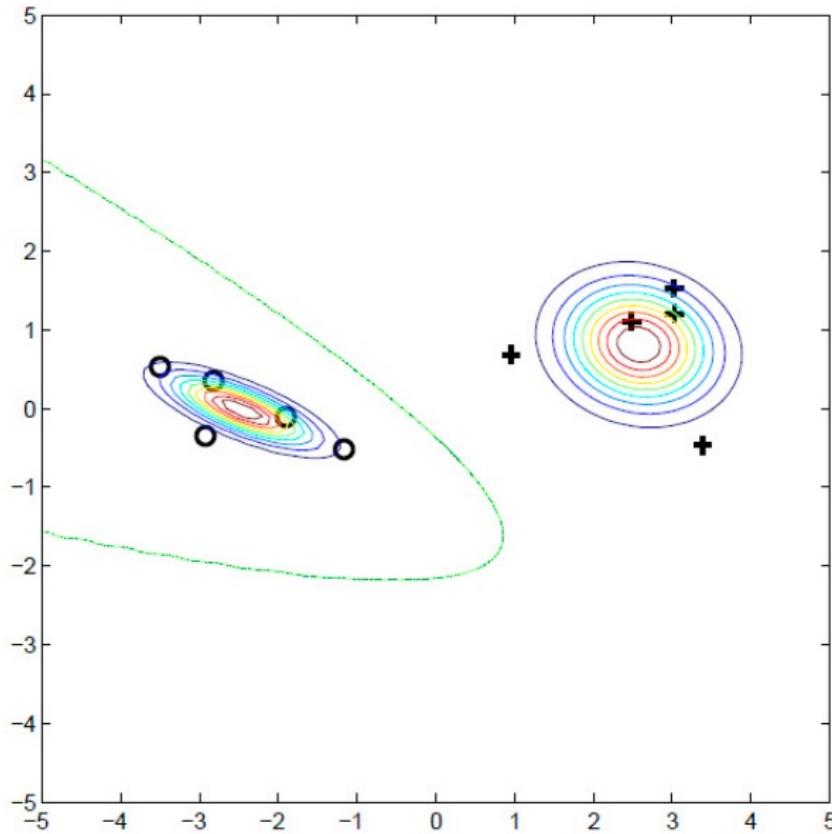
- Частичное обучение (semi-supervised learning, SSL):
построить алгоритм классификации $g: X \rightarrow Y$
- Трансдуктивное обучение (transductive learning):
зная **все** $\{x_{k+1}, \dots, x_\ell\}$, получить метки $\{a_{k+1}, \dots, a_\ell\}$.

Типичные приложения:

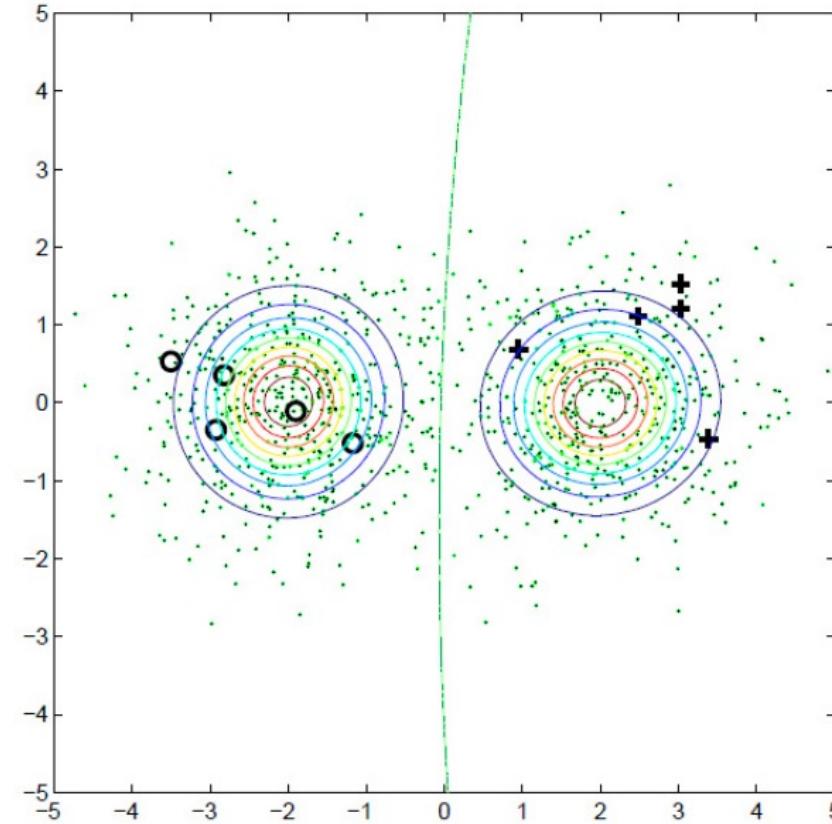
Классификация и каталогизация текстов, изображений и т.п.

