

TENSOR.BY

ML-course

# 5. Машинаное обучение без учителя

Александр Фридман (Data Scientist),  
[alexandef@epica.ai](mailto:alexandef@epica.ai)



# Задачи обучения без учителя

- Кластеризация
- Поиск ассоциативных правил
- Понижение размерности
- Выявление аномалий



# 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,

$a: X \rightarrow Y$  — алгоритм кластеризации,

такие, что:

- каждый кластер состоит из близких объектов;
- объекты разных кластеров существенно различны.

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

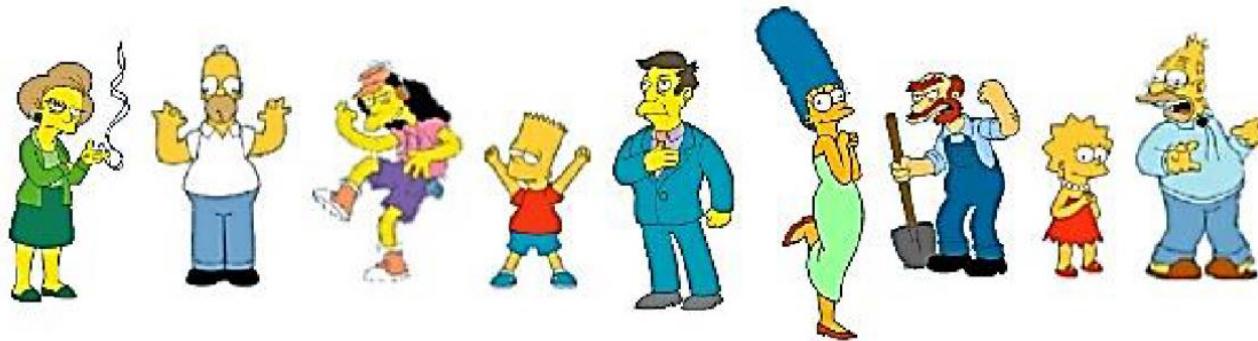


# Типы алгоритмов кластеризации

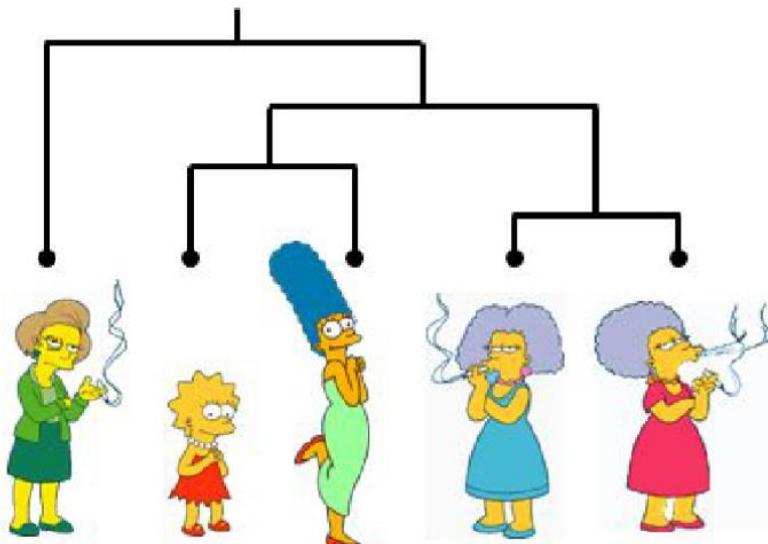
- **Жесткие (hard)**  
Элемент либо принадлежит кластеру, либо нет
- **Мягкие (soft)**  
Степень принадлежности элемента кластеру



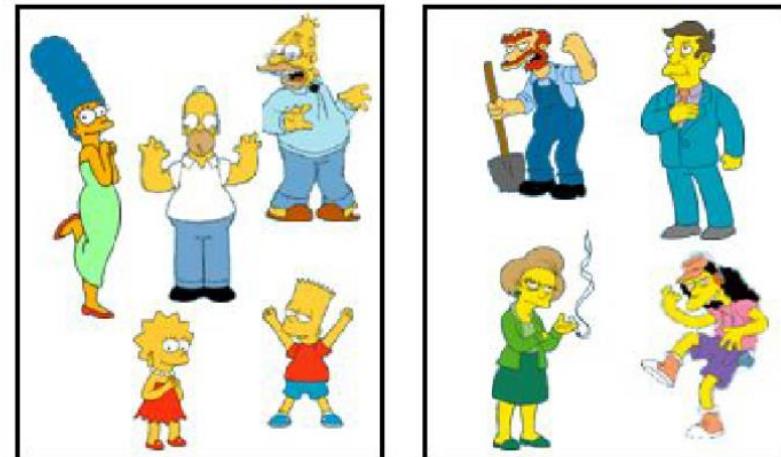
# Типы алгоритмов кластеризации



## Hierarchical



## Partitional





# 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  $\rho$ , выбор которой также является эвристикой.

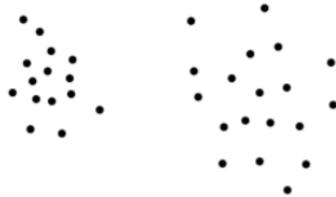


# Цели кластеризации

- Упростить дальнейшую обработку данных, разбить множество  $X^\ell$  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