SSL не сводится к классификации

Пример 1. плотности классов, восстановленные:
по размеченным данным X^k

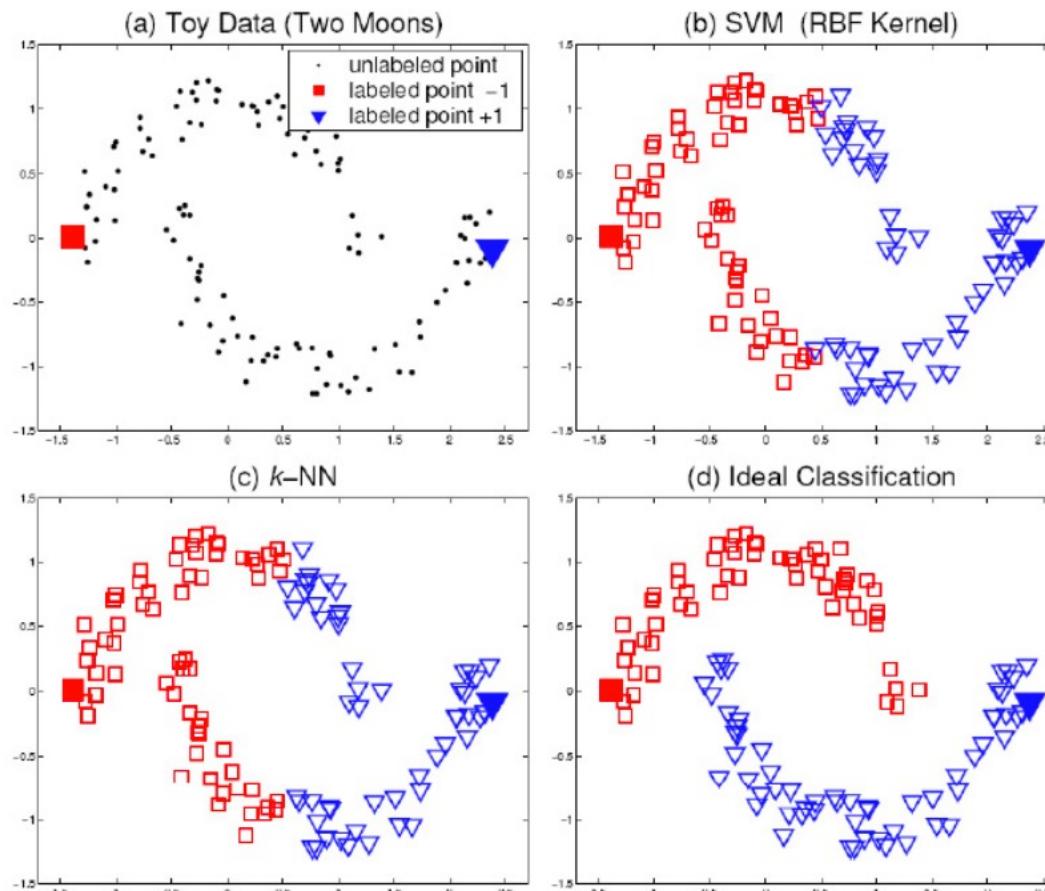


по полным данным X^ℓ



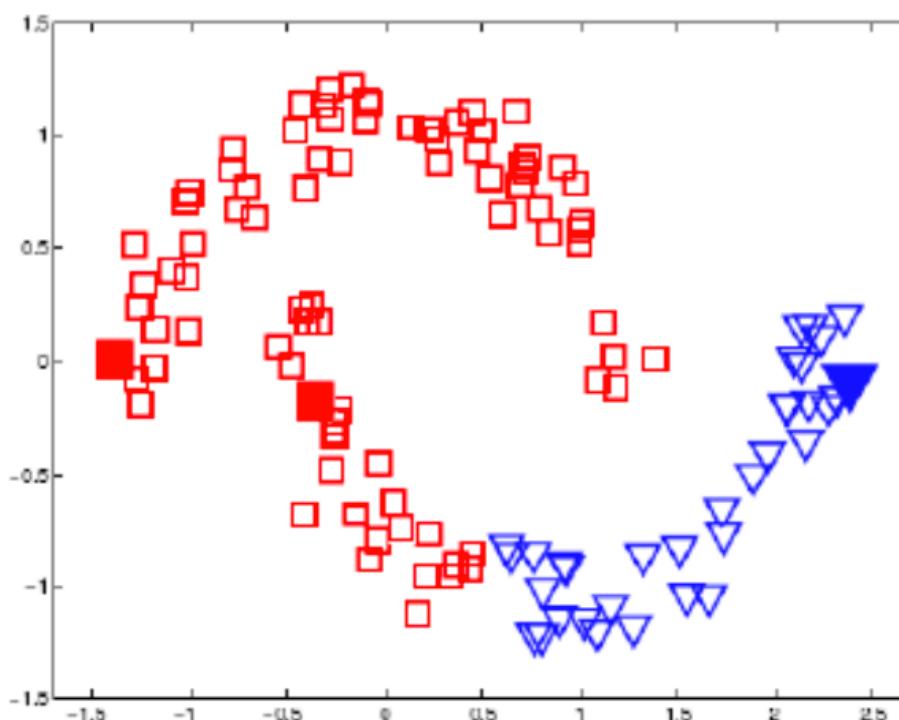
SSL не сводится к классификации

Пример 2. Методы классификации не учитывают кластерную структуру неразмеченных данных



Однако и к кластеризации SSL тоже не сводится

Пример 3. Методы кластеризации не учитывают приоритетность разметки над кластерной структурой.



Качество кластеризации в метрическом пространстве

Пусть известны только попарные расстояния между объектами.

$a_i = a(x_i)$ — кластеризация объекта x_i

- Среднее внутрикластерное расстояние:

$$F_0 = \frac{\sum_{i < j} [a_i = a_j] \rho(x_i, x_j)}{\sum_{i < j} [a_i = a_j]} \rightarrow \min .$$

- Среднее межкластерное расстояние:

$$F_1 = \frac{\sum_{i < j} [a_i \neq a_j] \rho(x_i, x_j)}{\sum_{i < j} [a_i \neq a_j]} \rightarrow \max .$$

- Отношение пары функционалов: $F_0/F_1 \rightarrow \min$.

Качество кластеризации в линейном векторном пространстве

Пусть объекты x_i задаются векторами $(f_1(x_i), \dots, f_n(x_i))$.

- Сумма средних внутрикластерных расстояний:

$$\Phi_0 = \sum_{a \in Y} \frac{1}{|X_a|} \sum_{i: a_i = a} \rho(x_i, \mu_a) \rightarrow \min,$$

$X_a = \{x_i \in X^\ell \mid a_i = a\}$ — кластер a ,
 μ_a — центр масс кластера a .

- Сумма межкластерных расстояний:

$$\Phi_1 = \sum_{a, b \in Y} \rho(\mu_a, \mu_b) \rightarrow \max.$$

- Отношение пары функционалов: $\Phi_0 / \Phi_1 \rightarrow \min$.

Коэффициент силуэта (анализ ошибок кластеризации)

Распределение качества кластеризации по объектам/кластерам

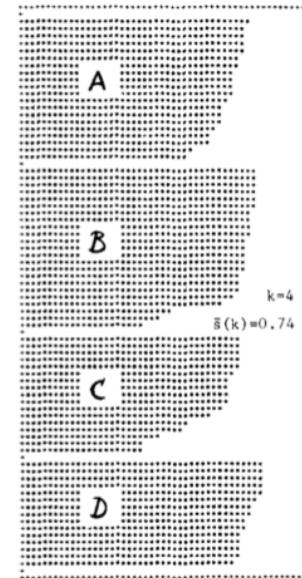
- Ср.расстояние до объектов своего кластера:

$$r_i = \frac{1}{|X_{a_i}| - 1} \sum_{x \in X_{a_i} \setminus x_i} \rho(x, x_i)$$

- Мин. ср.расстояние до чужого кластера:

$$R_i = \min_{a \in Y \setminus a_i} \frac{1}{|X_a|} \sum_{x \in X_a} \rho(x, x_i)$$

- Коэффициент силуэта объекта: $s(i) = \frac{R_i - r_i}{\max(R_i, r_i)} \in [-1, +1]$



Peter J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. 1987.

Точность и полнота кластеризации в сравнении с эталоном

$y_i \in Y_0$ — эталонная классификация объектов, $i = 1, \dots, \ell$

Y_0 может не совпадать с Y по мощности

$P_i = \{k : a_k = a_i\}$ — кластер объекта x_i

$Q_i = \{k : y_k = y_i\}$ — эталонный класс объекта x_i

BCubed-меры точности и полноты кластеризации:

$$\text{Precision} = \frac{1}{\ell} \sum_{i=1}^{\ell} \frac{|P_i \cap Q_i|}{|P_i|} \quad \text{— средняя точность}$$

$$\text{Recall} = \frac{1}{\ell} \sum_{i=1}^{\ell} \frac{|P_i \cap Q_i|}{|Q_i|} \quad \text{— средняя полнота}$$

$$F_1 = \frac{1}{\ell} \sum_{i=1}^{\ell} \frac{2|P_i \cap Q_i|}{|P_i| + |Q_i|} \quad \text{— средняя } F_1\text{-мера}$$

Алгоритмы кластеризации

K-средних, DBSCAN, иерархическая кластеризация, самоорганизующиеся карты Кохонена

Метод K -средних (K -means) для кластеризации

Минимизация суммы квадратов внутрикластерных расстояний:

$$\sum_{i=1}^{\ell} \|x_i - \mu_{a_i}\|^2 \rightarrow \min_{\{a_i\}, \{\mu_a\}}, \quad \|x_i - \mu_a\|^2 = \sum_{j=1}^n (f_j(x_i) - \mu_{aj})^2$$

Алгоритм Ллойда

вход: X^ℓ , $K = |Y|$; **выход:** центры кластеров μ_a , $a \in Y$;

$\mu_a :=$ начальное приближение центров, для всех $a \in Y$;

повторять

отнести каждый x_i к ближайшему центру:

$$a_i := \arg \min_{a \in Y} \|x_i - \mu_a\|, \quad i = 1, \dots, \ell;$$

вычислить новые положения центров:

$$\mu_a := \frac{\sum_{i=1}^{\ell} [a_i = a] x_i}{\sum_{i=1}^{\ell} [a_i = a]}, \quad a \in Y;$$

пока a_i не перестанут изменяться;

Метод K-средних (K-means) для частичного обучения

Модификация алгоритма Ллойда
при наличии размеченных объектов $\{x_1, \dots, x_k\}$

вход: X^ℓ , $K = |Y|$;

выход: центры кластеров μ_a , $a \in Y$;

$\mu_a :=$ начальное приближение центров, для всех $a \in Y$;

повторять

отнести каждый $x_i \in U$ к ближайшему центру:

$$a_i := \arg \min_{a \in Y} \|x_i - \mu_a\|, \quad i = k + 1, \dots, \ell;$$

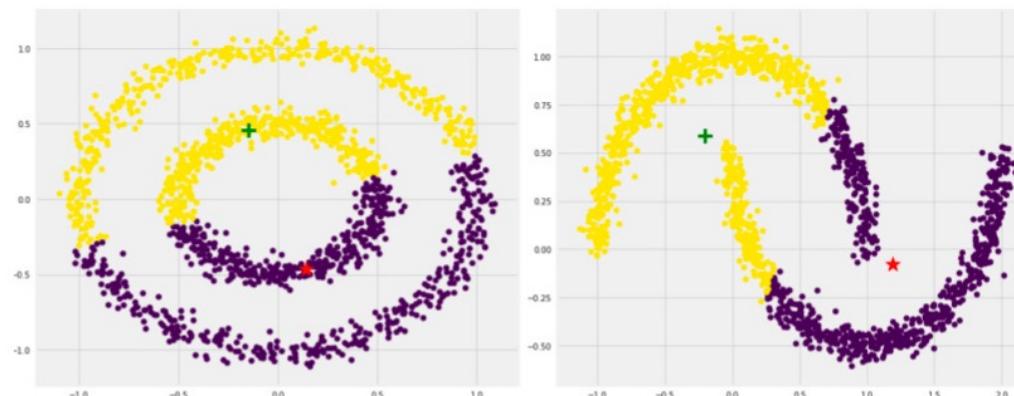
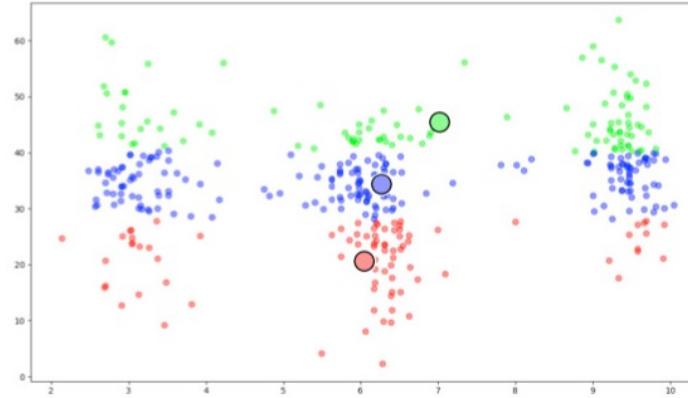
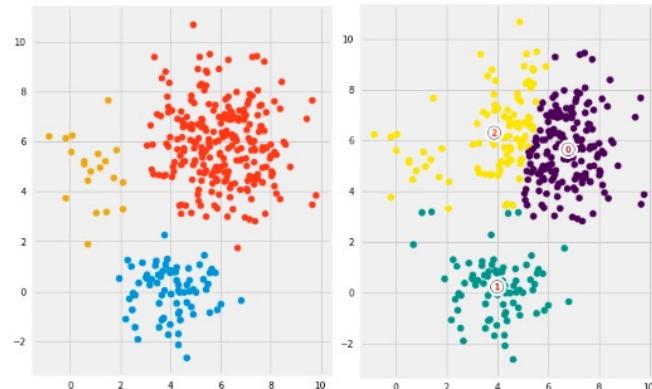
вычислить новые положения центров:

$$\mu_a := \frac{\sum_{i=1}^{\ell} [a_i = a] x_i}{\sum_{i=1}^{\ell} [a_i = a]}, \quad a \in Y;$$

пока a_i не перестанут изменяться;

Примеры неудачной кластеризации K-means

Причина — неудачное начальное приближение или
существенная негауссовость кластеров



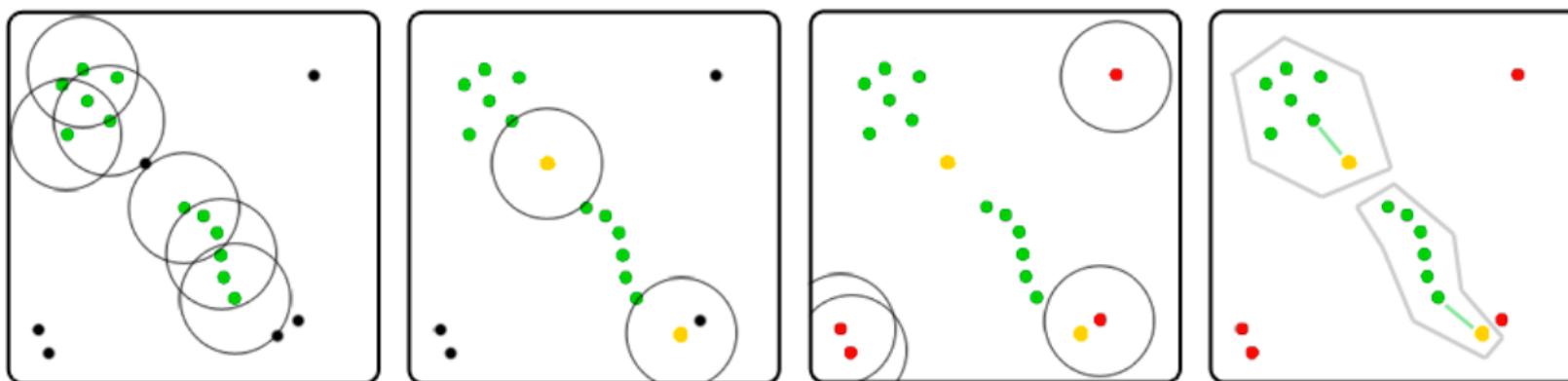
Алгоритм кластеризации DBSCAN

Density-Based Spatial Clustering of Applications with Noise

Объект $x \in U$, его ε -окрестность $U_\varepsilon(x) = \{u \in U: \rho(x, u) \leq \varepsilon\}$

Каждый объект может быть одного из трёх типов:

- корневой: имеющий плотную окрестность, $|U_\varepsilon(x)| \geq m$
- граничный: не корневой, но в окрестности корневого
- шумовой (выброс): не корневой и не граничный



Ester, Kriegel, Sander, Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. KDD-1996.