внутрикластерные расстояния, как правило,  
меньше межкластерных



ленточные кластеры



кластеры с центром



# Типы кластерных структур



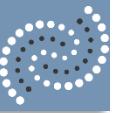
кластеры могут соединяться перемычками



кластеры могут накладываться на разреженный фон из редко расположенных объектов



кластеры могут перекрываться



# Типы кластерных структур



кластеры могут образовываться не по сходству, а по иным типам регулярностей

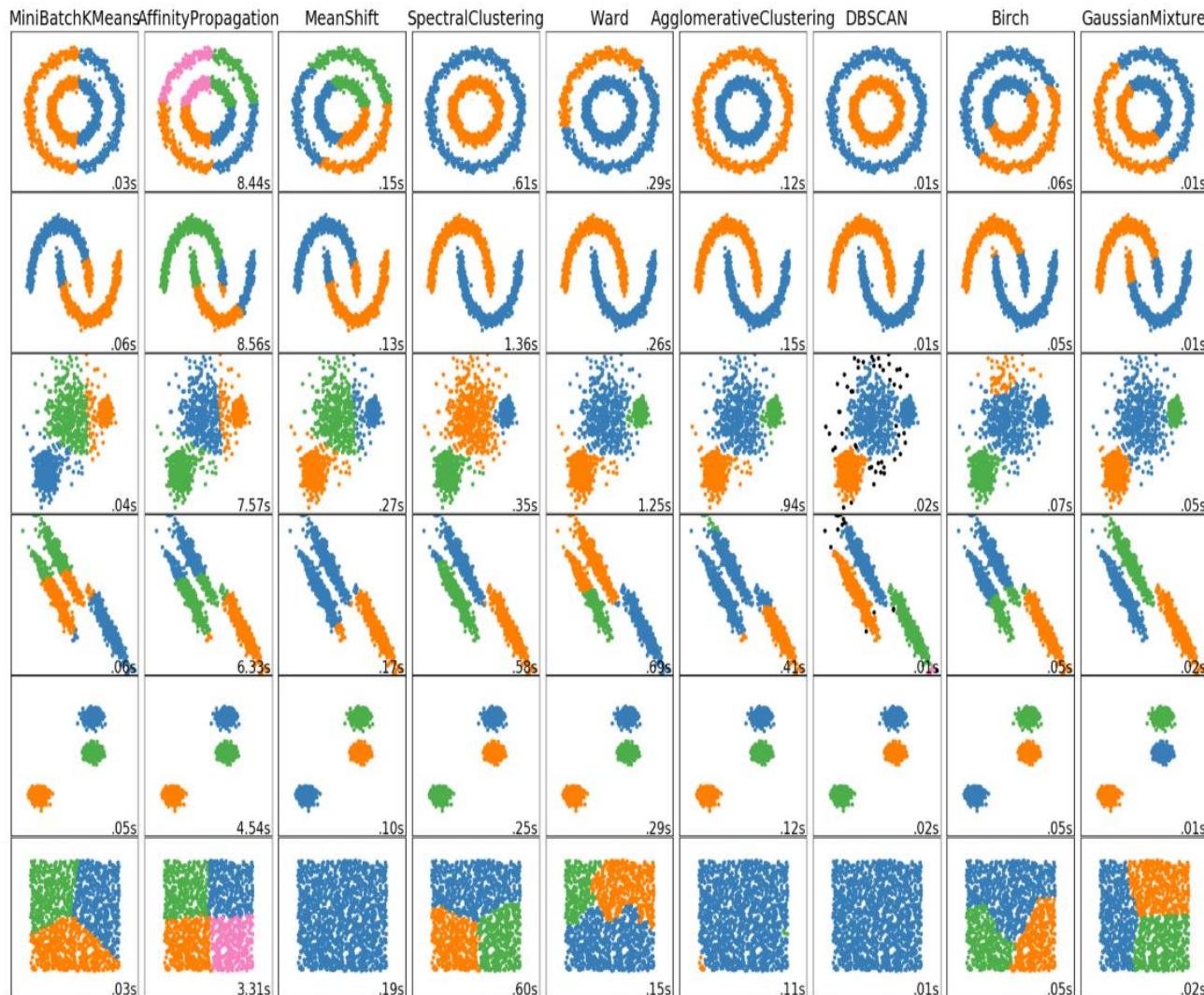


кластеры могут вообще отсутствовать

- 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.