Алгоритм кластеризации DBSCAN

вход: выборка $X^\ell = \{x_1, \dots, x_\ell\}$; параметры ε и m ;

выход: разбиение выборки на кластеры и шумовые выбросы;

$U := X^\ell$ — непомеченные; $a := 0$;

пока в выборке есть непомеченные точки, $U \neq \emptyset$:

 взять случайную точку $x \in U$;

если $|U_\varepsilon(x)| < m$ **то**

 пометить x как, возможно, шумовой;

иначе

 создать новый кластер: $K := U_\varepsilon(x)$; $a := a + 1$;

для всех $x' \in K$, не помеченных или шумовых

если $|U_\varepsilon(x')| \geq m$ **то** $K := K \cup U_\varepsilon(x')$;

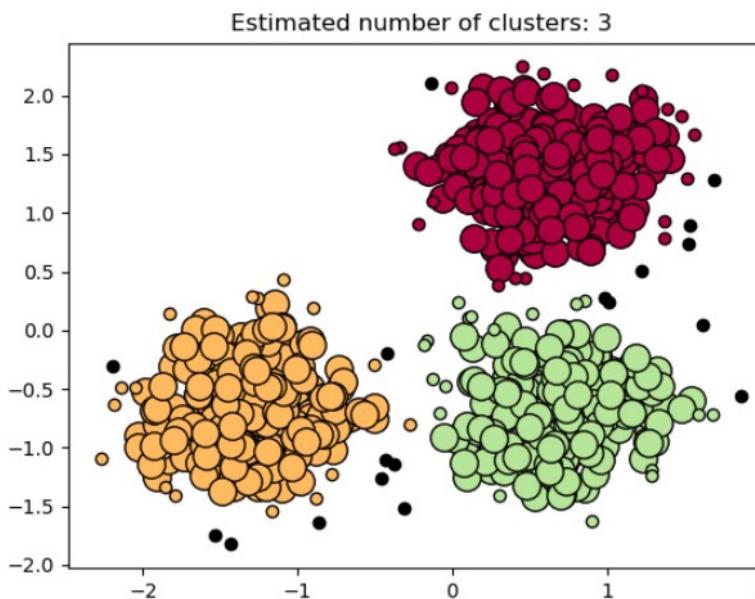
иначе пометить x' как граничный кластера K ;

$a_i := a$ для всех $x_i \in K$;

$U := U \setminus K$;

Преимущества алгоритма DBSCAN

- быстрая кластеризация больших данных:
 $O(\ell^2)$ в худшем случае,
 $O(\ell \ln \ell)$ при эффективной реализации $U_\varepsilon(x)$;
- кластеры произвольной формы (долой центры!);
- деление объектов на корневые, граничные, шумовые.



Агломеративная иерархическая кластеризация

Алгоритм иерархической кластеризации (Ланс, Уильямс, 1967):
итеративный пересчёт расстояний R_{UV} между кластерами U, V .

$C_1 := \{\{x_1\}, \dots, \{x_\ell\}\}$ — все кластеры 1-элементные;

$R_{\{x_i\}\{x_j\}} := \rho(x_i, x_j)$ — расстояния между ними;

для всех $t = 2, \dots, \ell$ (t — номер итерации):

найти в C_{t-1} пару кластеров (U, V) с минимальным R_{UV} ;

слить их в один кластер:

$W := U \cup V$;

$C_t := C_{t-1} \cup \{W\} \setminus \{U, V\}$;

для всех $S \in C_t$

вычислить R_{WS} по формуле Ланса-Уильямса:

$R_{WS} := \alpha_U R_{US} + \alpha_V R_{VS} + \beta R_{UV} + \gamma |R_{US} - R_{VS}|$;

Алгоритм Ланса-Уильямса для частичного обучения

Алгоритм иерархической кластеризации (Ланс, Уильямс, 1967):
итеративный пересчёт расстояний R_{UV} между кластерами U, V .

$C_1 := \{\{x_1\}, \dots, \{x_\ell\}\}$ — все кластеры 1-элементные;

$R_{\{x_i\}\{x_j\}} := \rho(x_i, x_j)$ — расстояния между ними;

для всех $t = 2, \dots, \ell$ (t — номер итерации):

найти в C_{t-1} пару кластеров (U, V) с минимальным R_{UV} ,

при условии, что в $U \cup V$ нет объектов с разными метками;

слить их в один кластер:

$W := U \cup V;$

$C_t := C_{t-1} \cup \{W\} \setminus \{U, V\};$

для всех $S \in C_t$

вычислить R_{WS} по формуле Ланса-Уильямса:

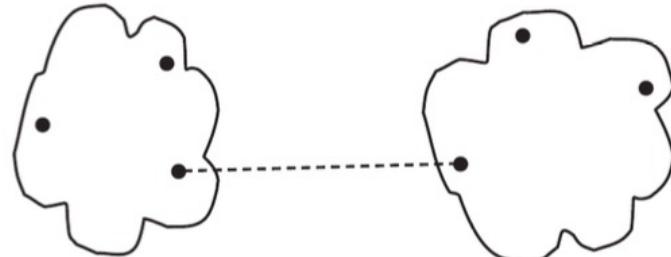
$R_{WS} := \alpha_U R_{US} + \alpha_V R_{VS} + \beta R_{UV} + \gamma |R_{US} - R_{VS}|;$

Частные случаи формулы Ланса-Уильямса

1. Расстояние ближнего соседа:

$$R_{WS}^6 = \min_{w \in W, s \in S} \rho(w, s);$$

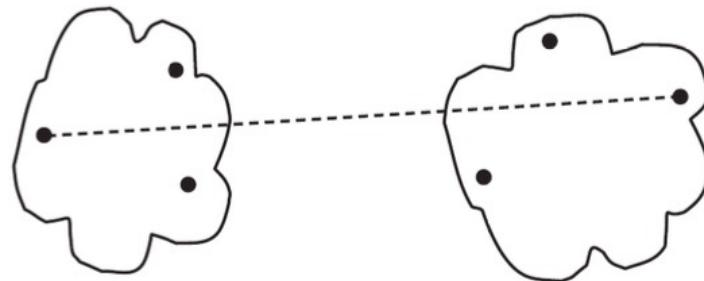
$$\alpha_U = \alpha_V = \frac{1}{2}, \quad \beta = 0, \quad \gamma = -\frac{1}{2}.$$



2. Расстояние дальнего соседа:

$$R_{WS}^d = \max_{w \in W, s \in S} \rho(w, s);$$

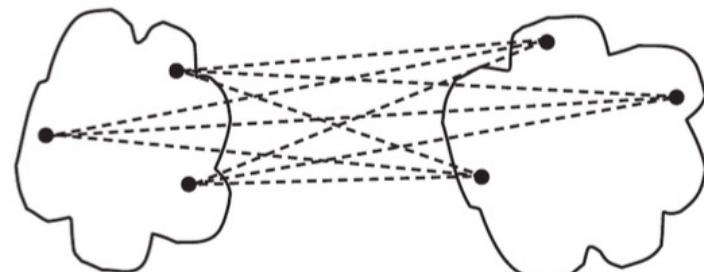
$$\alpha_U = \alpha_V = \frac{1}{2}, \quad \beta = 0, \quad \gamma = \frac{1}{2}.$$



3. Групповое среднее расстояние:

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

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



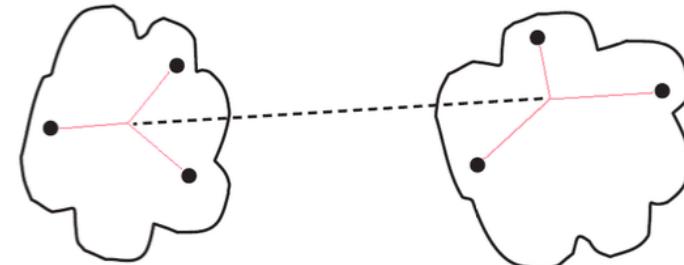
Частные случаи формулы Ланса-Уильямса

4. Расстояние между центрами:

$$R_{WS}^u = \rho^2 \left(\sum_{w \in W} \frac{w}{|W|}, \sum_{s \in S} \frac{s}{|S|} \right);$$

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

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



5. Расстояние Уорда:

$$R_{WS}^y = \frac{|S||W|}{|S|+|W|} \rho^2 \left(\sum_{w \in W} \frac{w}{|W|}, \sum_{s \in S} \frac{s}{|S|} \right);$$

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

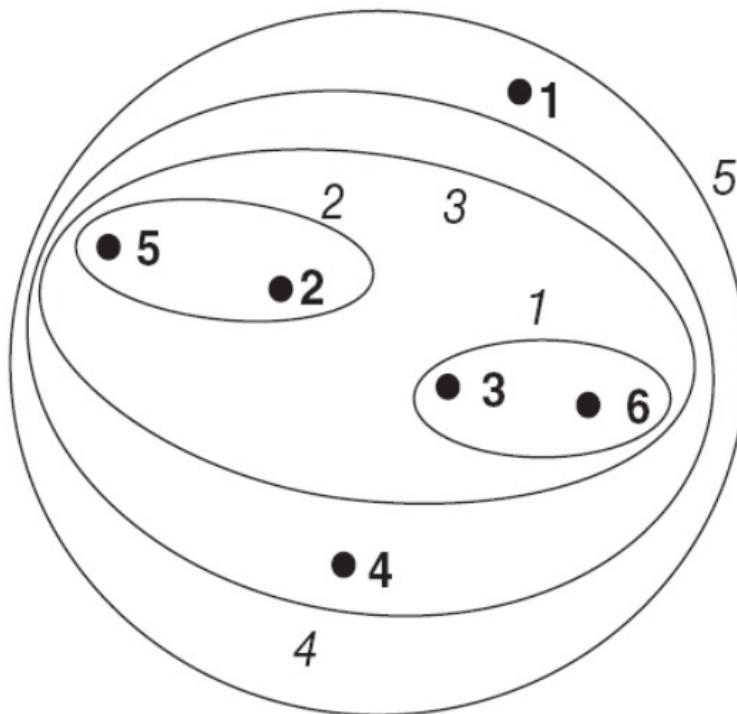
Проблема выбора

Какая функция расстояния лучше?

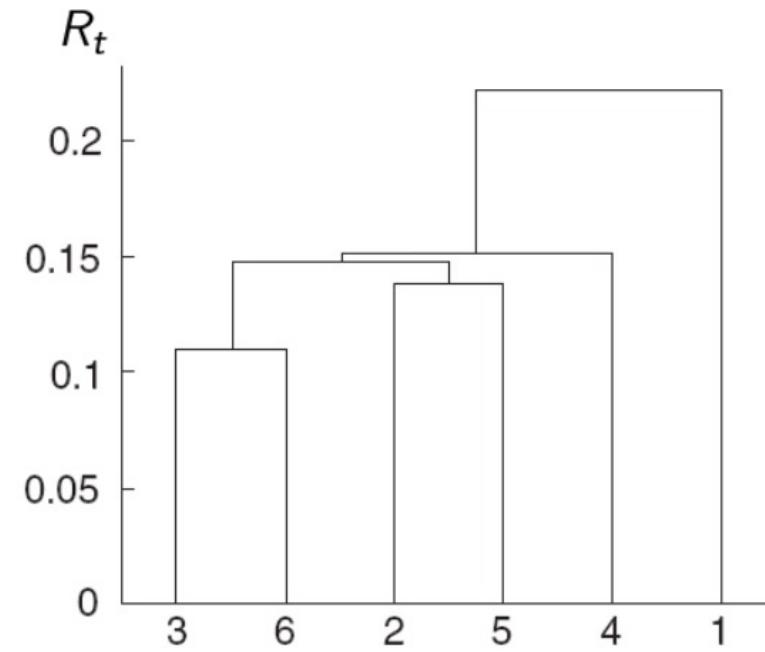
Визуализация кластерной структуры

1. Расстояние ближнего соседа:

Диаграмма вложения



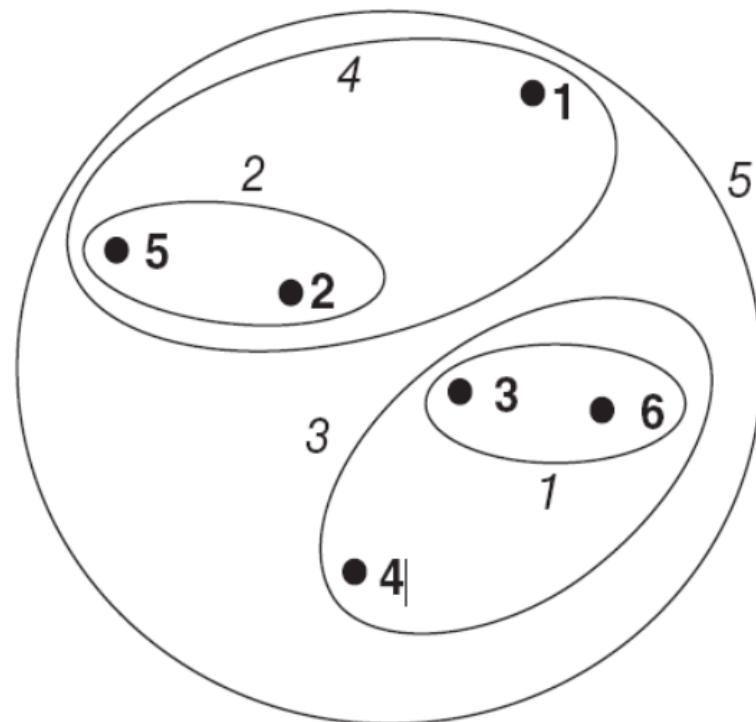
Дендрограмма



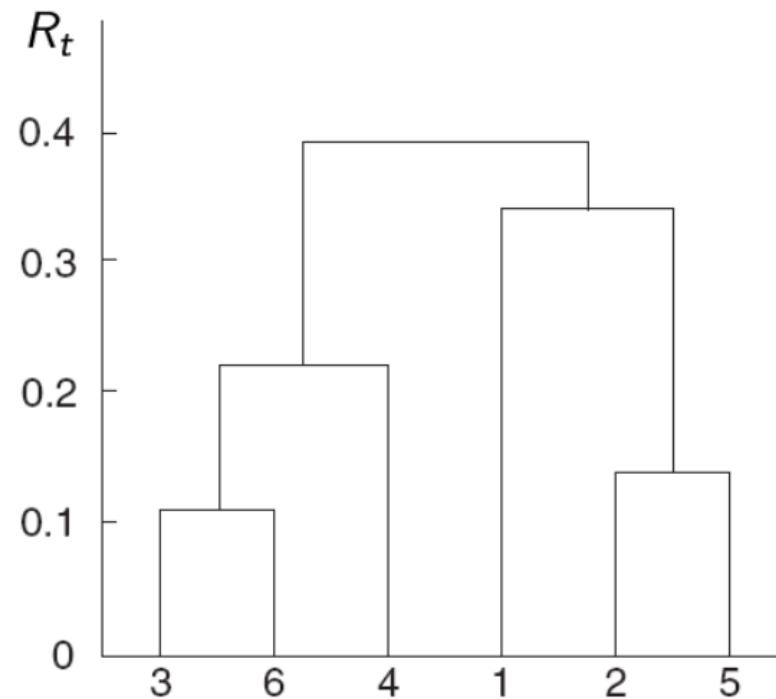
Визуализация кластерной структуры

2. Расстояние дальнего соседа:

Диаграмма вложения



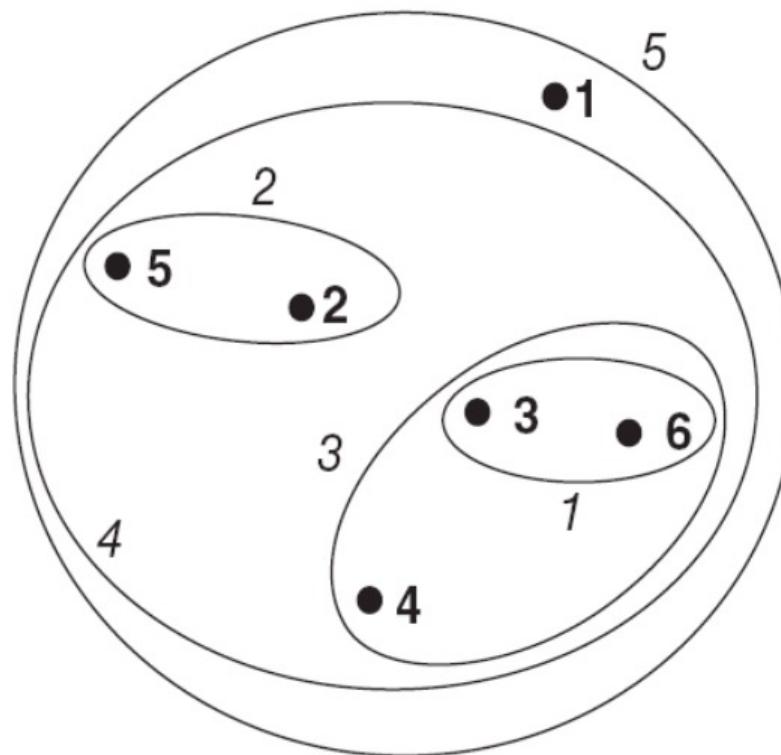
Дендрограмма



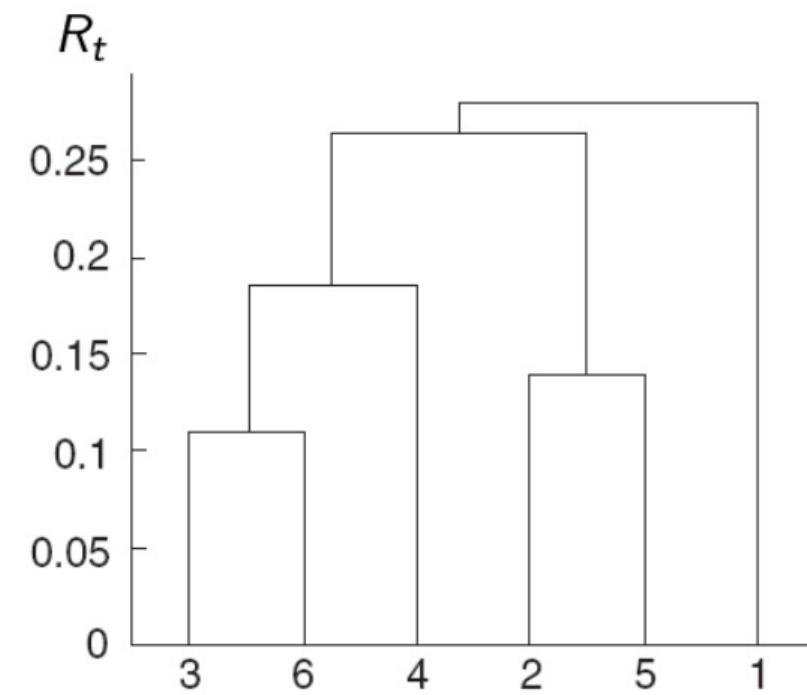
Визуализация кластерной структуры

3. Групповое среднее расстояние:

Диаграмма вложения



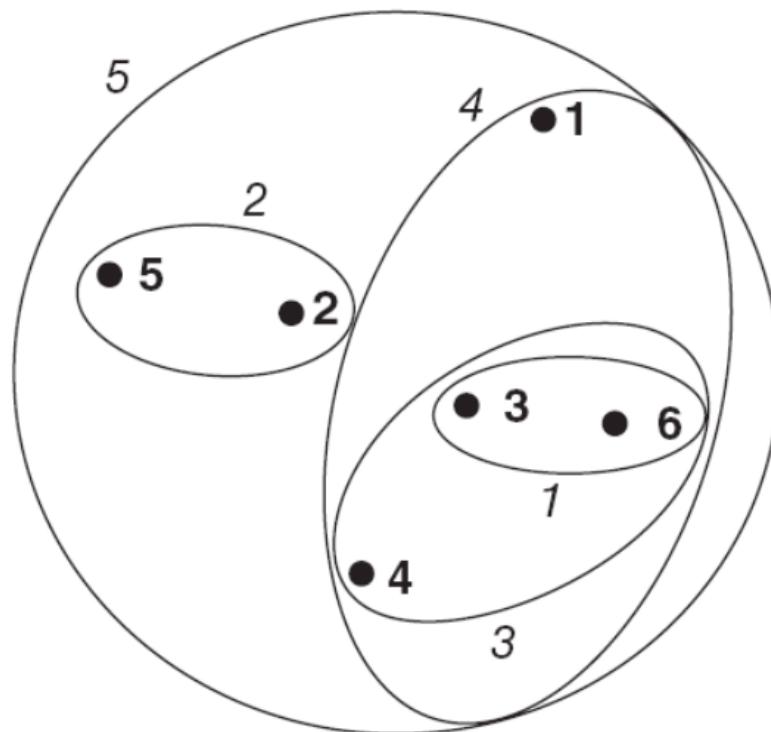
Дендрограмма



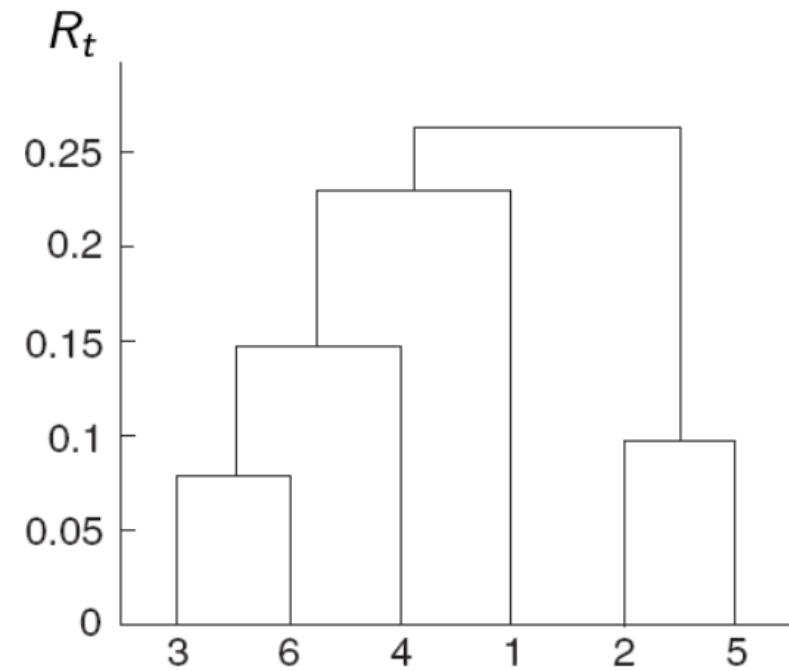
Визуализация кластерной структуры

5. Расстояние Уорда:

Диаграмма вложения

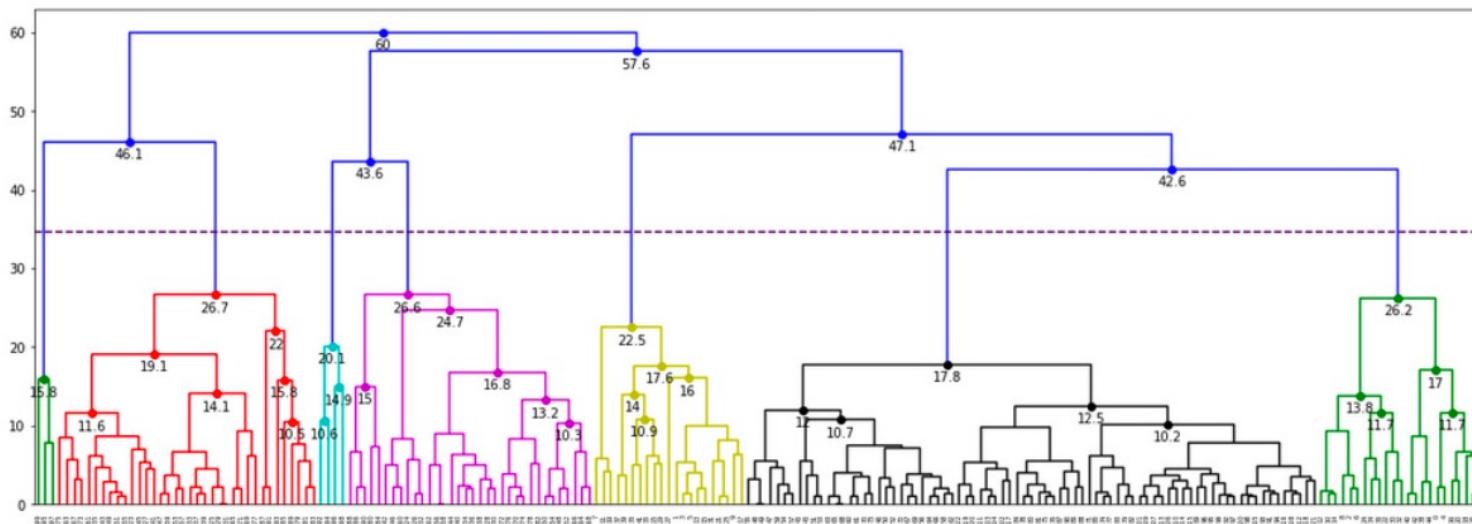


Дендрограмма



Дендрограмма – визуализация иерархической кластеризации

- Кластеры группируются вдоль горизонтальной оси
- По вертикальной оси откладываются расстояния R_t
- Расстояния возрастают, линии нигде не пересекаются
- Верхние уровни различимы лучше, чем нижние
- Уровень отсечения определяет число кластеров



Основные свойства иерархической кластеризации

- *Монотонность*: дендрограмма не имеет самопересечений, при каждом слиянии расстояние между объединяемыми кластерами только увеличивается: $R_2 \leq R_3 \leq \dots \leq R_\ell$.
- *Сжимающее расстояние*: $R_t \leq \rho(\mu_U, \mu_V), \forall t$.
- *Растягивающее расстояние*: $R_t \geq \rho(\mu_U, \mu_V), \forall t$

Теорема (Миллиган, 1979)

Кластеризация монотонна, если выполняются условия

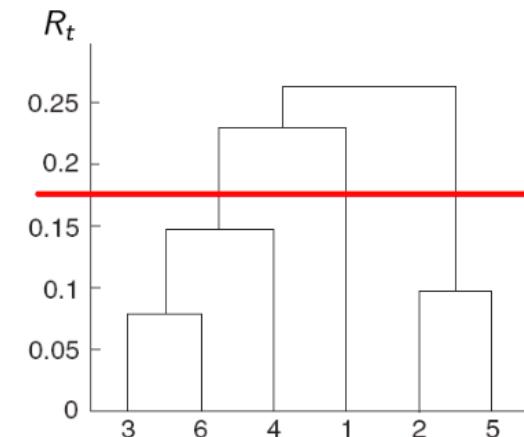
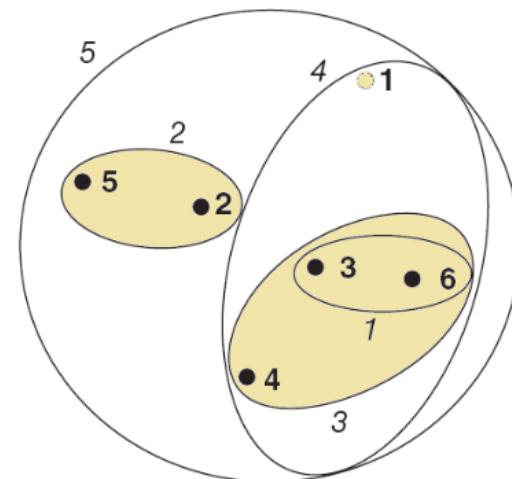
$$\alpha_U \geq 0, \alpha_V \geq 0, \alpha_U + \alpha_V + \beta \geq 1, \min\{\alpha_U, \alpha_V\} + \gamma \geq 0.$$

R^U не монотонно; R^B , R^A , R^Γ , R^Y — монотонны.

R^B — сжимающее; R^A , R^Y — растягивающие;

Рекомендации и выводы по иерархической кластеризации

- рекомендуется пользоваться расстоянием Уорда R^y ;
- 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 .



Карта Кохонена (Self Organizing Map, SOM)

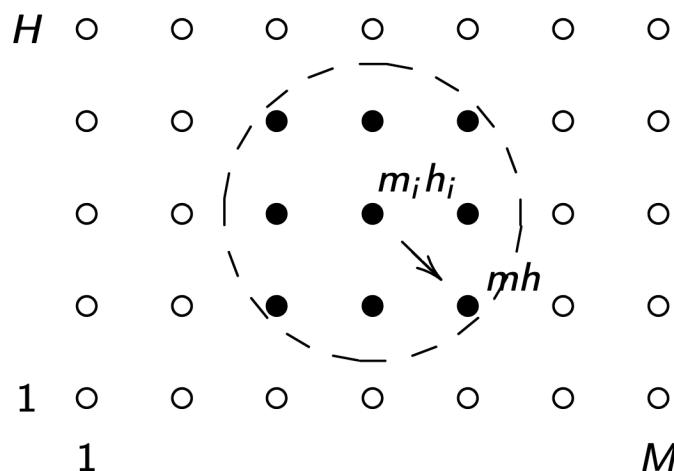
$Y = \{1, \dots, M\} \times \{1, \dots, H\}$ — прямоугольная сетка кластеров

Каждому узлу (m, h) приписан нейрон Кохонена $\theta_{mh} \in \mathbb{R}^n$

Наряду с метрикой $\rho(x_i, x)$ на X вводится метрика на сетке Y :

$$r((m_i, h_i), (m, h)) = \sqrt{(m - m_i)^2 + (h - h_i)^2}$$

Окрестность (m_i, h_i) :