# Возможности алгоритмов кластеризации

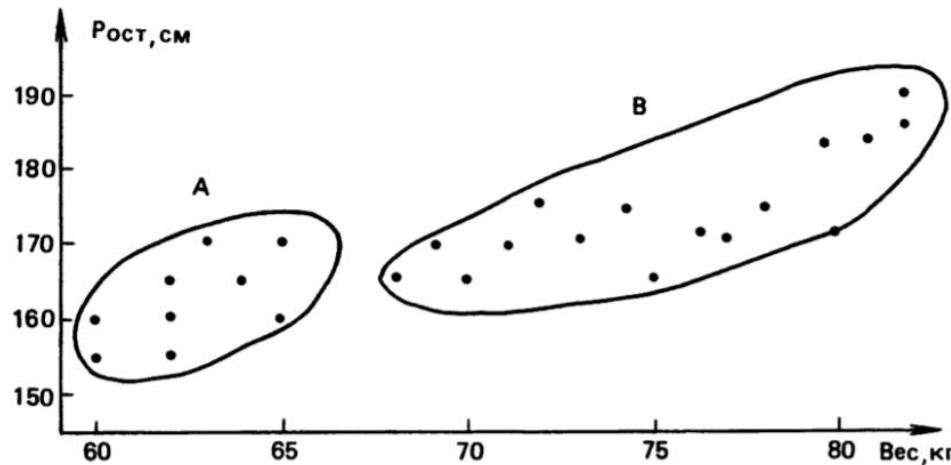


<https://scikit-learn.org/stable/modules/clustering.html#overview-of-clustering-methods>

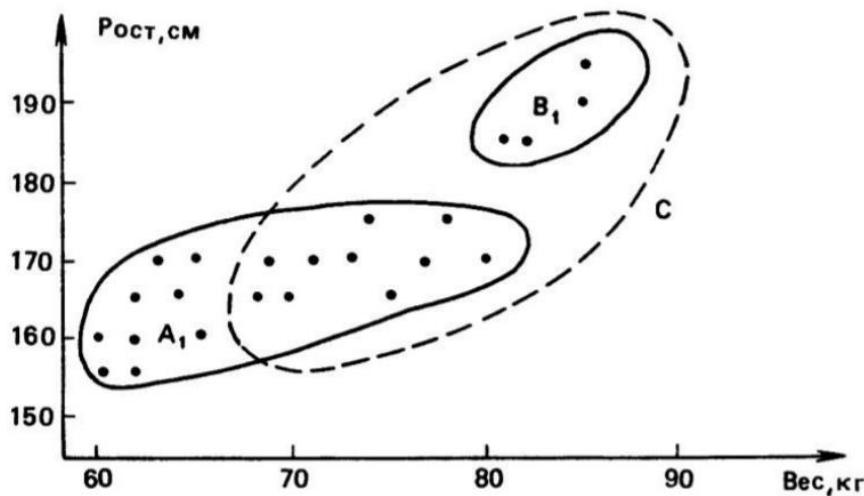


# Проблема чувствительности к выбору метрики

Результат зависит от нормировки признаков:



A — студентки,  
B — студенты



после перенормировки  
(сжали ось «вес» вдвое)



# Анализ качества решения задачи кластеризации

Известны лишь попарные расстояния между объектами

Пусть известны только попарные расстояния между объектами.

- 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:  $F_0/F_1 \rightarrow \min$ .

<https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation>



# Анализ качества решения задачи кластеризации

Известны признаковые описания объектов

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  $a$ ,

$\mu_a$  — центр масс кластера  $a$ .

- Сумма межкластерных расстояний:

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

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

<https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation>



# 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:**  $X^\ell$ ,  $K = |Y|$ . **Выход:** центры кластеров  $\mu_a$ ,  $a \in Y$

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

2: **повторять**

3: отнести каждый  $x_i$  к ближайшему центру:

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

4: вычислить новые положения центров:

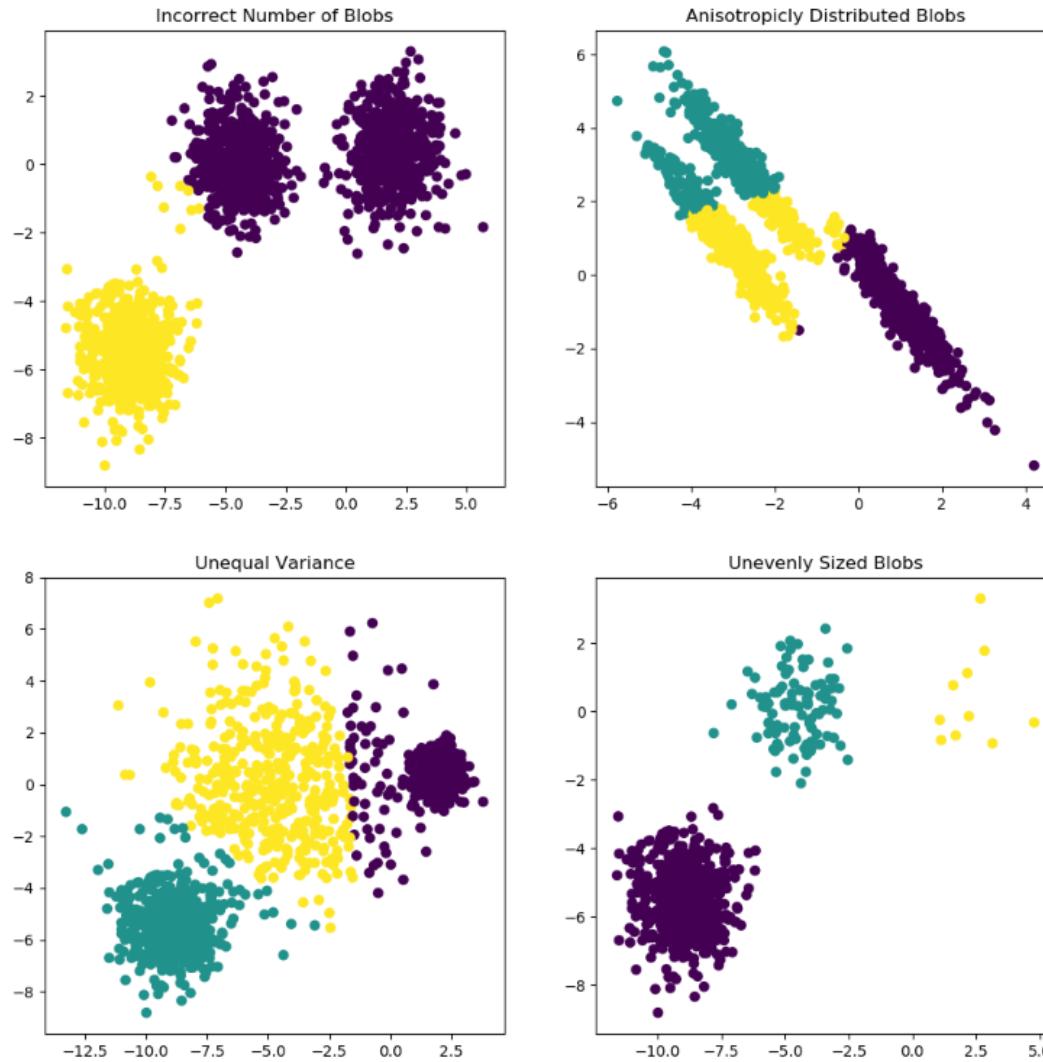
$$\mu_a := \frac{\sum_{i=1}^{\ell} [a_i = a] x_i}{\sum_{i=1}^{\ell} [a_i = a]}, \quad a \in Y;$$

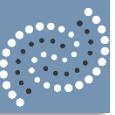
5: пока  $a_i$  не перестанут изменяться;

<https://www.naftaliharris.com/blog/visualizing-k-means-clustering/>



# Метод К-средних (K-means) для кластеризации





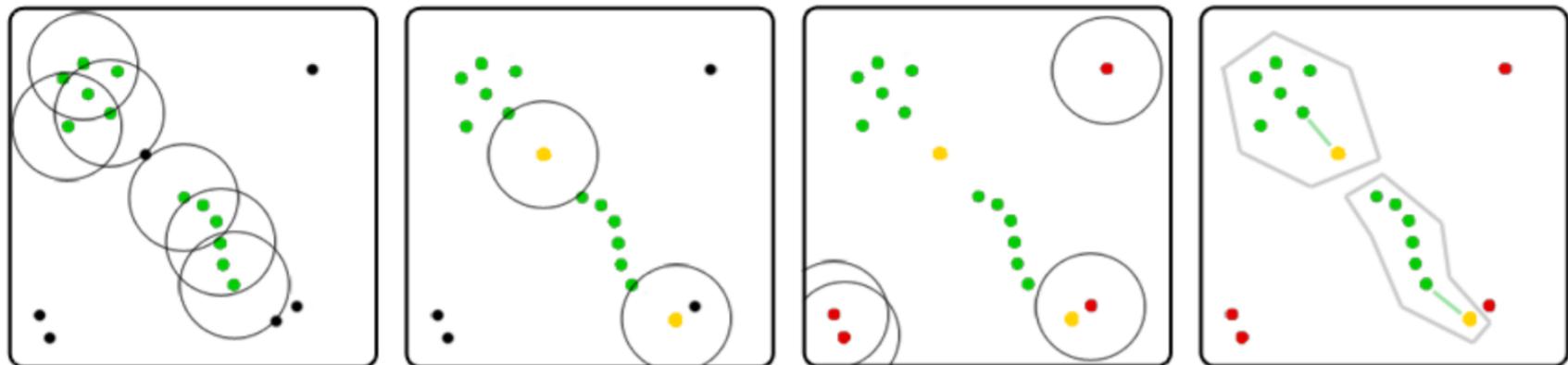
# Алгоритм кластеризации DBSCAN

(Density-Based Spatial Clustering of Applications With Noise)

Объект  $x \in U$ , его  $\varepsilon$ -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.*

<https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/>



# Алгоритм кластеризации DBSCAN

**Вход:** выборка  $X^\ell = \{x_1, \dots, x_\ell\}$ , параметры  $\varepsilon$  и  $m$ ;

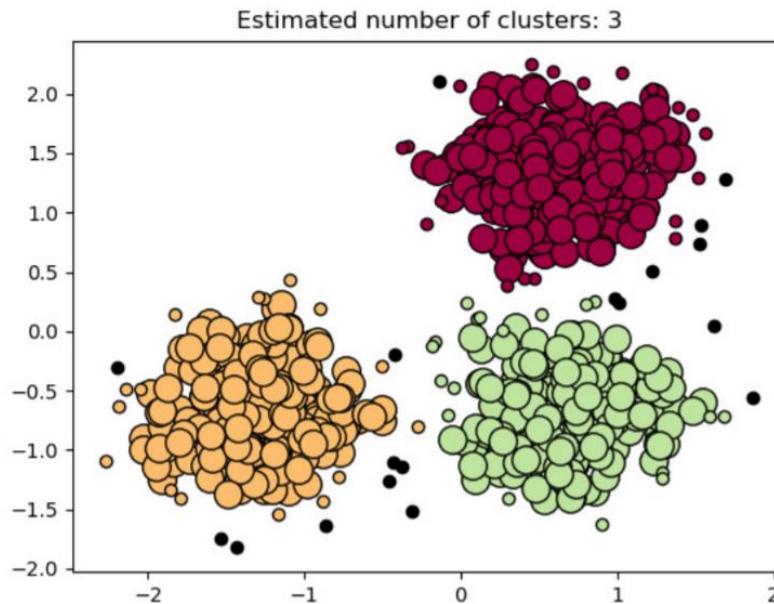
**Выход:** разбиение выборки на кластеры и шумовые выбросы  $N$ ;

- 1:  $U := X^\ell$  — некластеризованные;  $N := \emptyset$ ;  $a := 0$ ;
- 2: **пока** в выборке есть некластеризованные точки,  $U \neq \emptyset$ :
- 3:   взять случайную точку  $x \in U$ ;
- 4:   **если**  $|U_\varepsilon(x)| < m$  **то**
- 5:     пометить  $x$  как, возможно, шумовой;
- 6:   **иначе**
- 7:     создать новый кластер:  $K := U_\varepsilon(x)$ ;  $a := a + 1$ ;
- 8:     **для всех**  $x' \in K$
- 9:       **если**  $|U_\varepsilon(x')| \geq m$  **то**  $K := K \cup U_\varepsilon(x')$ ;
- 10:       **иначе** пометить  $x'$  как граничный кластера  $K$ ;
- 11:      $a_i := a$  для всех  $x_i \in K$ ;
- 12:      $U := U \setminus K$ ;



# 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.

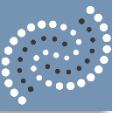




# Агломеративная иерархическая кластеризация

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

- 1:  $C_1 := \{\{x_1\}, \dots, \{x_\ell\}\}$  — все кластеры 1-элементные;  
 $R_{\{x_i\}\{x_j\}} := \rho(x_i, x_j)$  — расстояния между ними;
- 2: **для всех**  $t = 2, \dots, \ell$  ( $t$  — номер итерации):
- 3:   найти в  $C_{t-1}$  пару кластеров  $(U, V)$  с минимальным  $R_{UV}$ ;
- 4:   слить их в один кластер:  
 $W := U \cup V$ ;
- 5:   **для всех**  $S \in C_t$
- 6:     вычислить  $R_{WS}$  по формуле Ланса-Уильямса:  
 $R_{WS} := \alpha_U R_{US} + \alpha_V R_{VS} + \beta R_{UV} + \gamma |R_{US} - R_{VS}|$ ;



## Агломеративная иерархическая кластеризация. Рекомендации

- 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 :=  $C_t$ .