Обучение карты Кохонена

Вход: X^ℓ – обучающая выборка; η – темп обучения;

Выход: $\theta_{mh} \in \mathbb{R}^n$ – векторы весов, $m = 1..M$, $h = 1..H$;

$\theta_{mh} := \text{random}\left(-\frac{1}{2MN}, \frac{1}{2MN}\right)$ – инициализация весов;

повторять

выбрать объект x_i из X^ℓ случайным образом;

WTA: вычислить координаты кластера:

$$(m_i, h_i) := g(x_i) \equiv \arg \min_{(m,h) \in Y} \rho(x_i, \theta_{mh})$$

для всех $(m, h) \in \text{Окрестность}(m_i, h_i)$

WTM: сделать шаг градиентного спуска:

$$\theta_{mh} := \theta_{mh} + \eta(x_i - \theta_{mh})K(r((m_i, h_i), (m, h)))$$

пока кластеризация не стабилизируется;

Интерпретация карт Кохонена

Два типа графиков — цветных карт $M \times H$:

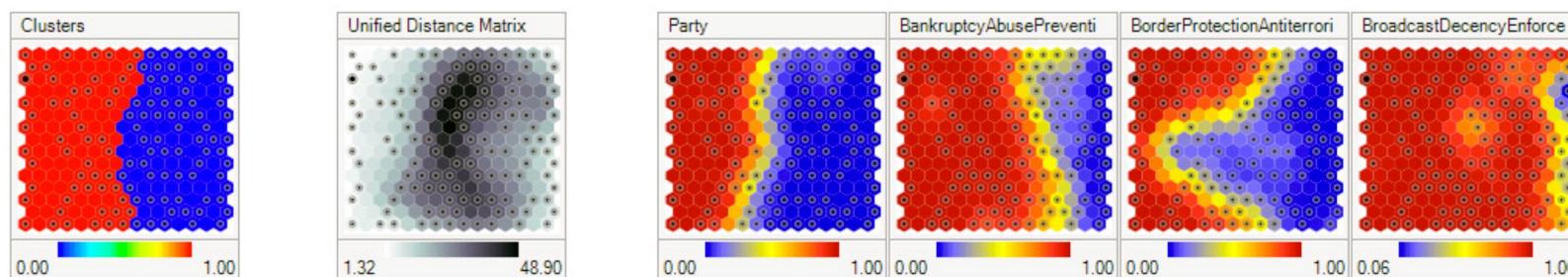
- Цвет узла (m, h) — локальная плотность в точке (m, h) — среднее расстояние до k ближайших точек выборки
- По одной карте на каждый признак:
цвет узла (m, h) — значение j -й компоненты вектора $w_{m,h}$

Пример: задача UCI house-votes (US Congress voting patterns)

Объекты — конгрессмены

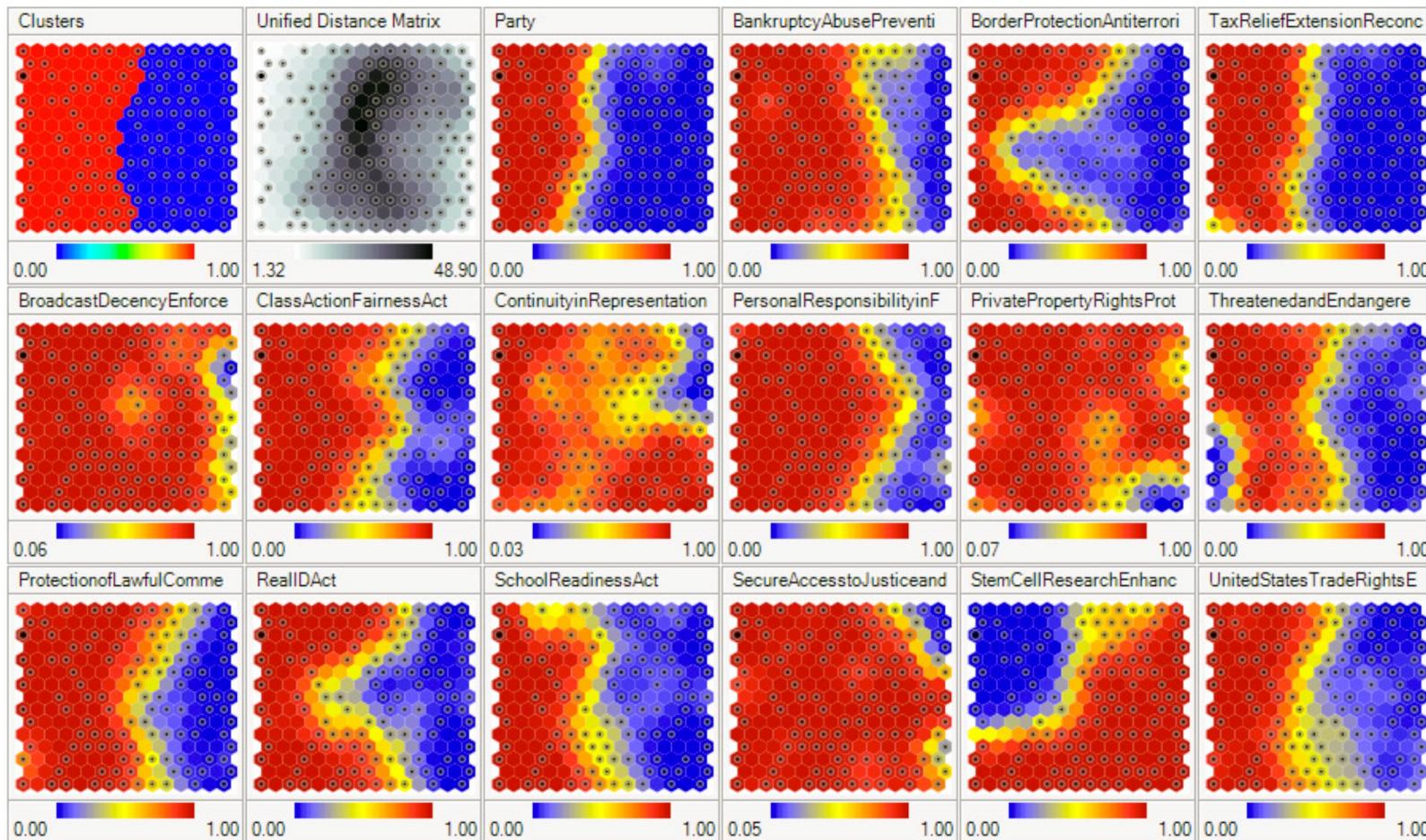
Признаки — результаты голосования по различным вопросам

Есть целевой признак «партия» $\in \{\text{демократ, республиканец}\}$



Интерпретация карт Кохонена (продолжение)

Пример: задача UCI house-votes (US Congress voting patterns)



Достоинства и недостатки карт Кохонена

Достоинства:

- Возможность визуального анализа многомерных данных
- Квантование выборки по кластерам,
с автоматическим определением числа непустых кластеров

Недостатки:

- **Субъективность.** Карта отражает не только кластерную структуру данных, но также зависит от...
 - свойств сглаживающего ядра;
 - (случайной) инициализации;
 - (случайного) выбора x_i в ходе итераций.
- **Искажения.** Близкие объекты исходного пространства могут переходить в далёкие точки на карте, и наоборот.

Рекомендуется только для разведочного анализа данных.