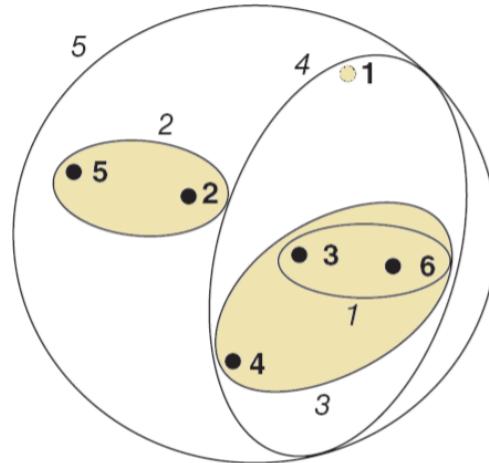
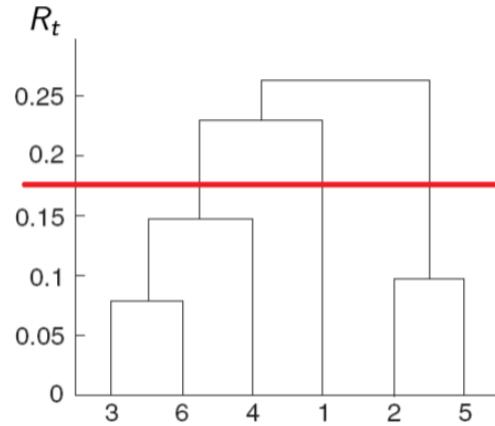


Диаграмма вложения



Дендрограмма



## Поиск ассоциативных правил

Ассоциативные правила позволяют находить закономерности между связанными событиями.

Примером такой закономерности служит правило, указывающее, что из события X следует событие Y с некоторой вероятностью.

Установление таких зависимостей дает возможность находить очень простые и интуитивно понятные правила.



# Примеры применения

- Анализ рыночной корзины (market basket analysis)

- оптимизировать размещение товаров на полках
- формировать персональные рекомендации
- планировать рекламные кампании (промо-акции)
- более эффективно управлять ценами и ассортиментом.

- Кросс-продажи (cross-sell) и продажи с повышением цены (up-selling)

выявление клиентов, склонных к откликам на продажам

персональные предложения и кросс-

- Директ майл (direct mail)

Для увеличения количества откликов на письма необходимо производить тщательный отбор объектов для рассылки посланий



# Определения

## Определение 1

Пусть дан контекст  $\mathbb{K} := (G, M, I)$ , где  $G$  — множество объектов,  $M$  — множество признаков (items),  $I \subseteq G \times M$

Ассоциативным правилом контекста  $\mathbb{K}$  называется выражение вида  $A \rightarrow B$ , где  $A, B \subseteq M$ .

## Определение 2

Поддержкой (*support*) ассоциативного правила  $A \rightarrow B$  называется величина  $supp(A \rightarrow B) = \frac{|(A \cup B)'|}{|G|}$ .

Значение  $supp(A \rightarrow B)$  показывает какая доля объектов  $G$  содержит  $A \cup B$ .

Часто поддержку выражают в %.

## Определение 3

Достоверностью (*confidence*) ассоциативного правила  $A \rightarrow B$  называется величина  $conf(A \rightarrow B) = \frac{|(A \cup B)'|}{|A'|}$ .

Значение  $conf(A \rightarrow B)$  показывает какая доля объектов обладающих  $A$  также содержит  $A \cup B$ . Величину поддержки также часто выражают в %.



# Пример

## Объектно-признаковая таблица транзакций

Покупатели/товары	Пиво	Пряники	Молоко	Мюсли	Чипсы
C <sub>1</sub>	1	0	0	0	1
C <sub>2</sub>	0	1	1	1	0
C <sub>3</sub>	1	0	1	1	1
C <sub>4</sub>	1	1	1	0	1
C <sub>5</sub>	0	1	1	1	1

- $supp(\{\text{Пиво, Чипсы}\}) = 3/5$
- $supp(\{\text{Пряники, Мюсли}\} \rightarrow \{\text{Молоко}\}) =$   
 $= \frac{|(\{\text{Пряники, Мюсли}\} \cup \{\text{Молоко}\})'|}{|G|} = \frac{|\{C_2, C_5\}|}{5} = 2/5$
- $conf(\{\text{Пряники, Мюсли}\} \rightarrow \{\text{Молоко}\}) =$   
 $= \frac{|(\{\text{Пряники, Мюсли}\} \cup \{\text{Молоко}\})'|}{|\{\text{Пряники, Мюсли}\}'|} = \frac{|\{C_2, C_5\}|}{|\{C_2, C_5\}'|} = 1$



# Постановка задачи

Требуется найти все ассоциативные правила контекста, для которых значения поддержки и достоверности превышают некоторые установленные значения, `min_supp` и `min_conf` соответственно



# Алгоритм FP-Growth. Пример

Элементы {B, J, P, M, E}

Min\_support = 40%

Min\_confidence = 80%

TID	Items Bought
1	{B, J, P}
2	{B, P}
3	{B, M, P}
4	{E, B}
5	{E, M}

<http://athena.ecs.csus.edu/~associationcw/FpGrowth.html>



# Алгоритм FP-Growth. Пример

Шаг 1: посчитаем частоты товаров

Items	Support	Support (in percentage) = (Support * 100) / No. of trans
B	4	$(4 * 100) / 5 = 80\%$
J	1	$(1 * 100) / 5 = 20\%$
P	3	$(3 * 100) / 5 = 60\%$
M	2	$(2 * 100) / 5 = 40\%$
E	2	$(2 * 100) / 5 = 40\%$

Шаг 1: исключим из рассмотрения товары, у которых support < min\_support (2)

Items	Support
B	4 (80%)
P	3 (60%)
M	2 (40%)
E	2 (40%)



# Алгоритм FP-Growth. Пример

Шаг 3: создадим F-список (список товаров отсортированных по убыванию их поддержки)

F - list	{ B , P , M , E }
----------	-------------------

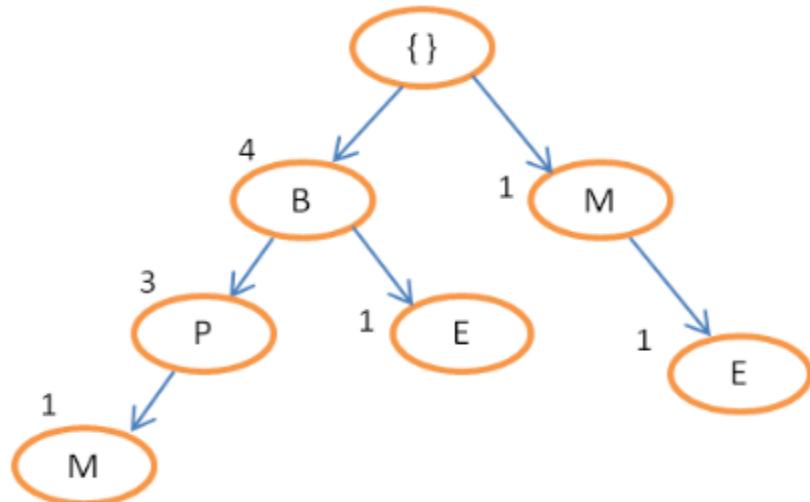
Шаг 4: отсортируем товары в транзакциях по убыванию их support

FPDB	
TID	Items Bought
1	{ B , P }
2	{ B , P }
3	{ B , P , M }
4	{ B , E }
5	{ M , E }



# Алгоритм FP-Growth. Пример

Шаг 5: построим FP-дерево





# Алгоритм FP-Growth. Пример

Шаг 6: построим ассоциативные правила

Item set	Transactions	Conditional FP Tree	Frequent Item set
E	TID-4 = { B } TID-5 = { M }	1 B      1 M	Nil
M	TID-3 = { B, P }	1 B 1 P	Nil
P	TID-1 = { B } TID-2 = { B } TID-3 = { B }	3 B	{ B, P }

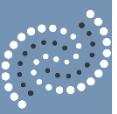
<http://athena.ecs.csus.edu/~associationcw/Algorithms/FpGrowth.html>



# Понижение размерности

## Цели:

- Упрощение последующей обработки данных
- Исключение скоррелированных признаков
- «Удаление» шума из данных
- Визуализация



# Матричные обозначения

Матрицы «объекты–признаки», старая и новая:

$$F_{\ell \times n} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}; \quad G_{\ell \times m} = \begin{pmatrix} g_1(x_1) & \dots & g_m(x_1) \\ \dots & \dots & \dots \\ g_1(x_\ell) & \dots & g_m(x_\ell) \end{pmatrix}.$$

Матрица линейного преобразования новых признаков в старые:

$$U_{n \times m} = \begin{pmatrix} u_{11} & \dots & u_{1m} \\ \dots & \dots & \dots \\ u_{n1} & \dots & u_{nm} \end{pmatrix}; \quad \hat{F} = GU^\top \stackrel{\text{хотим}}{\approx} F.$$

Найти: и новые признаки  $G$ , и преобразование  $U$ :

$$\sum_{i=1}^{\ell} \sum_{j=1}^n (\hat{f}_j(x_i) - f_j(x_i))^2 = \|GU^\top - F\|^2 \rightarrow \min_{G, U},$$

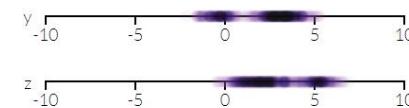
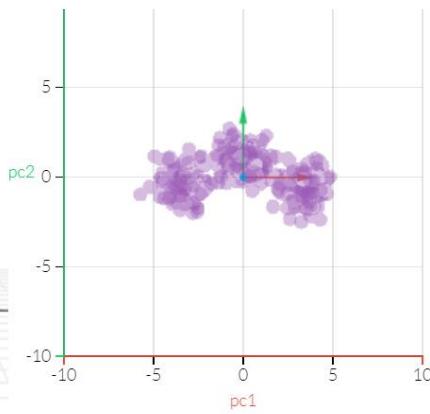
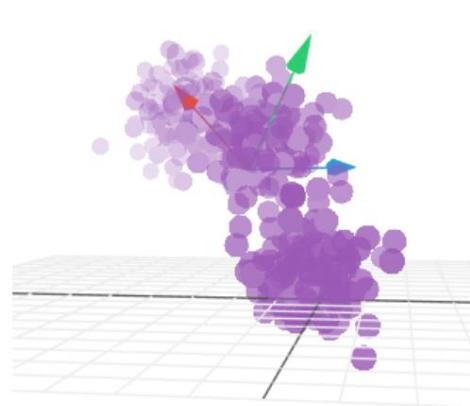
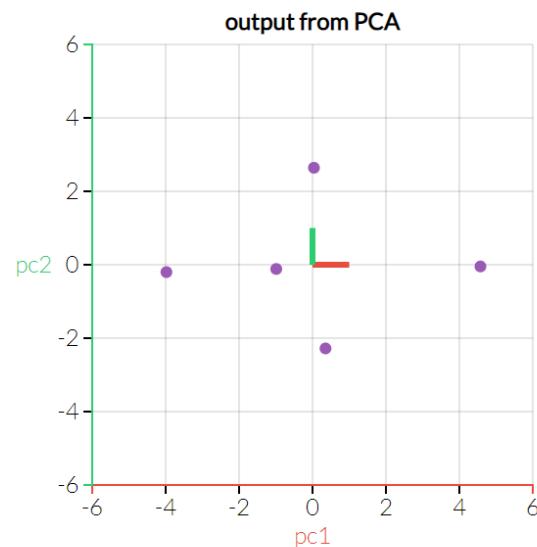
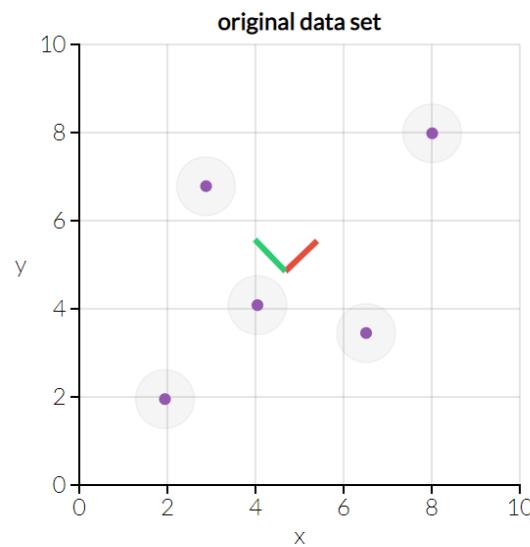


## Интуитивное объяснение

Хотим найти новое признаковое описание и некое линейное преобразование  $U$ , так, чтобы действуя на новое признаковое описание  $G$  мы получили данные близкие к исходным с заданной точностью.



# Визуальная демонстрация



show PCA

reset

<http://setosa.io/ev/principal-component-analysis/>



# Домашнее задание

- Ознакомиться с pros & cons различных алгоритмов и метриками оценки качества кластеризации по [ссылке](#).
- Послушать лекцию о кластеризации для закрепления ([лекция](#) Виктора Лавренко).
- Послушать лекцию о PCA для закрепления ([лекция](#) Виктора Лавренко).
- Пройти [часть Unsupervised курса DSND](#) (заполнить пропуски в ноутбуках, свериться с \*\_solution.ipynb).
- Ознакомиться с алгоритмом Gaussian Mixture Model ([лекция](#) Виктора Лавренко).
- Ознакомиться с алгоритмом X-Means ([статья](#), реализация доступна в пакете pyclustering).

# Q & A

# Thank you!